45

Stig Larsson
Vidar Thomée

TEXTS IN APPLIED MATHEMATICS

Partial Differential Equations with Numerical Methods

Springer

Texts in Applied Mathematics 45

Texts in Applied Mathematics

1. *Sirovich:* Introduction to Applied Mathematics.
2. *Wiggins:* Introduction to Applied Nonlinear Dynamical Systems and Chaos.
3. *Hale/Koçak:* Dynamics and Bifurcations.
4. *Chorin/Marsden:* A Mathematical Introduction to Fluid Mechanics, Third Edition.
5. *Hubbard/West:* Differential Equations: A Dynamical Systems Approach: Ordinary Differential Equations.
6. *Sontag:* Mathematical Control Theory: Deterministic Finite Dimensional Systems, Second Edition.
7. *Perko:* Differential Equations and Dynamical Systems, Third Edition.
8. *Seaborn:* Hypergeometric Functions and Their Applications.
9. *Pipkin:* A Course on Integral Equations.
10. *Hoppensteadt/Peskin:* Modeling and Simulation in Medicine and the Life Sciences, Second Edition.
11. *Braun:* Differential Equations and Their Applications, Fourth Edition.
12. *Stoer/Bulirsch:* Introduction to Numerical Analysis, Third Edition.
13. *Renardy/Rogers:* An Introduction to Partial Differential Equations.
14. *Banks:* Growth and Diffusion Phenomena: Mathematical Frameworks and Applications.
15. *Brenner/Scott:* The Mathematical Theory of Finite Element Methods, Second Edition.
16. *Van de Velde:* Concurrent Scientific Computing.
17. *Marsden/Ratiu:* Introduction to Mechanics and Symmetry, Second Edition.
18. *Hubbard/West:* Differential Equations: A Dynamical Systems Approach: Higher-Dimensional Systems.
19. *Kaplan/Glass:* Understanding Nonlinear Dynamics.
20. *Holmes:* Introduction to Perturbation Methods.
21. *Curtain/Zwart:* An Introduction to Infinite-Dimensional Linear Systems Theory.
22. *Thomas:* Numerical Partial Differential Equations: Finite Difference Methods.
23. *Taylor:* Partial Differential Equations: Basic Theory.
24. *Merkin:* Introduction to the Theory of Stability.
25. *Naber:* Topology, Geometry, and Gauge Fields: Foundations.
26. *Polderman/Willems:* Introduction to Mathematical Systems Theory: A Behavioral Approach.
27. *Reddy:* Introductory Functional Analysis: with Applications to Boundary Value Problems and Finite Elements.
28. *Gustafson/Wilcox:* Analytical and Computational Methods of Advanced Engineering Mathematics.
29. *Tveito/Winther:* Introduction to Partial Differential Equations: A Computational Approach.
30. *Gasquet/Witomski:* Fourier Analysis and Applications: Filtering, Numerical Computation, Wavelets.
31. *Brémaud:* Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues.
32. *Durran:* Numerical Methods for Wave Equations in Geophysical Fluid Dynamics.
33. *Thomas:* Numerical Partial Differential Equations: Conservation Laws and Elliptic Equations.

(continued after index)

Stig Larsson · Vidar Thomée

Partial Differential Equations with Numerical Methods

Springer

Stig Larsson
Vidar Thomée

Mathematical Sciences
Chalmers University of Technology
and University of Gothenburg
412 96 Göteborg
Sweden

stig@chalmers.se
thomee@chalmers.se

First softcover printing 2009

ISBN 978-3-540-88705-8 e-ISBN 978-3-540-88706-5

DOI 10.1007/978-3-540-88706-5

Texts in Applied Mathematics ISSN 0939-2475

Library of Congress Control Number: 2008940064

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

Coverdesign: WMXDesign GmbH, Heidelberg

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Series Preface

Mathematics is playing an ever more important role in the physical and biological sciences, provoking a blurring of boundaries between scientific disciplines and a resurgence of interest in the modern as well as the classical techniques of applied mathematics. This renewal of interest, both in research and teaching, has led to the establishment of the series Texts in Applied Mathematics (TAM).

The development of new courses is a natural consequence of a high level of excitement on the research frontier as newer techniques, such as numerical and symbolic computer systems, dynamical systems, and chaos, mix with and reinforce the traditional methods of applied mathematics. Thus, the purpose of this textbook series is to meet the current and future needs of these advances and to encourage the teaching of new courses.

TAM will publish textbooks suitable for use in advanced undergraduate and beginning graduate courses, and will complement the Applied Mathematical Sciences (AMS) series, which will focus on advanced textbooks and research-level monographs.

Pasadena, California — J.E. Marsden
New York, New York — L. Sirovich
College Park, Maryland — S.S. Antman

Preface

Our purpose in this book is to give an elementary, relatively short, and hopefully readable account of the basic types of linear partial differential equations and their properties, together with the most commonly used methods for their numerical solution. Our approach is to integrate the mathematical analysis of the differential equations with the corresponding numerical analysis. For the mathematician interested in partial differential equations or the person using such equations in the modelling of physical problems, it is important to realize that numerical methods are normally needed to find actual values of the solutions, and for the numerical analyst it is essential to be aware that numerical methods can only be designed, analyzed, and understood with sufficient knowledge of the theory of the differential equations, using discrete analogues of properties of these.

In our presentation we study the three major types of linear partial differential equations, namely elliptic, parabolic, and hyperbolic equations, and for each of these types of equations the text contains three chapters. In the first of these we introduce basic mathematical properties of the differential equation, and discuss existence, uniqueness, stability, and regularity of solutions of the various boundary value problems, and the remaining two chapters are devoted to the most important and widely used classes of numerical methods, namely finite difference methods and finite element methods.

Historically, finite difference methods were the first to be developed and applied. These are normally defined by looking for an approximate solution on a uniform mesh of points and by replacing the derivatives in the differential equation by difference quotients at the mesh-points. Finite element methods are based instead on variational formulations of the differential equations and determine approximate solutions that are piecewise polynomials on some partition of the domain under consideration. The former method is somewhat restricted by the difficulty of adapting the mesh to a general domain whereas the latter is more naturally suited for a general geometry. Finite element methods have become most popular for elliptic and also for parabolic problems, whereas for hyperbolic equations the finite difference method continues to dominate. In spite of the somewhat different philosophy underlying the two classes it is more reasonable in our view to consider the latter as further

developments of the former rather than as competitors, and we feel that the practitioner of differential equations should be familiar with both.

To make the presentation more easily accessible, the elliptic chapters are preceded by a chapter about the two-point boundary value problem for a second order ordinary differential equation, and those on parabolic and hyperbolic evolution equations by a short chapter about the initial value problem for a system of ordinary differential equations. We also include a chapter about eigenvalue problems and eigenfunction expansion, which is an important tool in the analysis of partial differential equations. There we also give some simple examples of numerical solution of eigenvalue problems.

The last chapter provides a short survey of other classes of numerical methods of importance, namely collocation methods, finite volume methods, spectral methods, and boundary element methods.

The presentation does not presume a deep knowledge of mathematical and functional analysis. In an appendix we collect some of the basic material that we need in these areas, mostly without proofs, such as elements of abstract linear spaces and function spaces, in particular Sobolev spaces, together with basic facts about Fourier transforms. In the implementation of numerical methods it will normally be necessary to solve large systems of linear algebraic equations, and these generally have to be solved by iterative methods. In a second appendix we therefore include an orientation about such methods.

Our purpose has thus been to cover a rather wide variety of topics, notions, and ideas, rather than to expound on the most general and far-reaching results or to go deeply into any one type of application. In the problem sections, which end the various chapters, we sometimes ask the reader to prove some results which are only stated in the text, and also to further develop some of the ideas presented. In some problems we propose testing some of the numerical methods on the computer, assuming that MATLAB or some similar software is available. At the end of the book we list a number of standard references where more material and more detail can be found, including issues concerned with implementation of the numerical methods.

This book has developed from courses that we have given over a rather long period of time at Chalmers University of Technology and Göteborg University originally for third year engineering students but later also in beginning graduate courses for applied mathematics students. We would like to thank the many students in these courses for the opportunities for us to test our ideas.

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

In the second printing 2005 we have corrected several misprints and minor inadequacies, and added a few problems. *SL & VT*

Contents

1 Introduction

In this first chapter we begin in Sect. 1.1 by introducing the partial differential equations and associated initial and boundary value problems that we shall study in the following chapters. The equations are classified into elliptic, parabolic, and hyperbolic equations, and we indicate the corresponding type of problems in physics that they model. We discuss briefly the concept of a well posed boundary value problem, and the various techniques used in our subsequent presentation. In Sect. 1.2 we introduce some notation and concepts that will be used throughout the text, and in Sect. 1.3 we include a detailed derivation of the heat equation from physical principles explaining the meaning of all terms that occur in the equation and the boundary conditions. In the problem section, Sect. 1.4, we add some further illustrative material.

1.1 Background

In this text we study boundary value and initial-boundary value problems for partial differential equations, that are significant in applications, from both a theoretical and a numerical point of view. As a typical example of such a boundary value problem we consider first Dirichlet's problem for Poisson's equation,

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

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

where $x = (x_1, \ldots, x_d)$, Δ is the Laplacian defined by $\Delta u = \sum_{j=1}^{d} \partial^2 u / \partial x_j^2$, and Ω is a bounded domain in d-dimensional Euclidean space $\mathbf{R}^d$ with boundary Γ. The given functions $f = f(x)$ and $g = g(x)$ are the *data* of the problem. Instead of Dirichlet's boundary condition (1.2) one can consider, for instance, Neumann's boundary condition

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

where $\partial u / \partial n$ denotes the derivative in the direction of the exterior unit normal n to Γ. Another choice is Robin's boundary condition

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

More generally, a linear second order elliptic equation is of the form

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

where $A(x) = (a_{ij}(x))$ is a sufficiently smooth positive definite matrix, and such an equation may also be considered in Ω together with various boundary conditions. In our treatment below we shall often restrict ourselves, for simplicity, to the isotropic case $A(x) = a(x)I$, where $a(x)$ is a smooth positive function and I the identity matrix.

Elliptic equations such as the above occur in a variety of applications, modeling, for instance, various potential fields (gravitational, electrostatic, magnetostatic, etc.), probability densities in random-walk problems, stationary heat flow, and biological phenomena. They are also related to important areas within pure mathematics, such as the theory of functions of a complex variable $z = x + iy$, conformal mapping, etc. In applications they often describe stationary, or time independent, physical states.

We also consider time dependent problems, and our two model equations are the heat equation,

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

and the wave equation,

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

These will be considered for positive time t, and for x varying either throughout $\mathbf{R}^d$ or in some bounded domain $\Omega \subset \mathbf{R}^d$, on the boundary of which boundary conditions are prescribed as for Poisson's equation above. For these time dependent problems, the value of the solution u has to be given at the initial time $t = 0$, and in the case of the wave equation, also the value of $\partial u/\partial t$ at $t = 0$. In the case of the unrestricted space $\mathbf{R}^d$ the respective problems are referred to as the pure *initial value problem* or *Cauchy problem* and, in the case of a bounded domain Ω, a mixed *initial-boundary value problem.*

Again, these equations, and their generalizations permitting more general elliptic operators than the Laplacian Δ, appear in a variety of applied contexts, such as, in the case of the heat equation, in the conduction of heat in solids, in mass transport by diffusion, in diffusion of vortices in viscous fluid flow, in telegraphic transmission in cables, in the theory of electromagnetic waves, in hydromagnetics, in stochastic and biological processes; and, in the case of the wave equation, in vibration problems in solids, in sound waves in

a tube, in the transmission of electricity along an insulated, low resistance cable, in long water waves in a straight canal, etc.

Some characteristics of equations of type (1.7) are shared with certain systems of first order partial differential equations. We shall therefore also have reason to study scalar linear partial differential equations of the form

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

and corresponding systems where the coefficients are matrices. Such systems appear, for instance, in fluid dynamics and electromagnetic field theory.

Applied problems often lead to partial differential equations which are nonlinear. The treatment of such equations is beyond the scope of this presentation. In many cases, however, it is useful to study linearized versions of these, and the theory of linear equations is therefore relevant also to nonlinear problems.

In applications, the equations used in the models normally contain physical parameters. For instance, in the case of the heat conduction problem, the temperature at a point of a homogeneous isotropic solid, extended over Ω, with the thermal conductivity k, density ρ, and specific heat capacity c, and with a heat source $f(x,t)$, satisfies

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

If ρ, c, and k are constant, this equation may be written in the form (1.6) after a simple transformation, but if they vary with x, a more general elliptic operator is involved.

In Sect. 1.3 below we derive the heat equation from physical principles and explain, in the context given, the physical meaning of all terms in the elliptic operator (1.5) as well as the boundary conditions (1.2), (1.3), and (1.4). A corresponding derivation or the wave equation is given in Problem 1.2. Boundary value problems for elliptic equations, or stationary problems, may appear as limiting cases of the evolution problems as $t \to \infty$.

One characteristic of mathematical modeling is that once the model is established, in our case as an initial or initial-boundary value problem for a partial differential equation, the analysis becomes purely mathematical and is independent of any specific application that the model describes. The results obtained are then valid for all the different examples of the model. We shall therefore not use much terminology from physics or other applied fields in our exposition, but invoke special applications in the exercises. It is often convenient to keep such examples in mind to enhance the intuitive understanding of a mathematical model.

The equations (1.1), (1.6), and (1.7) are said to be of elliptic, parabolic, and hyperbolic type, respectively. We shall return to the classification of

partial differential equations into different types in Chapt. 11 below, and note here only that a differential equation in two variables x and t of the form

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

is said to be *elliptic, hyperbolic* or *parabolic* depending on whether $\delta = ac - b^2$ is positive, negative, or zero. Here $\ldots$ stands for a linear combination of derivatives of orders at most 1. In particular,

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

and

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

are of these three types, respectively. Note that the conditions on the sign of δ are the same as those occurring in the classification of plane quadratic curves into ellipses, hyperbolas, and parabolas.

Together with the partial differential equations we also study numerical approximations by finite difference and finite element methods. For these problems, the continuous and the discretized equations, we prove results of the following types:

- *existence* of solutions,
- *uniqueness* of solutions,
- *stability*, or continuous dependence of solutions with respect to perturbations of data,
- *error estimates* (for numerical methods).

A boundary value problem that satisfies the three first of these conditions is said to be *well posed*. In order to prove such results we employ several techniques:

- *maximum principles*,
- *Fourier methods*; these are techniques that are based on the use of the Fourier transform, Fourier series expansion, or eigenfunction expansion,
- *energy estimates*,
- representation of solution operators by means of *Green's functions*.

1.2 Notation and Mathematical Preliminaries

In this section we briefly introduce some basic notation that will be used throughout the book. For more details on function spaces and norms we refer to App. A.

By $\mathbf{R}$ and $\mathbf{C}$ we denote the sets of real and complex numbers, respectively, and we write

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

A subset of $\mathbf{R}^d$ is called a domain if it is open and connected. By Ω we usually denote a bounded domain in $\mathbf{R}^d$, for $i = 1, 2$, or 3 (if $d = 1$, then Ω is a bounded open interval). Its boundary $\partial\Omega$ is usually denoted Γ. We assume throughout that Γ is either smooth or a polygon (if $d = 2$) or polyhedron (if $d = 3$). By $\bar{\Omega}$ we denote the closure of Ω, i.e., $\bar{\Omega} = \Omega \cup \Gamma$. The (length, area, or) volume of Ω is denoted by $|\Omega|$, the volume element in $\mathbf{R}^d$ is $\mathrm{d}x = \mathrm{d}x_1 \cdots \mathrm{d}x_d$, and $\mathrm{d}s$ denotes the element of arclength (if $d = 2$) or surface area (if $d = 3$) on Γ. For vectors in $\mathbf{R}^d$ we use the Euclidean inner product $x \cdot y = \sum_{i=1}^d x_i y_i$ and norm $|x| = \sqrt{x \cdot x}$.

Let u, v be scalar functions and $w = (w_1, \ldots, w_d)$ a vector-valued function of $x \in \mathbf{R}^d$. We define the gradient, the divergence, and the Laplace operator (Laplacian) by

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

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

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

We recall the *divergence theorem*

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

where $n = (n_1, \ldots, n_d)$ is the outward unit normal to Γ. Applying this to the product wv we obtain *Green's formula*:

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

When applied with $w = \nabla u$ the formula becomes

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

where $\partial u/\partial n = n \cdot \nabla u$ is the exterior normal derivative of u on Γ.

A *multi-index* $\alpha = (\alpha_1, \ldots, \alpha_d)$ is a d-vector where the α_i are non-negative integers. The *length* $|\alpha|$ of a multi-index α is defined by $|\alpha| = \sum_{i=1}^d \alpha_i$. Given a function $v : \mathbf{R}^d \to \mathbf{R}$ we may write its partial derivatives of order $|\alpha|$ as

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

A linear partial differential equation of order k in Ω can therefore be written

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

where the coefficients $a_\alpha(x)$ are functions of x in Ω. We also use subscripts to denote partial derivatives, e.g.,

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

For $M \subset \mathbf{R}^d$ we denote by $\mathcal{C}(M)$ the linear space of continuous functions on M, and for bounded continuous functions we define the maximum-norm

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

For example, this defines $\|v\|_{\mathcal{C}(\mathbf{R}^d)}$. When M is a bounded and closed set, i.e., a compact set, the supremum in (1.9) is attained and we may write

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

For a not necessarily bounded domain Ω and k a non-negative integer we denote by $\mathcal{C}^k(\Omega)$ the set of k times continuously differentiable functions in Ω. For a bounded domain Ω we write $\mathcal{C}^k(\bar{\Omega})$ for the functions $v \in \mathcal{C}^k(\Omega)$ such that $D^\alpha v \in \mathcal{C}(\bar{\Omega})$ for all $|\alpha| \le k$. For functions in $\mathcal{C}^k(\bar{\Omega})$ we use the norm

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

and the seminorm, including only the derivatives of highest order,

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

When we are working on a fixed domain Ω we often omit the set in the notation and write simply $\|v\|_{\mathcal{C}}$, $|v|_{\mathcal{C}^k}$, etc.

By $\mathcal{C}_0^k(\Omega)$ we denote the set of functions $v \in \mathcal{C}^k(\Omega)$ that vanish outside some compact subset of Ω, in particular, such functions satisfy $D^\alpha v = 0$ on the boundary of Ω for $|\alpha| \le k$. Similarly, $\mathcal{C}_0^\infty(\mathbf{R}^d)$ is the set of functions that have continuous derivatives of all orders and vanish outside some bounded set.

We say that a function is *smooth* if, depending on the situation, it has sufficiently many continuous derivatives.

We also frequently employ the space $L_2(\Omega)$ of square integrable functions with scalar product and norm

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

For Ω a domain we also employ the Sobolev space $H^k(\Omega)$, $k \geq 1$, of functions v such that $D^\alpha v \in L_2(\Omega)$ for all $|\alpha| \leq k$, equipped with the norm and seminorm

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

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

Additional norms are defined and used locally when the need arises.

We use the letters c, C to denote various positive constants that need not be the same at each occurrence.

1.3 Physical Derivation of the Heat Equation

Many equations in physics are derived by combining a conservation law with constitutive relations. A conservation law states that a physical quantity, such as energy, mass, or momentum, is conserved as the physical process develops in time. Constitutive relations express our assumptions about how the material behaves when the state variables change.

In this section we we consider the conduction of heat in a body $\Omega \subset \mathbf{R}^3$ with boundary Γ and derive the heat equation using conservation of energy together with linear constitutive relations.

Conservation of Energy

Consider the balance of heat in an arbitrary subset $\Omega_0 \subset \Omega$ with boundary Γ_0. The energy principle says that the rate of change of the total energy in Ω_0 equals the inflow of heat through Γ_0 plus the heat power produced by heat sources inside Ω_0. To express this in mathematical terms we introduce some physical quantities, each of which is followed, within brackets, by the associated standard unit of measurement.

With $e = e(x,t)$ `[J/m^3]` the *density of internal energy* at the point x `[m]` and time t `[s]`, the total amount of heat in Ω_0 is $\int_{\Omega_0} e\,\mathrm{d}x$ `[J]`. Further with the vector field $j = j(x,t)$ `[J/(m^2s)]` denoting the *heat flux* and n the exterior unit normal to Γ_0, the net outflow of heat through Γ_0 is $\int_{\Gamma_0} j \cdot n\,\mathrm{d}s$ `[J/s]`. Introducing also the power density of heat sources $p = p(x,t)$ `[J/(m^3s)]`, the energy principle then states that

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

Applying the divergence theorem we obtain

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

Since $\Omega_0 \subset \Omega$ is arbitrary this implies

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

Constitutive Relations

The internal energy density e depends on the absolute temperature T `[K]` and the spatial coordinates, and in our first constitutive relation we assume that e depends linearly on T near a suitably chosen reference temperature T_0, that is,

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

The coefficient $\sigma = \sigma(x)$ `[J/(m^3 K)]` is called the *specific heat capacity*. (It is usually expressed in the form $\sigma = \rho\, c$, where ρ `[kg/m^3]` is mass density and c `[J/(kg K)]` is the specific heat capacity per unit mass.)

According to *Fourier's law* the heat flux due to conduction is proportional to the temperature gradient, which gives a second constitutive relation,

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

The coefficient $\lambda = \lambda(x)$ `[J/(m K s)]` is called the *heat conductivity*. In some situations (e.g., gas in a porous medium, heat transport in a fluid) heat is also transported by convection with heat flux $v\, e$, where $v = v(x,t)$ `[m/s]` is the convective velocity vector field. The constitutive relation then reads

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

Substituting (1.11) and (1.12) into (1.10) we obtain

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

which is the *heat equation* with convection.

Boundary Conditions

In the modelling of heat conduction, the differential equation (1.13) is combined with an *initial condition* at time $t = 0$,

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

and a *boundary condition*, expressing that the heat flux through the boundary is proportional to the difference between the surface temperature and the ambient temperature, $j \cdot n = \kappa(\vartheta - \vartheta_a)$, where $\kappa = \kappa(x,t)$ [J/(m^2 s K)] is a heat transfer coefficient. Assuming that the material flow does not penetrate the boundary, i.e., $v \cdot n = 0$, we obtain from (1.12)

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

where $\partial\vartheta/\partial n = \nabla\vartheta \cdot n$ denotes the exterior normal derivative of ϑ. Therefore the boundary condition is *Robin's boundary condition*

$$\lambda \frac{\partial \vartheta}{\partial n} + \kappa(\vartheta - \vartheta_a) = 0 \quad \text{on } \Gamma. \tag{1.15}$$

The limit case $\kappa = 0$ means that the boundary surface is perfectly insulated, so that we have *Neumann's boundary condition*,

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

At the other extreme, dividing by κ in (1.15) and letting $\kappa \to \infty$, we obtain *Dirichlet's boundary condition*

$$\vartheta = \vartheta_a. \tag{1.16}$$

The limit case $\kappa = \infty$ thus means that the body is in perfect thermal contact with the surroundings, i.e., heat flows freely through the surface, so that the surface temperature of the body is equal to the ambient temperature.

Dimensionless Form

It is often useful to write the above equations in dimensionless form. Choosing reference constants L [m], τ [s], ϑ_f [K], σ_f [J/(m^3 K)], v_f [m/s], etc., we define dimensionless variables

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

In order to make the heat equation (1.13) dimensionless we divide it by $\lambda_f \vartheta_f / L^2$. Using the chain rule,

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

we get

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

where

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

It is natural to choose $\tau = L^2\sigma_\mathrm{f}/\lambda_\mathrm{f}$, so that $d = 1$ if $\sigma = \sigma_\mathrm{f}$ is constant. The dimensionless number $\mathrm{Pe} = v_\mathrm{f}\sigma_\mathrm{f}L/\lambda_\mathrm{f}$ that appears in the definition of b is called Peclet's number and measures the relative strengths of convection and conduction. Skipping the tilde from now on, we write (1.17) as

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

The boundary condition (1.15) and the initial condition (1.14) transform in a similar way to

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

and

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

Here $h = \mathrm{Bi}\,\kappa/\kappa_\mathrm{f}$, where $\mathrm{Bi} = L\kappa_\mathrm{f}/\lambda_\mathrm{f}$ is called the Biot number.

The partial differential equation (1.18) together with the initial condition (1.20) and the boundary condition (1.19) is called an *initial-boundary value problem*. The term $-\nabla\cdot(a\nabla u)$ is written in *divergence form*. This form arises naturally in the derivation of the equation, and it is convenient in much of the mathematical analysis, as we shall see below. However, we sometimes expand the derivative and write the equation in non-divergence form:

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

Some Simplified Problems

It is useful to study various simplifications of the above equations, because it may then be possible to carry the mathematical analysis further than in the general case. If we assume that the coefficients are constant, with $b = 0$, $c = 0$, then (1.18) reduces to (recall that $d = 1$ if σ is constant)

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

For $a = 1$ this is equation (1.6). If f and the boundary condition are independent of t, then u could be expected to approach a stationary state as t grows, i.e., $u(x, t) \to v(x)$ as $t \to \infty$, and since we should then have $\partial u/\partial t \to 0$, we find that v satisfies *Poisson's equation* (1.1). If in addition $f = 0$, we have *Laplace's equation*

$$-\Delta u = 0.$$

Solutions of Laplace's equation are called *harmonic functions*.

Another important kind of simplification is obtained by reduction of dimension. For example, consider stationary (time-independent) heat conduction in a (not necessarily circular) cylinder oriented along the x_1-axis with insulated mantle surface. If the coefficients a, b, c, f in (1.18) are independent of x_2 and x_3, then it is reasonable to assume that the solution u also depends only on one variable x_1, which we then denote by x, i.e., $u = u(x)$. The heat equation (1.18) then reduces to an ordinary differential equation

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

The boundary condition (1.19) becomes

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

We call this a *two-point boundary value problem*. Similar simplifications are obtained under cylindrical and spherical symmetry by writing the equations in cylindrical respectively spherical coordinates. If the coefficients are constant, then we can readily express the solution in terms of well-known special functions, see Problem 1.6.

Nonlinear Equations, Linearization

The coefficients in the heat equation (1.18) and in the boundary conditions often depend on the temperature u, which makes the equations nonlinear. Although the study of nonlinear equations is outside the scope of this book, we mention that the study of nonlinear equations often proceeds by *linearization*, i.e., by reduction to the study of related linear equations. We illustrate this in the case of the equation

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

which is of the form (1.18), and which is to be solved together with suitable initial and boundary conditions. One approach to such a problem is to use Newton's method, which produces a sequence of approximate solutions u^k from a starting guess u^0 in the following way: Given u^k we want to find an increment v^k such that $u^{k+1} = u^k + v^k$ is a better approximation of the exact solution than u^k. Approximating $F(u^{k+1}) = 0$ by $F(u^k) + F'(u^k)v^k = 0$, we obtain a linearized equation

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

which is solved together with an initial condition and linearized boundary conditions. This equation is a linear equation in v^k of the form (1.18), where the new coefficients $a(u^k(x,t))$, etc., depend on x and t.

1.4 Problems

Problem 1.1. (Derivation of the convection-diffusion equation.) Let $c = c(x,t)$ [mol/m^3] denote the concentration at the point x [m] and time t [s] of a substance that is being transported through a domain $x \in \Omega \subset \mathbf{R}^3$ by convection and diffusion. The flux due to convection is

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

where $v = v(x)$ [m/s] is the convective velocity field. The flux due to diffusion is (Fick's law)

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

where $D = D(x)$ [m^2/s] is the diffusion coefficient. Let r [mol/(m^{3}s)] denote the rate of creation/annihilation of material, e.g., by chemical reaction. The total mass of the substance within an arbitrary subdomain is $\int_{\Omega_0} c\,\mathrm{d}x$. Use the conservation of mass and the divergence theorem to derive the convection-diffusion equation

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

which is of the same mathematical form as (1.13). Derive a boundary condition of the form (1.15). Show that these equations can be written in the same dimensionless form as (1.18) and (1.19).

Problem 1.2. (Derivation of the wave equation.) Consider the longitudinal motion of an elastic bar of length L [m] and of constant cross-sectional area A [m^2] and with density ρ [kg/m^3]. Let $u = u(x,t)$ [m] denote the displacement at time t [s] of a cross-section originally located at $x \in [0, L]$. Newton's law of motion states that

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

where $\int_a^b pA\,\mathrm{d}x$ [kg m/s] is the total momentum of an arbitrary segment (a,b) and σ [N/m^2] is the stress (force per unit cross-sectional area). This leads to

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

For small displacements we have a linear relationship between the stress σ and the strain $\epsilon = \partial u/\partial x$, namely Hooke's law,

$$\sigma = E\epsilon,$$

where E [N/m^2] is the modulus of elasticity, and the momentum density is given by $p = \rho \partial u/\partial t$. Show that u satisfies the wave equation

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

Discuss various possible boundary conditions at the ends of the bar. For example, at $x = L$:

- fixed end, $u(L) = 0$,
- free end, $\sigma(L) = 0$, which leads to $u_x(L) = 0$,
- elastic support, $\sigma(L) = -ku(L)$, which leads to $Eu_x(L) + ku(L) = 0$.

Note that these are of the form (1.23).

Problem 1.3. (Elastic beam.) Consider the bending of an elastic beam that extends along the interval $0 \le x \le L$. At an arbitrary cross-section at a distance x from the left end we introduce the bending moment (torque) $M = M(x)$ [Nm], the transversal force $T = T(x)$ [N], and the external applied force $q = q(x)$ per unit length [N/m]. It can be shown that equilibrium of forces requires $M' = T$ and $T' = -q$. Let $u = u(x)$ [m] be the small transversal deflection of the beam. The bending angle is then approximately u'. The constitutive law is $M = -EIu''$, where E [$\mathrm{N/m^2}$] the modulus of elasticity and I [$\mathrm{m^4}$] is a moment of inertia of the cross-section of the beam. Show that this leads to the fourth order equation

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

Discuss various possible boundary conditions at the ends of the beam. For example, at $x = L$:

- clamped end, $u(L) = 0$, $u'(L) = 0$,
- free end, $M(L) = -(EIu'')(L) = 0$, $T(L) = -(EIu'')'(L) = 0$,
- hinge, $u'(L) = 0$, $M(L) = -(EIu'')(L) = 0$.

Problem 1.4. (The Laplace operator in spherical symmetry.) Introduce spherical coordinates (r, θ, ϕ) defined by $x_1 = r\sin\theta\cos\phi$, $x_2 = r\sin\theta\sin\phi$, $x_3 = r\cos\theta$. Assume that the function u does not depend on θ and ϕ, i.e., $u = u(r)$. Show that

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

Problem 1.5. (The Laplace operator in cylindrical symmetry.) Introduce cylindrical coordinates (ρ, φ, z) defined by $x_1 = \rho\cos\varphi$, $x_2 = \rho\sin\varphi$, $x_3 = z$. Assume that the function u does not depend on φ and z, i.e., $u = u(\rho)$. Show that

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

Problem 1.6. Let $\Omega = \{x \in \mathbf{R}^3 : |x| < 1\}$. Determine an explicit solution of the boundary value problem

$$-\Delta u + c^2 u = f \quad \text{in } \Omega, \quad \text{with } u = g \quad \text{on } \Gamma,$$

assuming spherical symmetry and that c, f, g are constants. That is, solve

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

Hint: Set $v(r) = ru(r)$.

2 A Two-Point Boundary Value Problem

For the purpose of preparing for the treatment of boundary value problems for elliptic partial differential equations we consider here a simple two-point boundary value problem for a second order linear ordinary differential equation. In the first section we derive a maximum principle for this problem, and use it to show uniqueness and continuous dependence on data. In the second section we construct a Green's function in a special case and show how this implies the existence of a solution. In the third section we write the problem in variational form, and use this together with simple tools from functional analysis to prove existence, uniqueness, and continuous dependence on data.

2.1 The Maximum Principle

We consider the boundary value problem

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

where the coefficients $a = a(x)$, $b = b(x)$, and $c = c(x)$ are smooth functions with

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

and where the function $f = f(x)$ and the numbers u_0, u_1 are given, cf. Sect. 1.3.

In the particular case that $a = 1,\ b = c = 0$, this reduces to

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

By integrating this differential equation twice we find that a solution must have the form

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

with the constants α, β to be determined. Setting $x = 0$ and $x = 1$ we find

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

Reversing the steps we find that (2.4), with these α, β, is the unique solution of (2.3).

In the special case $f = 0$ the solution of (2.3) is the linear function $u(x) = u_0(1-x) + u_1 x$. In particular, the values of this function lie between those at $x = 0$ and $x = 1$, and its maximum and minimum are thus located at the endpoints of the interval Ω. More generally, we have the following maximum (minimum) principle for (2.1).

Theorem 2.1. *Consider the differential operator $\mathcal{A}$ in* (2.1)*, and assume that $u \in \mathcal{C}^2 = \mathcal{C}^2(\bar{\Omega})$ and*

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

(i) *If $c = 0$, then*

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

(ii) *If $c \ge 0$ in Ω, then*

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

In case (i) we conclude that the maximum of u is attained at the boundary, i.e., at one of the endpoints of the interval Ω. In case (ii) we draw the same conclusion if the maximum is nonnegative. This does not exclude the possibility that the maximum is attained also in the interior of Ω. However, there is also a stronger form of the maximum principle, which in case (i) asserts that if (2.5) holds and u has a maximum at an interior point of Ω (in case (ii) a nonnegative interior maximum), then u is constant in $\bar{\Omega}$. We shall not prove this here, but we refer to Sect. 3.3 below for the corresponding result for harmonic functions. The variants within parentheses, with $\mathcal{A}u \ge 0$, may be described as a minimum principle; it is reduced to the maximum principle by looking at $-u$.

Proof. (i) Assume first, instead of (2.5), that $\mathcal{A}u < 0$ in Ω. If u has a maximum at an interior point $x_0 \in \Omega$, then at this point we have $u'(x_0) = 0$ and $u''(x_0) \le 0$, so that $\mathcal{A}u(x_0) \ge 0$, which contradicts our assumption. Hence u cannot have an interior maximum point and (2.6) follows.

Assume now that we only know that $\mathcal{A}u \le 0$ in Ω. Let ϕ be a function such that $\phi \ge 0$ in $\bar{\Omega}$ and $\mathcal{A}\phi < 0$ in Ω. For example, we may use the function $\phi(x) = \mathrm{e}^{\lambda x}$ with λ so large that $\mathcal{A}\phi = (-a\lambda^2 + (b - a')\lambda)\phi < 0$ in $\bar{\Omega}$. Assume now that u attains its maximum at an interior point x_0 but not at $x = 0$ or $x = 1$. Then for $\epsilon > 0$ sufficiently small this is true also for $v = u + \epsilon\phi$. But $\mathcal{A}v = \mathcal{A}u + \epsilon\mathcal{A}\phi < 0$ in $\bar{\Omega}$, which contradicts the first part of the proof.

(ii) If $u \le 0$ in Ω, then (2.7) holds trivially. Otherwise assume that $\max_{\bar\Omega} u = u(x_0) > 0$ and $x_0 \ne 0, 1$. Let (α, β) be the largest subinterval of Ω containing x_0 in which $u > 0$. We now have $\tilde{\mathcal{A}}u := \mathcal{A}u - cu \le 0$ in (α, β). Part (i), applied with the operator $\tilde{\mathcal{A}}$ in the interval (α, β), therefore implies $u(x_0) = \max\{u(\alpha), u(\beta)\}$. But then α and β could not both be interior points of Ω, for then either $u(\alpha)$ or $u(\beta)$ would be positive, and the interval (α, β) would not be as large as possible with $u > 0$. This implies $u(x_0) = \max\{u(0), u(1)\}$ and hence (2.7). □

As a consequence of this theorem we have the following stability estimate with respect to the maximum-norm, where we use the notation of Sect. 1.2.

Theorem 2.2. *Let $\mathcal{A}$ be as in* (2.1) *and* (2.2). *If $u \in \mathcal{C}^2$, then*

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

The constant C depends on the coefficients of $\mathcal{A}$ but not on u.

Proof. We shall bound the maxima of $\pm u$. We set $\phi(x) = \mathrm{e}^{\lambda} - \mathrm{e}^{\lambda x}$ and define the two functions

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

Since $\phi \ge 0$ in Ω and $\mathcal{A}\phi = c\mathrm{e}^{\lambda} + (a\lambda^2 + (a' - b)\lambda - c)\mathrm{e}^{\lambda x} \ge 1$ in $\bar\Omega$, if $\lambda > 0$ is chosen sufficiently large, we have, with such a choice of 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) therefore yields

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

because $v_{\pm}(x) \le \pm u(x)$ for all x. Hence,

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

which completes the proof. □

From Theorem 2.2 we immediately conclude the uniqueness of a solution of (2.1). In fact, if u and v were two solutions, then their difference $w = u - v$ would satisfy $\mathcal{A}w = 0$, $w(0) = w(1) = 0$, and hence $\|w\|_{\mathcal{C}} = 0$, so that $u = v$.

More generally, if u and v are two solutions of (2.1) with right hand sides f and g and boundary values u_0, u_1 and v_0, v_1, respectively, then

$$\|u - v\|_{\mathcal{C}} \le \max\big\{|u_0 - v_0|, |u_1 - v_1|\big\} + C\|f - g\|_{\mathcal{C}}.$$

Thus the problem (2.1) is stable, i.e., a small change in data does not cause a big change in the solution.

As another application of the maximum principle we note that if all the data of the boundary value problem (2.1) are nonpositive, then the solution is nonpositive. That is, if $f \le 0$ and $u_0, u_1 \le 0$, then $u \le 0$. By means of the stronger variant of the maximum principle mentioned after Theorem 2.1, we may even conclude that $u < 0$ in Ω unless $u(x) \equiv 0$. More generally, we have the following *monotonicity property*: If

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

and if $f \le g$, $u_0 \le v_0$, and $u_1 \le v_1$, then $u \le v$.

2.2 Green's Function

We now consider the problem (2.1) with $b = 0$ and with boundary values $u_0 = u_1 = 0$. We shall derive a representation of a solution in terms of a so-called Green's function $G(x, y)$. For this purpose, let U_0 and U_1 be two solutions of the homogeneous equation such that

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

To see that such solutions exist, we note that by the standard theory of ordinary differential equations the initial value problem for $\mathcal{A}u = 0$ with $u(0) = 0$, $u'(0) = 1$ has a unique solution, and that $u(1) \neq 0$ for this solution, since otherwise $u(x) \equiv 0$ in Ω by Theorem 2.2. By multiplication of this solution by an appropriate constant we obtain the desired function U_1. The function U_0 is constructed similarly, starting at $x = 1$. By Theorem 2.1 U_0 and U_1 are nonnegative. We refer to Problem 2.5 for the case when $b \neq 0$.

Theorem 2.3. *Let $b = 0$ and let U_0, U_1 be as described above. Then a solution of (2.1) with $u_0 = u_1 = 0$ is given by*

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

where

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

and

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

Proof. We begin by showing that κ is constant: Since $(a\,U_j')' = c\,U_j$, we have

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

Setting $x = 0$ we find $\kappa = a(0)\,U_1'(0) \neq 0$, because otherwise $U_1(0) = U_1'(0) = 0$ and hence $U_1(x) \equiv 0$. Since U_1 is nonnegative we have $U_1'(0)$ nonnegative and hence it follows that $\kappa > 0$.

Clearly u as defined in (2.8) satisfies the homogeneous boundary conditions. To show that it is a solution of the differential equation we write

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

Hence, by differentiation,

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

where the terms involving $f(x)$ cancel. Multiplying by $-a(x)$ and differentiating we thus obtain, using $(a\,U_j')' = c\,U_j$ and (2.9),

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

which completes the proof. □

In particular, this theorem shows the existence of a solution of the problem considered. We already know from Sect. 2.1 that the solution is unique. The representation of the solution as an integral in terms of the Green's function can also be used to obtain additional information about the solution. As a simple example we have the maximum-norm estimate

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

which gives a more precise value of the constant in Theorem 2.2. Here we have used the fact that U_0 and U_1, and hence G, are nonnegative by Theorem 2.1.

Theorem 2.3 may also be used to show the existence of a solution for general boundary values u_0 and u_1. In fact, if $\bar{u}(x) = u_0(1-x) + u_1 x$, and if v is a solution of

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

then $u = v + \bar{u}$ satisfies $\mathcal{A}u = f$ and $u(0) = u_0,\ u(1) = u_1$.

2.3 Variational Formulation

We shall now treat our two-point boundary value problem within the framework of the Hilbert space $L_2 = L_2(\Omega)$, and derive a so-called variational formulation. We refer to App. A for the functional analytic concepts used.

We consider the boundary value problem (2.1) with homogeneous boundary conditions, i.e.,

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

We assume that the coefficients a, b, and c are smooth and, instead of (2.2), that

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

Multiplying the differential equation by a function $\varphi \in \mathcal{C}_0^1 = \mathcal{C}_0^1(\Omega)$, and integrating over the interval Ω, we obtain

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

or, after integration by parts, using $\varphi(0) = \varphi(1) = 0$,

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

which we refer to as the *variational* or *weak formulation* of (2.11).

Introducing the bilinear form

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

and the linear functional

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

and using the fact that $\mathcal{C}_0^1$ is dense in $H_0^1 = H_0^1(\Omega)$, we may write the equation (2.14) as

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

We say that u is a *weak solution* of (2.11) if $u \in H_0^1$ and (2.16) holds. Thus we do not require a weak solution to be twice differentiable. However, if a weak solution belongs to $\mathcal{C}^2$, then it is actually a classical solution of (2.11). In fact, by integration by parts in (2.14) we conclude that (2.13) holds, i.e.,

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

This immediately implies $\mathcal{A}u = f$ in Ω, and since $u \in H_0^1$ we also have $u(0) = u(1) = 0$. This calculation can also be performed if $u \in H^2 \cap H_0^1$, in which case we say that u is a *strong solution* of (2.11).

We note that, with the notation of Sect. 1.2,

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

In fact, by the Cauchy-Schwarz inequality we have for all $x \in \Omega$,

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

from which (2.17) follows by integration. This is a special case of Poincaré's inequality, which has a counterpart also for functions of several variables, see Theorem A.6. It follows at once that

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

which shows that $\|v\|_1$ and $|v|_1 = \|v'\|$ are equivalent norms.

Using our assumption (2.12), we find that

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

Hence, from (2.12) and (2.18) it follows that the bilinear form $a(v, w)$ has the property

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

The inequality (2.19) expresses that the bilinear form $a(\cdot, \cdot)$ is *coercive* in H_0^1, see (A.12). Setting $\varphi = u$ in (2.16) and using (2.19) and (2.17), we find

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

so that

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

The bilinear form $a(v,w)$ is also bounded on H_0^1 in the sense that (cf. (A.9))

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

For, estimating the coefficients in (2.15) by their maxima and using the Cauchy-Schwarz inequality, we have

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

We now turn to the question of existence of a solution of the variational equation (2.16).

Theorem 2.4. *Assume that* (2.12) *holds and let* $f \in L_2$. *Then there exists a unique solution* $u \in H_0^1$ *of* (2.16). *This solution satisfies* (2.20).

Proof. The proof is based on the Lax-Milgram lemma, Theorem A.3. We already checked that $a(\cdot,\cdot)$ is coercive and bounded in H_0^1. The linear functional $L(\cdot)$ is also bounded in H_0^1, because

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

Hence the assumptions of the Lax-Milgram lemma are satisfied and it follows that there exists a unique $u \in H_0^1$ satisfying (2.16). Together with (2.20) this completes the proof. □

We remark that when $b = 0$ the bilinear form $a(\cdot,\cdot)$ is symmetric positive definite and thus an inner product, with the associated norm equivalent to $\|\cdot\|_1$. The existence of a unique solution then follows from the more elementary Riesz representation theorem, Theorem A.1.

In the symmetric case when $b = 0$, the solution of (2.16) may also be characterized as the minimizer of a certain quadratic functional, see Theorem A.2. This is a special case of the famous Dirichlet principle.

Theorem 2.5. *Assume that* (2.2) *holds and that* $b = 0$. *Let* $f \in L_2$ *and* $u \in H_0^1$ *be the solution of* (2.16), *and set*

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

Then $F(u) \le F(\varphi)$ *for all* $\varphi \in H_0^1$, *with equality only for* $\varphi = u$.

The weak solution u of (2.16) obtained in Theorem 2.4 is actually more regular than stated there. Using our definitions one may, in fact, show that u'' exists as a weak derivative (cf. (A.21)), and that $au'' = -f + (b - a')u' + cu \in L_2$. It follows that $u \in H^2$ and that

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

Together with (2.20) this implies the *regularity estimate*

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

We conclude that the weak solution of (2.1) found in Theorem 2.4 is actually a strong solution. The proof of H^2-regularity uses the assumption that a is smooth and $f \in L_2$. With a less smooth, or with f only in H^{-1}, see (A.30), we still obtain a weak solution in H_0^1, but then it may not belong to H^2, see Problem 2.8.

2.4 Problems

Problem 2.1. Determine explicit solutions of the boundary value problem

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

where c, f, g are constants. Use this to illustrate the maximum principle.

Problem 2.2. Determine Green's functions for the following problems:

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

where c is a positive constant.

Problem 2.3. Consider the nonlinear boundary value problem

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

Use the maximum principle to show that all solutions are nonnegative, i.e., $u(x) \ge 0$ for all $x \in \bar{\Omega}$. Use the strong version of the maximum principle to show that all solutions are positive, i.e., $u(x) > 0$ for all $x \in \Omega$.

Problem 2.4. Assume that $b = 0$ as in Theorem 2.3 and let $G(x,y)$ be the Green's function defined there.

(a) Prove that G is symmetric, $G(x,y) = G(y,x)$.
(b) Prove that

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

This means that $\mathcal{A}G(x,\cdot) = \delta_x$, where δ_x is Dirac's delta at x, defined as the linear functional $\delta_x(\phi) = \phi(x)$ for all $\phi \in \mathcal{C}_0^0$, see Problem A.9.

Problem 2.5. In the unsymmetric case when $b \neq 0$, Green's function is defined in a similar way as in Theorem 2.3:

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

The main difference is that κ is no longer constant. The functions U_0 and U_1 are linearly independent, and hence it follows from the theory of ordinary differential equations that their Wronski determinant $U_0U_1' - U_0'U_1$ does not vanish. As before we may then conclude that $\kappa(x) > 0$ in $\bar{\Omega}$. Repeat the steps of the proof Theorem 2.3 in this case.

Problem 2.6. Give variational formulations and prove existence of solutions of

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

with the following boundary conditions

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

Problem 2.7. Consider the "beam equation" from Problem 1.3,

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

together with the boundary conditions

(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) $u(0) = u'(0) = u(1) = u'''(1) = 0$.

Give variational formulations and investigate existence and uniqueness of solutions of these problems. Give mechanical interpretations of the boundary conditions.

Problem 2.8. Find an explicit solution of (2.11) with $a = 1$, $b = c = 0$, and $f(x) = 1/x$. Recall from Problem A.11 that $f \in H^{-1}$ but $f \notin L_2$. Check that $u \in H_0^1$ but $u \notin H^2$. Hint: $u(x) = -x \log x$.

3 Elliptic Equations

In this chapter we study boundary value problems for elliptic partial differential equations. As we have seen in Chapt. 1 such equations are central in both theory and application of partial differential equations; they describe a large number of physical phenomena, particularly modelling stationary situations, and are stationary limits of evolution equations. After some preliminaries in Sect. 3.1 we begin by showing a maximum principle in Sect. 3.2. In the same way as for the two-point boundary value problem in Chapt. 2 this may be used to show uniqueness and continuous dependence on data for boundary value problems. In the following Sect. 3.3 we show the existence of a solution of Dirichlet's problem for Poisson's equation in a disc with homogeneous boundary conditions, using an integral representation in terms of Poisson's kernel. In Sect. 3.4 similar ideas are employed to introduce fundamental solutions of elliptic equations, and we illustrate their use by constructing a Green's function. Another important approach, presented in Sect. 3.5, is based on a variational formulation of the boundary value problem and simple functional analytic tools. In Sect. 3.6 we discuss briefly the Neumann problem, and in Sect. 3.7 we describe some regularity results.

3.1 Preliminaries

Rather than considering a general second order elliptic equation of the form (1.5) we shall restrict ourselves, for the sake of simplicity, to the special case when the matrix $A = (a_{ij})$ in (1.5) reduces to a scalar multiple $a\,I$ of the identity matrix, where a is a smooth function.

We consider first the Dirichlet problem

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

where $\Omega \subset \mathbf{R}^d$ is a domain with appropriately smooth boundary Γ, where the coefficients $a = a(x)$, $b = b(x)$, $c = c(x)$ are smooth and such that

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

and where f and g are given functions on Ω and Γ, respectively. This is the stationary case of the heat equation (1.18).

The particular case $a = 1$, $b = 0$, $c = 0$ is Poisson's equation, i.e.,

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

When $f = 0$ this equation is referred to as Laplace's equation and its solutions are called harmonic functions.

We note that if v and w are solutions of the two problems

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

then $u = v + w$ is a solution of (3.1). It is therefore sometimes convenient to consider separately the homogeneous equation with given boundary values and the inhomogeneous equation with vanishing boundary values.

One may also study the partial differential equation in (3.1) together with Robin's boundary condition

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

where the coefficient $h = h(x)$ is positive and n is the outward unit normal to Γ. The Dirichlet boundary condition used in (3.1) may be formally obtained as the extreme case $h = \infty$ of (3.4). At the other extreme, $h = 0$, we obtain Neumann's boundary condition

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

Sometimes one considers mixed boundary conditions in which, e.g., Dirichlet boundary conditions are given on one part of the boundary and Neumann conditions on the remaining part. A function $u \in \mathcal{C}^2(\bar{\Omega})$ that satisfies the differential equation and the boundary condition in (3.1) is called a *classical solution* of this boundary value problem.

3.2 A Maximum Principle

We begin our study of the Dirichlet problem (3.1) by showing a maximum principle analogous to that of Theorem 2.1.

Theorem 3.1. *Consider the differential operator $\mathcal{A}$ in* (3.1)*, and assume that $u \in \mathcal{C}^2 = \mathcal{C}^2(\bar{\Omega})$ and*

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

(i) *If $c = 0$, then*

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

(ii) *If $c \geq 0$ in Ω, then*

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

Proof. (i) Let ϕ be a function such that $\phi \geq 0$ in $\bar\Omega$ and $\mathcal{A}\phi < 0$ in Ω. Such a function is, e.g., $\phi(x) = \mathrm{e}^{\lambda x_1}$ for λ so large that $\mathcal{A}\phi = (-a\,\lambda^2 + (b_1 - \partial a/\partial x_1)\lambda)\mathrm{e}^{\lambda x_1} < 0$ in $\bar\Omega$. Assume now that u attains its maximum at an interior point x_0 in Ω but not on Γ. Then for ϵ sufficiently small this is true also for $v = u + \epsilon\phi$. But $\mathcal{A}v = \mathcal{A}u + \epsilon\mathcal{A}\phi < 0$ in $\bar\Omega$. On the other hand, if the maximum of v is $v(\bar x_0)$, then $\nabla v(\bar x_0) = 0$ and hence $\mathcal{A}v(\bar x_0) = -a(\bar x_0)\Delta v(\bar x_0) \geq 0$, which is a contradiction, and thus shows our claim.

(ii) If $u \leq 0$ in Ω, then (3.7) holds trivially. Otherwise assume that $\max_{\bar\Omega} u = u(x_0) > 0$ and $x_0 \in \Omega$. Let Ω_0 be the largest open connected subset of Ω containing x_0 in which $u > 0$. We now have $\tilde{\mathcal{A}}u := \mathcal{A}u - cu \leq 0$ in Ω_0. Part (i), applied with the operator $\tilde{\mathcal{A}}$ in Ω_0, therefore implies $u(x_0) = \max_{\Gamma_0} u$, where Γ_0 is the boundary of Ω_0. But then Γ_0 could not lie completely in the open set Ω, for then there would be a point on Γ_0 where u were positive, and Ω_0 would not be as large as possible with $u > 0$. This shows (3.7). □

Theorem 3.1 implies stability with respect to the maximum-norm.

Theorem 3.2. *Let $u \in \mathcal{C}^2(\bar\Omega)$. Then there is a constant C such that*

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

Proof. Let ϕ be a function such that $\phi \geq 0$ and $\mathcal{A}\phi \leq -1$ in Ω, e.g., a suitable multiple of the function ϕ in the proof of Theorem 3.1. We now define two functions $v_\pm(x) = \pm u(x) + \|\mathcal{A}u\|_{\mathcal{C}(\bar\Omega)}\phi(x)$. Then

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

Therefore both functions $v_\pm$ take their maxima on Γ, so that

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

Since $\pm u(x) \leq v_\pm(x)$ this proves the theorem. □

In the same way as for the two-point boundary value problem it follows that there is at most one solution of our Dirichlet problem (3.1), and that, if $u_j,\ j = 1, 2$, are solutions of (3.1) with $f = f_j,\ g = g_j,\ j = 1, 2$, then

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

3.3 Dirichlet's Problem for a Disc. Poisson's Integral

In this section we study the Dirichlet problem to find a harmonic function in a disc $\Omega = \{x \in \mathbf{R}^2 : |x| < R\}$ with given boundary values, i.e.,

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

In the following theorem a solution of (3.8) is given as an integral over the boundary of the disc.

Theorem 3.3. (Poisson's integral formula.) *Let $P_R(r,\varphi)$ denote the Poisson kernel*

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

Then, using polar coordinates $x = (r\cos\varphi, r\sin\varphi)$, the function defined by

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

is a solution of (3.8) *for g appropriately smooth,*

Proof. We first note that, for each $n \ge 0$, $v(x) = r^n \mathrm{e}^{\pm \mathrm{i}n\varphi}$ is a harmonic function. In fact, we have

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

It follows, for c_n bounded, say, that the series

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

is harmonic in Ω. We assume now that $g(\varphi)$ has a Fourier series

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

which is absolutely convergent. The function $u(x)$ in (3.10) with the coefficients c_n is a then solution of (3.8), and u is continuous in $\bar{\Omega}$. The latter means that $u(r\mathrm{e}^{\mathrm{i}\psi}) \to g(\mathrm{e}^{\mathrm{i}\varphi})$ when $r \to R$, $\psi \to \varphi$. To see that this holds, we choose N so large that $\sum_{|n|>N} |c_n| < \epsilon/3$ and write

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

Here obviously the first term on the right tends to 0 when $r \to R,\ \psi \to \varphi$, and hence becomes smaller than $\epsilon/3$, which shows our claim.

Recall that the Fourier coefficients of g are given by

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

Formally we thus have

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

which is of the form (3.9) with

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

Setting $z = (r/R)\mathrm{e}^{\mathrm{i}\varphi}$ we have

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

which completes the proof. □

One consequence of the theorem is that if u is a harmonic function in Ω, $\tilde{x}$ is any point in Ω, and if the disc $\{x : |x - \tilde{x}| \le R\}$ is contained in Ω, then

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

since $P_R(0,\varphi) = 1$. Hence $u(\tilde{x})$ is the average of the values of $u(x)$ with $|x - \tilde{x}| = R$. Thus the value of u at the center of a disc equals the average of its boundary values. We say that u satisfies the meanvalue property. This proves a special case of the strong maximum principle we have mentioned earlier: If a harmonic function u takes its maximum value at an interior point of Ω, then it is constant. In fact, if $\tilde{x}$ is an interior point of Ω where u attains its maximum, then by (3.11) $u(x) = u(\tilde{x})$ for all x with $\{x : |x - \tilde{x}| = R\} \subset \Omega$, and since R is arbitrary and Ω connected it follows easily that u takes the constant value $u(\tilde{x})$ in $\bar{\Omega}$. In particular, the maximum is also attained on Γ.

3.4 Fundamental Solutions. Green's Function

Let u be a solution of the inhomogeneous equation

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

where $\mathcal{A}$ is as in (3.1), with $b = 0$. Multiplying by $\varphi \in \mathcal{C}_0^\infty(\mathbf{R}^d)$, integrating over $\mathbf{R}^d$, and integrating by parts twice, we obtain

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

We say that U is a fundamental solution of (3.12) if U is smooth for $x \neq 0$, has a singularity at $x = 0$ such that $U \in L_1(B)$, where $B = \{x \in \mathbf{R}^d : |x| < 1\}$, and

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

and if

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

This means that, in the sense of weak derivative (see (A.21)),

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

where δ is Dirac's delta, defined in Problem A.9.

We now use the fundamental solution to construct a solution to (3.12).

Theorem 3.4. *If U is a fundamental solution of* (3.12) *and if $f \in \mathcal{C}_0^1(\mathbf{R}^d)$, then*

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

is a solution of (3.12).

Proof. We have, by (3.15),

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

Hence, if $u = U * f$, then, by changing the order of integration,

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

Since $f \in \mathcal{C}_0^1$ it follows that $u \in \mathcal{C}^2$ because with $D_i = \partial/\partial x_i$ we have $D_i D_j u(x) = (D_i U * D_j f)(x)$ (cf. App. A.3) and $D_i U \in L_1(\mathbf{R}^d)$ and $D_j f \in \mathcal{C}_0(\mathbf{R}^d)$. Thus we may integrate by parts in (3.16) to obtain (cf. (3.13))

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

from which we conclude that $\mathcal{A}u = f$. □

In the next theorem we determine fundamental solutions for Poisson's equation in two and three dimensions.

Theorem 3.5. *Let*

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

Then U is a fundamental solution for Poisson's equation (3.3).

Proof. We carry out the proof for $d = 2$; the proof for $d = 3$ is similar. By differentiation we find, for $x \neq 0$,

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

so that, in particular, $-\Delta U = 0$ for $x \neq 0$. Similarly, (3.14) holds.

Let $\varphi \in \mathcal{C}_0^\infty(\mathbf{R}^2)$. We have by Green's formula, with $n = x/|x|$,

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

Note that n points inwards here. The first term on the right side vanishes. Further, since

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

we have

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

Also,

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

Hence

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

□

We may now construct a Green's function for the boundary value problem

$$-\Delta u = f \quad \text{in } \Omega, \quad \text{with } u = 0 \quad \text{on } \Gamma, \tag{3.17}$$

namely a function $G(x, y)$ defined for $x, y \in \Omega$ such that the solution of (3.17) may be represented as

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

Let

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

where U is the fundamental solution for $-\Delta$ from Theorem 3.5 and, for fixed $y \in \Omega$, let v_y be the solution of

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

In the next section we shall show that this problem has a solution. The Green's function thus has the singularity of the fundamental solution and vanishes for $x \in \Gamma$, and it is easily seen that the function defined by (3.18) is therefore a solution of (3.17). It is also the only solution, because we have already proved uniqueness in Sect. 3.2. Note that $G(x, y)$ consists of a singular part, $U(x - y)$ with a singularity at $x = y$, and a smooth part, $v_y(x)$.

3.5 Variational Formulation of the Dirichlet Problem

We first consider the Dirichlet problem with homogeneous boundary conditions

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

where the coefficients a, b, and c are smooth functions in $\bar{\Omega}$ which satisfy

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

and where f is a given function. In the classical formulation of this problem one looks for a function $u \in \mathcal{C}^2 = \mathcal{C}^2(\bar{\Omega})$ which satisfies (3.20). In this section we shall reformulate (3.20) in variational form and seek a solution in the larger class H_0^1. In some cases it is then possible to prove such regularity for this solution that it is indeed a classical solution.

Assuming first that u is a solution in $\mathcal{C}^2$, we multiply (3.20) by $v \in \mathcal{C}_0^1$ and integrate over Ω. By Green's formula and since $v = 0$ on Γ, we find that

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

and then also, since $\mathcal{C}_0^1$ is dense in H_0^1,

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

The variational problem corresponding to (3.20) is thus to find $u\in H_0^1$ such that (3.23) holds. It will be shown below, by means of the Lax-Milgram lemma, that this problem admits a unique solution for $f\in L_2$. We say that this solution is a *weak* or *variational solution* of (3.20).

We have thus seen that a classical solution is also a weak solution. Conversely, suppose that $u\in H_0^1$ is a weak solution, i.e., u satisfies (3.23). If *in addition* we know that $u\in\mathcal{C}^2$, then by Green's formula we have from (3.23)

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

i.e.,

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

If $f\in\mathcal{C}$ we have $\mathcal{A}u - f\in\mathcal{C}$, and therefore this relation implies

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

Because $u\in H_0^1$ we also have $u=0$ on Γ, and it follows that u is a classical solution of (3.20). A weak solution which is smooth enough is thus also a classical solution. However, depending on the data f and the domain Ω, a weak solution may or may not be smooth enough to be a classical solution and the weak formulation (3.23) therefore really constitutes an extension of the classical formulation. Note that the weak formulation (3.23) is meaningful for any $f\in L_2$, so that, e.g., f may be discontinuous, while the classical formulation (3.20) requires f to be continuous. If $f\in L_2$ and $u\in H^2\cap H_0^1$ satisfies (3.20), then we say that u is a *strong solution*. Clearly, a classical solution is also a strong solution, and a strong solution is a weak solution. Further a weak solution that belongs to H^2 is a strong solution. We shall return below to the problem of the regularity of weak solutions.

We are now ready to show the existence of a weak solution. We use our standard notation from Sect. 1.2.

Theorem 3.6. *Assume that* (3.21) *holds and let $f\in L_2$. Then the boundary value problem* (3.20) *admits a unique weak solution, i.e., there exists a unique $u\in H_0^1$ which satisfies* (3.23). *Moreover, there exists a constant C independent of f such that*

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

Proof. We apply the Lax-Milgram lemma, Theorem A.3, in the Hilbert space $V = H_0^1$ equipped with the norm $|\cdot|_1$, and with

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

Clearly the bilinear form $a(\cdot,\cdot)$ is bounded in H_0^1 and it is coercive when (3.21) holds, since

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

Further $L(\cdot)$ is a bounded linear functional on H_0^1, since by Poincaré's inequality, Theorem A.6,

$$|L(v)| \leq \|f\|\,\|v\| \leq C\|f\|\,|v|_1.$$

This implies that $\|L\|_{V^*} \leq C\|f\|$ and the statement of the theorem thus follows directly from Theorem A.3. □

We observe that when $b = 0$, (3.21) reduces to (3.2), and the bilinear form $a(\cdot,\cdot)$ is an inner product on H_0^1. The theorem can then be proved by means of the Riesz representation theorem. In this case Theorem A.2 shows that the weak solution of (3.20) may also be characterized as follows:

Theorem 3.7. (Dirichlet's principle.) *Assume that* (3.2) *holds and that* $b = 0$. *Let* $f \in L_2$ *and* $u \in H_0^1$ *be the solution of* (3.23), *and set*

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

Then $F(u) \leq F(v)$ *for all* $v \in H_0^1$, *with equality only for* $v = u$.

Remark 3.1. If (3.20) is considered, e.g., to be a model of an elastic membrane fixed at its boundary, then $F(v)$ as defined by (3.26) is the *potential energy* associated with the deflection v; the first term in $F(v)$ corresponds to the *internal elastic energy* and the second term is a *load potential* (analogous interpretations can be made for other problems in mechanics and physics that are modeled by (3.20)). Dirichlet's principle in this case corresponds to the *Principle of Minimum Potential Energy* in mechanics and (3.23) to the *Principle of Virtual Work*.

We now consider the boundary value problem with inhomogeneous boundary condition,

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

where we assume that $f \in L_2$ and $g \in L_2(\Gamma)$. The weak formulation of this problem is then to find $u \in H^1$ such that, with $a(\cdot,\cdot)$ and $L(\cdot)$ as in (3.25),

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

where $\gamma : H^1 \to L_2(\Gamma)$ is the trace operator, cf. Theorem A.4. For the existence of a solution, we assume that the given function g on Γ is the trace of some function $u_0 \in H^1$, i.e., $g = \gamma u_0$. Setting $w = u - u_0$, we then seek $w \in H_0^1$ satisfying

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

The right hand side is a bounded linear functional on H_0^1 and hence it follows by the Lax-Milgram lemma that there exists a unique $w \in H_0^1$ satisfying (3.29). Clearly, $u = u_0 + w$ satisfies (3.28) and $\gamma u = g$. This solution is unique, for if (3.27) had two weak solutions u_1, u_2 with the same data f, g, then their difference $u_1 - u_2 \in H_0^1$ would be a weak solution of (3.20) with $f = 0$, and hence the stability estimate (3.24) would imply $u_1 - u_2 = 0$, i.e., $u_1 = u_2$. Hence, (3.27) has a unique weak solution. In particular, the solution u is independent of the choice of extension u_0 of the boundary values g.

When $b = 0$, the weak solution $u \in H^1$ can equivalently be characterized as the unique solution of the minimization problem

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

3.6 A Neumann Problem

We now consider the Neumann problem

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

where we now in addition to (3.2) require $c(x) \geq c_0 > 0$ in Ω, and where $f \in L_2$. (The case $c = 0$ is discussed in Problem 3.9.) For a variational formulation of (3.30) we multiply the differential equation in (3.30) by $v \in \mathcal{C}^1$ (note that we do not require v to satisfy any boundary conditions), and integrate over Ω using Green's formula, to obtain

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

so that since $\partial u / \partial n = 0$ on Γ,

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

Conversely, if $u \in \mathcal{C}^2$ satisfies (3.31), then by Green's formula we have

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

If we first let v vary only over $\mathcal{C}_0^1$, we see that u must satisfy the differential equation in (3.30). Thus, the first term on the left-hand side of (3.32) vanishes, and by varying v on Γ, we see that u also satisfies the boundary condition in (3.30).

We are thus led to the following variational formulation of (3.30): Find $u \in H^1$ such that

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

where $a(\cdot, \cdot)$ and $L(\cdot)$ are as in (3.25) with $b = 0$.

We have seen that if u is a classical solution of (3.30), then u satisfies (3.33). Conversely, if u satisfies (3.33) and in addition $u \in \mathcal{C}^2$, then u is a classical solution of (3.30).

By the Riesz representation theorem we have this time the following existence, uniqueness, and stability result. Note that since $c(x) \geq c_0 > 0$ the bilinear form $a(\cdot, \cdot)$ is an inner product on H^1.

Theorem 3.8. *If $f \in L_2$, then the Neumann problem* (3.30) *admits a unique weak solution, i.e., there is a unique function $u \in H^1$ that satisfies* (3.33)*. Moreover,*

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

Remark 3.2. Note that the Neumann boundary condition $\partial u/\partial n = 0$ on Γ is not enforced explicitly in the variational formulation (3.33); the function u is just required to belong to H^1. The boundary condition is implicitly contained in (3.33), since the test function v may be an arbitrary function in H^1. Such a boundary condition, which does not have to be enforced explicitly, is called a *natural boundary condition*. In contrast, a boundary condition, such as the Dirichlet condition $u = g$ on Γ, which is imposed explicitly as part of the variational formulation, is said to be an *essential boundary condition*.

Remark 3.3. The problem

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

where $f \in L_2(\Omega)$ and $g \in L_2(\Gamma)$ can be given the variational formulation: Find $u \in H^1$ such that

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

where $a(\cdot, \cdot)$ is as in (3.25) with $b = 0$ and

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

By the Cauchy-Schwarz inequality and the trace inequality (Theorem A.4) we have

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

and thus $L(\cdot)$ is a bounded linear form on H^1. The Riesz representation theorem therefore yields the existence and uniqueness of a function $u \in H^1$ satisfying (3.35). See also Problem 3.7.

3.7 Regularity

We have learned in Theorem 3.6 that for any $f \in L_2$ the Dirichlet problem (3.20) has unique weak solution $u \in H_0^1$. It can be proved that if Γ is smooth, or if Γ is a convex polygon, then, in fact, $u \in H^2$, and there is a constant C independent of f such that

$$\|u\|_2 \le C\|f\|.$$

Since $f = \mathcal{A}u$, this may also be expressed as

$$\|u\|_2 \le C\|\mathcal{A}u\|, \quad \forall u \in H^2 \cap H_0^1. \tag{3.36}$$

Note that, when applied with, e.g., $\mathcal{A} = -\Delta$, this inequality means that it is possible to estimate the L_2-norm of *all* second order derivatives of a function u, which vanishes on Γ, in terms of the L_2-norm of the special combination of second derivates of u given by the Laplacian $-\Delta$. We refer to Problem 3.10 for an example, with Ω neither smooth nor convex, for which the regularity estimate (3.36) does not hold.

The inequality (3.36) shows that u and its first and second order derivatives depend continuously on f in the sense that if u_1 and u_2 satisfy

$$-\mathcal{A}u_i = f_i \quad \text{in } \Omega, \quad \text{with } u_i = 0 \quad \text{on } \Gamma, \quad \text{for } i = 1, 2,$$

then

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

If Γ is smooth, then (3.36) can be generalized as follows. For any integer $k \ge 0$ there is a constant C independent of f such that if u is the weak solution of (3.20) with $f \in H^k$, then $u \in H^{k+2} \cap H_0^1$ and

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

In particular, in view of Sobolev's inequality, Theorem A.5, this implies that if $k > d/2$, then $u \in \mathcal{C}^2$ and thus u is also a classical solution of (3.20).

When Γ is a polygon the situation is not so favorable. In fact, if $\mathcal{A} = -\Delta$ and $\Omega \subset \mathbf{R}^2$ has a corner with interior angle ω, then using polar coordinates (r, φ) centered at the corner, with $\varphi = 0$ corresponding to one of the edges, one can show that the solution of (3.20) behaves as $u(r, \varphi) = c\, r^\beta \sin(\beta\varphi)$ near

the corner, with $\beta = \pi/\omega$. For such a function to have H^k-regularity near the corner, it is necessary that $(\partial/\partial r)^k u(r,\varphi) \in L_2(\Omega_0)$, where $\Omega_0 \subset \Omega$ contains a neighborhood of the corner under consideration, but no other corners. But this requires that

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

for b sufficiently small, or that $2(\beta-k)+1 \geq -1$ (note that $\beta-k+1=0$ when $2(\beta-k)+1=-1$). This in turn means that $\omega \leq \pi/(k-1)$. For $k=2$ all angles thus have to be $\leq \pi$, i.e., Ω has to be convex. For $k=3$ all angles have to be $\leq \pi/2$, which is a serious restriction. We refer to Problem 3.10 for an example that illustrates this.

3.8 Problems

Problem 3.1. Give a variational formulation and prove the existence and uniqueness of a weak solution of the Dirichlet problem

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

where $a_{jk}(x)$ and $a_0(x)$ are functions in $\mathcal{C}(\bar{\Omega})$ such that $a_0(x) \geq 0$ and the matrix $(a_{jk}(x))$ is symmetric and uniformly positive definite in Ω, so that $a_{jk}(x) = a_{kj}(x)$ and

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

Problem 3.2. Show that if u satisfies $-\Delta u = f$ in $\Omega, u = 0$ on Γ, where $f \in L_2$, then $p = \nabla u$ is the solution to the minimization problem

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

where

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

Problem 3.3. Consider two bounded domains Ω_1 and Ω_2 with a common boundary S and let $\Gamma_i = \partial\Omega_i \setminus S$, where $\partial\Omega_i$ is the boundary of $\Omega_i,\ i = 1, 2$, see Fig. 3.1.

Give a variational formulation of the following problem: Find u_i defined in $\Omega_i,\ i = 1, 2$, such that

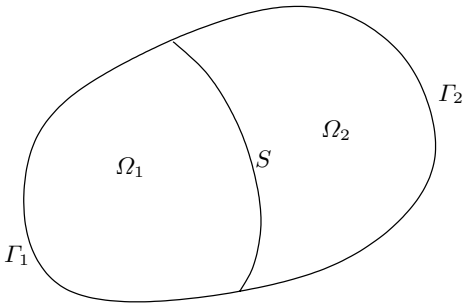

Fig. 3.1. Domain with interface.

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

and

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

where $f_i \in L_2(\Omega_i)$, $a_i > 0$ is a constant, for $i = 1, 2$, and n is a unit normal to S. Prove existence and uniqueness of a solution. Give an interpretation from physics.

Problem 3.4. Prove *Friedrichs' inequality*

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

where Ω is a bounded domain in $\mathbf{R}^d$ with boundary Γ. Hint: Integrate by parts in the identity $\int_\Omega v^2 \,\mathrm{d}x = \int_\Omega v^2\Delta\phi \,\mathrm{d}x$, where $\phi(x) = \frac{1}{2d}|x|^2$.

Problem 3.5. Prove

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

where Ω is the unit square in $\mathbf{R}^2$. The inequality holds also when Ω is a bounded domain in $\mathbf{R}^d$. Hint: $v(x) = v(y) + \int_{y_1}^{x_1} D_1 v(s, x_2)\,\mathrm{d}s + \int_{y_2}^{x_2} D_2 v(y_1, s)\,\mathrm{d}s$.

Problem 3.6. Give a variational formulation of the problem

$$-\Delta u = f \quad \text{in } \Omega, \quad \text{with } \frac{\partial u}{\partial n} + u = g \quad \text{on } \Gamma,$$

where $f \in L_2(\Omega)$ and $g \in L_2(\Gamma)$. Prove existence and uniqueness of a weak solution. Give an interpretation of the boundary condition in connection with some problem in mechanics or physics. Hint: See Problem 3.4.

Problem 3.7. Prove the stability estimate

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

for the solution of (3.34).

Problem 3.8. Give a variational formulation of the problem

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

where $f \in L_2(\Omega)$, $g, k \in L_2(\Gamma)$, and the coefficients a, c, h are smooth and such that

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

Prove existence and uniqueness of a weak solution. Prove the stability estimate

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

Hint: Use Problem 3.4.

Problem 3.9. Consider the Neumann problem

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

(a) Assume that $f \in L_2(\Omega)$ and show that the condition

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

is necessary for the existence of a solution.

(b) Notice that if u satisfies (3.38), then so does $u + c$ for any constant c. To obtain uniqueness, we add the extra condition

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

requiring the mean value of u to be zero. Give this problem a variational formulation using the space

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

Prove that there is a unique weak solution. Hint: See Problem 3.5.

(c) Show that if the weak solution $u \in V$ belongs to H^2, then it solves

$$-\Delta u = f - \int_\Omega f\,\mathrm{d}x \quad \text{in } \Omega, \quad \text{with } \frac{\partial u}{\partial n} = 0 \quad \text{on } \Gamma.$$

Problem 3.10. Let Ω be a sector with angle $\omega = \pi/\beta$:

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

where r, φ are polar coordinates in the plane. Let $v(r,\varphi) = r^\beta \sin(\beta\varphi)$. Verify that v is harmonic, i.e., $\Delta v = 0$, by computing

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

(This also follows immediately by noting that v is the imaginary part of the complex analytic function z^β.) Set $u(r,\varphi) = (1-r^2)v(r,\varphi)$. Then $u = 0$ on Γ. Show that u satisfies $-\Delta u = f$ with $f = 4(1+\beta)v$. Hence $f \in H^1(\Omega)$. Then compute $\|\partial^2 u/\partial r^2\|_{L_2(\Omega)}$ and conclude that $u \notin H^2(\Omega)$ if $\beta < 1$, i.e., if Ω is non-convex or $\omega > \pi$. Show in a similar way that $u \notin H^3(\Omega)$ if $\omega > \pi/2$. Hint: The most singular term in u_{rr} is $\beta(\beta-1)r^{\beta-2}\sin(\beta\varphi)$.

Problem 3.11. (Elliptic regularity for a rectangle.) Assume that $\Omega \subset \mathbf{R}^2$ is a rectangle and that u is a smooth function with $u = 0$ on Γ. Prove that

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

Use this to prove (3.36) for $\mathcal{A} = -\Delta$.

Hint: Recall that

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

and integrate by parts in $\int_\Omega \big(\frac{\partial^2 u}{\partial x_1 \partial x_2}\big)^2\,\mathrm{d}x$. Then recall the definition $\|u\|_2 = \big(\|u\|^2 + |u|_1^2 + |u|_2^2\big)^{1/2}$ and prove that $\|u\| \le C|u|_1$ and $|u|_1 \le \big(\|u\|\,|u|_2\big)^{1/2}$.

For arbitrary convex domains one can prove $|u|_2 \le \|\Delta u\|$ by a slightly more complicated argument based on the same idea.

Problem 3.12. Replace the boundary condition in Problem 3.11 by the Neumann condition $\partial u/\partial n = 0$ on Γ. Prove that $|u|_2 = \|\Delta u\|$.

Problem 3.13. (Stability with respect to the coefficient.) Let u_i, $i = 1, 2$, be the weak solutions of the problems

$$-\nabla \cdot \big(a_i \nabla u_i\big) = f \quad \text{in } \Omega, \quad \text{with } u_i = 0 \quad \text{on } \Gamma,$$

where $\Omega \subset \mathbf{R}^d$ is a domain with appropriately smooth boundary Γ, $f \in L_2(\Omega)$, and the coefficients $a_i(x)$ are smooth and such that

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

Prove the stability estimate

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

4 Finite Difference Methods for Elliptic Equations

The early development of numerical analysis of partial differential equations was dominated by finite difference methods. In such a method an approximate solution is sought at the points of a finite grid of points, and the approximation of the differential equation is accomplished by replacing derivatives by appropriate difference quotients. This reduces the differential equation problem to a finite linear system of algebraic equations. In this chapter we illustrate this for a two-point boundary value problem in one dimension and for the Dirichlet problem for Poisson's equation in the plane. The analysis is based on discrete versions of the maximum principles of the previous two chapters.

4.1 A Two-Point Boundary Value Problem

We consider the two-point boundary value problem

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

where the coefficients $a = a(x)$, $b = b(x)$, and $c = c(x)$ are smooth functions satisfying $a(x) > 0$ and $c(x) \geq 0$ in $\bar{\Omega}$, and where the function $f = f(x)$ and the numbers u_0, u_1 are given.

For the purpose of numerical solution of (4.1) we introduce $M+1$ mesh-points $0 = x_0 < x_1 < \cdots < x_M = 1$ by setting $x_j = jh$, $j = 0, \ldots, M$, where $h = 1/M$. We denote the approximation of $u(x_j)$ by U_j and define the following finite difference approximations of derivatives,

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

With the notation of Sect. 1.2 we have with $\mathcal{C}^j = \mathcal{C}^j(\bar{\Omega})$ (see Problem 4.1),

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

Setting also $a_j = a(x_j)$, $b_j = b(x_j)$, $c_j = c(x_j)$, $f_j = f(x_j)$, we now define a finite difference approximation of (4.1) by

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

The equation at the interior point x_j may be written

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

Our discrete problem (4.3) may thus be put in matrix form as

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

where $U = (U_1, \dots, U_{M-1})^{\mathrm{T}}$ and where the first and last components of the vector $g = (g_1, \dots, g_{M-1})^{\mathrm{T}}$ contain contributions from the boundary values u_0, u_1 as well as f_1 and f_{M-1}, respectively. The $(M-1) \times (M-1)$ matrix A is tridiagonal and diagonally dominant for h sufficiently small, i.e., the sum of the absolute values of the off-diagonal elements in one row is bounded by the diagonal element in that row, see Problem 4.2.

For our analysis we first show a discrete maximum principle similar to that in the continuous case, cf. Theorem 2.1.

Lemma 4.1. *Assume that h is so small that $a_j \pm \frac{1}{2} h b_j \geq 0$ and that U satisfies $\mathcal{A}_h U_j \leq 0$ ($\mathcal{A}_h U_j \geq 0$).*
(i) *If $c = 0$, then*

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

(ii) *If $c \geq 0$, then*

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

Proof. (i) In view of (4.4) we have, since $c = 0$ and $\mathcal{A}_h U_j \leq 0$,

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

Assume now that U has an interior maximum U_j. Then if either $U_{j+1} < U_j$ or $U_{j-1} < U_j$ this would contradict (4.6), since the coefficients on the right are nonnegative and add up to 1. Hence, $U_j = U_{j-1} = U_{j+1}$ and the latter values are also maxima. Continuing in this way we conclude that if the maximum is attained in the interior, then U is constant, and the maximum is thus also attained at the endpoints. This proves (i). Case (ii) is treated in the same way as case (ii) of Theorem 2.1. The versions with minimum are shown by considering $-U_j$. □

In the same way as for the continuous problem the maximum principle leads to a stability estimate in the discrete maximum-norm, as we shall now demonstrate. We assume for simplicity that $b = 0$. In this chapter we shall write for mesh-functions,

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

Lemma 4.2. *Let $\mathcal{A}_h$ be as in* (4.3)*, with $b = 0$. Then we have, for any mesh-function U,*

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

The constant C depends on the coefficients of $\mathcal{A}$ but not on h or U.

Proof. Let $w(x) = x - x^2 = \frac{1}{4} - (x - \frac{1}{2})^2$ and $W_j = w(x_j)$. Then, with $\underline{a} = \min_{\bar{\Omega}} a(x)$,

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

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

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

so that we may apply Lemma 4.1 (note that the condition on h required is automatically satisfied when $b = 0$). Since $W_0 = W_M = 0$ we obtain

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

Since $W_j \le \frac{1}{4}$ this shows the lemma with $C = (8\underline{a})^{-1}$, □

Lemma 4.2 immediately shows the existence and uniqueness of the solution of (4.3) when $b = 0$. For uniqueness it suffices to note that if $\mathcal{A}_h U = 0$ and $U_0 = U_M = 0$, then $U = 0$, and the uniqueness implies the existence of a solution, since we are in a finite dimensional situation. For the case when $b \neq 0$ we refer to Problem 4.3.

We are now ready for an error estimate, which again for simplicity we demonstrate for $b = 0$ only.

Theorem 4.1. *Let $b = 0$, and let U and u be the solutions of* (4.3) *and* (4.1)*. Then*

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

Proof. We have for the error $z_j = U_j - u(x_j)$ at the interior mesh-points

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

By (4.2) we have for the *truncation error*

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

so that the result follows by Lemma 4.2, since $z_0 = z_M = 0$. □

4.2 Poisson's Equation

We consider the Dirichlet problem for Poisson's equation,

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

where Ω is a domain in $\mathbf{R}^2$ with boundary Γ. To begin with we assume that Ω is a square, $\Omega = (0,1) \times (0,1) = \{x = (x_1, x_2),\ 0 < x_l < 1,\ l = 1,2\}$.

To define a finite difference approximation we write $j = (j_1, j_2)$, where j_1, j_2 are integers and consider the mesh-points $x_j = jh$, with $h = 1/M$ the mesh-width, and mesh-functions U, with $U_j = U(x_j)$. With $e_1 = (1,0),\ e_2 = (0,1)$ we use the difference quotients

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

Setting $f_j = f(x_j)$ we then replace (4.9) by

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

The difference equation at the interior mesh-points in Ω may be written

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

which is the famous 5-point approximation of Poisson's equation. The problem (4.11) may thus be written in matrix form as $AU = g$, where A is a symmetric $(M-1)^2 \times (M-1)^2$ matrix whose elements are 4, -1, or 0, with 0 the most common occurrence, and $\bar{U}$ the vector of interior nodal values.

We have the following discrete maximum principle.

Lemma 4.3. *If U is such that $-\Delta_h U_j \le 0$ ($-\Delta_h U_j \ge 0$) for $x_j \in \Omega$, then U attains its maximum (minimum) for some $x_j \in \Gamma$.*

Proof. We may write, at the interior mesh-points,

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

so that $-\Delta_h U_j \le 0$ implies $U_j \le \frac{1}{4}(U_{j+e_1} + U_{j-e_1} + U_{j+e_2} + U_{j-e_2})$. If U_j is an interior maximum, then $U_j \ge \frac{1}{4}(U_{j+e_1} + U_{j-e_1} + U_{j+e_2} + U_{j-e_2})$. Therefore equality holds, and the maximum value is taken also at all the neighboring points $x_{j \pm e_l}$, which are therefore also maximum points. Continuing in the same way we conclude that if the maximum is attained in the interior, then U is constant. This proves the lemma. □

As before the maximum principle implies a stability estimate. Using again the notation (4.7) we have the following.

Lemma 4.4. *With Δ_h defined in* (4.11) *we have, for any mesh-function U,*

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

Proof. The proof is analogous to that of Theorem 3.2. With $\bar x = (\frac12, \frac12)$ and $x = (x_1, x_2)$ we set $w(x) = \frac12 - |x - \bar x|^2 = x_1 + x_2 - x_1^2 - x_2^2$ and define the mesh-function $W_j = w(x_j)$. Then $W_j \ge 0$ in Ω and $-\Delta_h W_j = 4$. Setting $V_j^\pm = \pm U_j - \frac14 |\Delta_h U|_\Omega W_j$ we conclude that

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

and, since $W_j \ge 0$ for $x_j \in \Gamma$, it follows from Lemma 4.3 that $V_j^\pm \le |U|_\Gamma$. Since $W_j \le \frac12$ in Ω this implies our statement with $C = 1/8$. □

In particular, Lemma 4.4 implies the uniqueness of the solution of (4.11), and hence also the existence of a solution. In the same way as for the two-point boundary value problem the lemma also implies an error estimate.

Theorem 4.2. *Let U and u be the solutions of* (4.11) *and* (4.9). *Then*

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

Proof. The error $z_j = U_j - u(x_j)$ satisfies, at the interior mesh-points,

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

where τ is the truncation error, which may easily be estimated as in (4.2) by

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

The result therefore follows by application of Lemma 4.4 to z_j, since $z_j = 0$ for $x_j \in \Gamma$. □

The above analysis uses the fact that all the neighbors of the interior mesh-points in Ω are either interior mesh-points or belong to Γ. In the case of a curved boundary this cannot be achieved. We shall briefly discuss such a situation.

We assume for simplicity that Ω is a convex plane domain with a smooth boundary Γ. We denote by Ω_h those interior mesh-points x_j for which all four neighbors of x_j are also in $\bar\Omega$. (For the above case of a square, Ω_h simply consists of all interior mesh-points.) Let now ω_h be the mesh-points in Ω that are not in Ω_h. For each $x_j \in \omega_h$ we may then select a (not necessarily unique) neighbor $x_i \in \Omega_h \cup \omega_h$ such that the horizontal or vertical line through x_j

and x_i cuts Γ at a point $\bar{x}_j$, which is not a mesh-point (see Fig. 4.1). For this $x_j \in \omega_h$, we then define the linear interpolation operator

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

Denoting by Γ_h the points of Γ which are either neighbors of points in Ω_h or points $\bar{x}_j$ associated with points in ω_h, we now pose the problem

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

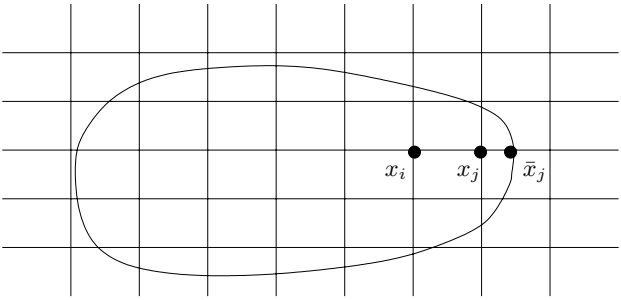

Fig. 4.1. Interpolation near the boundary.

This time we have the following stability estimate.

Lemma 4.5. *With Δ_h defined in* (4.11) *and ℓ_h in* (4.14)*, we have, for any mesh-function U,*

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

Proof. Similarly to the proof of Lemma 4.4 we obtain

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

Here, for $x_j \in \omega_h$, we have

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

and hence

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

Together these estimates show

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

which completes the proof. □

Again this shows uniqueness and existence of a solution of (4.15). Note that in this case the corresponding matrix A is nonsymmetric as, for instance, the elements a_{ij} and a_{ji} corresponding to the points x_i and x_j in Fig. 4.1 are different. We conclude with the following error estimate.

Theorem 4.3. *Let U and u be the solutions of* (4.15) *and* (4.9). *Then*

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

Proof. As in the proof of Theorem 4.2 we consider $z_j = U_j - u(x_j)$, and now apply Lemma 4.5. The only new term is $\ell_h z_j = -\ell_h u(x_j)$ and hence $|\ell_h z_j| \leq Ch^2 |u|_{\mathcal{C}^2}$, which completes the proof. □

The above method of interpolation near the boundary is attributed to L. Collatz. It is also possible to use a five point finite difference approximation of $-\Delta$ based on nonuniform spacing on ω_h, the so-called Shortley-Weller approximation. This also yields an $O(h^2)$ error estimate.

4.3 Problems

Problem 4.1. Prove (4.2) and (4.13) by means of Taylor's formula.

Problem 4.2. Derive (4.5) and show that the matrix A is tridiagonal and (row-wise) diagonally dominant, i.e., $\sum_{j \neq i} |a_{ij}| \leq a_{ii}$, if h is sufficiently small. Hint: assume $a_j \pm \frac{1}{2} h b_j \geq 0$.

Problem 4.3. Show that the conclusion of Lemma 4.2 (and hence that of Theorem 4.1) holds also when $b \neq 0$, if h is sufficiently small and if we have at our disposal a mesh-function W such that $\mathcal{A}_h W_j \geq 1$ for $x_j \in \Omega$ and $W_j \geq 0$ for $x_j \in \bar{\Omega}$. Construct such a function. (Hint: use the function $w(x) = \mathrm{e}^{\lambda} - \mathrm{e}^{\lambda x}$ with λ suitably chosen.)

Problem 4.4. (Computer exercise.) Consider the two-point boundary value problem

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

Apply the finite difference method (4.3) with $h = 1/10,\ 1/20$. Find the exact solution and compute the maximum of the error at the mesh-points.

Problem 4.5. (Computer exercise.) Consider the Dirichlet problem (4.9) with

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

in $\Omega = (0,1) \times (0,1)$. Compute the approximate solution by the finite difference method (4.11) with $h = 1/10,\ 1/20$, and find the error at (0.5, 0.5), using that the exact solution is

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

5 Finite Element Methods for Elliptic Equations

Over the last decades the *finite element method*, which was introduced by engineers in the 1960s, has become the perhaps most important numerical method for partial differential equations, particularly for equations of elliptic and parabolic types. This method is based on the variational form of the boundary value problem and approximates the exact solution by a piecewise polynomial function. It is more easily adapted to the geometry of the underlying domain than the finite difference method, and for symmetric positive definite elliptic problems it reduces to a finite linear system with a symmetric positive definite matrix.

We first introduce this method in Sect. 5.1 for the case of a two-point boundary value problem and show a number of error estimates. In Sect. 5.2 we then formulate the method for a two-dimensional model problem. Here the piecewise polynomial approximations are defined on triangulations of the spatial domain, and in the following Sect. 5.3 we study such approximation in more detail. In Sect. 5.4 we show basic error estimates for the finite element method for the model problem, using piecewise linear approximating functions. All error bounds derived up to this point contain a norm of the unknown exact solution and are therefore often referred to as *a priori* error estimates. In Sect. 5.5 we show a so-called *a posteriori* error estimate in which the error bound is expressed in terms of the data of the problem and the computed solution. In Sect. 5.6 we analyze the effect of numerical integration, which is often used when the finite element equations are assembled in a computer program. In Sect. 5.7 we briefly describe a so-called *mixed finite element method*.

5.1 A Two-Point Boundary Value Problem

We consider the special case $b = 0$ of the two-point boundary value problem treated in Sect. 2.3,

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

where $a = a(x),\ c = c(x)$ are smooth functions with $a(x) \geq a_0 > 0,\ c(x) \geq 0$ in $\bar{\Omega}$, and $f \in L_2 = L_2(\Omega)$. We recall that the variational formulation of this problem is to find $u \in H_0^1$ such that

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

where

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

and that this problem has a unique solution $u \in H^2$.

For the purpose of finding an approximate solution of (5.2) we introduce a partition of Ω,

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

and set

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

The discrete solution will be sought in the finite-dimensional space of functions

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

(By a linear function we understand a function of the form $f(x) = \alpha x + \beta$; strictly speaking such a function is called an affine function when $\beta \neq 0$.) It is easy to see that $S_h \subset H_0^1$. The set $\{\Phi_i\}_{i=1}^{M-1} \subset S_h$ of *hat functions* defined by

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

see Fig. 5.1, is a basis for S_h, and any $v \in S_h$ may be written as

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

We now pose the finite-dimensional problem to find $u_h \in S_h$ such that

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

In terms of the basis $\{\Phi_i\}_{i=1}^{M-1}$ we write $u_h(x) = \sum_{j=1}^{M-1} U_j \Phi_j(x)$ and insert this into (5.3) to find that this equation is equivalent to

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

This linear system of equations may be expressed in matrix form as

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

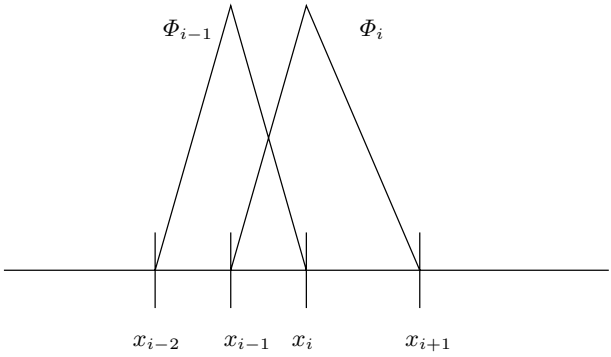

Fig. 5.1. Hat functions.

where $U = (U_i)$, $A = (a_{ij})$ is the *stiffness matrix* with elements $a_{ij} = a(\Phi_j, \Phi_i)$, and $b = (b_i)$ the *load vector* with elements $b_i = (f, \Phi_i)$. The matrix A is symmetric and positive definite, because for $V = (V_i)$ and $v(x) = \sum_{i=1}^{M-1} V_i \Phi_i(x)$ we have

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

and hence $V^{\mathrm{T}} A V = 0$ implies $v' = 0$, so that v is constant $= 0$ because $v(0) = 0$, and thus $V = 0$. It follows that (5.5), and therefore also (5.3), has a unique solution, which is the *finite element solution* of (5.1). The matrix A is tridiagonal since $a_{ij} = 0$ when x_i and x_j are not neighbors, i.e., when $|i - j| \geq 2$, and the system (5.5) is therefore easy to solve, see App. B.1.

We note that when $\mathcal{A}u = -u''$ and the meshsize is constant, i.e., when $h_j = h = 1/M$ for $j = 1, \ldots, M$, then, with the notation of Sect. 4.1, the equation (5.4) may be written

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

(cf. Problem 5.2). The finite element method thus coincides with the finite difference equation (4.3), except that an average of f over $(x_j - h, x_j + h)$ is now used instead of the point-values $f_j = f(x_j)$.

The idea of replacing the space H_0^1 in (5.2) by a finite-dimensional subspace and to determine the coefficients of the corresponding approximate solution as in (5.4) is referred to as Galerkin's method. The finite element method is thus Galerkin's method, applied with a special choice of the finite-dimensional subspace, namely, in this case, the space of continuous, piecewise linear functions. The intervals K_j, together with the restriction of these functions to K_j, are then thought of as the finite elements.

Before we analyze the error in the finite element solution u_h, we discuss some approximation properties of the space S_h. We define the piecewise linear interpolant $I_h v \in S_h$ of a function $v \in \mathcal{C} = \mathcal{C}(\bar{\Omega})$ with $v(0) = v(1) = 0$ by

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

Recall that $H_0^1 \subset \mathcal{C}$ in one dimension by Sobolev's inequality, Theorem A.5, so that $I_h v$ is defined for $v \in H_0^1$. It may be shown, which we leave as an exercise, see Problem 5.1, that, with $\|v\|_{K_j} = \|v\|_{L_2(K_j)}$ and $|v|_{2,K_j} = |v|_{H^2(K_j)}$,

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

and

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

It follows that

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

and similarly

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

We now turn to the task of estimating the error in the finite element approximation u_h defined by (5.3). Since $a(\cdot, \cdot)$ is symmetric positive definite, it is an inner product on H_0^1, and the corresponding norm is the energy norm

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

Theorem 5.1. *Let u_h and u be the solutions of* (5.3) *and* (5.2). *Then*

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

and

$$\|u_h' - u'\| \leq C h \|u\|_2. \tag{5.13}$$

Proof. Since $S_h \subset H_0^1$ we may take $\varphi = \chi \in S_h$ in (5.2) and subtract it from (5.3) to obtain

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

This equation means that the finite element solution u_h may be described as the orthogonal projection of the exact solution u onto S_h with respect to

the inner product $a(\cdot,\cdot)$. This also immediately implies that u_h is the best approximation of u in S_h with respect to the energy norm, and hence that (5.12) holds. This can be seen directly as follows: Using (5.14) we have, for any $\chi \in S_h$,

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

which shows (5.12) after cancellation of a factor $\|u_h - u\|_a$. By our assumptions we have, with C independent of h,

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

where the first inequality is obvious by (5.11) and the second follows from (2.17). Hence, (5.12) implies

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

Taking $\chi = I_h u$ and using the interpolation error bound in (5.10), we obtain (5.13), and the proof is complete. □

Our next result concerns the L_2-norm of the error.

Theorem 5.2. *Let u_h and u be the solutions of* (5.3) *and* (5.2). *Then*

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

Proof. We use a duality argument based on the auxiliary problem

$$\mathcal{A}\phi = e \quad \text{in } \Omega, \quad \text{with } \phi(0) = \phi(1) = 0, \quad \text{where } e = u_h - u. \tag{5.17}$$

Its weak formulation is to find $\phi \in H_0^1$ such that

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

We put the test function w on the left side, because (5.18) plays the role of the adjoint (or dual) problem to (5.2). Of course, this makes no difference here since $a(\cdot,\cdot)$ is symmetric, but is important in the case of a nonsymmetric differential operator $\mathcal{A}$, see Problem 5.7. By the regularity estimate (2.22) we have

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

Taking $w = e$ in (5.18) and using (5.14) and (5.10), we therefore obtain

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

Cancelling one factor $\|e\|$ we see that we have gained one factor h over the error estimate for e',

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

and the proof may now be completed by using (5.13). □

Remark 5.1. We note that the above error estimates contain the norm of the second order derivative, while the fourth order derivative was needed in the corresponding result for the finite difference method in Theorem 4.1. This is related to the fact that in the finite element method the load term f enters via averages rather than through point-values as in the finite difference scheme. This will be discussed further in Sect. 5.6 below.

Remark 5.2. The solution of the very special equation (5.6) agrees with the nodal values of the exact solution of the corresponding two-point boundary value problem. In fact, with $u = u(x)$ the exact solution, we have by Taylor's formula

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

Thus the finite element solution u_h is identical to the interpolant $I_h u$ of the exact solution. For a discussion of this based on the Green's function, see Problem 5.4.

In the above analysis we could have considered a more general finite element space, consisting of piecewise polynomials of degree $r-1$, where r is an integer ≥ 2, with the above piecewise linear case included for $r = 2$, thus with

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

where Π_k denotes the space of polynomials of degree $\leq k$. In addition to the hat functions above we may then associate with each interval K_i the basis functions $\Phi_{ij} \in \Pi_{r-1}$ on K_i for $j = 1, \ldots, r-2$, and vanishing outside K_i, defined by

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

Using also these additional nodal points in the definition of the interpolant $I_h v$ one may show the local estimates

$$\|I_h v - v\|_{K_j} \leq Ch_j^r\|v^{(r)}\|_{K_j} \quad \text{and} \quad \|(I_h v - v)'\|_{K_j} \leq Ch_j^{r-1}\|v^{(r)}\|_{K_j},$$

and consequently the global estimates

$$\|I_h v - v\| \leq Ch^r\|v\|_r \quad \text{and} \quad \|(I_h v - v)'\| \leq Ch^{r-1}\|v\|_r, \quad \forall v \in H^r. \tag{5.21}$$

For the finite element solution one then obtains, in the same way as above,

$$\|u_h - u\| \leq Ch^r\|u\|_r \quad \text{and} \quad \|u_h' - u'\| \leq Ch^{r-1}\|u\|_r. \tag{5.22}$$

These inequalities thus require $v, u \in H^r$. Using that the interpolant $I_h v$ is well defined for $v \in H_0^1$, one can show that they also hold with r replaced by any s with $1 \leq s \leq r$. The case $s = 2$ of the second estimate in (5.21) is needed in the proof of the $O(h^r)$ estimate in (5.22) by duality.

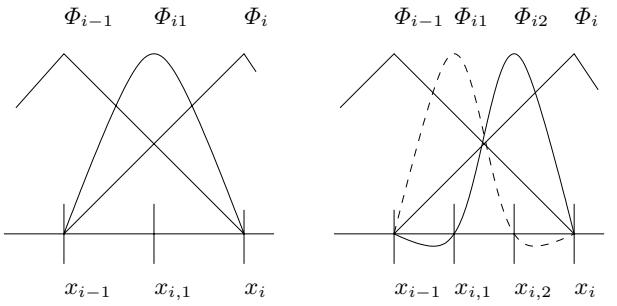

Fig. 5.2. Global basis functions for $r = 3$ and 4.

5.2 A Model Problem in the Plane

Let now Ω be a polygonal domain in $\mathbf{R}^2$, i.e., a domain whose boundary Γ is a polygon, and consider the simple model problem

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

We assume that the coefficient $a = a(x)$ is smooth with $a(x) \geq a_0 > 0$ in $\bar{\Omega}$ and that $f \in L_2$.

We recall from Sect. 3.5 that the variational formulation of (5.23) is to find $u \in H_0^1$ such that

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

where

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

and that this problem has a unique solution in H_0^1. Moreover, if Ω is assumed to be convex, then the regularity estimate in (3.36) implies that $u \in H^2$ and

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

Our discussion of the approximation of (5.23) follows similar lines as for the two-point boundary value problem above. This time we divide the polygonal domain Ω into triangles. More precisely, let $\mathcal{T}_h = \{K\}$ be a set of closed triangles K, a *triangulation* of Ω, such that

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

The vertices P of the triangles $K \in \mathcal{T}_h$ are called the nodes of the triangulation $\mathcal{T}_h$. We require that the intersection of any two triangles of $\mathcal{T}_h$ is either empty, a node, or a common edge, and that no node is located in the interior of an edge of $\mathcal{T}_h$, see Fig. 5.3.

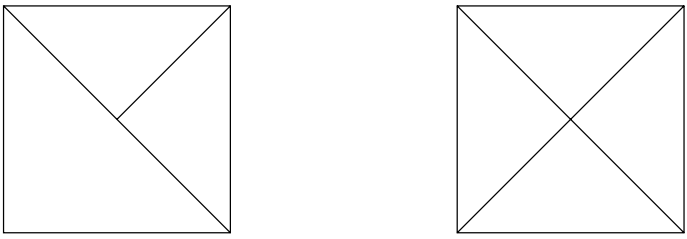

Fig. 5.3. Invalid (left) and valid (right) triangulation.

With the triangulation $\mathcal{T}_h$ we associate the function space S_h consisting of continuous, piecewise linear functions on $\mathcal{T}_h$, vanishing on Γ, i.e.,

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

Using our above assumptions on $\mathcal{T}_h$ it is not difficult to verify that $S_h \subset H_0^1$. Let $\{P_i\}_{i=1}^{M_h}$ be the set of interior nodes, i.e., those that do not lie on Γ. A function in S_h is then uniquely determined by its values at the P_j, and the set of *pyramid functions* $\{\Phi_i\}_{i=1}^{M_h} \subset S_h$, defined by

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

forms a basis for S_h. If $v \in S_h$ we thus have $v(x) = \sum_{i=1}^{M_h} v_i \Phi_i(x)$, where the $v_i = v(P_i)$ are the nodal values of v. It follows that S_h is a finite-dimensional subspace of the Hilbert space H_0^1.

The finite element approximation of the problem (5.24) is then to find $u_h \in S_h$ such that

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

Using the basis $\{\Phi_i\}_{i=1}^{M_h}$ we write $u_h(x) = \sum_{i=1}^{M_h} U_i \Phi_i(x)$, which, inserted into (5.26), gives a linear system of equations for the determination of the U_j,

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

This may be written in matrix form as $AU = b$, where $U = (U_i)$, $A = (a_{ij})$ is the stiffness matrix with elements $a_{ij} = a(\Phi_j, \Phi_i)$, and $b = (b_i)$ the load vector with elements $b_i = (f, \Phi_i)$. The matrix A is symmetric and positive definite as in Sect. 5.1, so that (5.27) and hence (5.26) has a unique solution in S_h. Moreover, the matrix A is large and sparse if the mesh is fine, i.e., a large portion of its elements are zero because each Φ_i vanishes except in the union of the triangles that contain the node P_i, so that $a_{ij} = a(\Phi_j, \Phi_i) = 0$

unless P_i and P_j are neighbors. This property is important for the efficient solution of the linear system, cf. App. B. This time the finite elements are the triangles $K \in \mathcal{T}_h$ together with the restrictions to the K of the functions in S_h.

More generally, given the triangulation $\mathcal{T}_h$ we can take S_h to be the functions on Ω, which reduce to polynomials of degree $r-1$ on the triangles $K \in \mathcal{T}_h$, where r is a fixed integer ≥ 2. It may be shown that such a function χ is uniquely determined by its values at a certain finite number of nodes in each K, which can be chosen in different ways. In the case $r = 3$, i.e., when S_h consists of piecewise quadratic functions, these points may be taken to be the vertices of $\mathcal{T}_h$ together with the midpoints of the edges in $\mathcal{T}_h$, altogether 6 points for each $K \in \mathcal{T}_h$. For piecewise cubics, i.e., when $r = 4$, we may take the vertices of $\mathcal{T}_h$, two interior points on each edge of $\mathcal{T}_h$, and the barycenter of each $K \in \mathcal{T}_h$, thus using 10 points for each $K \in \mathcal{T}_h$, see Fig. 5.4. Note that a polynomial in two variables of second and third degree is determined uniquely by the values of 6 and 10 coefficients, respectively, and to determine these we need to require this number of linear conditions, or degrees of freedom, as they are referred to in this context.

The finite element space S_h thus defined is still a finite dimensional subspace of H_0^1, and one basis function $\Phi_j \in S_h$ may be associated with each of the nodes described. The finite element problem (5.26) and its matrix formulation (5.5) remain of the same form as before.

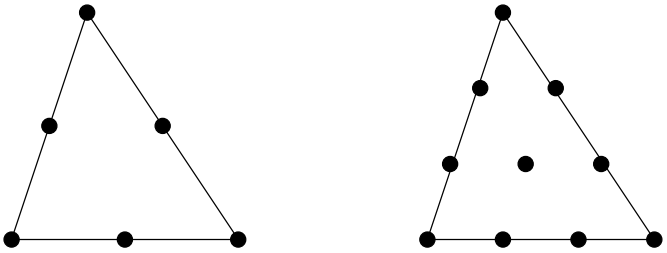

Fig. 5.4. Triangles with 6 and 10 nodes.

If the boundary Γ of Ω is not a polygon but a smooth curve, then a triangulation of the above type will not fit Ω exactly. If Ω is convex it is possible to choose the triangulation in such a way that the union Ω_h of the triangles still approximates Ω, by choosing the boundary vertices of Ω_h on Γ, so that the set $\Omega \setminus \Omega_h$ of points in Ω not covered by the triangulation has a width of order $O(h^2)$, see Fig. 5.5. Defining the functions in S_h to vanish on $\Omega \setminus \Omega_h$, a finite element solution u_h can be defined as above. It turns out that for S_h consisting of piecewise linear functions nothing is lost by this extension, see Sect. 5.3 below, but for piecewise polynomials of higher degree the situation is not so favorable. Various modifications of the methods

have then been devised to deal with the approximation near Γ. We shall not go into details but remark that at any rate a triangulation provides a more flexible way of approximating a domain Ω than is possible with the square mesh used in the finite difference method, and that this is a useful property of the finite element method.

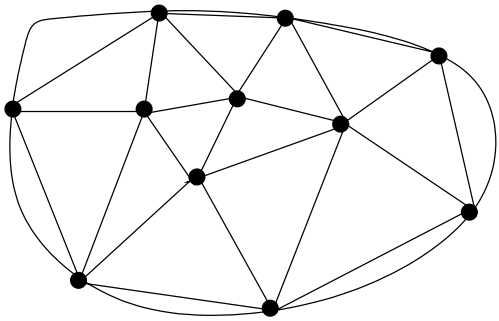

Fig. 5.5. Smooth convex domain with triangulation.

In the sequel we take the point of view that we consider not only one triangulation $\mathcal{T}_h$ and associated function space S_h, but a whole family of triangulations $\{\mathcal{T}_h\}_{0<h<1}$ and associated finite element spaces $\{S_h\}_{0<h<1}$. One important task is to estimate how fast the error $u_h - u$ tends to zero as h tends to zero.

5.3 Some Facts from Approximation Theory

Let $\tilde{S}_h$ denote the continuous piecewise linear functions on the triangulation $\mathcal{T}_h$, where the functions are not required to vanish on Γ. With $\{P_j\}_{j=1}^{N_h}$ denoting all the nodes of $\mathcal{T}_h$, including those on Γ for $M_h + 1 \leq j \leq N_h$, and $\{\Phi_j\}_{j=1}^{N_h}$ the corresponding pyramid functions, we define the interpolation operator $I_h : \mathcal{C}(\bar{\Omega}) \to \tilde{S}_h$ by

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

The interpolant $I_h v$ thus agrees with v at the nodes P_j, or

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

and if v vanishes on Γ, then $I_h v$ belongs to the finite element space S_h introduced in the previous section. An analogous definition may be used also

in the more general case of piecewise polynomials of degree $r-1$ described above.

In the piecewise linear case one can prove the following local error estimates, with $|v|_K = \|v\|_{L_2(K)}$, $|v|_{2,K} = \|v\|_{H^2(K)}$,

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

and

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

The proofs are based on the Bramble-Hilbert lemma and are left as an exercise, see Problem 5.12.

In what follows we impose the restriction on the family $\{\mathcal{T}_h\}_{0<h<1}$ of triangulations that the angles of all triangles K belonging to all members of the family $\{\mathcal{T}_h\}$ are bounded below, independently of h. It is then possible to prove that the constants C_K are uniformly bounded, so that we have the global estimates

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

and similarly

$$|I_h v - v|_1 \le Ch\|v\|_2, \quad \forall v \in H^2. \tag{5.32}$$

For $\tilde{S}_h$ consisting instead of piecewise polynomials of degree $r-1$ the corresponding results may be expressed locally by

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

and globally as

$$\|I_h v - v\| \le Ch^r \|v\|_r, \quad |I_h v - v|_1 \le Ch^{r-1}\|v\|_r, \quad \text{for } v \in H^r. \tag{5.34}$$

We note that since we simply bound h_K by h in (5.31) and (5.34), these bounds are crude if the meshsize h_K varies significantly over the triangulation. For instance, if we refine the triangulation by subdividing some triangles K, then the sum over K in (5.31) becomes smaller, but the global bound does not change if $h = \max_K h_K$ remains the same.

We observe that the interpolant $I_h v$ is well defined only for continuous functions v, because it uses the values of v at the nodes. Since functions in H^2 are continuous by Sobolev's inequality (Theorem A.5), $I_h v$ is thus defined for $v \in H^r$ with $r \ge 2$, but a function in H^1 does not have to be continuous and therefore the point-values are not well defined. If v is not smooth enough

to belong to H^r but $v \in H^s$ for some s with $2 \le s \le r$ then instead of (5.34) one may use the estimates

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

for $2 \le s \le r$. Thus the order of approximation of $I_h v$ depends on the regularity of the function v.

Consider now the case when Ω is convex and the boundary Γ is a smooth curve rather than a polygon. Let Ω_h be the polygonal domain covered by the triangles of $\mathcal{T}_h$ as described at the end of the previous section, and recall that the set $\Omega \setminus \Omega_h$ has a width of order $O(h^2)$. If $v = 0$ on Γ, then the interpolation error in $\Omega \setminus \Omega_h$ equals v, since $I_h v = 0$ there. For smooth functions v vanishing on Γ we have $v = O(h^2)$ in $\Omega \setminus \Omega_h$ and hence its contribution to the interpolation error is also of this order, more precisely

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

To show the latter inequality we integrate the trace inequality $\|w\|^2_{L_2(\gamma)} \le C\|w\|^2_{H^1(\Omega)}$ over a family of curves γ parallel to Γ and covering $\Omega \setminus \Omega_h$. Since the width of $\Omega \setminus \Omega_h$ is $O(h^2)$ this yields $\|w\|_{\Omega\setminus\Omega_h} \le Ch\|w\|_1$, which is then applied to $w = \nabla v$.

The gradient of v does not vanish on Γ and thus does not have to be small in $\Omega \setminus \Omega_h$, and one may therefore only show

$$\|\nabla(I_h v - v)\|_{\Omega\setminus\Omega_h} = \|\nabla v\|_{\Omega\setminus\Omega_h} \le Ch\|v\|_2. \tag{5.37}$$

Thus, for $r = 2$, including the contributions from $\Omega \setminus \Omega_h$ to the interpolation error, (5.31) and (5.32) remain valid. However, if $r > 2$, the contributions in (5.36) and (5.37) are the best one can expect, and therefore the first inequality in (5.34) holds with $r = 2$ and 3, but the second only for $r = 2$.

We close with a remark about the orthogonal projection $P_h = P_{S_h}$ of the Hilbert space L_2 onto the finite-dimensional subspace S_h, which is defined by

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

Since $P_h v$ is the best approximation of v in S_h with respect to the L_2-norm, and hence by the above, in the case of a polygonal domain,

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

Here we use the notation $H^r \cap H_0^1$ for the space of functions that belong to H^r and vanish on Γ. The requirement that $v \in H^r \cap H_0^1$ is a rather strong one and not normally satisfied for solutions of our elliptic problem when $r > 2$ because of the singularities at the corners of the domain, which we discussed at the end of Sect. 3.7. For a convex domain with smooth boundary the regularity is not a problem, but without further modification of the method near the boundary, we only know (5.39) to hold for $r = 2$ and 3.

5.4 Error Estimates

We return to the task of estimating the error in the finite element approximation u_h of the solution u of our Dirichlet problem. Since the bilinear form $a(\cdot,\cdot)$ is an inner product in H_0^1 it is natural to use the energy norm

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

Theorem 5.3. *Let u_h and u be the solutions of* (5.26) *and* (5.24)*. Then*

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

and

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

Proof. Since $S_h \subset H_0^1$ we may take $v = \chi \in S_h$ in (5.24) and subtract it from (5.26) to obtain

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

which means that u_h is the orthogonal projection of u onto S_h with respect to the inner product $a(\cdot,\cdot)$. The equality (5.40) hence follows in the same way as (5.12). In view of our assumptions on a we have, with C and c independent of h,

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

Hence, (5.40) implies

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

Taking $\chi = I_h u$ and using the interpolation error bound in (5.32), this proves (5.41). □

For the analogous result in the case of a nonsymmetric elliptic operator, see Problems 5.6 and 5.7.

The equality (5.40) means that u_h is the best, or optimal, approximation of u in S_h with respect to the energy norm, and (5.44) shows that it is an almost best, or quasi-optimal, approximation in the standard Sobolev norm in H_0^1. Note that the energy norm is a weighted norm in H_0^1; in order to take full advantage of the best approximation property (5.40) one would need to prove a weighted variant of the interpolation error bound (5.32). This can be done but we will not pursue it here. Of course, these norms coincide when $a = 1$.

For (5.41) to be of interest it is necessary that $u \in H^2$. In the case that Ω is convex we know from Sect. 3.7 that such regularity follows from $f \in L_2$, and that (5.25) holds. From (5.41) we therefore conclude that

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

where the constant is the product of those in (5.41) and (5.25). If Ω is nonconvex, then the solution u will generally have such singularities at the corners of Γ that will make (5.25) invalid, and this will result in lower order of convergence. Note that (5.40) still holds in this case.

Our next result concerns the L_2-norm of the error. Here we need the regularity estimate (5.25) and therefore assume that Ω is convex.

Theorem 5.4. *Let Ω be convex and let u_h and u be the solutions of* (5.26) *and* (5.24). *Then*

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

Proof. The proof proceeds as for the two-point boundary value problem in Theorem 5.2 by duality, using the auxiliary problem

$$\mathcal{A}\phi = e \quad \text{in } \Omega, \quad \text{with } \phi = 0 \quad \text{on } \Gamma, \quad \text{where } e = u_h - u. \tag{5.46}$$

We have as in (5.25)

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

and this is used as in Theorem 5.2 to show

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

By Theorem 5.3 this completes the proof. □

The previous theorems show the same error bounds for u_h as for the interpolant $I_h u$ in (5.31) and (5.32), except that the constants may be different. Note that we have only used (5.32) and not (5.31) in the proofs.

Let $R_h : H_0^1 \to S_h$ be the orthogonal projection with respect to the energy inner product, so that

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

The operator R_h is called the *Ritz projection* (or *elliptic projection*). It follows from (5.42) that the finite element solution u_h is exactly the Ritz projection of the exact solution u of (5.24), i.e., $u_h = R_h u$. Our previous error estimates for the finite element solution may be expressed as follows in terms of the operator R_h, which will be convenient when we discuss parabolic finite element problems later.

Theorem 5.5. *Let Ω be convex. Then we have, for $s = 1, 2$,*

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

Proof. The case $s = 2$ is contained in Theorems 5.3 and 5.4. For the case $s = 1$ we first note that since R_h is the orthogonal projection with respect to $a(\cdot, \cdot)$, we have $\|R_h v\|_a \le \|v\|_a$. Hence $|R_h v|_1 \le C|v|_1$ and $|R_h v - v|_1 \le C|v|_1$. Finally, using (5.48) we obtain

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

which completes the proof. □

Formally the above error analysis extends immediately to finite elements of higher order $r > 2$. In the argument in Theorem 5.4 we simply use the second interpolation error estimate in (5.34) instead of (5.32), together with the case $s = 2$ of (5.35). We then find, for $2 \le s \le r$,

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

These estimates thus show a reduced convergence rate $O(h^s)$ if $v \in H^s$ with $s < r$. As we pointed out at the end of Sect. 5.3, the regularity assumption $v \in H^r$ with $r > 2$ is somewhat unrealistic for solutions of our elliptic problem in a polygonal domain. For a domain Ω with a smooth boundary Γ the regularity is not a problem but special considerations for handling the boundary layer $\Omega \setminus \Omega_h$ are then needed to attain high accuracy.

Because the variational formulation of our discrete problem is based on L_2 inner products, the most natural error estimates are also expressed in such L_2 based norms, and therefore measure certain averages of the error. It is, of course, also of interest to derive error bounds in the maximum-norm, which express uniform error bounds over Ω. We first note that the error in the interpolant introduced above satisfies

$$\|I_h v - v\|_{\mathcal{C}(K)} \le Ch_K^2 \|v\|_{\mathcal{C}^2(K)}, \quad \forall K \in \mathcal{T}_h,$$

and thus, also in the case of a smooth boundary Γ, since then $\|v\|_{\mathcal{C}(\Omega \setminus \Omega_h)} \le Ch^2 \|v\|_{\mathcal{C}^1}$, we have

$$\|I_h v - v\|_{\mathcal{C}} \le Ch^2 \|v\|_{\mathcal{C}^2}. \tag{5.51}$$

Under the additional assumption that the family of triangulations $\{\mathcal{T}_h\}$ is *quasi-uniform*, i.e., that

$$h_K \ge ch \tag{5.52}$$

for some positive c independent of h, it is also possible, but not easy, to show that, for our elliptic problem we have

$$\|u_h - u\|_{\mathcal{C}} \le Ch^2 \log(1/h)\|u\|_{\mathcal{C}^2}, \quad \text{for } h \text{ small}, \tag{5.53}$$

see Problem 5.4. Compared to the L_2-norm estimate of Theorem 5.4 this estimate contains an additional factor $\log(1/h)$, which is not present in the interpolation error estimate (5.51), and it may be shown that this factor cannot be removed.

5.5 An A Posteriori Error Estimate

The error bounds of the previous section contain norms of the exact, unkown, solution. Using the regularity estimate (5.25) these error bounds may also be expressed in terms of the data f of (5.23), and, if the constants entering are known, stringent bounds for the error are obtained. However, as we have seen in Sect. 5.3, these bounds could be pessimistic, particularly when the triangulations are far from uniform, in which case, for instance, the inequality (5.32) could be very crude. These estimates involving $h = \max_K h_K$ should therefore be interpreted as asymptotic estimates, showing the rate of convergence of the error as $h \to 0$. Thus, for example, Theorem 5.4 shows that $\|u_h - u\| = O(h^2)$ as $h \to 0$ if $u \in H^2$.

Since these bounds do not depend on the computed solution, they are often referred to as *a priori bounds*; they may be stated before the computation has been carried out. In the next theorem we shall give an example of an *a posteriori error estimate*, which is expressed in terms of the computed solution and the data.

Theorem 5.6. *Assume that Ω is a convex polygonal domain in the plane. Let u_h and u be the solutions of* (5.26) *and* (5.24)*, respectively. Then*

$$\|u_h - u\| \le C\Big(\sum_{K\in\mathcal{T}_h} R_K^2\Big)^{1/2},$$

where

$$R_K = h_K^2\|\mathcal{A}u_h - f\|_K + h_K^{3/2}\|a[n\cdot\nabla u_h]\|_{\partial K\setminus\Gamma},$$

and $[n\cdot\nabla u_h]$ denotes the jump across ∂K in the normal derivative $n\cdot\nabla u_h$.

Proof. We use the duality argument from the proof of Theorem 5.4. Set $e = u_h - u$ and let ϕ be the solution of (5.46). Then, with $(v,w)_K = \int_K v\,w\,\mathrm{d}x$, $\|v\|_K = \|v\|_{L_2(K)}$, and $|v|_{2,K} = |v|_{H^2(K)}$,

$$\begin{aligned}
\|e\|^2 &= a(e,\phi) = a(u_h - u,\phi) = a(u_h,\phi) - (f,\phi)\\
&= \sum_K \Big((a\nabla u_h,\nabla\phi)_K - (f,\phi)_K\Big)\\
&= \sum_K \Big((\mathcal{A}u_h - f,\phi)_K + (an\cdot\nabla u_h,\phi)_{\partial K}\Big)\\
&= \sum_K \Big((\mathcal{A}u_h - f,\phi)_K - \tfrac{1}{2}(a[n\cdot\nabla u_h],\phi)_{\partial K\setminus\Gamma}\Big),
\end{aligned}$$

where the factor $1/2$ in the last term appears because the term occurs twice in the sum. Since $a(e,\chi)=0$ for $\chi\in S_h$, we may replace ϕ in the above by $\phi-\chi$ to obtain

$$\begin{aligned}\|e\|^2 &= |a(e,\phi-\chi)|\\ &\le \sum_K \Big(\|\mathcal{A}u_h - f\|_K\,\|\phi-\chi\|_K + \tfrac12\|a[n\cdot\nabla u_h]\|_{\partial K\setminus\Gamma}\,\|\phi-\chi\|_{\partial K\setminus\Gamma}\Big).\end{aligned}$$

We now choose $\chi = I_h\phi$ and recall (5.29), (5.30), and also the scaled trace inequality, obtained by transformation of the trace inequality (A.26) from a reference triangle $\hat K$ of unit size to the small triangle K, see Problem A.15,

$$\|w\|_{\partial K} \le C\Big(h_K^{-1/2}\|w\|_K + h_K^{1/2}\|\nabla w\|_K\Big). \tag{5.54}$$

Hence we obtain

$$\|\phi - I_h\phi\|_{\partial K} \le Ch_K^{3/2}|\phi|_{2,K}, \tag{5.55}$$

and, in view of the regularity estimate (5.47), we may conclude

$$\begin{aligned}\|e\|^2 = a(e,\phi) &\le C\sum_K R_K|\phi|_{2,K} \le C\Big(\sum_K R_K^2\Big)^{1/2}\Big(\sum_K |\phi|_{2,K}^2\Big)^{1/2}\\ &\le C\Big(\sum_K R_K^2\Big)^{1/2}\|\phi\|_2 \le C\Big(\sum_K R_K^2\Big)^{1/2}\|e\|,\end{aligned}$$

which completes the proof. □

If $\mathcal{A}=-\Delta$, i.e., if $a=1$, then $\mathcal{A}u_h=0$ in K, and since $n\cdot\nabla u_h$ is constant along ∂K, we have $R_K = h_K^2\big(\|f\|_K + \big|[n\cdot\nabla u_h]|_{\partial K\setminus\Gamma}\big|\big)$, so that the computed solution only enters in the second term. The *a posteriori* error estimate does not by itself imply that the error converges to zero with h, but this follows from the $O(h^2)$ *a priori* error estimate as shown before.

The *a posteriori* error estimate also suggests an approach to adaptive error control, namely to refine the mesh by subdividing those triangles K for which R_K is large compared to some tolerance. We will not go into the details.

5.6 Numerical Integration

An important feature of the finite element method is that the equations (5.27) can be generated automatically by a computer program. This procedure, which is called *assembly* of the equations, is based on an elementwise computation of the stiffness matrix and the load vector,

$$a(\Phi_j,\Phi_i)=\sum_{K\in\mathcal{T}_h}\int_K a\nabla\Phi_j\cdot\nabla\Phi_i\,\mathrm{d}x,\quad (f,\Phi_i)=\sum_{K\in\mathcal{T}_h}\int_K f\Phi_i\,\mathrm{d}x. \tag{5.56}$$

In practice, the integrals in these sums are seldom computed exactly even if analytical expressions for a and f are available. Instead they are approximated by numerical integration through a quadrature formula of the form

$$\int_K \phi\,\mathrm{d}x\approx q_K(\phi):=\sum_{l=1}^{L}\omega_{l,K}\phi(b_{l,K}). \tag{5.57}$$

The numbers $\omega_{l,K}$ are called the weights and the points $b_{l,K}$ the nodes of the quadrature formula.

If the equations are assembled by numerical integration, then instead of (5.26) we solve a modified finite element problem, which is to find $u_h\in S_h$ such that

$$a_h(u_h,\chi)=(f,\chi)_h,\quad \forall\chi\in S_h, \tag{5.58}$$

where

$$a_h(v,w)=\sum_{K\in\mathcal{T}_h}q_K(a\nabla v\cdot\nabla w),\quad (f,w)_h=\sum_{K\in\mathcal{T}_h}q_K(fw). \tag{5.59}$$

The quadrature formula q_K in (5.57) should be chosen in such a way that the error in u_h is of the same order as in the original finite element solution. An example of such a quadrature formula is the *barycentric quadrature rule*

$$q_K(\phi)=|K|\phi(P_K),\quad \text{where } |K|=\mathrm{area}(K),\quad P_K=\tfrac{1}{3}\sum_{l=1}^{3}P_{l,K}, \tag{5.60}$$

with $P_{l,K}$ and P_K the vertices and the barycenter of the triangle K. This quadrature rule is exact for linear functions, i.e.,

$$\int_K\phi\,\mathrm{d}x=|K|\phi(P_K),\quad\forall\phi\in\Pi_1. \tag{5.61}$$

This implies that the rule is accurate of order 2 so that (Problem 5.13)

$$\Big|q_K(\phi)-\int_K\phi\,\mathrm{d}x\Big|\le Ch_K^2|\phi|_{W_1^2(K)}, \tag{5.62}$$

where, with $D_{ij}=\partial^2/\partial x_i\partial x_j$,

$$|v|_{W_1^2(M)}=\sum_{i,j=1}^{2}\|D_{ij}v\|_{L_1(M)},\quad \|v\|_{L_1(M)}=\int_M|v|\,\mathrm{d}x.$$

Hence, the global quadrature error is bounded by

$$\Big|\sum_{K\in\mathcal{T}_h} q_K(\phi) - \int_\Omega \phi\,\mathrm{d}x\Big| \le Ch^2 \sum_{K\in\mathcal{T}_h} |\phi|_{W_1^2(K)}. \tag{5.63}$$

From (5.61) we conclude that $a_h(u_h,\chi)$ and $(f,\chi)_h$ are exact, for example, when a and f are constant.

Another example of a quadrature formula, which is exact for linear functions, is provided by the *nodal quadrature rule*, see Problem 5.15,

$$q_K(\phi) = \tfrac{1}{3}|K| \sum_{l=1}^{3} \phi(P_{l,K}). \tag{5.64}$$

In the following lemma we collect the properties of $a_h(\cdot,\cdot)$ and $(\cdot,\cdot)_h$ that we need in order to prove an error estimate for the modified problem (5.58).

Lemma 5.1. *If $a_h(\cdot,\cdot)$ and $(\cdot,\cdot)_h$ in* (5.59) *are computed by the quadrature formula* (5.60) *or* (5.64)*, then*

$$a_0|\chi|_1^2 \le a_h(\chi,\chi) \le C|\chi|_1^2, \quad \forall \chi \in S_h, \tag{5.65}$$

and

$$|a_h(\psi,\chi) - a(\psi,\chi)| \le Ch^2\|a\|_{\mathcal{C}^2}\,|\psi|_1\,|\chi|_1, \qquad \forall \psi,\chi \in S_h, \tag{5.66}$$

$$|(f,\chi)_h - (f,\chi)| \le Ch^2\|f\|_2\,|\chi|_1, \qquad \forall \chi \in S_h. \tag{5.67}$$

Proof. We carry out the proof for the quadrature rule (5.60); the proof for (5.64) is analogous.

Since $\nabla\chi$ is constant on K and $a_0 \le a(x) \le C$, we have

$$a_h(\chi,\chi) = \sum_K a(P_K)|\nabla\chi(P_K)|^2|K| \ge a_0 \sum_K \int_K |\nabla\chi|^2\,\mathrm{d}x = a_0|\chi|_1^2.$$

The estimate from above is derived in the same way, which shows (5.65). Using (5.63) with $\phi = a\nabla\psi\cdot\nabla\chi$ we get

$$|a_h(\psi,\chi) - a(\psi,\chi)| \le Ch^2 \sum_K |a\nabla\psi\cdot\nabla\chi|_{W_1^2(K)}.$$

To bound the right hand side, we have for ψ, $\chi \in S_h$,

$$\|D_{ij}(a\nabla\psi\cdot\nabla\chi)\|_{L_1(K)} = \|(D_{ij}a)\nabla\psi\cdot\nabla\chi\|_{L_1(K)} \le \|a\|_{\mathcal{C}^2}\,\|\nabla\psi\|_K\,\|\nabla\chi\|_K.$$

Invoking the Cauchy-Schwarz inequality for sums, we conclude

$$\sum_K |a\nabla\psi\cdot\nabla\chi|_{W_1^2(K)} \le C\|a\|_{\mathcal{C}^2}\,|\psi|_1\,|\chi|_1,$$

which proves (5.66). Similarly, since $D_{ij}\chi = 0$ on K, we have

$$\|D_{ij}(f\chi)\|_{L_1(K)} = \|D_{ij}f\,\chi + D_i f D_j\chi + D_j f D_i\chi\|_{L_1(K)} \le C\|f\|_{2,K}\,\|\chi\|_{1,K},$$

so that

$$\sum_K |f\chi|_{W_1^2(K)} \le C\|f\|_2\,\|\chi\|_1 \le C\|f\|_2\,|\chi|_1,$$

which proves (5.67). □

The inequality (5.65) shows that the symmetric bilinear form $a_h(\cdot,\cdot)$ is an inner product on S_h and that the corresponding norm is equivalent to $|\cdot|_1$, *uniformly with respect to* h. From (5.67) we deduce that the linear form $L_h(\chi) = (f,\chi)_h$ is bounded on S_h with respect to $|\cdot|_1$, again uniformly with respect to h, because

$$\begin{aligned}|(f,\chi)_h| &\le |(f,\chi)| + |(f,\chi)_h - (f,\chi)|\\ &\le \|f\|\,\|\chi\| + Ch^2\|f\|_2\,|\chi|_1 \le C\|f\|_2\,|\chi|_1.\end{aligned}$$

By the Riesz representation theorem we may therefore conclude that (5.58) has a unique solution and that it satisfies the stability estimate

$$|u_h|_1 \le C\|f\|_2, \tag{5.68}$$

see Problem 5.14. This stability of the modified finite element problem is used together with the consistency error bounds (5.66), (5.67) in the proof of the following error estimate.

Theorem 5.7. *Assume that* $a_h(\cdot,\cdot)$ *and* $(\cdot,\cdot)_h$ *in* (5.59) *are computed by the quadrature formula* (5.60) *or* (5.64)*. Let* u_h *and* u *be the solutions of* (5.58) *and* (5.24)*, respectively. Then*

$$|u_h - u|_1 \le Ch\|u\|_2 + Ch^2\Big(\|a\|_{\mathcal{C}^2}\,\|u\|_2 + \|f\|_2\Big). \tag{5.69}$$

Proof. We write $u_h - u = (u_h - I_h u) + (I_h u - u) = \theta + \rho$. Using (5.24) and (5.58), we get, for any $\chi \in S_h$,

$$\begin{aligned}a_h(\theta,\chi) &= a_h(u_h,\chi) - a_h(I_h u,\chi) + \Big(a(u,\chi) - (f,\chi)\Big)\\ &\quad - \Big(a_h(u_h,\chi) - (f,\chi)_h\Big) + a(I_h u,\chi) - a(I_h u,\chi)\\ &= -a(\rho,\chi) - \Big(a_h(I_h u,\chi) - a(I_h u,\chi)\Big) + \Big((f,\chi)_h - (f,\chi)\Big).\end{aligned}$$

Since $\theta \in S_h$ we may choose $\chi = \theta$. In view of (5.65), the boundedness of the bilinear form $a(\cdot,\cdot)$, and the error estimates (5.66) and (5.67), this implies

$$a_0|\theta|_1^2 \le a_h(\theta,\theta) \le \Big(C|\rho|_1 + Ch^2\|a\|_{\mathcal{C}^2}\,|I_h u|_1 + Ch^2\|f\|_2\Big)|\theta|_1.$$

Hence,

$$|u_h - u|_1 \le |\theta|_1 + |\rho|_1 \le C|\rho|_1 + Ch^2\Big(\|a\|_{C^2}\,|I_h u|_1 + \|f\|_2\Big).$$

Using the interpolation error estimate (5.32), we have

$$|\rho|_1 \le Ch\|u\|_2, \quad |I_h u|_1 \le |u|_1 + |\rho|_1 \le |u|_1 + Ch\|u\|_2 \le C\|u\|_2.$$

Together these estimates prove (5.69). □

The first term on the right side of (5.69) is (essentially) the same as in (5.41), whereas the remaining terms estimate the effect of numerical integration. Note that this result requires more regularity than (5.41). For example, we need $f \in H^2$, which (at least formally) implies that $u \in H^4$. This is consistent with the result for the finite difference scheme in Theorem 4.2, where u is required to have four derivatives, see also Remark 5.1. We may think of the finite element method with numerical integration as a finite difference scheme on a non-uniform mesh.

The original finite element method (5.26) is *conforming* in the sense that $S_h \subset H_0^1$ and the forms $a(\cdot,\cdot)$ and $L(\cdot) = (f,\cdot)$ are the same as in the continuous problem (5.24). In the modified finite element method (5.58) we still have $S_h \subset H_0^1$, but the forms are different. It is therefore called *non-conforming*. Other non-conforming finite element methods violate the inclusion $S_h \subset H_0^1$, for example, by the use of discontinuous piecewise polynomials. They can sometimes be analyzed in a similar way. The argument used in the proof of Theorem 5.7 is based on what is known as Strang's first lemma in the literature on non-conforming finite element methods.

5.7 A Mixed Finite Element Method

In some situations it is the flux, $-a\nabla u$, of the solution u that is of primary interest. However, in the standard finite element method the derivatives, and hence the flux, are approximated to lower order $O(h)$ rather than the $O(h^2)$ approximation of the solution. We shall now briefly describe a finite element method for our model problem (5.23), which is based on a so called mixed formulation of this problem, and which does not have this disadvantage. Here the flux of the solution u is introduced as a separate dependent variable whose approximation is sought in a different finite element space than the solution itself. This may be done in such a way that the flux is approximated to the same order of accuracy as u. For simplicity we assume that $a = 1$ in (5.23). With $\sigma = \nabla u$ as a separate two-dimensional variable, this equation may then be formulated as the system

$$\begin{aligned} -\nabla\cdot\sigma &= f \qquad \text{in } \Omega, \\ \sigma &= \nabla u \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \partial\Omega. \end{aligned} \tag{5.70}$$

With $H = \{\omega = (\omega_1, \omega_2) \in L_2 \times L_2 : \nabla \cdot \omega \in L_2\}$ we note that the solution $(u, \sigma) \in L_2 \times H$ also solves the variational problem

$$\begin{aligned} (\nabla \cdot \sigma, \varphi) + (f, \varphi) &= 0, \qquad \forall \varphi \in L_2, \\ (\sigma, \omega) + (u, \nabla \cdot \omega) &= 0, \qquad \forall \omega \in H, \end{aligned} \tag{5.71}$$

where the $(\cdot, \cdot)$ denotes the appropriate L_2 inner products, and a smooth solution of (5.71) satisfies (5.70). Setting $L(v, \mu) = \frac{1}{2}\|\mu\|^2 + (\nabla \cdot \mu + f, v)$, one may show that the solution (u, σ) of (5.70) can be characterized as the saddle-point satisfying

$$L(v, \sigma) \le L(u, \sigma) \le L(u, \mu), \quad \forall v \in L_2,\ \mu \in H \tag{5.72}$$

(see Problem 5.16), and the key to the existence of a solution is the inequality

$$\inf_{v \in L_2} \sup_{\mu \in H} \frac{(v, \nabla \cdot \mu)}{\|v\|\, \|\mu\|_H} \ge c > 0, \quad \text{where } \|\mu\|_H^2 = \|\mu\|^2 + \|\nabla \cdot \mu\|^2. \tag{5.73}$$

With S_h and H_h certain finite-dimensional subspaces of L_2 and H we shall consider the discrete analogue of (5.71), which is to find $(u_h, \sigma_h) \in S_h \times H_h$ such that

$$\begin{aligned} (\nabla \cdot \sigma_h, \chi) + (f, \chi) &= 0, \qquad \forall \chi \in S_h, \\ (\sigma_h, \psi) + (u_h, \nabla \cdot \psi) &= 0, \qquad \forall \psi \in H_h. \end{aligned} \tag{5.74}$$

As in the continuous case this problem is equivalent to the discrete analogue of the saddle point problem (5.72), and in order for this discrete problem to have a solution with the desired properties, the choice of spaces $S_h \times H_h$ must be such that the analogue of (5.73) holds, in this context referred to as the Babuška-Brezzi inf-sup condition. More precisely,

$$\inf_{v \in S_h} \sup_{\mu \in H_h} \frac{(v, \nabla \cdot \mu)}{\|v\|\, \|\mu\|_H} \ge c > 0, \tag{5.75}$$

must hold uniformly with respect to h.

An example of a pair of spaces which satisfy the inf-sup condition, introduced by Raviart and Thomas, is as follows: With $\mathcal{T}_h$ a quasi-uniform family of triangulation of Ω, which we assume here to be polygonal, we set

$$S_h = \big\{\chi \in L_2 : \chi|_K \text{ linear}, \forall K \in \mathcal{T}_h\big\},$$

with no continuity required across inter-element boundaries. We also define

$$H_h = \big\{\psi = (\psi_1, \psi_2) \in H : \psi|_K \in H(K),\ \forall K \in \mathcal{T}_h\big\},$$

where $H(K)$ denotes affine maps of quadratics on a reference triangle $\hat{K}$ of the form $(l_1(\xi) + \alpha\xi_1(\xi_1 + \xi_2),\ l_2(\xi) + \beta\xi_2(\xi_1 + \xi_2))$, with $l_1(\xi), l_2(\xi)$ linear,

$\alpha, \beta \in \mathbf{R}$. Since each of the functions $l_j(\xi)$ has three parameters we have $\dim H(K) = 8$. The space H_h thus consists of piecewise quadratics on the triangulation $\mathcal{T}_h$, which are of the specific form implied by the definition of $H(K)$. As degrees of freedom for H_h one may use the values of $\psi \cdot n$ at two points on each side of K (6 conditions) and in addition the mean-values of ψ_1 and ψ_2 over K (2 conditions). We note that the condition $\psi \in H$ in the definition of H_h requires that $\nabla \cdot \psi \in L_2$, which is equivalent to the continuity of $\chi \cdot n$ across inter-element boundaries. For the solutions of (5.74) and (5.70) one may show

$$\|u_h - u\| \le Ch^2 \|u\|_2 \quad \text{and} \quad \|\sigma_h - \sigma\| \le Ch^s \|u\|_{s+1}, \quad s = 1, 2.$$

Thus, the flux σ is approximated to the same order $O(h^2)$ as u.

5.8 Problems

Problem 5.1. Prove (5.7) and (5.8).

Hint for (5.8): $(I_h v)'(x) - v'(x) = h_j^{-1} \int_{K_j} \big(v'(y) - v'(x)\big)\, \mathrm{d}y$ for $x \in K_j$.

Hint for (5.7): Let $Q_1 v$ the polynomial of degree 1 obtained from Taylor's formula for v at x_{j-1}. Note that $I_h(Q_1 v) = Q_1 v$ and $\|I_h v\|_{\mathcal{C}(K_j)} \le \|v\|_{\mathcal{C}(K_j)}$, so that $\|I_h v - v\|_{\mathcal{C}(K_j)} = \|I_h(v - Q_1 v) + (Q_1 v - v)\|_{\mathcal{C}(K_j)} \le 2\|v - Q_1 v\|_{\mathcal{C}(K_j)}$. Estimate the remainder: $\|v - Q_1 v\|_{\mathcal{C}(K_j)} \le \max_{x \in K_j} \int_{K_j} |x - y|\, |v''(y)|\, \mathrm{d}y$. Conclude $\|I_h v - v\|_{\mathcal{C}(K_j)} \le 2h_j \int_{K_j} |v''(y)|\, \mathrm{d}y$, which implies (5.7). This proof can be generalized to functions of two variables, see (5.29), the main difference is that it is more difficult to estimate the remainder in Taylor's formula.

Problem 5.2. Find the elements of the matrix A in (5.5) when $h_j = h =$ constant.

Problem 5.3. Use the basis $\{\Phi_i\}_{i=1}^{M_h}$ to show that (5.38) can be written in matrix form as $BV = b$, where the matrix B (the so-called mass matrix) is symmetric, positive definite, and sparse if M_h is large.

Problem 5.4. Consider the situation in Sect. 5.1 with a piecewise linear finite element space S_h.

(a) Use the Green's function in Theorem 2.3 to prove that $u_h = I_h u$ when $a = 1$, $c = 0$ in (5.1), cf. Remark 5.2. Hint: Use the results from Problem 2.4, Problem 2.2 (a), and the fact that $G(x_j, \cdot) \in S_h$ if x_j is a node.

(b) In the case of variable coefficients prove that

$$|u_h(x_j) - u(x_j)| \le Ch^2 \|u\|_2.$$

Hint: Show that $e(x_j) = a(e, G(x_j, \cdot) - I_h G(x_j, \cdot))$ and use an interpolation error estimate on the intervals $(0, x_j)$, $(x_j, 1)$, where $G(x_j, \cdot)$ is smooth.

(c) Conclude that $\|u_h - u\|_{\mathcal{C}} \le Ch^2\|u\|_{\mathcal{C}^2}$, which is (5.53) in this simple special case. Hint: $\|u_h - I_h u\|_{\mathcal{C}} = \max_j |u_h(x_j) - u(x_j)|$.

This is the basic idea behind one approach to maximum-norm estimates for elliptic problems in several variables. However, the stronger singularity of the Green's function, see Sect. 3.4, makes the analysis much more difficult.

Problem 5.5. Prove that, under the assumptions of Theorem 5.3,

$$|u_h - u|_1 \le C\Big(\sum_K h_K^2 |u|_{2,K}^2\Big)^{1/2}.$$

Problem 5.6. (Galerkin's method.) Let $a(\cdot,\cdot)$ and $L(\cdot)$ satisfy the assumptions of the Lax-Milgram lemma, i.e.,

$$\begin{aligned} |a(v,w)| &\le C_1\|v\|_V\,\|w\|_V, && \forall v, w \in V,\\ a(v,v) &\ge C_2\|v\|_V^2, && \forall v \in V,\\ |L(v)| &\le C_3\|v\|_V, && \forall v \in V. \end{aligned}$$

Let $u \in V$ be the solution of

$$a(u,v) = L(v), \quad \forall v \in V.$$

Let $\tilde{V} \subset V$ be a finite-dimensional subspace and let $\tilde{u} \in \tilde{V}$ be determined by Galerkin's method:

$$a(\tilde{u},v) = L(v), \quad \forall v \in \tilde{V}.$$

Prove that (note that $a(\cdot,\cdot)$ may be non-symmetric)

$$\|\tilde{u} - u\|_V \le \frac{C_1}{C_2} \min_{\chi\in\tilde{V}} \|\chi - u\|_V.$$

Prove that, if $a(\cdot,\cdot)$ is symmetric and $\|v\|_a = a(v,v)^{1/2}$, then

$$\|\tilde{u} - u\|_a = \min_{\chi\in\tilde{V}} \|\chi - u\|_a \quad \text{and} \quad \|\tilde{u} - u\|_V \le \sqrt{\frac{C_1}{C_2}} \min_{\chi\in\tilde{V}} \|\chi - u\|_V.$$

Problem 5.7. Consider the problem

$$-\nabla\cdot\big(a\nabla u\big) + b\cdot\nabla u + cu = f \quad \text{in } \Omega, \quad \text{with } u = 0 \quad \text{on } \Gamma,$$

from Sect. 3.5. Note that the presence of the convection term $b\cdot\nabla u$ makes the bilinear form non-symmetric.

(a) Formulate a finite element method for this problem and prove an error bound in the H^1-norm. Hint: See Problem 5.6.

(b) Prove an error bound in the L_2-norm. Hint: Modify the proof of Theorem 5.4 by using the auxiliary problem

$$\mathcal{A}^*\phi := -\nabla \cdot (a\nabla\phi) - b \cdot \nabla\phi + (c - \nabla \cdot b)\phi = e \quad \text{in } \Omega, \quad \phi = 0 \quad \text{on } \Gamma,$$

instead of (5.46). The operator $\mathcal{A}^*$ is the *adjoint* of $\mathcal{A}$, defined by $(\mathcal{A}v, w) = a(v, w) = (v, \mathcal{A}^*w)$ for all $v, w \in H^2 \cap H_0^1$.

Problem 5.8. Formulate a finite element problem corresponding to the nonhomogeneous Dirichlet problem (3.27). Prove error estimates. Hint: With the notation of Sect. 5.3, $u_h(x) = \sum_{j=1}^{M_h} U_j\Phi_j(x) + \sum_{j=M_h+1}^{N_h} g(P_j)\Phi_j(x)$.

Problem 5.9. Formulate a finite element problem corresponding to the Neumann problem (3.30). Prove error estimates.

Problem 5.10. Formulate a finite element problem corresponding to the nonhomogeneous Neumann problem (3.34). Prove error estimates.

Problem 5.11. Formulate a finite element problem corresponding to the Robin problem in Problem 3.6. Prove error estimates.

Problem 5.12. The following important result is called the Bramble-Hilbert lemma. *Let F be a nonnegative functional on $W_p^m = W_p^m(\Omega)$, where Ω is a bounded convex domain in $\mathbf{R}^d$, and assume that*

$$\begin{aligned} &F(v + w) \le F(v) + F(w), && \forall v, w \in W_p^m, \\ &F(v) \le C\|v\|_{W_p^m}, && \forall v \in W_p^m, \\ &F(v) = 0, && \forall v \in \Pi_{m-1}. \end{aligned}$$

Then F is bounded with respect to the corresponding seminorm, i.e., there is a constant $C = C(\Omega, m, p)$ such that

$$F(v) \le C|v|_{W_p^m}, \quad \forall v \in W_p^m.$$

(a) Prove the Bramble-Hilbert lemma for $d = 1$, $\Omega = (0, 1)$, and $m = p = 2$. Hint: As in Problem 5.1 show that $\|v - Q_1v\|_2 \le C|v|_2$. Then $F(v) \le F(v - Q_1v) + F(Q_1v) = F(v - Q_1v) \le C\|v - Q_1v\|_2 \le C|v|_2$.

(b) Use the Bramble-Hilbert lemma to show (5.29) and (5.30). Hint: Do this first for a fixed triangle $\widehat{K}$ of unit size and then make an affine transformation of this triangle to a small triangle K, see Problem A.14. For (5.29) choose $W_p^m(\Omega) = H^2(\widehat{K})$, $F(v) = \|I_hv - v\|_{L_2(\widehat{K})}$ and estimate the nodal values by Sobolev's inequality $|v(\widehat{P}_j)| \le C\|v\|_{H^2(\widehat{K})}$.

Problem 5.13. Prove (5.62) by using the Bramble-Hilbert lemma with $m = 2$, $p = 1$.

Problem 5.14. Prove (5.68).

Problem 5.15. Prove an analogue of Lemma 5.1 for the nodal quadrature formula (5.64).

Problem 5.16. Prove that any solution of (5.71) satisfies (5.72).

Problem 5.17. (Computer exercise.) Consider the same two-point boundary value problem as in Problem 4.4. Apply the finite element method (5.3) based on piecewise linear approximating functions on the same partition as in Problem 4.4 with $h = 1/10,\ 1/20$. Find the exact solution and compute the maximum of the error at the mesh-points.

Problem 5.18. (Computer exercise.) Consider the same boundary value problem as in Problem 4.5. Solve it by the finite element method (5.26) based on piecewise linear approximating functions on the same partition as in Problem 4.5, divided into triangles by inserting a diagonal with positive slope into each mesh-square, with $h = 1/10,\ 1/20$. Recall the exact solution and compute the L_2-norm of the error. Use the barycentric quadrature rule to compute the stiffness matrix, the load vector, and the L_2-norm.

6 The Elliptic Eigenvalue Problem

Eigenvalue problems are important in the mathematical analysis of partial differential equations, and occur, e.g., in the modelling of vibrating membranes and other applications. In our study of time-dependent partial differential equations it will be important to develop functions in eigenfunction expansions, and we therefore discuss such expansions in Sect. 6.1 below. In the following Sect. 6.2 we present some simple approaches and results for the numerical solution of eigenvalue problems.

6.1 Eigenfunction Expansions

We shall first consider the eigenvalue problem corresponding to the symmetric case of the two-point boundary value problem in Chapt. 2, to find a number λ and a function φ, which is not identically zero, such that

$$\mathcal{A}\varphi := -(a\varphi')' + c\varphi = \lambda\varphi \quad \text{in } \Omega = (0,1), \quad \text{with } \varphi(0) = \varphi(1) = 0, \tag{6.1}$$

where a and c are smooth functions such that $a(x) \geq a_0 > 0$ and $c(x) \geq 0$ on $\bar{\Omega}$. Such a number λ is called an eigenvalue and φ is the corresponding eigenfunction.

Recall that the two-point boundary value problem

$$\mathcal{A}u = f \quad \text{in } \Omega, \quad \text{with } u(0) = u(1) = 0, \tag{6.2}$$

may be written in weak form as: Find $u \in H_0^1 = H_0^1(\Omega)$ such that

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

where the bilinear form and the inner product are defined by

$$a(u,v) = \int_0^1 (a\,u'v' + c\,uv)\,\mathrm{d}x \quad \text{and} \quad (u,v) = \int_0^1 uv\,\mathrm{d}x, \tag{6.3}$$

respectively. With this notation the eigenvalue problem (6.1) may be stated: Find a number λ and a function $\varphi \in H_0^1$, $\varphi \neq 0$, such that

$$a(\varphi,v) = \lambda\,(\varphi,v), \quad \forall v \in H_0^1. \tag{6.4}$$

We shall also consider the Dirichlet eigenvalue problem to find a number λ and a function φ, not identically zero, such that

$$-\Delta\varphi = \lambda\varphi \quad \text{in } \Omega, \quad \text{with } \varphi = 0 \quad \text{on } \Gamma, \tag{6.5}$$

where Ω is a bounded domain in $\mathbf{R}^d$ with smooth boundary Γ. Here the associated Dirichlet boundary value problem is

$$-\Delta u = f \quad \text{in } \Omega, \quad \text{with } u = 0 \quad \text{on } \Gamma, \tag{6.6}$$

or, in variational form, to find $u \in H_0^1 = H_0^1(\Omega)$ such that

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

where now

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x = (\nabla u, \nabla v) \quad \text{and} \quad (u, v) = \int_\Omega uv \, \mathrm{d}x. \tag{6.7}$$

The variational form of the eigenvalue problem corresponding to (6.5) is again to find a number λ and a function $\varphi \in H_0^1$, $\varphi \neq 0$, such that (6.4) holds.

Recall that, if u is a solution of (6.6) and $f \in H^k$, then $u \in H^{k+2} \cap H_0^1$, see Sect. 3.7. Hence we may conclude at once that a possible eigenfunction is smooth: Since $\varphi \in L_2$, elliptic regularity implies that $\varphi \in H^2 \cap H_0^1$, which in turn shows $\varphi \in H^4 \cap H_0^1$, and so on. The corresponding observation applies also to the simpler eigenvalue problem (6.1).

Both eigenvalue problems (6.1) and (6.6) thus have the variational formulation (6.4), and this would also be the case if instead of the Laplacian we used the more general elliptic operator $\mathcal{A}u = -\nabla \cdot (a\nabla u) + c\, u$ together with suitable homogeneous boundary conditions of Dirichlet, Neumann, or Robin type, as described in Chapt. 3. This would lead to the following more general eigenvalue problem. Let $H = L_2 = L_2(\Omega)$ and V be a linear subspace of $H^1 = H^1(\Omega)$, with $\Omega \subset \mathbf{R}^d$, and assume that the bilinear form $a(\cdot, \cdot)$ is *symmetric* and *coercive on* V, i.e.,

$$a(v, v) \geq \alpha \|v\|_1^2, \quad \forall v \in V, \quad \text{with } \alpha > 0. \tag{6.8}$$

Find $\varphi \in V$, $\varphi \neq 0$, and a number λ such that

$$a(\varphi, v) = \lambda(\varphi, v), \quad \forall v \in V, \tag{6.9}$$

where $(\cdot, \cdot)$ denotes the inner product in $H = L_2$. Note that $a(\cdot, \cdot)$ is an inner product in V.

For simplicity we shall consider the concrete case of (6.5), with $a(\cdot, \cdot)$ defined by (6.7), and we invite the reader to check that the theory that we shall present actually applies, with minor notational changes, to the more general eigenvalue problem (6.9).

We begin with some general simple properties of the eigenvalues and eigenfunctions.

Theorem 6.1. *The eigenvalues of* (6.5) *are real and positive. Two eigenfunctions corresponding to different eigenvalues are orthogonal in* L_2 *and* H_0^1.

Proof. Let λ be an eigenvalue and φ the corresponding eigenfunction. Then

$$\lambda\|\varphi\|^2 = \lambda\,(\varphi,\varphi) = a(\varphi,\varphi),$$

which together with (6.8) implies that $\lambda > 0$. Let λ_1 and λ_2 be two different eigenvalues and φ_1 and φ_2 the corresponding eigenfunctions. Then

$$\lambda_1\,(\varphi_1,\varphi_2) = a(\varphi_1,\varphi_2) = a(\varphi_2,\varphi_1) = \lambda_2\,(\varphi_2,\varphi_1) = \lambda_2\,(\varphi_1,\varphi_2),$$

so that

$$(\lambda_1 - \lambda_2)(\varphi_1,\varphi_2) = 0.$$

Since $\lambda_1 \neq \lambda_2$ it follows that $(\varphi_1,\varphi_2) = 0$, and hence also $a(\varphi_1,\varphi_2) = 0$. □

As a first step we shall show that there exists an eigenvalue, in fact, a smallest eigenvalue. This eigenvalue will be characterized by

$$\lambda_1 = \inf\Big\{a(v,v):\ v \in H_0^1,\ \|v\| = 1\Big\}, \quad \text{where } a(v,v) = \|\nabla v\|^2. \tag{6.10}$$

The equality (6.10) may also be written (this is referred to as the Rayleigh-Ritz characterization of the principal eigenvalue)

$$\lambda_1 = \inf_{v\neq 0} \frac{\|\nabla v\|^2}{\|v\|^2},$$

which follows from $\|\nabla(\alpha v)\|^2 = \alpha^2\|\nabla v\|^2$. Since, for an arbitrary eigenvalue λ and corresponding eigenfunction φ,

$$\|\nabla\varphi\|^2 = a(\varphi,\varphi) = \lambda\,(\varphi,\varphi) = \lambda\|\varphi\|^2,$$

we conclude that $\lambda \geq \lambda_1$, so that λ_1 is a lower bound for the eigenvalues.

Theorem 6.2. *The infimum in* (6.10) *is attained by a function* $\varphi_1 \in H_0^1$. *This function is an eigenfunction of* (6.5) *and* λ_1 *the corresponding eigenvalue.*

Proof. We shall postpone the proof of the first statement of the theorem till the end of this section and assume that the infimum is attained by $\varphi_1 \in H_0^1$, i.e., $\lambda_1 = \|\nabla\varphi_1\|^2$ and $\|\varphi_1\| = 1$. We now show that φ_1 is an eigenfunction of (6.5) corresponding to the eigenvalue λ_1, that is,

$$a(\varphi_1, v) = \lambda_1\,(\varphi_1, v), \quad \forall v \in H_0^1. \tag{6.11}$$

Note that, for α an arbitrary real number,

$$a(\varphi_1 + \alpha v, \varphi_1 + \alpha v) = \lambda_1 + 2\alpha\, a(\varphi_1, v) + \alpha^2 a(v, v)$$

and

$$\|\varphi_1 + \alpha v\|^2 = 1 + 2\alpha\,(\varphi_1, v) + \alpha^2 \|v\|^2.$$

Since the ratio of the two norms is bounded below by λ_1 for all α, we have

$$\lambda_1 + 2\alpha a(\varphi_1, v) + \alpha^2 a(v, v) \geq \lambda_1 + 2\lambda_1 \alpha\,(\varphi_1, v) + \lambda_1 \alpha^2 \|v\|^2,$$

or

$$2\alpha\big(a(\varphi_1, v) - \lambda_1(\varphi_1, v)\big) + \alpha^2\big(a(v, v) - \lambda_1 \|v\|^2\big) \geq 0.$$

Suppose now that (6.11) does not hold, so that the coefficient of α is $\neq 0$. Then choosing $|\alpha|$ small and with a sign such that the first term is negative, we have a contradiction. □

From Theorem 6.2 we thus know that at least one eigenfunction φ_1 exists. We now repeat the considerations above in the subspace V_1 of $V = H_0^1$ consisting of functions which are orthogonal to φ_1 with respect to $(\cdot, \cdot)$. Note that these functions are then orthogonal to φ_1 also with respect to $a(\cdot, \cdot)$ since $a(v, \varphi_1) = \lambda_1(v, \varphi_1) = 0$. We consider thus

$$\begin{aligned} \lambda_2 &= \inf\Big\{a(v, v) : v \in V,\ \|v\| = 1,\ (v, \varphi_1) = 0\Big\} \\ &= \inf\Big\{\|\nabla v\|^2 : v \in H_0^1,\ \|v\| = 1,\ (v, \varphi_1) = 0\Big\}. \end{aligned} \tag{6.12}$$

Clearly $\lambda_2 \geq \lambda_1$, since the infimum here is taken over a smaller set of functions v than in (6.10). In the same way as above we may show that the infimum is attained and we call the minimizing function φ_2 (we shall return to the question of existence of φ_2 later), which then satisfies

$$a(\varphi_2, \varphi_2) = \|\nabla \varphi_2\|^2 = \lambda_2, \quad \|\varphi_2\| = 1, \quad (\varphi_1, \varphi_2) = 0.$$

To show that φ_2 is an eigenfunction, we first show, exactly as above that

$$a(\varphi_2, v) = \lambda_2\,(\varphi_2, v), \quad \text{for all } v \in H_0^1 \text{ with } (v, \varphi_1) = 0.$$

To see that the equation holds for all $v \in H_0^1$ and not just for those orthogonal to φ_1, we note that any $v \in H_0^1$ may be written as

$$v = \alpha\varphi_1 + w, \quad \text{with } \alpha = (v, \varphi_1) \text{ and } (w, \varphi_1) = 0.$$

It therefore remains only to show that $a(\varphi_2, \varphi_1) = \lambda_2\,(\varphi_2, \varphi_1)$. But this follows at once since $(\varphi_2, \varphi_1) = 0$ and $a(\varphi_2, \varphi_1) = 0$.

Continuing in this way we find a nondecreasing sequence of eigenvalues $\{\lambda_j\}_{j=1}^{\infty}$ and a corresponding sequence of eigenfunctions $\{\varphi_j\}_{j=1}^{\infty}$, which are mutually orthogonal and have L_2-norm 1, such that

$$\begin{aligned}\lambda_n &= a(\varphi_n, \varphi_n) \\ &= \inf\Big\{a(v,v):\ v \in H_0^1,\ \|v\| = 1,\ (v,\varphi_j) = 0,\ j = 1,\dots,n-1\Big\}.\end{aligned} \tag{6.13}$$

Note that the process does not stop after a finite number of steps. For, if $(v,\varphi_j) = 0$, $j = 1,\dots,n-1$, were to imply $v = 0$, then L_2 would be finite-dimensional, which it is not. One may show the following.

Theorem 6.3. *With λ_n the n^{th} eigenvalue of* (6.5) *we have that $\lambda_n \to \infty$ as $n \to \infty$.*

The proof of this theorem will also be postponed till later.

As a result of this theorem a number in the nondecreasing sequence $\{\lambda_j\}_{j=1}^{\infty}$ can only be repeated a finite number of times. If $\lambda_{n-1} < \lambda_n = \lambda_{n+1} = \cdots = \lambda_{n+m-1} < \lambda_{n+m}$, then we say that the eigenvalue λ_n has multiplicity m. The set E_n of linear combinations of $\varphi_n,\dots,\varphi_{n+m-1}$ is then a finite-dimensional linear space of dimension m, the eigenspace corresponding to λ_n. For $v \in E_n$ we thus have $-\Delta v = \lambda_n v$.

We remark that the first, or principal, eigenvalue λ_1 is a simple eigenvalue, i.e., that $\lambda_2 > \lambda_1$, and that the corresponding principal eigenfunction φ_1 may be chosen to be positive in Ω, after a possible change of sign. To indicate a proof of this, we write $\varphi_1 = \varphi = \varphi^+ - \varphi^-$, where $\varphi^\pm = \max(\pm\varphi, 0)$. One may show that $\varphi^\pm \in H_0^1$ and that $\nabla\varphi^+ = \nabla\varphi$ when $\varphi \geq 0$ and $\nabla\varphi^+ = 0$ when $\varphi < 0$ so that $a(\varphi^+,\varphi^-) = (\nabla\varphi^+,\nabla\varphi^-) = 0$. We then have that $\|\nabla\varphi^\pm\|^2 = \lambda_1\|\varphi^\pm\|^2$, since otherwise

$$\lambda_1 = \|\nabla\varphi\|^2 = \|\nabla\varphi^+\|^2 + \|\nabla\varphi^-\|^2 > \lambda_1\big(\|\varphi^+\|^2 + \|\varphi^-\|^2\big) = \lambda_1\|\varphi\|^2 = \lambda_1,$$

which is a contradiction. Hence $\varphi^\pm$ both satisfy $-\Delta\varphi^\pm = \lambda_1\varphi^\pm$. But then $-\Delta\varphi^+ = \lambda_1\varphi^+ \geq 0$, and hence the strong maximum principle shows that $\varphi^+ > 0$ in Ω or $\varphi^+ = 0$ in Ω, so that $\varphi_1 > 0$ in Ω or $\varphi_1 < 0$ in Ω. This also implies that λ_1 is a simple eigenvalue, since there cannot exist two orthogonal eigenfunctions with constant sign.

We now turn to the question of how eigenfunctions may be used for series expansions of other functions and consider the case of a general Hilbert space H. Let $\{\varphi_j\}_{j=1}^{\infty}$ be an orthonormal sequence, i.e., one for which

$$(\varphi_i,\varphi_j) = \delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases}$$

Such a sequence is called an *orthonormal basis* (or a complete orthonormal set) if any v in H can be approximated arbitrarily well by a linear combination of elements from the sequence, i.e., if for any $\epsilon > 0$ there exist an integer N and real numbers $\alpha_1,\dots,\alpha_N$ such that

$$\Big\|v - \sum_{j=1}^{N} \alpha_j \varphi_j\Big\| < \epsilon.$$

Note that it suffices to show this for v in a dense subset $\mathcal{M}$ of H. In fact, let $v \in H$. That $\mathcal{M}$ is a dense subset of H means that one may find $w \in \mathcal{M}$ such that $\|v - w\| < \epsilon/2$. Therefore it suffices to find a linear combination such that

$$\Big\|w - \sum_{j=1}^{N} \alpha_j \varphi_j\Big\| < \epsilon/2,$$

since then

$$\Big\|v - \sum_{j=1}^{N} \alpha_j \varphi_j\Big\| \le \|v - w\| + \Big\|w - \sum_{j=1}^{N} \alpha_j \varphi_j\Big\| < \epsilon.$$

Lemma 6.1. *Let $\{\varphi_j\}_{j=1}^{\infty}$ be an orthonormal set in H. Then the best approximation of $v \in H$ by a linear combination of the first N functions φ_j is $v_N = \sum_{j=1}^{N} (v, \varphi_j)\varphi_j$.*

Proof. We have, for arbitrary $\alpha_1, \ldots, \alpha_N$,

$$\begin{aligned}\Big\|v - \sum_{j=1}^{N} \alpha_j \varphi_j\Big\|^2 &= \|v\|^2 - 2\sum_{j=1}^{N} \alpha_j (v, \varphi_j) + \sum_{j=1}^{N} a_j^2 \\ &= \|v\|^2 + \sum_{j=1}^{N} \big(\alpha_j - (v, \varphi_j)\big)^2 - \sum_{j=1}^{N} (v, \varphi_j)^2,\end{aligned}$$

from which the result follows at once. □

Since the left-hand side is nonnegative, we obtain, in particular, for $\alpha_j = (v, \varphi_j)$,

$$\sum_{j=1}^{N} (v, \varphi_j)^2 \le \|v\|^2.$$

Since this holds for each N we infer *Bessel's inequality*

$$\sum_{j=1}^{\infty} (v, \varphi_j)^2 \le \|v\|^2.$$

If $\{\varphi_j\}_{j=1}^{\infty}$ is an orthonormal *basis*, then the error in the best approximation has to tend to zero as N tends to infinity, so that

$$\Big\|v - \sum_{j=1}^{N} (v, \varphi_j)\varphi_j\Big\|^2 = \|v\|^2 - \sum_{j=1}^{N} (v, \varphi_j)^2 \to 0, \quad \text{as } N \to \infty. \tag{6.14}$$

This is equivalent to *Parseval's relation*

$$\sum_{j=1}^{\infty}(v,\varphi_j)^2 = \|v\|^2.$$

Thus the orthonormal set $\{\varphi_j\}_{j=1}^{\infty}$ is an orthonormal basis of H if and only if Parseval's relation holds for all v in H (or for all v in a dense subset of H).

We return to the concrete case of $H = L_2$. We then have the following.

Theorem 6.4. *The eigenfunctions $\{\varphi_j\}_{j=1}^{\infty}$ of (6.5) form an orthonormal basis for L_2. The series $\sum_{j=1}^{\infty}\lambda_j(v,\varphi_j)^2$ is convergent if and only if $v \in H_0^1$. Moreover,*

$$\|\nabla v\|^2 = a(v,v) = \sum_{j=1}^{\infty}\lambda_j(v,\varphi_j)^2, \quad \textit{for all } v \in H_0^1. \tag{6.15}$$

Proof. By our above discussion it follows that for the first statement it suffices to show (6.14) for all v in H_0^1, which is a dense subspace of L_2. We shall demonstrate that

$$\Big\|v - \sum_{j=1}^{N}(v,\varphi_j)\varphi_j\Big\| \le C\lambda_{N+1}^{-1/2}, \quad \text{for all } v \in H_0^1, \tag{6.16}$$

which then implies (6.14) in view of Theorem 6.3.

To prove (6.16), set $v_N = \sum_{j=1}^{N}(v,\varphi_j)\varphi_j$ and $r_N = v - v_N$. Then $(r_N,\varphi_j) = 0$ for $j = 1,\dots,N$, so that

$$\frac{\|\nabla r_N\|^2}{\|r_N\|^2} \ge \inf\Big\{\|\nabla v\|^2 : v \in H_0^1, \|v\| = 1, (v,\varphi_j) = 0, j = 1,\dots,N\Big\} = \lambda_{N+1},$$

and hence

$$\|r_N\| \le \lambda_{N+1}^{-1/2}\|\nabla r_N\|.$$

It now suffices to show that the sequence $\|\nabla r_N\|$ is bounded. We first recall from Theorem 6.1 that $a(\varphi_i,\varphi_j) = 0$ for $i \ne j$, so that $a(r_N, v_N) = 0$. Hence $a(v,v) = a(v_N,v_N) + 2a(v_N,r_N) + a(r_N,r_N) = a(v_N,v_N) + a(r_N,r_N)$ and

$$\|\nabla r_N\|^2 = a(r_N,r_N) = a(v,v) - a(v_N,v_N) \le a(v,v) = \|\nabla v\|^2,$$

which completes the proof of (6.16).

For the proof of the second statement, we first note that, for $v \in H_0^1$,

$$\sum_{j=1}^{N}\lambda_j(v,\varphi_j)^2 = a(v_N,v_N) = a(v,v) - a(r_N,r_N) \le a(v,v),$$

and we conclude that $\sum_{j=1}^{\infty}\lambda_j(v,\varphi_j)^2 < \infty$. Conversely, we assume that $v \in L_2$ and $\sum_{j=1}^{\infty}\lambda_j(v,\varphi_j)^2 < \infty$. We already know that $v_N \to v$ in L_2 as

$N \to \infty$. To obtain convergence in H^1 we note that, with $M > N$, in view of (6.8),

$$\alpha \|v_N - v_M\|_1^2 \le \|\nabla(v_N - v_M)\|^2 = \sum_{j=N+1}^{M} \lambda_j (v, \varphi_j)^2 \to 0 \text{ as } N \to \infty.$$

Hence, v_N is a Cauchy sequence in H^1 and converges to a limit in H^1. Clearly, this limit is the same as v. By the trace theorem (Theorem A.4) v_N is also a Cauchy sequence in $L_2(\Gamma)$, and since $v_N = 0$ on Γ we conclude that $v = 0$ on Γ. Hence, $v \in H_0^1$. Finally, (6.15) is obtained by letting $N \to \infty$ in $a(v_N, v_N) = \sum_{j=1}^{N} \lambda_j (v, \varphi_j)^2$. □

It follows immediately from (6.13) that

$$\lambda_n = \min_{\substack{(v,\varphi_j)=0,\\ j=1,\dots,n-1}} \frac{a(v,v)}{\|v\|^2}. \tag{6.17}$$

This in turn implies the following *min-max principle.*

Theorem 6.5. *We have*

$$\lambda_n = \min_{V_n} \max_{v \in V_n} \frac{a(v,v)}{\|v\|^2}, \tag{6.18}$$

where V_n varies over all subspaces of H_0^1 of finite dimension n.

Proof. Let E_n denote the n-dimensional subspace of linear combinations $v = \sum_{j=1}^{n} \alpha_j \varphi_j$ of the eigenfunctions $\varphi_1, \dots, \varphi_n$. Then clearly

$$\max_{v \in E_n} \frac{a(v,v)}{\|v\|^2} = \max_{\alpha_1,\dots,\alpha_n} \frac{\sum_{j=1}^{n} \alpha_j^2 \lambda_j}{\sum_{j=1}^{n} \alpha_j^2} = \lambda_n,$$

where the maximum is attained by φ_n. It therefore remains to show that for any V_n of dimension n,

$$\max_{v \in V_n} \frac{a(v,v)}{\|v\|^2} \ge \lambda_n.$$

To see this we choose $w \in V_n$ such that

$$(w, \varphi_j) = 0, \quad \text{for } j = 1, \dots, n-1.$$

If $\{\psi_j\}_{j=1}^{n}$ is a basis for V_n, then such a $w = \sum_{j=1}^{n} \alpha_j \psi_j$ may be determined from the linear system of equations

$$(w, \varphi_j) = \sum_{l=1}^{n} \alpha_l (\psi_l, \varphi_j) = 0, \quad j = 1, \dots, n-1,$$

which has a nonzero solution since the number of equations is smaller than n. By (6.17) it follows that

$$\frac{a(w,w)}{\|w\|^2} \geq \lambda_n,$$

which thus completes the proof of (6.18). □

One consequence of this result is that the eigenvalues depend monotonously on the underlying domain. More precisely, if $\Omega \subset \tilde{\Omega}$ and the corresponding eigenvalues are $\lambda_n(\Omega)$ and $\lambda_n(\tilde{\Omega})$, then we have $\lambda_n(\tilde{\Omega}) \leq \lambda_n(\Omega)$ for all n. In fact, by extending the functions in $H_0^1(\Omega)$ by zero in $\tilde{\Omega} \setminus \Omega$, we have $H_0^1(\Omega) \subset H_0^1(\tilde{\Omega})$, and hence the minimum in the expression for $\lambda_n(\Omega)$ in (6.18) is taken over a smaller set of n-dimensional spaces than for $\lambda_n(\tilde{\Omega})$, and hence the latter minimum is at least as small.

We now return to the more subtle mathematical points that we postponed earlier. For their treatment we shall need to use the concept of compactness, which we first briefly discuss.

We say that a set $\mathcal{M}$ in a Hilbert space H (with norm $\|\cdot\|$) is pre-compact, if every infinite sequence $\{u_n\}_{n=1}^{\infty} \subset \mathcal{M}$ contains a convergent subsequence, i.e., there is a subsequence $\{u_{n_j}\}_{j=1}^{\infty}$ and an element $\bar{u} \in H$ such that

$$\|u_{n_j} - \bar{u}\| \to 0, \quad \text{as } j \to \infty. \tag{6.19}$$

As an example we recall from elementary calculus that a bounded infinite sequence of real numbers is pre-compact (the Bolzano-Weierstrass theorem). The set $\mathcal{M}$ is called compact, if it is also a closed set, i.e., if the limit $\bar{u}$ in (6.19) always belongs to $\mathcal{M}$. Below we shall need the following result for $H_0^1 = H_0^1(\Omega)$ with $\Omega \subset \mathbf{R}^d$, the proof of which is beyond the scope of this book.

Lemma 6.2. (Rellich's lemma.) *A bounded subset $\mathcal{M}$ of H^1 is pre-compact in L_2.*

Thus if $\{u_n\}_{n=1}^{\infty} \subset H^1$ and $\|u_n\|_1 \leq C$ for $n \geq 1$, then there exists a subsequence $\{u_{n_j}\}_{j=1}^{\infty}$ and $\bar{u} \in L_2$ such that (6.19) holds in L_2-norm.

We shall now use this to prove the first statement of Theorem 6.2, that the infimum in (6.10) is attained in H_0^1. For this, we take a sequence $\{u_n\}_{n=1}^{\infty}$ such that

$$\|\nabla u_n\|^2 = a(u_n, u_n) \to \lambda_1 \quad \text{and} \quad \|u_n\| = 1, \quad \text{as } n \to \infty, \tag{6.20}$$

which is possible by the definition of the infimum. Then clearly $\{u_n\}_{n=1}^{\infty}$ is bounded in H^1, and, by Lemma 6.2, we may therefore take a subsequence, which converges to an element $\varphi_1 \in L_2$. By changing the notation if necessary, we may assume that $\{u_n\}_{n=1}^{\infty}$ itself is this subsequence, so that $\|u_n - \varphi_1\| \to 0$.

We now want to show that $\{u_n\}_{n=1}^{\infty}$ converges in H_0^1. By a simple calculation we have

$$\|\nabla(u_n - u_m)\|^2 = 2\|\nabla u_n\|^2 + 2\|\nabla u_m\|^2 - 4\|\tfrac{1}{2}\nabla(u_n + u_m)\|^2$$

(the parallellogram law), and by the definition of λ_1,

$$\|\tfrac{1}{2}\nabla(u_n + u_m)\|^2 \geq \lambda_1 \|\tfrac{1}{2}(u_n + u_m)\|^2.$$

Hence

$$\|\nabla(u_n - u_m)\|^2 \leq 2\|\nabla u_n\|^2 + 2\|\nabla u_m\|^2 - 4\lambda_1\|\tfrac{1}{2}(u_n + u_m)\|^2. \tag{6.21}$$

It is clear that $\|\frac{1}{2}(u_n + u_m)\| \to \|\varphi_1\|$ as $n, m \to \infty$, and since $\|u_n\| = 1$, we have $\|\varphi_1\| = 1$. Thus, by (6.20) the right-hand side of (6.21) tends to zero, so that $\{u_n\}_{n=1}^{\infty}$ is a Cauchy sequence in H_0^1, i.e., $\|\nabla(u_n - u_m)\| \to 0$ as $m, n \to \infty$. Since H_0^1 is a Hilbert space, the sequence thus converges to an element of H_0^1, which then has to be the same as the limit in L_2, i.e., φ_1. In particular,

$$\|\nabla\varphi_1\|^2 = \lambda_1,$$

which shows that φ_1 realizes the minimum in (6.10).

The proof that the infimum is attained in (6.12) is analogous. In fact, if $\{u_n\}_{n=1}^{\infty}$ is a minimizing sequence that converges to some φ_2 in L_2 and satisfies the side conditions in (6.12), then, since $(\frac{1}{2}(u_n + u_m), \varphi_1) = 0$, we have now (6.21) with λ_1 replaced by λ_2, and we conclude that u_n converges in H_0^1 to φ_2, and that

$$\|\varphi_2\| = 1, \quad (\varphi_2, \varphi_1) = 0, \quad \|\nabla\varphi_2\|^2 = \lambda_2.$$

We finally give the proof of Theorem 6.3. Assume then that the result is not valid, so that

$$\|\nabla\varphi_n\|^2 = \lambda_n \leq C, \quad \text{for } n \geq 1.$$

But then, by compactness, $\{\varphi_n\}_{n=1}^{\infty}$ contains a subsequence $\{\varphi_{n_j}\}_{j=1}^{\infty}$, which converges in L_2. But since $\{\varphi_n\}_{n=1}^{\infty}$ is orthonormal we have

$$\|\varphi_i - \varphi_j\|^2 = \|\varphi_1\|^2 + \|\varphi_2\|^2 = 2, \quad \text{for } i \neq j,$$

so that no convergent subsequence can exist.

As we mentioned before, the theory for the more general eigenvalue problem (6.9) is analogous. For example, for the eigenvalue problem related to the Neumann problem in Sect. 3.6, our theorems hold with $a(v, w) = \int_\Omega \left(a\,\nabla v \cdot \nabla w + c\,vw\right) \mathrm{d}x$ and H_0^1 replaced by $V = H^1$. In particular,

$$\lambda_1 = \min_{v \in V} \frac{\displaystyle\int_\Omega \left(a\,|\nabla v|^2 + c\,v^2\right) \mathrm{d}x}{\displaystyle\int_\Omega v^2\,\mathrm{d}x}.$$

We close by two examples where we can solve the eigenvalue problem explicitly.

Example 6.1. Let $\Omega = (0, b) \subset \mathbf{R}$. The problem (6.5) then reduces to

$$-u'' = \lambda u \quad \text{in } \Omega, \quad \text{with } u(0) = u(b) = 0. \tag{6.22}$$

Here we may easily determine the eigenfunctions and eigenvalues explicitly. In fact, the general solution of the differential equation in (6.22) is

$$u = C_1 \sin(\sqrt{\lambda}x) + C_2 \cos(\sqrt{\lambda}x), \quad \text{with } \lambda > 0,$$

and the boundary conditions show that $C_2 = 0$ and $\sqrt{\lambda}b = n\pi$. Hence the eigenfunctions are $\{\sin(n\pi x/b)\}_{n=1}^{\infty}$ and the corresponding eigenvalues $\lambda_n = n^2\pi^2/b^2$. After normalization we thus find $\varphi_n(x) = \sqrt{2/b}\sin(n\pi x/b)$, $n = 1, 2, \ldots$, for our orthonormal basis of eigenfunctions in $L_2(\Omega)$. Note in particular that the eigenvalues decrease with increasing b.

Example 6.2. Let $\Omega = (0, b) \times (0, b)$ and consider the eigenvalue problem

$$-\Delta u = \lambda u \quad \text{in } \Omega, \quad \text{with } u = 0 \quad \text{on } \Gamma.$$

Then with λ_n and φ_n as in Example 6.1 it is easy to check that the products $\phi_{ml}(x) = \varphi_m(x_1)\varphi_l(x_2)$, $m, l = 1, 2, \ldots$, are eigenfunctions corresponding to the eigenvalues $\lambda_{ml} = (m^2 + l^2)\pi^2/b^2$. To see that these are all the eigenfunctions, it suffices to show Parseval's relation $\sum_{m,l=1}^{\infty}(\phi_{ml}, v)^2 = \|v\|^2$, i.e.,

$$\sum_{m,l=1}^{\infty} \Big(\int_{\Omega} v(x)\varphi_m(x_1)\varphi_l(x_2)\,\mathrm{d}x \Big)^2 = \int_{\Omega} v(x)^2\,\mathrm{d}x. \tag{6.23}$$

But, using Parseval's relation in x_2 we have

$$\int_0^b v(x_1, x_2)^2\,\mathrm{d}x_2 = \sum_{l=1}^{\infty} w_l(x_1)^2, \quad w_l(x_1) = \int_0^b v(x_1, x_2)\varphi_l(x_2)\,\mathrm{d}x_2. \tag{6.24}$$

Applying Parseval's relation to $w_l(x_1)$ we find

$$\begin{aligned} \int_0^b w_l(x_1)^2\,\mathrm{d}x_1 &= \sum_{m=1}^{\infty} (w_l, \varphi_m)^2 \\ &= \sum_{m=1}^{\infty} \Big(\int_0^b \int_0^b v(x_1, x_2)\varphi_m(x_1)\varphi_l(x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 \Big)^2. \end{aligned} \tag{6.25}$$

We now integrate (6.24) in x_1 and insert (6.25) to obtain (6.23).

This shows that the eigenvalues are the numbers $\lambda_{ml} = (m^2 + l^2)\pi^2/b^2$, arranged in increasing order, and with multiple eigenvalues repeated. To determine the rate of growth of the eigenvalues λ_n as n increases, we observe

that the number of eigenvalues with $\lambda_n \le \rho^2$ is equal to the number of mesh-points $(m\pi/b, l\pi/b)$ in the disc $D_\rho = \{x_1^2 + x_2^2 \le \rho^2\}$. Since the number N_ρ of such mesh-points equals the number of mesh-squares with area π^2/b^2 that can be fitted into D_ρ, we have $N_\rho \approx \rho^2 b^2/\pi$. Hence, for λ_n corresponding to λ_{ml}, we have $\lambda_n = \lambda_{ml} \approx \rho^2 \approx \pi N_\rho/b^2 \approx \pi n/b^2$.

Since any domain $\Omega \subset \mathbf{R}^2$ contains a square and is contained in another square, it follows by the monotonicity of the eigenvalues that for any domain Ω, there are positive constants c and C such that $cn \le \lambda_n \le Cn$. In d dimensions the corresponding inequality is $cn^{2/d} \le \lambda_n \le Cn^{2/d}$.

6.2 Numerical Solution of the Eigenvalue Problem

We shall first consider the one-dimensional eigenvalue problem (6.1), i.e.,

$$\mathcal{A}\varphi := -(a\varphi')' + c\varphi = \lambda\varphi \quad \text{in } \Omega = (0,1), \quad \text{with } \varphi(0) = \varphi(1) = 0, \tag{6.26}$$

where a and c are smooth functions such that $a(x) \ge a_0$ and $c(x) \ge 0$ on $\bar{\Omega}$. To formulate a finite difference discretization we use the notation of Sect. 4.1 based on the mesh-points $x_j = jh$, $j = 0, \dots, M$, where $h = 1/M$, with $U_j \approx u(x_j)$, and consider the finite-dimensional eigenvalue problem

$$\begin{aligned} \mathcal{A}_h U_j &:= -\bar{\partial}(a_{j+1/2}\partial U_j) + c_j U_j = \Lambda U_j, \quad j = 1, \dots, M-1, \\ U_0 &= U_M = 0, \end{aligned} \tag{6.27}$$

where $c_j = c(x_j)$ and $a_{j+1/2} = a(x_j + h/2)$. The equation at the interior mesh-point x_j may then be written

$$-\big(a_{j+1/2}U_{j+1} - (a_{j+1/2} + a_{j-1/2})U_j + a_{j-1/2}U_{j-1}\big)/h^2 + c_j U_j = \Lambda U_j,$$

and thus, with A an $(M-1) \times (M-1)$ tridiagonal matrix and $\bar{U} = (U_1, \dots, U_{M-1}) \in \mathbf{R}^{M-1}$ the vector corresponding to the interior mesh-points, (6.27) may be expressed as the matrix eigenvalue problem

$$A\bar{U} = \Lambda\bar{U}.$$

For the analysis we introduce a discrete inner product and a norm by

$$(V, W)_h = h\sum_{j=0}^{M} V_j W_j \quad \text{and} \quad \|V\|_h = (V, V)_h^{1/2},$$

respectively. The operator $\mathcal{A}_h$ is easily seen to be symmetric with respect to this inner product, and positive definite since

$$(\mathcal{A}_h U, U)_h = h\sum_{j=1}^{M-1} \mathcal{A}_h U_j \, U_j = h\sum_{j=0}^{M-1} a_{j+1/2}(\partial U_j)^2 + h\sum_{j=1}^{M-1} c_j U_j^2.$$

It would be natural to expect the eigenvalues of the discrete problem (6.27) (or of the matrix A) to approximate those of the continuous eigenvalue problem (6.26). We shall show this for the principal eigenvalue Λ_1 only.

Theorem 6.6. *Let Λ_1 and λ_1 be the smallest eigenvalues of* (6.27) *and* (6.26). *Then*

$$|\Lambda_1 - \lambda_1| \le Ch^2.$$

Proof. We carry out the proof for $c = 0$ only and leave the general case to Problem 6.4. Let $U \in \mathbf{R}^{M+1}$ be arbitrary with $U_0 = U_M = 0$, and let $\tilde{u} = I_h U \in \mathcal{C}(\bar{\Omega})$ be the associated piecewise linear interpolant. Then

$$\lambda_1 \le \frac{a(\tilde{u}, \tilde{u})}{\|\tilde{u}\|^2}. \tag{6.28}$$

Since $\tilde{u}' = \partial U_j$ in (x_j, x_{j+1}) we have, using $a \ge a_0 > 0$ in Ω, that

$$\begin{aligned} a(\tilde{u}, \tilde{u}) &= \sum_{j=0}^{M-1} \int_{x_j}^{x_{j+1}} a \,\mathrm{d}x \, (\partial U_j)^2 \le h \sum_{j=0}^{M-1} (a_{j+1/2} + Ch^2)(\partial U_j)^2 \\ &\le (\mathcal{A}_h U, U)_h (1 + Ch^2). \end{aligned}$$

Further, by simple calculations,

$$\begin{aligned} \|\tilde{u}\|^2 &= \sum_{j=0}^{M-1} h^{-2} \int_{x_j}^{x_{j+1}} \big((x_{j+1} - x)U_j + (x - x_j)U_{j+1}\big)^2 \,\mathrm{d}x \\ &= \tfrac{1}{3} h \sum_{j=0}^{M-1} (U_j^2 + U_j U_{j+1} + U_{j+1}^2) \end{aligned}$$

and, since $U_0 = U_M = 0$,

$$\|U\|_h^2 = \tfrac{1}{2} h \sum_{j=0}^{M-1} (U_j^2 + U_{j+1}^2). \tag{6.29}$$

Hence, since $a_{j+1/2} \ge a_0$, $c_j \ge 0$,

$$\|U\|_h^2 - \|\tilde{u}\|^2 = \tfrac{1}{6} h \sum_{j=0}^{M-1} (U_j - U_{j+1})^2 = \tfrac{1}{6} h^3 \sum_{j=0}^{M-1} (\partial U_j)^2 \le Ch^2 (\mathcal{A}_h U, U)_h,$$

or

$$\|U\|_h^2 \le \|\tilde{u}\|^2 + Ch^2 (\mathcal{A}_h U, U)_h.$$

Hence, if U is chosen as an eigenvector corresponding to Λ_1, with $\|U\|_h = 1$, then we have for small h,

$$\frac{a(\tilde{u}, \tilde{u})}{\|\tilde{u}\|^2} \le \frac{(\mathcal{A}_h U, U)_h (1 + Ch^2)}{\|U\|_h^2 - Ch^2 (\mathcal{A}_h U, U)_h} = \frac{\Lambda_1 (1 + Ch^2)}{1 - C\Lambda_1 h^2} \le \Lambda_1 + Ch^2,$$

so that, by (6.28),

$$\lambda_1 \le \Lambda_1 + Ch^2. \tag{6.30}$$

To show the converse inequality, note that for u smooth and U defined as the restriction of u to the mesh-points, we have

$$(\mathcal{A}_h U, U)_h = a(u,u) + O(h^2) \quad \text{and} \quad \|U\|_h^2 = \|u\|^2 + O(h^2), \quad \text{as } h \to 0.$$

For the second relation we have used that, since $U_0 = U_M = 0$, (6.29) is the second order trapezoidal quadrature rule for $\int_0^1 u^2 \, \mathrm{d}x$. In particular, with $u = \varphi_1$, the principal eigenfunction of the continuous problem, we have

$$\Lambda_1 \le \frac{(\mathcal{A}_h U, U)_h}{\|U\|_h^2} \le \frac{a(\varphi_1,\varphi_1) + Ch^2}{\|\varphi_1\| - Ch^2} \le \lambda_1 + Ch^2, \quad \text{for } h \text{ small},$$

Together with (6.30) this completes the proof. □

Consider now the eigenvalue problem

$$-\Delta u = \lambda u \quad \text{in } \Omega, \quad \text{with } u = 0 \quad \text{on } \Gamma, \tag{6.31}$$

where $\Omega \subset \mathbf{R}^2$, and let λ_n be the n^{th} eigenvalue and φ_n the corresponding eigenfunction.

For the case that Ω is the square $(0,1) \times (0,1)$ we may use the five-point approximation $-\Delta_h$ defined in Chapt. 4 and pose the discrete eigenvalue problem

$$-\Delta_h u = \Lambda u \quad \text{in } \Omega, \quad \text{with } u = 0 \quad \text{on } \Gamma.$$

This may be treated as our above one-dimensional problem, but this is not so interesting, since the eigenvalues of (6.31) can be determined directly, as we have seen in Example 6.2. When Ω is a more general domain with a curved smooth boundary Γ, we encounter, as for the Dirichlet problem discussed in Chapt. 4, the difficulties caused by the fact that the uniform mesh does not fit the domain. The analysis therefore becomes involved and we shall not pursue it here.

In this case, the finite element method, with its greater flexibility, is better suited, and we shall give an elementary presentation of some simple results. We assume thus for simplicity that $\Omega \subset \mathbf{R}^2$ is a convex domain with smooth boundary Γ and denote by $\{S_h\}$ a family of spaces of continuous piecewise linear functions based on regular triangulations $\mathcal{T}_h$. The corresponding discrete eigenvalue problem is then

$$a(u_h, \chi) = \lambda(u_h, \chi), \quad \forall \chi \in S_h, \quad \text{where } a(v,w) = (\nabla v, \nabla w). \tag{6.32}$$

Using the basis $\{\Phi_i\}_{i=1}^{M_h}$ of pyramid functions from Sect. 5.2, and the positive definite matrices A and B with elements $a_{ij} = (\nabla\Phi_i, \nabla\Phi_j)$ and $b_{ij} = (\Phi_i, \Phi_j)$, respectively, this problem may be written in matrix form as

$$A\,U = \lambda B\,U. \tag{6.33}$$

Note that in contrast to the finite difference case, the matrix B is not diagonal. Nevertheless, the eigenvalue problem (6.32) or (6.33) has positive eigenvalues $\{\lambda_{n,h}\}_{n=1}^{M_h}$ and orthonormal eigenfunctions $\{\varphi_{n,h}\}_{n=1}^{M_h}$. In this case we have first the following error estimates for the eigenvalues.

Theorem 6.7. *Let $\lambda_{n,h}$ and λ_n be the n^{th} eigenvalues of* (6.32) *and* (6.31)*, respectively. Then there are constants C and h_0 (depending on n) such that*

$$\lambda_n \le \lambda_{n,h} \le \lambda_n + Ch^2, \quad \text{for } h \le h_0. \tag{6.34}$$

Proof. By the min-max principle in Theorem 6.5

$$\lambda_n = \min_{V_n \subset H_0^1} \max_{v \in V_n} \frac{\|\nabla v\|^2}{\|v\|^2}, \quad \dim V_n = n.$$

Similarly,

$$\lambda_{n,h} = \min_{V_n \subset S_h} \max_{\chi \in V_n} \frac{\|\nabla \chi\|^2}{\|\chi\|^2}, \quad \dim V_n = n. \tag{6.35}$$

Since $S_h \subset H_0^1$, the minimum in the latter expression is taken over a smaller set of subspaces than the former, and hence is at least as large, which shows the first inequality in the theorem.

To show the second inequality we note that, with E_n the space spanned by $\varphi_1, \ldots, \varphi_n$ and $E_{n,h} = R_h E_n$, where R_h is the Ritz projection,

$$\lambda_{n,h} \le \max_{\chi \in E_{n,h}} \frac{\|\nabla \chi\|^2}{\|\chi\|^2} = \max_{v \in E_n} \frac{\|\nabla R_h v\|^2}{\|R_h v\|^2} \le \max_{v \in E_n} \frac{\|\nabla v\|^2}{\|R_h v\|^2}, \tag{6.36}$$

since $\|\nabla R_h v\| \le \|\nabla v\|$. To estimate the denominator, we have

$$\|R_h v\| \ge \|v\| - \|R_h v - v\|.$$

Here, for $v \in E_n$, using Theorem 5.5 and the regularity estimate (3.36),

$$\|R_h v - v\| \le Ch^2 \|v\|_2 \le Ch^2 \|\Delta v\| \le Ch^2 \lambda_n \|v\| \le Ch^2 \|v\|,$$

where we have used that n is fixed. Hence

$$\|R_h v\| \ge \|v\|(1 - Ch^2),$$

and it follows from (6.36), for h small,

$$\lambda_{n,h} \le \max_{v \in E_n} \frac{\|\nabla v\|^2}{\|v\|^2}(1 + Ch^2) \le \lambda_n + Ch^2,$$

which completes the proof. □

A property that is sometimes used for finite element spaces $\{S_h\}$ is the so-called *inverse inequality*

$$\|\nabla\chi\| \le Ch^{-1}\|\chi\| \quad \text{for } \chi \in S_h. \tag{6.37}$$

In particular, this is valid for piecewise linear finite element spaces based on a quasi-uniform family of triangulations $\{\mathcal{T}_h\}$, see Problem 6.6. When this holds it follows immediately from (6.35) that the largest eigenvalue satisfies

$$\lambda_{M_h,h} = \max_{\chi\in S_h} \frac{\|\nabla\chi\|^2}{\|\chi\|^2} \le Ch^{-2}. \tag{6.38}$$

One may also derive error estimates for the eigenfunctions. We do this only for the first eigenvalue, because we want to avoid the complications that arise for multiple eigenvalues.

Theorem 6.8. *Let $\varphi_{1,h}$ and φ_1 be normalized eigenfunctions corresponding to the principal eigenvalues of* (6.32) *and* (6.31)*, respectively. Then*

$$\|\varphi_{1,h} - \varphi_1\| \le Ch^2 \tag{6.39}$$

and

$$\|\nabla\varphi_{1,h} - \nabla\varphi_1\| \le Ch. \tag{6.40}$$

Proof. We expand $R_h\varphi_1$ in discrete eigenfunctions,

$$R_h\varphi_1 = \sum_{j=1}^{M_h} a_j\varphi_{j,h}, \quad \text{where } a_j = (R_h\varphi_1, \varphi_{j,h}),$$

and conclude by Parseval's relation

$$\|R_h\varphi_1 - a_1\varphi_{1,h}\|^2 = \sum_{j=2}^{M_h} a_j^2. \tag{6.41}$$

Using (6.32) we find

$$\lambda_{j,h}a_j = \lambda_{j,h}(R_h\varphi_1, \varphi_{j,h}) = a(R_h\varphi_1, \varphi_{j,h}) = a(\varphi_1, \varphi_{j,h}) = \lambda_1(\varphi_1, \varphi_{j,h}),$$

and hence

$$(\lambda_{j,h} - \lambda_1)a_j = \lambda_1(\varphi_1 - R_h\varphi_1, \varphi_{j,h}).$$

Using the first inequality in (6.34) and the fact that λ_1 is a simple eigenvalue, we have $\lambda_{jh} - \lambda_1 \ge \lambda_2 - \lambda_1 > 0$ for $j \ge 2$, and we may conclude

$$\sum_{j=2}^{M_h} a_j^2 \le \sum_{j=2}^{M_h} \Big(\frac{\lambda_1}{\lambda_{j,h} - \lambda_1}\Big)^2 (\varphi_1 - R_h\varphi_1, \varphi_{j,h})^2 \le C\|R_h\varphi_1 - \varphi_1\|^2 \le Ch^4,$$

so that by (6.41)

$$\|R_h\varphi_1 - a_1\varphi_{1,h}\| \le Ch^2.$$

We therefore have,

$$\|a_1\varphi_{1,h} - \varphi_1\| \le \|R_h\varphi_1 - \varphi_1\| + \|R_h\varphi_1 - a_1\varphi_{1,h}\| \le Ch^2, \tag{6.42}$$

and it thus remains to bound $\|a_1\varphi_{1,h} - \varphi_{1,h}\| = |a_1 - 1|$. We may assume that the sign of $\varphi_{1,h}$ is chosen so that $a_1 \ge 0$. Then, by the triangle inequality and (6.42),

$$|a_1 - 1| = \big|\|a_1\varphi_{1,h}\| - \|\varphi_1\|\big| \le \|a_1\varphi_{1,h} - \varphi_1\| \le Ch^2,$$

which completes the proof of (6.39).

We now turn to the error in the gradient. We have, using (6.32) and the error bounds already derived,

$$\begin{aligned}\|\nabla\varphi_{1,h} - \nabla\varphi_1\|^2 &= \|\nabla\varphi_{1,h}\|^2 - 2(\nabla\varphi_{1,h}, \nabla\varphi_1) + \|\nabla\varphi_1\|^2 \\ &= \lambda_{1,h} - 2\lambda_1(\varphi_{1,h}, \varphi_1) + \lambda_1 = \lambda_{1,h} - \lambda_1 + \lambda_1\|\varphi_{1,h} - \varphi_1\|^2 \le Ch^2,\end{aligned}$$

which shows (6.40) and thus completes the proof of the theorem. □

6.3 Problems

Problem 6.1. Consider the problem (6.1).

(a) Show that if the functions $a(x)$ and $c(x)$ are increased then all the corresponding eigenvalues increase.
(b) Find the eigenvalues when $a(x)$ and $c(x)$ are constant on Ω.
(c) Show that for given $a(x)$ and $c(x)$ there are constants k_1 and k_2 such that

$$0 < k_1 n^2 \le \lambda_n \le k_2 n^2.$$

Problem 6.2. Consider the Laplacian in spherical symmetry, see Problem 1.4. Then the corresponding eigenvalue problem is

$$-\frac{1}{r^2}\frac{d}{dr}\Big(r^2\frac{d\varphi}{dr}\Big) = \lambda\varphi \quad \text{for } 0 < r < 1, \quad \text{with } \varphi(1) = 0,\ \varphi(0) \text{ finite.}$$

Prove that the eigenfunctions φ_j of (6.2), corresponding to different eigenvalues λ_i and λ_j, satisfy

$$\int_0^1 \varphi_i(r)\varphi_j(r)\, r^2\, \mathrm{d}r = \int_0^1 \varphi_i'(r)\varphi_j'(r)\, r^2\, \mathrm{d}r = 0,$$

i.e., $\{\varphi_i\}_{i=1}^{\infty}$ is an orthogonal set in $L_2((0,1); r^2\,\mathrm{d}r)$, the set of functions that are square integrable on $(0,1)$ with respect to the measure $r^2\,\mathrm{d}r$. Prove also that, properly normalized, the functions φ_i form an orthonormal basis for $L_2((0,1); r^2\,\mathrm{d}r)$.

Problem 6.3. Assume that Ω is such that (3.36) holds. (a) Use an argument similar to that of Theorem 6.4 to show that

$$v \in H^2 \cap H_0^1 \quad \text{if and only if} \quad \sum_{i=1}^{\infty} \lambda_i^2 (v, \varphi_i)^2 < \infty.$$

(b) Show that

$$-\Delta v = \sum_{i=1}^{\infty} \lambda_i (v, \varphi_i) \varphi_i, \quad \|\Delta v\|^2 = \sum_{i=1}^{\infty} \lambda_i^2 (v, \varphi_i)^2, \quad \text{for } v \in H^2 \cap H_0^1.$$

Problem 6.4. Prove Theorem 6.6 in the general case when the function $c(x) \geq 0$ does not necessarily vanish.

Problem 6.5. Show that the largest eigenvalue of (6.27) satisfies

$$\Lambda_{M-1} \leq CM^2,$$

with C independent of M.

Problem 6.6. Show the inverse inequality (6.37) for piecewise linear finite element functions based on a family $\{\mathcal{T}_h\}$ of quasi-uniform triangulations of a plane domain, see (5.52). Hint: Make an affine transformation $x = A\hat{x} + b$ from the small triangle K to a fixed reference triangle $\hat{K}$ of unit size, see Problem A.15, and use the fact that the norms $\|\cdot\|_{L_2(\hat{K})}$ and $\|\cdot\|_{H^1(\hat{K})}$ are equivalent on the finite-dimensional space Π_1.

Problem 6.7. Let G the Green's function in (3.18) of Sect. 3.4 and let $\{\lambda_j\}_{j=1}^{\infty}$ and $\{\varphi_j\}_{j=1}^{\infty}$ be the eigenvalues and normalized eigenfunctions of (6.5) as in Theorem 6.4. Show that

$$G(x, y) = \sum_{j=1}^{\infty} \lambda_j^{-1} \varphi_j(x) \varphi_j(y).$$

Problem 6.8. Discuss the eigenvalue problem related to the Neumann problem in Problem 3.9. Hint: The smallest eigenvalue is $\lambda_1 = 0$.

7 Initial-Value Problems for Ordinary Differential Equations

As a preparation for our study of initial-value problems for parabolic and hyperbolic differential equations we shall review in this chapter some facts about linear systems of ordinary differential equations and their numerical solution. We start with the continuous problem in Sect. 7.1 and continue in Sect. 7.2 with the numerical solution of such problems by time stepping.

7.1 The Initial Value Problem for a Linear System

We first consider the initial-value problem for the first order scalar linear ordinary differential equation

$$u' + au = f(t), \quad \text{for } t > 0, \quad \text{with } u(0) = v, \tag{7.1}$$

where a is a constant, $f(t)$ a given smooth function, and v a given number. We recall from elementary calculus that this problem may be solved by multiplication by the integrating factor e^{at}, which gives

$$(\mathrm{e}^{at}u)' = \mathrm{e}^{at}f(t),$$

from which

$$\mathrm{e}^{at}u(t) = v + \int_0^t \mathrm{e}^{as}f(s)\,\mathrm{d}s,$$

or

$$u(t) = \mathrm{e}^{-at}v + \int_0^t \mathrm{e}^{-a(t-s)}f(s)\,\mathrm{d}s. \tag{7.2}$$

We consider now the corresponding problem for a system of equations

$$\begin{aligned} u_i' + \sum_{j=1}^{N} a_{ij}u_j &= f_i(t), \qquad i = 1, \ldots, N, \text{ for } t > 0, \\ u_i(0) &= v_i, \qquad\qquad\quad i = 1, \ldots, N. \end{aligned}$$

Introducing the column vector $u = (u_1, \ldots, u_N)^{\mathrm{T}}$, and similarly for $f(t)$ and v, and also the matrix $A = (a_{ij})$, this may be written

$$u' + Au = f(t), \quad \text{for } t > 0, \quad \text{with } u(0) = v. \tag{7.3}$$

We now want to generalize the above solution method in the scalar case to the system (7.3). For this purpose we first define the exponential of an $N \times N$ matrix $B = (b_{ij})$ by means of the power series

$$\mathrm{e}^B = \exp(B) = \sum_{j=0}^{\infty} \frac{1}{j!} B^j,$$

where $B^0 = I$, the identity matrix. This definition is based on the Maclaurin expansion of e^x, and it is easily shown that the series converges for any matrix B. We note that if B_1 and B_2 are two $N \times N$ matrices which commute, i.e., such that $B_1 B_2 = B_2 B_1$, then

$$\mathrm{e}^{B_1+B_2} = \mathrm{e}^{B_1}\mathrm{e}^{B_2} = \mathrm{e}^{B_2}\mathrm{e}^{B_1} \tag{7.4}$$

In fact, since B_1 and B_2 commute, we have

$$(B_1 + B_2)^j = \sum_{l=0}^{j} \binom{j}{l} B_1^l B_2^{j-l},$$

and hence, formally,

$$\begin{aligned}\mathrm{e}^{B_1+B_2} &= \sum_{j=0}^{\infty} \frac{1}{j!} \sum_{l=0}^{j} \binom{j}{l} B_1^l B_2^{j-l} = \sum_{j=0}^{\infty} \sum_{l=0}^{j} \frac{1}{l!(j-l)!} B_1^l B_2^{j-l} \\ &= \sum_{l=0}^{\infty} \sum_{m=0}^{\infty} \frac{1}{l!\,m!} B_1^l B_2^m = \sum_{l=0}^{\infty} \frac{1}{l!} B_1^l \sum_{m=0}^{\infty} \frac{1}{m!} B_2^m = \mathrm{e}^{B_1}\mathrm{e}^{B_2}.\end{aligned}$$

Note that if B_1 and B_2 do not commute, then we have, for instance,

$$(B_1 + B_2)^2 = B_1^2 + B_1 B_2 + B_2 B_1 + B_2^2 \neq B_1^2 + 2B_1 B_2 + B_2^2.$$

Considering the matrix e^{-tA} we have for its derivative

$$\frac{d}{dt}\mathrm{e}^{-At} = \frac{d}{dt} \sum_{j=0}^{\infty} \frac{1}{j!} t^j (-A)^j = \sum_{j=1}^{\infty} \frac{1}{(j-1)!} t^{j-1} (-A)^j = -A\mathrm{e}^{-tA},$$

and hence $u(t) = \mathrm{e}^{-tA} v$ satisfies

$$u' + Au = 0, \quad \text{for } t > 0, \quad \text{with } u(0) = v. \tag{7.5}$$

Multiplication by e^{-tA} may thus be thought of as an operator $E(t)$, the solution operator of (7.5), that takes the initial data v of this problem into the solution at time t, so that $u(t) = E(t)v = \mathrm{e}^{-tA} v$. Note that, by (7.4),

$$E(t+s) = \mathrm{e}^{-(t+s)A} = \mathrm{e}^{-tA}\mathrm{e}^{-sA} = E(t)E(s), \quad \text{for } s,t \geq 0$$

which is referred to as the semigroup property of $E(t)$. This expresses the fact that the solution of (7.5) at time $t+s$ may be obtained by using the solution at time s as the initial value for (7.5) and then looking at its solution at time t. Note also that it follows from the above that A commutes with $E(t)$, so that $AE(t) = E(t)A$.

To solve the inhomogeneous equation (7.3) analogously to the above we multiply the equation by e^{tA} to obtain

$$\mathrm{e}^{tA}(u' + Au) = \mathrm{e}^{tA}f(t), \quad \text{for } t > 0.$$

This may be written

$$\frac{d}{dt}(\mathrm{e}^{tA}u) = \mathrm{e}^{tA}f(t),$$

and hence by integration

$$\mathrm{e}^{tA}u(t) = v + \int_0^t \mathrm{e}^{sA}f(s)\,\mathrm{d}s.$$

Multiplying by e^{-tA} and using (7.4), we obtain the formula analogous to (7.2),

$$u(t) = \mathrm{e}^{-tA}v + \int_0^t \mathrm{e}^{-(t-s)A}f(s)\,\mathrm{d}s. \tag{7.6}$$

In terms of the solution operator introduced above this may also be expressed as

$$u(t) = E(t)v + \int_0^t E(t-s)f(s)\,\mathrm{d}s. \tag{7.7}$$

We note that the integrand $E(t-s)f(s)$ is the solution at $t-s$ of the homogeneous equation in (7.5) with initial data $f(s)$. The integral may therefore be interpreted as the superposition of the solutions of these initial value problems, and (7.7) is often referred to as Duhamel's principle.

In some cases the system (7.3) may be reduced to a finite number of independent scalar equations of type (7.1). To see this, we assume that A is such that there is a diagonal matrix Λ and a nonsingular matrix P such that $A = P\Lambda P^{-1}$. We may then introduce the new dependent variable $w = P^{-1}u$ and the source term $g = P^{-1}f$ to find, after multiplication by P^{-1}, that the equation (7.3) may be written

$$w' + \Lambda w = g(t), \quad \text{for } t > 0, \quad \text{with } w(0) = P^{-1}v.$$

Denoting the diagonal elements of Λ by λ_i we may write this as

$$w_i' + \lambda_i w_i = g_i(t), \quad i = 1, \dots, N, \text{ for } t > 0,$$

These equations may now be solved individually and we find the solution of our original problem by taking $u = Pw$.

The assumption that A may be transformed as above to a diagonal matrix is satisfied, for example, if A is symmetric (or selfadjoint), i.e., if $a_{ij} = a_{ji}$ for all i, j, in which case $A = P\Lambda P^{\mathrm{T}}$, where P is an orthogonal matrix, so that $P^{\mathrm{T}} = P^{-1}$. In any case, the elements of Λ are the eigenvalues of A, and the method applies when A has N linearly independent eigenvectors. For large N this is not necessarily a good method for practical calculations as the diagonalization of A could be costly.

We shall now briefly study how the solutions behave for large t, and restrict ourselves to the case that A is symmetric. Let thus $A = P\Lambda P^{\mathrm{T}}$, where P is an orthogonal matrix and Λ is the diagonal matrix whose diagonal entries are the eigenvalues λ_j of A, which are real. Recall that the j^{th} column of P is the eigenvector corresponding to λ_j. Then, since $P^{\mathrm{T}}P = PP^{\mathrm{T}} = I$,

$$\mathrm{e}^{-tA} = \sum_{j=0}^{\infty} \frac{1}{j!}(-P\Lambda P^{\mathrm{T}})^j t^j = P\mathrm{e}^{-t\Lambda}P^{\mathrm{T}},$$

where $\mathrm{e}^{-t\Lambda}$ is the diagonal matrix with elements $\mathrm{e}^{-t\lambda_j}$. Recall that for a symmetric matrix the matrix norm subordinate to the Euclidean norm $|v| = (\sum_{i=1}^{N} v_i^2)^{1/2}$ of $v \in \mathbf{R}^N$ we have

$$|A| = \sup_{|v|=1} |Av| = \max_j |\lambda_j|.$$

We conclude, since $|P| = |P^{\mathrm{T}}| = 1$, that, with λ_1 the smallest eigenvalue of A,

$$|E(t)| = |\mathrm{e}^{-tA}| = \max_j \mathrm{e}^{-t\lambda_j} = \mathrm{e}^{-t\lambda_1}.$$

In particular, if all $\lambda_j \geq 0$, i.e., if A is positive semidefinite, we find from (7.7) the stability estimate

$$|u(t)| \leq |v| + \int_0^t |f(s)|\,\mathrm{d}s, \quad \text{for } t \geq 0.$$

Similarly, if A is positive definite, so that $\lambda_1 > 0$, we have

$$|u(t)| \leq \mathrm{e}^{-t\lambda_1}|v| + \int_0^t \mathrm{e}^{-(t-s)\lambda_1}|f(s)|\,\mathrm{d}s, \quad \text{for } t \geq 0.$$

We say that the system (7.5) is *stable*, or *asymptotically stable*, in these two cases, respectively. If A has a negative eigenvalue, however, we have $|\mathrm{e}^{-tA}| \to \infty$ as $t \to \infty$, and we then say that the system is *unstable*.

In the stable case, the difference between two solutions $u_1(t)$ and $u_2(t)$ remains small, if the initial data v_1 and v_2 and the source terms $f_1(t)$ and $f_2(t)$ are close. More precisely, since the difference $u_1 - u_2$ is a solution of the system with right hand side $f_1 - f_2$ and initial value $v_1 - v_2$, we have

$$|u_1(t) - u_2(t)| \le |v_1 - v_2| + \int_0^t |f_1(s) - f_2(s)| \, \mathrm{d}s, \quad \text{for } t \ge 0.$$

In the asymptotically stable case we have similarly

$$|u_1(t) - u_2(t)| = \mathrm{e}^{-t\lambda_1}|v_1 - v_2| + \int_0^t \mathrm{e}^{-(t-s)\lambda_1}|f_1(s) - f_2(s)| \, \mathrm{d}s, \quad \text{for } t \ge 0,$$

which shows, in particular, that the influence of the initial data and the value of the source terms at time s decreases exponentially as $t \to \infty$.

The above analysis does not apply if the matrix A in (7.3) depends on t. To illustrate this we consider the scalar equation

$$u' + a(t)u = f(t) \quad \text{for } t > 0, \quad \text{with } u(0) = v.$$

Let $\widetilde{a}(t) = \int_0^t a(s) \, \mathrm{d}s$ so that $\widetilde{a}'(t) = a(t)$. Then following the same steps as above, we have

$$u(t) = \mathrm{e}^{-\widetilde{a}(t)}v + \int_0^t \mathrm{e}^{-(\widetilde{a}(t)-\widetilde{a}(s))} f(s) \, \mathrm{d}s,$$

but, since in general $\widetilde{a}(t) - \widetilde{a}(s) = \int_s^t a(\tau) \, \mathrm{d}\tau \ne \int_0^{t-s} a(\tau) \, \mathrm{d}\tau = \widetilde{a}(t-s)$, the analogue of (7.7) does not hold. Instead, this time we may write

(7.8)
$$u(t) = E(t,0)v + \int_0^t E(t,s) f(s) \, \mathrm{d}s, \text{ with } E(t,s) = \exp\Big(- \int_s^t a(\tau) \, \mathrm{d}\tau \Big).$$

For the initial-value problem for the linear system

$$u' + A(t)u = f(t), \quad \text{for } t > 0,$$

where $A(t)$ is a matrix, it can be shown that the solution may again be written as in (7.8), but the matrix $E(t,s)$ will then in general have a more complicated form. It may be thought of as the operator that takes the value of the solution of the homogeneous equation $u' + A(t)u = 0$ from time s to time t, so that $u(t) = E(t,s)u(s)$. If $A(t) = A$ is independent of t, then $E(t,s)$ depends only on the difference $t - s$ and $E(t,s) = E(t-s) = e^{-(t-s)A}$.

We shall give a glimpse of the general theory for ordinary differential equations by considering the possibly nonlinear scalar initial-value problem

(7.9) $$u' = f(t,u), \quad \text{for } t > 0, \quad \text{with } u(0) = v,$$

where f is now a smooth function of t and u. The equation gives the direction of the tangent of a solution curve at any point, where the curve is defined by the points $(t, u(t)) \in \mathbf{R}^2$. To show that there exists a solution starting at $u(0) = v$, i.e., a solution curve $u(t)$ which passes through $(0, v)$, one may use *Euler's method*, which consists in approximating the solution by a polygonal curve as follows: Let k be a small time step and set $t_n = nk$, $n = 0, 1, \ldots$. Then the approximation U^n to $u(t_n)$ is defined successively by

$$\frac{U^n - U^{n-1}}{k} = f(t_{n-1}, U^{n-1}), \quad \text{for } n \geq 1, \tag{7.10}$$

or

$$U^n = U^{n-1} + kf(t_{n-1}, U^{n-1}), \quad \text{for } n \geq 1, \quad \text{with } U^0 = v.$$

This means that starting at the point (t_{n-1}, U^{n-1}), we follow the tangent direction defined by the differential equation in (7.9), and take the value at $t = t_n$ as the approximation of u at that point. The approximate solution is then the continuous, piecewise linear function which takes the value U^n at t_n. One may show by that the curves thus defined tend to a limit curve as $k \to 0$, and that this is our desired solution of (7.9). We refer to a book on ordinary differential equations for details. Another method for solving (7.9), Picard's method, is discussed in Problem 7.4.

We shall now briefly look at second order systems and begin with the simple scalar problem

$$u'' + au = 0, \quad \text{for } t > 0, \quad \text{with } u(0) = v, \quad u'(0) = w, \tag{7.11}$$

where a is a positive number. As is well-known, and easily checked, the solution of this problem is

$$u(t) = \cos(\sqrt{a}t)v + \frac{1}{\sqrt{a}} \sin(\sqrt{a}t)w, \quad \text{for } t \geq 0.$$

We next turn to the corresponding system

$$u'' + Au = 0, \quad \text{for } t > 0, \quad \text{with } u(0) = v, \quad u'(0) = w, \tag{7.12}$$

where now u is an N-vector and A is a symmetric positive definite $N \times N$ matrix. Letting $A = P\Lambda P^{\mathrm{T}}$ we may define $\sqrt{A}$ to be the positive definite matrix $P\sqrt{\Lambda}P^{\mathrm{T}}$, where $\sqrt{\Lambda}$ is the diagonal matrix with the positive square roots of the eigenvalues of A as its diagonal elements. Note that $\sqrt{A}$ has the same eigenvectors as A. Using the Euler formulas to define $\cos(B)$ and $\sin(B)$, where B is an $N \times N$ matrix, i.e.,

$$\cos B = \frac{1}{2}\big(\mathrm{e}^{iB} + \mathrm{e}^{-iB}\big), \quad \sin B = \frac{1}{2i}\big(\mathrm{e}^{iB} - \mathrm{e}^{-iB}\big),$$

we then easily find that the solution of (7.12) is

$$u(t) = \cos(t\sqrt{A})v + (\sqrt{A})^{-1}\sin(t\sqrt{A})w, \quad \text{for } t \geq 0. \tag{7.13}$$

We note that if $\{\varphi_j\}_{j=1}^N$ are the normalized eigenvectors of A corresponding to the eigenvalues $\{\lambda_j\}_{j=1}^N$, and if $v_j = (v, \varphi_j)$ and $w_j = (w, \varphi_j)$ are the components of v and w in the direction of φ_j (here $(v, w) = v^{\mathrm{T}}w$), then

$$u_j(t) = (u(t), \varphi_j) = \cos(\sqrt{\lambda_j}t)v_j + \frac{1}{\sqrt{\lambda_j}}\sin(\sqrt{\lambda_j}t)w_j, \quad \text{for } j = 1, \ldots, N.$$

These components thus vary periodically as t grows. In particular, $u(t)$ does not tend to zero as t tends to ∞, in contrast to the situation for (7.5) with A symmetric positive definite.

Another way to treat a second order system is to reduce it to first order by the introduction of a new dependent variable. Thus, in the case of (7.12), we now set $U = (U_1, U_2)^{\mathrm{T}} = (u, u')^{\mathrm{T}}$ and obtain the first order system

$$\begin{aligned} U_1' - U_2 &= 0, \\ U_2' + AU_1 &= 0, \end{aligned} \quad \text{for } t > 0, \quad \text{with } U(0) = \begin{bmatrix} v \\ w \end{bmatrix}.$$

The solution is

$$U(t) = \exp\left(t \begin{bmatrix} 0 & I \\ -A & 0 \end{bmatrix}\right) \begin{bmatrix} v \\ w \end{bmatrix}, \quad \text{for } t \geq 0. \tag{7.14}$$

It is easy to see that this implies (7.13), see Problem 7.7.

7.2 Numerical Solution of ODEs

The Euler method just described may also be used for the numerical solution of the initial value problem (7.3). Note that even for a system of ordinary differential equations with constant coefficients we may need numerical methods, because e^{-tA} may not be very easy to compute if the dimension N is large.

Let us begin with the model problem

$$u' + au = 0, \quad \text{for } t > 0, \quad \text{with } u(0) = v.$$

In this case Euler's method (7.10) gives, for the approximate solution U^n at $t_n = nk$,

$$U^n = (1 - ak)U^{n-1} = (1 - ak)^n v.$$

(In numerical analysis this method is referred to as the *forward Euler method*, since the derivative at t_{n-1} is replaced by the forward difference quotient $(U^n - U^{n-1})/k$.) We find, for $t = t_n$ a fixed time, that

$$U^n = \left(1 - \frac{t}{n}a\right)^n v \to \mathrm{e}^{-at}v, \quad \text{as } n \to \infty,$$

so that the numerical solution converges to the exact solution as $k \to 0$ in such a way that $nk = t$ is kept constant.

We shall now discuss the size of the error. Assume $a \geq 0$, i.e., that we are in the stable case for the differential equation. Now take k so small that $1 - ak \geq -1$, or $k \leq 2/a$. Then

$$|U^n| = |(1 - ak)^n v| \leq |v|, \quad \text{for } n \geq 0.$$

so that the numerical solution is also stable. Note that the requirement $ak \leq 2$ means that for large a, the time step k has to be chosen $\leq 2/a$. If k is larger, then U^n grows with n, in contrast to the behavior of the exact solution of the differential equation. We have

$$\begin{aligned} U^n - u(t_n) &= (1 - ak)^n v - (\mathrm{e}^{-ak})^n v \\ &= \big((1 - ak) - \mathrm{e}^{-ak}\big) \sum_{j=0}^{n-1} (1 - ak)^j \mathrm{e}^{-(n-1-j)ak} v. \end{aligned}$$

We find easily by Maclaurin's formula

$$|1 - x - \mathrm{e}^{-x}| \leq \tfrac{1}{2} x^2, \quad \text{for } x \geq 0,$$

and hence

$$|U^n - u(t_n)| \leq \tfrac{1}{2} a^2 k^2 \sum_{j=0}^{n-1} |v| = \tfrac{1}{2} nka^2 k|v| = (\tfrac{1}{2} t_n a^2) k|v| = C(t_n, a)k|v|,$$

so that the error is $O(k)$ as $k \to 0$ on any finite interval in time.

Recalling that the previous result is valid only under the stability condition $ak \leq 2$, we shall now consider an alternative method which does not have the latter disadvantage, namely the *backward Euler method*, in which the difference quotient is taken in the backward direction, so that U^n is defined by

$$\frac{U^n - U^{n-1}}{k} + aU^n = 0 \quad \text{for } n \geq 1, \quad \text{with } U^0 = v.$$

This time

$$U^n = \frac{1}{1 + ak} U^{n-1} = \frac{1}{(1 + ak)^n} v,$$

and, if $a \geq 0$, the stability bound $|U^n| \leq |v|$ holds for $n \geq 0$, independently of the sizes of k and a. Now

$$U^n - u(t_n) = \Big(\frac{1}{1 + ak} - \mathrm{e}^{-ak}\Big) \sum_{j=0}^{n-1} \frac{1}{(1 + ak)^j} \mathrm{e}^{-(n-1-j)ak} v. \tag{7.15}$$

Here

$$\left|\frac{1}{1+x} - \mathrm{e}^{-x}\right| \le 2x^2, \quad \text{for } x \ge 0, \tag{7.16}$$

so that now, without any restriction on k,

$$|U^n - u(t_n)| \le 2t_n a^2 k\,|v| = C(t_n, a)k|v|.$$

For numerical purposes it would be desirable to have a higher power of k than the first in the error bound. This motivates the *Crank-Nicolson method*,

$$\frac{U^n - U^{n-1}}{k} + a\frac{U^n + U^{n-1}}{2} = 0, \quad \text{for } n \ge 1, \quad \text{with } U^0 = v,$$

which implies

$$U^n = \frac{1 - \frac{1}{2}ak}{1 + \frac{1}{2}ak}U^{n-1} = \Big(\frac{1 - \frac{1}{2}ak}{1 + \frac{1}{2}ak}\Big)^n v.$$

Here again, for any k and n, we have the stability property $|U^n| \le |v|$ for $n \ge 0$. Since

$$\left|\frac{1 - \frac{1}{2}x}{1 + \frac{1}{2}x} - \mathrm{e}^{-x}\right| \le x^3, \quad \text{for } x \ge 0,$$

we now have

$$\begin{aligned} |U^n - u(t_n)| &= \left|\Big(\frac{1 - \frac{1}{2}ak}{1 + \frac{1}{2}ak} - \mathrm{e}^{-ak}\Big)\sum_{j=0}^{n-1}\Big(\frac{1 - \frac{1}{2}ak}{1 + \frac{1}{2}ak}\Big)^j \mathrm{e}^{-(n-1-j)ak}v\right| \\ &\le a^3k^3\sum_{j=0}^{n-1}|v| = t_n a^3 k^2|v| = C(t_n, a)k^2|v|. \end{aligned}$$

The error thus tends to zero as $O(k^2)$ rather than $O(k)$.

In all the above error estimates the constants on the right grow with a. We shall now demonstrate that if the backward Euler rule is used, then one may show an error bound which is independent of a. This is convenient if a is allowed to become very large. We shall show

$$|U^n - u(t_n)| \le Ct_n^{-1}k|v|, \tag{7.17}$$

where C is independent of a and t_n. For fixed $t_n = t$ positive, this thus shows $O(k)$ convergence, uniformly in a. To prove (7.17), consider first $ak \ge 1$, say. Then

$$|U^n| = \frac{1}{(1+ak)^n}|v| \le 2^{-n}|v|. \tag{7.18}$$

But, for a suitable C_1, we have

$$2^{-n} \le C_1/n = C_1 t_n^{-1}k.$$

Further

$$|u(t_n)| = \mathrm{e}^{-nak}|v| \le \mathrm{e}^{-n}|v| \le C_2 n^{-1}|v| = C_2 t_n^{-1} k|v|.$$

so that (7.17) holds by the triangle inequality.

In order to treat the case $ak \le 1$, we note that, for suitable γ with $0 < \gamma \le 1$,

$$\frac{1}{1+x} \le \mathrm{e}^{-\gamma x}, \quad \text{for } 0 \le x \le 1,$$

so that

$$\frac{1}{(1+ak)^j} \le \mathrm{e}^{-\gamma jak}.$$

Hence, using (7.15) and (7.16), we obtain

$$\begin{aligned} |U^n - u(t_n)| &\le 2a^2k^2 \sum_{j=0}^{n-1} \mathrm{e}^{-\gamma jak}\mathrm{e}^{-\gamma(n-1-j)ak} = 2a^2k^2 n\mathrm{e}^{-\gamma(n-1)ak} \\ &\le 2\mathrm{e}^{\gamma} a^2 t_n \mathrm{e}^{-t_n\gamma a} k \le C_3 t_n^{-1} k, \quad \text{where } C_3 = 2\ \mathrm{e}^{\gamma} \sup_{x\ge 0} x^2\mathrm{e}^{-\gamma x}. \end{aligned}$$

Together our estimates complete the proof of (7.17).

This property is not valid for the Crank-Nicolson method, because the analogue of (7.18) does not hold, since $|(1 - \frac{1}{2}ak)/(1 + \frac{1}{2}ak)|$ tends to 1 as ak tends to ∞.

The strong stability property just described for the backward Euler method is useful when treating systems of the form

$$u' + Au = f(t), \quad \text{for } t > 0,$$

where A is a symmetric positive definite $N \times N$ matrix, which is not well conditioned, i.e., for which the ratio between the largest and smallest eigenvalues is large. Such a system is said to be a *stiff* system of ordinary differential equations. The backward Euler method in this case is defined by the equations

$$(I + kA)U^n = U^{n-1} + kf(t_n), \quad \text{for } n \ge 1, \quad \text{with } U^0 = v, \tag{7.19}$$

and we thus have to solve a system of equations at each time step. We therefore say that this method is *implicit*. This is in contrast to the *explicit* forward Euler method

$$U^n = (I - kA)U^{n-1} + kf(t_{n-1}), \quad \text{for } n \ge 1, \quad \text{with } U^0 = v, \tag{7.20}$$

which, however, has the drawbacks concerning stability described above.

The system (7.19) may be written

$$U^n = (I + kA)^{-1}U^{n-1} + k(I + kA)^{-1}f(t_n), \quad \text{for } n \ge 1,$$

and we note that when A is symmetric positive definite we have

$$|(I + kA)^{-1}| = \max_j \frac{1}{1 + k\lambda_j} = \frac{1}{1 + k\lambda_1} < 1, \quad \text{for any } k > 0, \tag{7.21}$$

where λ_1 is the smallest eigenvalue of A.

Thus, for the homogeneous system, i.e., when $f = 0$, we have $|U^n| = |(I + kA)^{-n}v| \to 0$ as $n \to \infty$. The numerical solution therefore reproduces the asymptotic behavior of the differential equations. On the other hand, for the matrix occurring in (7.20), we have

$$|I - kA| = \max_j |1 - k\lambda_j|,$$

which is less than 1 only if $1 - k\lambda_N > -1$, i.e., if $k < 2/\lambda_N$, which could be a very restrictive condition, where λ_N is the largest eigenvalue of A.

In the case of a system the Crank-Nicolson method may be written

$$(I + \tfrac{1}{2}kA)U^n = (I - \tfrac{1}{2}kA)U^{n-1} + kf(t_{n-1/2}),$$

where $t_{n-1/2} = (n - 1/2)k$, or

$$U^n = (I + \tfrac{1}{2}kA)^{-1}(I - \tfrac{1}{2}kA)U^{n-1} + k(I + \tfrac{1}{2}kA)^{-1}f(t_{n-1/2}),$$

and it is thus an implicit method. Here

$$|(I + \tfrac{1}{2}kA)^{-1}(I - \tfrac{1}{2}kA)| = \max_j \left|\frac{1 - \frac{1}{2}k\lambda_j}{1 + \frac{1}{2}k\lambda_j}\right| < 1,$$

for all k. As pointed out previously, for fixed k, the norm tends to 1 if $\lambda_{\max} \to \infty$, which is less satisfactory than (7.21).

We close with a short discussion of numerical methods for the initial value problem for the second order scalar equation (7.11). We first replace the second derivative by a symmetric difference quotient and obtain

$$\frac{U^{n+1} - 2U^n + U^{n-1}}{k^2} + aU^n = 0, \quad \text{for } n \geq 1,$$

with, for instance, the initial conditions

$$U^0 = v, \ \frac{U^1 - U^0}{k} = w.$$

The difference equation may also be written

$$U^{n+1} - 2\mu U^n + U^{n-1} = 0, \quad \text{where } \mu = 1 - ak^2/2, \tag{7.22}$$

and this difference equation has the characteristic equation

$$\tau^2 - 2\mu\tau + 1 = 0.$$

If its roots $\tau_{1,2}$ are distinct, then the solution of (7.22) is of the form

$$U^n = c_1\tau_1^n + c_2\tau_2^n, \tag{7.23}$$

with c_1 and c_2 determined by the initial conditions. For $|\mu| < 1$, i.e., for $ak^2 < 4$, the roots are distinct and $|\tau_1| = |\tau_2| = 1$ so that stability holds,

$$|U^n| \le C(|v| + |w|), \quad \text{for } n \ge 0.$$

On the other hand, if $|\mu| > 1$, or $ak^2 > 4$, then we have $|\tau_1| > 1$ and $|\tau_2| < 1$. In this case the general solution of (7.22) increases exponentially. If instead we consider the implicit method,

$$\frac{U^{n+1} - 2U^n + U^{n-1}}{k^2} + a\,U^{n+1} = 0, \quad \text{for } n \ge 1,$$

then the characteristic equation becomes

$$\nu\tau^2 - 2\tau + 1 = 0, \quad \text{where } \nu = 1 + ak^2.$$

The roots are now less than one in modulus for any choice of k and a and the stability is secured. However, this method is only first order accurate because of the lack of symmetry in the difference approximation. The method

$$\frac{U^{n+1} - 2U^n + U^{n-1}}{k^2} + a\left(\frac{1}{4}U^{n+1} + \frac{1}{2}U^n + \frac{1}{4}U^{n-1}\right)U^{n+1} = 0, \quad \text{for } n \ge 1,$$

is second order accurate and stable for any k and a because the characteristic equation

$$\tau^2 - 2\kappa\tau + 1 = 0, \quad \text{where } \kappa = (1 - \tfrac{1}{4}ak^2)/(1 + \tfrac{1}{4}ak^2).$$

has distinct roots with $|\tau_1| = |\tau_2| = 1$.

7.3 Problems

Problem 7.1. Solve the initial value problem

$$u'(t) = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} u(t), \quad \text{for } t > 0, \quad \text{with } u(0) = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

Problem 7.2. (Computer exercise.) Find an approximate solution at $t = 1$ of Problem 7.1 by the forward and backward Euler methods and by the Crank-Nicolson method for $k = 1/10,\ 1/100$. Compare with the exact solution.

Problem 7.3. Solve the initial value problem

$$u'(t) = \begin{bmatrix} 1 & 2t \\ 2t & 1 \end{bmatrix} u(t), \quad \text{for } t > 0, \quad \text{with } u(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Problem 7.4. (Picard's method.) Prove the existence of solutions to (7.9) as follows: Show first that a solution of the initial value problem (7.9) satisfies the integral equation

$$u(t) = v + \int_0^t f(s, u(s))\,\mathrm{d}s =: T(u)(t),$$

and, conversely, that a solution of this integral equation is a solution of (7.9). Assume that the continuous function $f(t,u)$ satisfies a global Lipschitz condition with respect to the second variable, i.e.,

$$|f(t,v) - f(t,w)| \le K|v-w|, \quad \forall v, w \in \mathbf{R},\ 0 \le t \le a.$$

Show that the sequence u_n, $n = 0, 1, \ldots$, defined by

$$u_0(t) = v, \quad u_{n+1}(t) = T(u_n)(t), \quad \text{for } n \ge 0,$$

satisfies

$$|u_{n+1}(t) - u_n(t)| \le CK^n a^{n+1}/(n+1)!,$$

that this implies that $\sum_{n=0}^{\infty}(u_{n+1}(t) - u_n(t))$ converges uniformly to $u(t) - v$ for $0 \le t \le a$, that $u \in C([0,a])$, and that $u = T(u)$. In particular, this implies that u satisfies (7.9). Show finally that $f(t,u)$ satisfies a Lipschitz condition with respect to the second variable if $\partial f/\partial u$ is bounded.

Problem 7.5. Prove a uniqueness result for (7.9), when $f(t,u)$ is a continuous function which satisfies a Lipschitz condition with respect to the second variable (cf. Problem 7.4). Hint: Assume that u_1 and u_2 are two solutions, which both satisfy the integral equation of Problem 7.4, and use the Lipschitz condition to derive an inequality which shows that $u_1 - u_2 = 0$.

Problem 7.6. (a) (Gronwall's lemma.) Suppose that φ is a nonnegative continuous function such that

$$\varphi(t) \le a + b\int_0^t \varphi(s)\,\mathrm{d}s, \quad \text{for } t > 0,$$

where a and b are nonnegative constants. Prove that

$$\varphi(t) \le a\,\mathrm{e}^{bt}, \quad \text{for } t > 0.$$

(b) Use Gronwall's lemma to show that the solution of (7.3) satisfies

$$|u(t)| \le \mathrm{e}^{|A|T}\Big(|v| + \int_0^T |f(s)|\,\mathrm{d}s\Big), \quad \text{for } 0 \le t \le T.$$

Show that this implies uniqueness of the solution.

Problem 7.7. Show that (7.13) and (7.14) are equivalent.

Problem 7.8. Prove that the general solution of (7.22) is (7.23), if $\tau_1 \neq \tau_2$. Show also that

$$|\mu| < 1 \text{ implies } |\tau_1| = |\tau_2| = 1,$$
$$|\mu| > 1 \text{ implies } |\tau_1| < 1,\ |\tau_2| > 1.$$

What is the general form of the solution if $\tau_1 = \tau_2$?

8 Parabolic Equations

In this chapter we study both the pure initial value problem and the mixed initial-boundary value problem for the model heat equation, using Fourier techniques as well as energy arguments. In Sect. 8.1 we analyze the solution of the pure initial value problem for the homogeneous heat equation by means of a representation in terms of the Gauss kernel, and use it to investigate properties of the solution. In the remainder of the chapter we consider the initial-boundary value problem in a bounded spatial domain. In Sect. 8.2 we solve the homogeneous equation by means of eigenfunction expansions, and apply Duhamel's principle to find a solution of the inhomogeneous equation. In Sect. 8.3 we introduce the variational formulation of the problem and give examples of the use of energy arguments, and in Sect. 8.4 we show and apply the maximum principle.

8.1 The Pure Initial Value Problem

We begin our study of parabolic equations by considering the pure initial value problem (or the Cauchy problem) for the heat equation, which is to find $u(x,t)$ such that

$$\begin{aligned} &\frac{\partial u}{\partial t} - \Delta u = 0, && \text{in } \mathbf{R}^d \times \mathbf{R}_+, \\ &u(\cdot,0) = v, && \text{in } \mathbf{R}^d. \end{aligned} \tag{8.1}$$

We shall employ the Fourier transform of u with respect to x, cf. App. A.3,

$$\hat{u}(\xi,t) = \mathcal{F}u(\cdot,t)(\xi) = \int_{\mathbf{R}^d} u(x,t)\mathrm{e}^{-\mathrm{i}x\cdot\xi}\,\mathrm{d}x, \quad \text{for } \xi \in \mathbf{R}^d.$$

If u and its derivatives are small enough for large $|x|$, then we have

$$\big(\mathcal{F}\Delta u(\cdot,t)\big)(\xi) = \int_{\mathbf{R}^d} \Delta u(x,t)\mathrm{e}^{-\mathrm{i}x\cdot\xi}\,\mathrm{d}x = -|\xi|^2\hat{u}(\xi,t)$$

and, with $u_t = \partial u/\partial t$,

$$\big(\mathcal{F}u_t(\cdot,t)\big)(\xi) = \frac{d\hat{u}}{dt}(\xi,t).$$

Hence we conclude from (8.1) that $\hat{u}$ satisfies

$$\begin{aligned} \frac{d\hat{u}}{dt} &= -|\xi|^2\hat{u}, \qquad \text{for } \xi \in \mathbf{R}^d,\ t > 0, \\ \hat{u}(\xi, 0) &= \hat{v}(\xi), \qquad \text{for } \xi \in \mathbf{R}^d. \end{aligned}$$

This is a simple initial value problem for a first order linear ordinary differential equation, with ξ as a parameter, and its solution is

$$\hat{u}(\xi, t) = \hat{v}(\xi)\mathrm{e}^{-t|\xi|^2}. \tag{8.2}$$

Recalling that $w(x) = \mathrm{e}^{-|x|^2}$ has the Fourier transform

$$\hat{w}(\xi) = \pi^{d/2}\mathrm{e}^{-|\xi|^2/4}$$

(cf. Problem A.19), we conclude from (A.34) that $\mathrm{e}^{-t|\xi|^2}$ is the Fourier transform of the *Gauss kernel*

$$U(x, t) = (4\pi t)^{-d/2}\mathrm{e}^{-|x|^2/4t},$$

and hence we obtain formally from (8.2) that

$$u(x, t) = \big(U(\cdot, t) * v\big)(x) = (4\pi t)^{-d/2} \int_{\mathbf{R}^d} v(y)\mathrm{e}^{-|x-y|^2/4t}\,\mathrm{d}y. \tag{8.3}$$

The function $U(x, t)$ is a *fundamental solution* of the initial value problem. We shall now show that the function defined in (8.3) is, in fact, a solution of (8.1) under a weak assumption on the initial function. Note that $U(x, t)$ and $u(x, t)$ in (8.3) are only defined for $t > 0$.

Theorem 8.1. *If v is a bounded continuous function on $\mathbf{R}^d$, then the function $u(x, t)$ defined by* (8.3) *is a solution of the heat equation for $t > 0$, and tends to the initial data v as t tends to* 0.

Proof. We first note that for $t > 0$ we may differentiate the integral in (8.3) with respect to x and t under the integral sign, and show directly that this function satisfies the heat equation in (8.1). To see that $u(x, t)$ tends to the desired initial values as $t \to 0$ we let $x_0 \in \mathbf{R}^d$ be arbitrary and show that

$$u(x, t) \to v(x_0), \qquad \text{as } \ (x, t) \to (x_0, 0).$$

In fact, using the transformation $\eta = (y - x)/\sqrt{4t}$, and the formula

$$\pi^{-d/2} \int_{\mathbf{R}^d} \mathrm{e}^{-|x|^2}\,\mathrm{d}x = 1, \tag{8.4}$$

we may write

$$u(x,t) - v(x_0) = (4\pi t)^{-d/2} \int_{\mathbf{R}^d} v(y) \mathrm{e}^{-|x-y|^2/4t} \, \mathrm{d}y - v(x_0)$$
$$= \pi^{-d/2} \int_{\mathbf{R}^d} \big(v(x + \sqrt{4t}\eta) - v(x_0)\big) \, \mathrm{e}^{-|\eta|^2} \, \mathrm{d}\eta.$$

Let $\epsilon > 0$ be arbitrary and let δ be so small that

$$|v(z) - v(x_0)| < \epsilon, \quad \text{if } |z - x_0| < \delta. \tag{8.5}$$

Let $M = \|v\|_{\mathcal{C}} = \|v\|_{\mathcal{C}(\mathbf{R}^d)}$. For any $\omega > 0$, we have

$$|u(x,t) - v(x_0)| \le 2M\pi^{-d/2} \int_{|y|>\omega} \mathrm{e}^{-|y|^2} \, \mathrm{d}y$$
$$+ \pi^{-d/2} \int_{|y|<\omega} \big|v(x + \sqrt{4t}y) - v(x_0)\big| \, \mathrm{e}^{-|y|^2} \, \mathrm{d}y = I + II.$$

We now fix ω so large that $I < \epsilon$, which is possible in view of (8.4). Then, with ω fixed, we obtain, using (8.5) and (8.4),

$$II \le \sup_{|y|<\omega} |v(x + \sqrt{4t}y) - v(x_0)| < \epsilon, \quad \text{if } |x - x_0| + \sqrt{4t}\omega < \delta.$$

Hence, for these x, t we have

$$|u(x,t) - v(x_0)| < 2\epsilon,$$

which completes the proof. □

Theorem 8.1 thus shows that the initial value problem (8.1) admits a solution, and is therefore an existence theorem. We shall show that this solution depends continuously on the initial data v.

We write (8.3) in the form

$$u(x,t) = (E(t)v)(x) = (4\pi t)^{-d/2} \int_{\mathbf{R}^d} v(y) \mathrm{e}^{-|x-y|^2/4t} \, \mathrm{d}y, \tag{8.6}$$

where we may think of $E(t)$ as defining a linear operator, the solution operator of (8.1), which takes the given initial data into the solution at time t.

Note that by (8.4), with $\|v\|_{\mathcal{C}} = \|v\|_{\mathcal{C}(\mathbf{R}^d)}$,

$$|u(x,t)| \le (4\pi t)^{-d/2} \int_{\mathbf{R}^d} \mathrm{e}^{-|x-y|^2/4t} \, \mathrm{d}y \, \|v\|_{\mathcal{C}} = \|v\|_{\mathcal{C}},$$

so that

$$\|u(\cdot,t)\|_{\mathcal{C}} \le \|v\|_{\mathcal{C}}, \quad \text{for } t > 0.$$

This shows that the operator $E(t)$ is bounded with respect to the maximum-norm, with operator norm 1, which is the first part of the following result.

Theorem 8.2. *The solution operator $E(t)$ defined by* (8.6) *is bounded in $\mathcal{C}$, and*

$$\|E(t)v\|_{\mathcal{C}} \le \|v\|_{\mathcal{C}}, \quad \text{for } t \ge 0. \tag{8.7}$$

If v_1 and v_2 are two bounded continuous functions on $\mathbf{R}^d$ and u_1 and u_2 are the corresponding solutions of the initial value problem (8.1)*, then*

$$\|u_1(t) - u_2(t)\|_{\mathcal{C}} \le \|v_1 - v_2\|_{\mathcal{C}}, \quad \text{for } t \ge 0. \tag{8.8}$$

Proof. It remains only to show the second part of the theorem. But, since $E(t)$ is a linear operator,

$$u_1(t) - u_2(t) = E(t)v_1 - E(t)v_2 = E(t)(v_1 - v_2),$$

and hence (8.8) follows at once from (8.7). □

By using a maximum principle we shall prove in Sect. 8.4 the corresponding uniqueness result, i.e., that there exists at most one bounded solution of (8.1) and thus (8.3) is the only one.

Together the existence, uniqueness, and continuous dependence properties make the problem (8.1) a *well posed problem*. In particular, the continuous dependence property is important in applications. It shows that a small change in the data of the problem has only a small effect on the solution.

Not all problems which admit solutions have this continuous dependence property. Consider for example the initial value problem

$$\begin{aligned} &u_t + u_{xx} = 0, && \text{in } \mathbf{R} \times \mathbf{R}_+, \\ &u(x,0) = v_n(x) = n^{-1}\sin(nx), && \text{for } x \in \mathbf{R}, \end{aligned} \tag{8.9}$$

which has the solution

$$u_n(x,t) = n^{-1}\mathrm{e}^{n^2 t}\sin(nx).$$

Here

$$\|v_n\|_{\mathcal{C}} = n^{-1} \to 0, \quad \text{as } n \to \infty,$$

whereas, for any $t > 0$,

$$\|u_n(t)\|_{\mathcal{C}} = n^{-1}\mathrm{e}^{n^2 t} \to \infty, \quad \text{as } n \to \infty.$$

Hence, although the initial value v_n is close to 0, the solution is not for $t > 0$.

The differential equation in (8.9) is the heat equation with the sign for the time derivative reversed, i.e., the *backward heat equation*. The result therefore means that the problem of determining an earlier distribution of heat in a body from the present one is *ill posed*.

We have already noted above that the representation of $u(x,t)$ in terms of v in (8.3) allows differentiation with respect to both x and t under the

integral sign for $t > 0$, even without regularity assumptions on v. In fact, this differentiation can be carried out an arbitrary number of times so that u is infinitely differentiable, $u \in \mathcal{C}^\infty$, for $t > 0$. Using the multi-index notation from (1.8), one finds easily

$$\begin{aligned} |D_t^j D^\alpha U(x,t)| &\le t^{-j-|\alpha|/2-d/2} P(|x|/\sqrt{4t}) \mathrm{e}^{-|x|^2/4t} \\ &\le C t^{-j-|\alpha|/2-d/2} \mathrm{e}^{-|x|^2/8t}, \end{aligned}$$

where $P(y)$ is a polynomial in y, and where we have used the fact that for any polynomial P there is a C such that

$$|P(y)\mathrm{e}^{-y^2}| \le C\mathrm{e}^{-y^2/2}, \qquad \text{for } y > 0.$$

Hence

$$\begin{aligned} \sup_{x\in\mathbf{R}^d} |D_t^j D^\alpha u(x,t)| &\le C t^{-j-|\alpha|/2-d/2} \sup_{x\in\mathbf{R}^d} \int_{\mathbf{R}^d} |v(y)| \mathrm{e}^{-|x-y|^2/8t} \, \mathrm{d}y \\ &\le C t^{-j-|\alpha|/2} \sup_{y\in\mathbf{R}^d} |v(y)|, \end{aligned}$$

or

$$\|D_t^j D^\alpha E(t) v\|_{\mathcal{C}} \le C t^{-j-|\alpha|/2} \|v\|_{\mathcal{C}}, \qquad \text{for } t > 0,$$

which shows that the operator $E(t)$ has a *smoothing property*: The solution of (8.1) is smooth for $t > 0$ even if v is nonsmooth. However, the bounds for the derivatives then grow as t tends to zero.

On the other hand, if the initial values are smooth then the derivatives of the solution are bounded uniformly down to $t = 0$: We have from (8.6), after the change of variables $z = x - y$,

$$\begin{aligned} \big(D^\alpha E(t) v\big)(x) = D_x^\alpha u(x,t) &= (4\pi t)^{-d/2} D_x^\alpha \int_{\mathbf{R}^d} v(x-z) \mathrm{e}^{-|z|^2/4t} \, dz \\ &= (4\pi t)^{-d/2} \int_{\mathbf{R}^d} D_x^\alpha v(x-z) \mathrm{e}^{-|z|^2/4t} \, dz = (E(t) D^\alpha v)(x), \end{aligned}$$

and hence, by (8.1) and (8.7),

$$\|D_t^j D^\alpha E(t) v\|_{\mathcal{C}} = \|\Delta^j D^\alpha E(t) v\|_{\mathcal{C}} = \|E(t) \Delta^j D^\alpha v\|_{\mathcal{C}} \le \|\Delta^j D^\alpha v\|_{\mathcal{C}}.$$

It can be shown that the solution of the initial value problem for the inhomogeneous heat equation,

$$\begin{aligned} u_t - \Delta u &= f, && \text{in } \mathbf{R}^d \times \mathbf{R}_+, \\ u(\cdot, 0) &= v, && \text{in } \mathbf{R}^d, \end{aligned}$$

where $f = f(x,t)$ is given, may be represented in the form

$$u(x,t) = \int_{\mathbf{R}^d} v(y)U(x-y,t)\,\mathrm{d}y + \int_0^t \int_{\mathbf{R}^d} f(y,s)U(x-y,t-s)\,\mathrm{d}y\,\mathrm{d}s$$
$$= E(t)v + \int_0^t E(t-s)f(\cdot,s)\,\mathrm{d}s,$$

provided, e.g., that v, f, and ∇f are continuous and bounded.

8.2 Solution of the Initial-Boundary Value Problem by Eigenfunction Expansion

We shall first consider the mixed initial-boundary value problem for the homogeneous heat equation: Find $u(x,t)$ such that

$$\begin{aligned} &u_t - \Delta u = 0 && \text{in } \Omega \times \mathbf{R}_+, \\ &u = 0, && \text{on } \Gamma \times \mathbf{R}_+, \\ &u(\cdot,0) = v && \text{in } \Omega, \end{aligned} \tag{8.10}$$

where Ω is a bounded domain in $\mathbf{R}^d$ with smooth boundary Γ, $u_t = \partial u/\partial t$, and v is a given function in $L_2 = L_2(\Omega)$. We shall now solve this problem by using eigenfunction expansions. We denote by $(\cdot,\cdot)$ and $\|\cdot\|$ the inner product and norm in $L_2 = L_2(\Omega)$, respectively.

We recall from Chapt. 6 that there exists an orthonormal basis $\{\varphi_i\}_{i=1}^\infty$ in L_2 of smooth eigenfunctions φ_i and corresponding eigenvalues $\{\lambda_i\}_{i=1}^\infty$ satisfying

$$-\Delta\varphi_i = \lambda_i\varphi_i \quad \text{in } \Omega, \quad \text{with } \varphi_i = 0 \quad \text{on } \Gamma, \tag{8.11}$$

or, equivalently, with our usual notation

$$a(\varphi_i, v) = \int_\Omega \nabla\varphi_i \cdot \nabla v\,\mathrm{d}x = \lambda_i(\varphi_i, v), \quad \forall v \in H_0^1,$$

Recall that $0 < \lambda_1 < \lambda_2 \le \cdots \le \lambda_i \le \cdots$, that $\lambda_i \to \infty$ as $i \to \infty$, and that (with Kronecker's symbol $\delta_{ij} = 1$ for $j = i$ and 0 otherwise)

$$a(\varphi_i, \varphi_j) = \lambda_i\delta_{ij}.$$

We now seek a solution to (8.10) of the form

$$u(x,t) = \sum_{i=1}^\infty \hat{u}_i(t)\varphi_i(x), \tag{8.12}$$

where the $\hat{u}_i : \mathbf{R}_+ \to \mathbf{R}$ are coefficients to be determined. Because this is a sum of products of functions of x and t this approach is also called the method of *separation of variables*. Inserting (8.12) into the differential equation in (8.10) and using (8.11) we obtain formally

$$\sum_{i=1}^{\infty} \big(\hat{u}_i'(t) + \lambda_i \hat{u}_i(t)\big)\varphi_i(x) = 0, \quad \text{for } x \in \Omega,\ t \in \mathbf{R}_+,$$

and hence, since the φ_i form a basis,

$$\hat{u}_i'(t) + \lambda_i \hat{u}_i(t) = 0, \quad \text{for } t \in \mathbf{R}_+, \quad i = 1, 2, \ldots,$$

so that

$$\hat{u}_i(t) = \hat{u}_i(0)\mathrm{e}^{-\lambda_i t}.$$

Moreover, from the initial condition in (8.10) it follows that

$$u(\cdot, 0) = \sum_{i=1}^{\infty} \hat{u}_i(0)\varphi_i = v = \sum_{i=1}^{\infty} \hat{v}_i \varphi_i, \quad \text{where } \hat{v}_i = (v, \varphi_i) = \int_\Omega v\, \varphi_i \, \mathrm{d}x.$$

We thus see that, at least formally, the solution of (8.10) has to be

$$u(x,t) = \sum_{i=1}^{\infty} \hat{v}_i \mathrm{e}^{-\lambda_i t}\varphi_i(x), \tag{8.13}$$

where by Parseval's relation, with $\|\cdot\| = \|\cdot\|_{L_2}$,

$$\|u(\cdot,t)\|^2 = \sum_{i=1}^{\infty} \big(\hat{v}_i \mathrm{e}^{-\lambda_i t}\big)^2 \le \mathrm{e}^{-2\lambda_1 t} \sum_{i=1}^{\infty} \hat{v}_i^2 = \mathrm{e}^{-2\lambda_1 t}\|v\|^2 < \infty,$$

Thus $u(\cdot,t) \in L_2$ for $t \ge 0$, and its L_2-norm decreases exponentially as $t \to \infty$. Although this decay is important in some situations, for simplicity we shall refrain from keeping track of the behavior of $u(\cdot,t)$ for large t in the sequel and content ourselves with the conclusion that

$$\|u(\cdot,t)\| \le \|v\|, \quad \text{for } t \in \mathbf{R}_+.$$

We now show that for $t > 0$ the function $u(\cdot,t)$ defined in (8.13) is smooth and satisfies the differential equation and the boundary condition in (8.10) in the classical sense, and that the initial condition holds in the sense that

$$\|u(\cdot,t) - v\| \to 0, \quad \text{as } t \to 0. \tag{8.14}$$

We first note that for any $k \ge 0$ there is a constant C_k such that $s^k \mathrm{e}^{-s} \le C_k$ for $s \ge 0$, Using this with $k = 1$ we have

$$|u(\cdot,t)|_1^2 = \sum_{i=1}^{\infty} \lambda_i \big(\hat{v}_i \mathrm{e}^{-\lambda_i t}\big)^2 = t^{-1} \sum_{i=1}^{\infty} \hat{v}_i^2 (\lambda_i t)\ \mathrm{e}^{-2\lambda_i t} \le C_1 t^{-1}\|v\|^2,$$

so that

$$|u(\cdot,t)|_1 \le C t^{-1/2}\|v\|, \quad \text{for } t > 0. \tag{8.15}$$

Thus $u(\cdot,t) \in H_0^1$ for $t > 0$, by Theorem 6.4, and, in particular, $u(\cdot,t)$ satisfies the boundary condition in (8.10). Now, applying $(-\Delta)^k$ to each term in (8.13), we obtain since $-\Delta\varphi_i = \lambda_i\varphi_i$

$$(-\Delta)^k u(x,t) = \sum_{i=1}^{\infty} \hat{v}_i \lambda_i^k \mathrm{e}^{-\lambda_i t}\varphi_i(x), \tag{8.16}$$

and hence, for $t > 0$,

$$\|\Delta^k u(\cdot,t)\|^2 = \sum_{i=1}^{\infty} \left(\hat{v}_i\lambda_i^k \mathrm{e}^{-\lambda_i t}\right)^2 \le C_k^2 t^{-2k} \sum_{i=1}^{\infty} \hat{v}_i^2 = C_k^2 t^{-2k}\|v\|^2 < \infty.$$

In the same way as in (8.15), we also have

$$|\Delta^k u(\cdot,t)|_1 \le C_k t^{-k-1/2}\|v\| < \infty, \quad \text{for } t > 0,$$

and thus $\Delta^k u(\cdot,t) = 0$ on Γ for any $k \ge 0$ when $t > 0$. We may also apply D_t^m to each term in (8.16), and since $D_t \mathrm{e}^{-\lambda_i t} = -\lambda_i \mathrm{e}^{-\lambda_i t}$, we obtain

$$|D_t^m \Delta^k u(\cdot,t)|_\delta \le C t^{-m-k-\delta/2}\|v\| < \infty, \quad \text{for } t > 0, \quad \delta = 0, 1.$$

Recall from the theory of elliptic equations the regularity estimate (3.37),

$$\|w\|_s \le C\|\Delta w\|_{s-2}, \quad \forall w \in H^s \cap H_0^1, \quad \text{for } s \ge 2.$$

By repeated application of this we obtain, again for $\delta = 0$ or 1,

$$\|w\|_{2k+\delta} \le C\|\Delta^k w\|_\delta, \quad \forall w \in H^{2k+\delta}, \quad \text{if } \Delta^j w = 0 \text{ on } \Gamma \text{ for } j < k,$$

and we finally conclude that, for any nonnegative integers s and m,

$$\|D_t^m u(\cdot,t)\|_s \le C t^{-m-s/2}\|v\|, \quad \text{for } t > 0. \tag{8.17}$$

It follows by Sobolev's inequality, Theorem A.5, that $D_t^m u(\cdot,t) \in \mathcal{C}^p$ for $t > 0$, for any $p \ge 0$.

Thus $u(x,t)$ is a smooth function of x and t for $t > 0$ even though we have not assumed the initial data v to be smooth, and $u(\cdot,t)$ therefore satisfies the heat equation in the classical sense. By above we also know that the boundary condition is satisfied, and finally we obtain (8.14) by showing that

$$\|u(\cdot,t) - v\|^2 = \sum_{i=1}^{\infty} \left(\mathrm{e}^{-\lambda_i t} - 1\right)^2 \hat{v}_i^2 \to 0, \quad \text{as } t \to 0.$$

To prove this we let $\epsilon > 0$ be arbitrarily small and choose N large enough that $\sum_{i=N+1}^{\infty} \hat{v}_i^2 < \epsilon$. Then

$$\|u(\cdot,t) - v\|^2 \le \sum_{i=1}^{N} \left(\mathrm{e}^{-\lambda_i t} - 1\right)^2 \hat{v}_i^2 + \epsilon.$$

Since each of the terms of the sum tends to zero as $t \to 0$, we conclude that

$$\|u(\cdot,t) - v\|^2 < 2\epsilon, \quad \text{for } t \text{ small enough.}$$

We collect these results in the following theorem.

Theorem 8.3. *For any $v \in L_2$ the function $u(x,t)$ defined by* (8.13) *is a classical solution of the heat equation in* (8.10)*, vanishes on Γ for $t > 0$, and satisfies the initial condition in the sense of* (8.14)*. Moreover, the smoothness estimate* (8.17) *holds.*

Since the factor $t^{-m-s/2}$ on the right in (8.17) tends to infinity as t tends to zero, the smoothness of the solution is not guaranteed uniformly down to $t = 0$. If the initial function is smoother, then better results are possible in this regard. We have, for instance, the following result in H_0^1.

Theorem 8.4. *Assume that $v \in H_0^1$. Then the solution $u(x,t)$ of* (8.10) *determined in Theorem 8.3 satisfies*

$$|u(\cdot,t)|_1 \le |v|_1 \quad \textit{for } t \ge 0.$$

Proof. We have by Theorem 6.4

$$|u(\cdot,t)|_1^2 = \sum_{i=1}^{\infty} \lambda_i \hat{v}_i^2 \mathrm{e}^{-2\lambda_i t} \le \sum_{i=1}^{\infty} \lambda_i \hat{v}_i^2 = |v|_1^2,$$

which shows our claim. □

We note that this result requires not only that the initial data are in H^1 but also that they vanish on Γ. This means that the initial data have to be compatible with the boundary data on $\Gamma \times \mathbf{R}_+$, which is obviously required for the solution to be continuous at $t = 0$. For higher order regularity further compatibility conditions are needed.

In the same way as in Sect. 8.1 we may think of the solution at time t as the result of a solution operator $E(t)$ acting on the initial data v, and thus write $u(t) = E(t)v$. By (8.13) this operator satisfies the stability estimate

$$\|E(t)v\| \le \|v\|, \quad \text{for } t > 0,$$

and the estimate (8.17) may be expressed as

$$\|D_t^m E(t)v\|_s \le Ct^{-m-s/2}\|v\|, \quad \text{for } t > 0, \quad m, s \ge 0, \tag{8.18}$$

which expresses a smoothing property of the solution operator.

The following simple example illustrates the above solution method.

Example 8.1. The solution of the spatially one-dimensional problem

$$\begin{aligned} &u_t - u_{xx} = 0, && \text{in } \Omega \times \mathbf{R}_+, \\ &u(0,\cdot) = u(\pi,\cdot) = 0, && \text{in } \mathbf{R}_+, \\ &u(\cdot,0) = v, && \text{in } \Omega, \end{aligned} \tag{8.19}$$

with $\Omega = (0,\pi)$ and $v \in L_2(\Omega)$, is given by

$$u(x,t) = \sum_{j=1}^{\infty} \hat{v}_j \mathrm{e}^{-j^2 t} \sin(jx), \quad \text{where } \hat{v}_j = \frac{2}{\pi} \int_0^{\pi} v(x) \sin(jx)\, \mathrm{d}x. \tag{8.20}$$

In this case the associated eigenvalue problem (8.11) reduces to (6.22) with $b = \pi$, and the result thus follows from Theorem 8.3 and the results obtained in Sect. 6.1, except that the eigenfunctions are not normalized here. Note that in (8.20) the coefficient $\hat{v}_j \mathrm{e}^{-j^2 t}$ of the eigenfunction $\sin(jx)$ is obtained by multiplying the corresponding coefficient $\hat{v}_j$ in the expansion of the initial function v by the factor $\mathrm{e}^{-j^2 t}$. If j is large, then $\sin(jx)$ is rapidly oscillating and the factor $\mathrm{e}^{-j^2 t}$ rapidly becomes very small as t increases from 0. Thus, the components of the solution $u(x,t)$ corresponding to the eigenfunctions $\sin(jx)$ with j large are strongly damped as t grows. This means that rapid variations or oscillations in the initial function v, such as, for instance, in the case of a discontinuity (jump), are smoothed out as t increases. This is thus a special case of the smoothing property of the solution operator discussed above, which is typical for parabolic problems.

The solution operator $E(t)$ introduced above is convenient to use in the study of the boundary value problem for the inhomogeneous equation,

$$\begin{aligned} &u_t - \Delta u = f, && \text{in } \Omega \times \mathbf{R}_+, \\ &u = 0, && \text{on } \Gamma \times \mathbf{R}_+, \\ &u(\cdot,0) = v, && \text{in } \Omega, \end{aligned} \tag{8.21}$$

In fact, as we shall see, the solution of this problem may be expressed as

$$u(t) = E(t)v + \int_0^t E(t-s) f(s)\, \mathrm{d}s, \tag{8.22}$$

where we write $u(t)$ for $u(\cdot,t)$ and similarly for $f(s)$. This formula represents the solution of the inhomogeneous equation as a superposition of solutions of homogeneous equations, and is referred to as Duhamel's principle.

Clearly, since $E(t)$ is bounded in L_2-norm, the right hand side of (8.22) is well defined. The first term is the solution of (8.1), the second term vanishes for $t = 0$, and both terms vanish on $\Gamma \times \mathbf{R}_+$. Therefore, in order to show that u in (8.21) is a solution of (8.22), we need to demonstrate that

$$D_t F(t) - \Delta F(t) = f(t), \quad \text{where } F(t) = \int_0^t E(t-s) f(s)\,\mathrm{d}s. \tag{8.23}$$

Formally, by differentiation of the integral, we have

$$D_t F(t) - \Delta F(t) = f(t) + \int_0^t D_t E(t-s) f(s)\,\mathrm{d}s - \int_0^t \Delta E(t-s) f(s)\,\mathrm{d}s, \tag{8.24}$$

and since $D_t E(t-s) = \Delta E(t-s)$ the integrals should cancel. However, if we require only $f(s) \in L_2$ for $s \in (0,t)$, then (8.18) indicates a singularity of order $O((t-s)^{-1})$ in the integrands, so that the integrals are not necessarily well defined. For this reason we now assume that $\|D_t f(t)\|$ is bounded for $t \in [0,T]$ with arbitrary $T > 0$, and write, after replacing $t-s$ by s in the last term,

$$F(t) = \int_0^t E(t-s)(f(s) - f(t))\,\mathrm{d}s + \int_0^t E(s) f(t)\,\mathrm{d}s.$$

By differentiation with respect to t we obtain

$$D_t F(t) = \int_0^t D_t E(t-s)(f(s) - f(t))\,\mathrm{d}s + E(t) f(t), \tag{8.25}$$

where the integrand is now bounded since $\|f(s) - f(t)\| \le C|s-t|$. Similarly, since $\Delta E(t-s) = D_t E(t-s)$,

$$\Delta F(t) = \int_0^t \Delta E(t-s)(f(s) - f(t))\,\mathrm{d}s + (E(t) - I) f(t). \tag{8.26}$$

Taking the difference between (8.25) and (8.26) now shows (8.23).

Another way to deal with the singularities in the integrands in (8.24) would be to use regularity of $f(s)$ in the spatial variable, e.g., through the inequality $\|\Delta E(t-s) f(s)\| \le \|\Delta f(s)\|$. However, in addition to regularity of $f(s)$ this would require the unnatural boundary condition $f(s) = 0$ on Γ.

By (8.22) we obtain at once the stability estimate

$$\|u(t)\| \le \|v\| + \int_0^t \|f(s)\|\,\mathrm{d}s. \tag{8.27}$$

In the standard way this may be used to show uniqueness of the solution of (8.21) as well as the continuous dependence of the solution on the data. For example, if u_1 and u_2 are solutions corresponding to the right-hand sides f_1 and f_2 and initial values v_1 and v_2, then we have

$$\|u_1(t) - u_2(t)\| \le \|v_1 - v_2\| + \int_0^t \|f_1(s) - f_2(s)\|\,\mathrm{d}s, \quad \text{for } t \in \mathbf{R}_+. \tag{8.28}$$

8.3 Variational Formulation. Energy Estimates

We shall now write the initial-boundary value problem (8.21) in variational, or weak form, and use this to derive some estimates for its solution. Although we shall not pursue this here, variational methods may be used to prove existence and uniqueness of solutions of parabolic problems which are considerably more general than (8.21), such as problems with time-dependent coefficients or non-selfadjoint elliptic operator, problems with inhomogeneous boundary conditions, and also some nonlinear problems. For such problems the method of eigenfunction expansion of the previous section is difficult or impossible to use. Moreover, the variational formulation is the basis for the finite element method for parabolic problems, which we shall study in Chapt. 10.

For the variational formulation we multiply the heat equation in (8.21) by a smooth function $\varphi = \varphi(x)$, which vanishes on Γ and find, after integration over Ω and using Green's formula, that

$$(u_t, \varphi) + a(u, \varphi) = (f, \varphi), \quad \forall \varphi \in H_0^1, \quad t \in \mathbf{R}_+, \tag{8.29}$$

with our standard notation

$$a(v, w) = \int_\Omega \nabla v \cdot \nabla w \, \mathrm{d}x, \quad (v, w) = \int_\Omega vw \, \mathrm{d}x.$$

The variational problem may then be formulated: Find $u = u(x, t) \in H_0^1$, thus vanishing on Γ, for $t > 0$, such that (8.29) holds and such that

$$u(\cdot, 0) = v \quad \text{in } \Omega. \tag{8.30}$$

By taking the above steps in the opposite order it is easy to see that if u is a sufficiently smooth solution of this problem, it is also a solution of (8.21). In fact, by integration by parts in (8.29) we obtain

$$(u_t - \Delta u - f, \varphi) = 0, \quad \forall \varphi \in H_0^1, \quad t \in \mathbf{R}_+,$$

or, for any $t \in \mathbf{R}_+$,

$$\int_\Omega \rho(\cdot, t) \, \varphi \, \mathrm{d}x = 0, \quad \forall \varphi \in H_0^1, \quad \text{where } \rho = u_t - \Delta u - f.$$

We conclude, in the same way as for the stationary problem, that this is possible only if $\rho = 0$.

The following result shows some bounds in various natural norms for the solution of our above problem in terms of its data. We proceed formally and refrain from precise statements about the regularity requirements needed. We write $u(t)$ for $u(\cdot, t)$ and similarly for $f(t)$.

Theorem 8.5. *Let $u(t)$ satisfy (8.29) and (8.30), vanish on Γ, and be appropriately smooth for $t \geq 0$. Then there is a constant C such that, for $t \geq 0$,*

$$\|u(t)\|^2 + \int_0^t |u(s)|_1^2 \,\mathrm{d}s \le \|v\|^2 + C\int_0^t \|f(s)\|^2 \,\mathrm{d}s \tag{8.31}$$

and

$$|u(t)|_1^2 + \int_0^t \|u_t(s)\|^2 \,\mathrm{d}s \le |v|_1^2 + \int_0^t \|f(s)\|^2 \,\mathrm{d}s. \tag{8.32}$$

Proof. Taking $\varphi = u$ in (8.29) we obtain

$$(u_t, u) + a(u, u) = (f, u), \quad \text{for } t > 0. \tag{8.33}$$

Here

$$(u_t, u) = \int_\Omega u_t u \,\mathrm{d}x = \int_\Omega \tfrac{1}{2}(u^2)_t \,\mathrm{d}x = \tfrac{1}{2}\frac{d}{dt}\|u\|^2.$$

Applying Poincaré's inequality, Theorem A.6, i.e.,

$$\|\varphi\| \le C|\varphi|_1, \quad \text{for } \varphi \in H_0^1,$$

we have, using also the inequality $2ab \le a^2 + b^2$, that

$$|(f, u)| \le \|f\| \, \|u\| \le C\|f\| \, |u|_1 \le \tfrac{1}{2}|u|_1^2 + \tfrac{1}{2}C^2\|f\|^2.$$

Recalling $a(u, u) = |u|_1^2$, we thus obtain from (8.33) that

$$\tfrac{1}{2}\frac{d}{dt}\|u\|^2 + |u|_1^2 \le \tfrac{1}{2}|u|_1^2 + \tfrac{1}{2}C^2\|f\|^2,$$

or, with a new C,

$$\frac{d}{dt}\|u\|^2 + |u|_1^2 \le C\|f\|^2.$$

By integration over $(0, t)$ this yields

$$\|u(t)\|^2 + \int_0^t |u(s)|_1^2 \,\mathrm{d}s \le \|v\|^2 + C\int_0^t \|f\|^2 \,\mathrm{d}s,$$

which is (8.31).

To prove (8.32) we now choose $\varphi = u_t$ in (8.29) and obtain

$$\|u_t\|^2 + a(u, u_t) = (f, u_t) \le \tfrac{1}{2}\|f\|^2 + \tfrac{1}{2}\|u_t\|^2.$$

Here

$$a(u, u_t) = \int_\Omega \nabla u \cdot \nabla u_t \,\mathrm{d}x = \int_\Omega \tfrac{1}{2}(|\nabla u|^2)_t \,\mathrm{d}x = \tfrac{1}{2}\frac{d}{dt}|u|_1^2,$$

so that we may conclude,

$$\|u_t\|^2 + \frac{d}{dt}|u|_1^2 \le \|f\|^2,$$

whence, by integration over $(0, t)$,

$$|u(t)|_1^2 + \int_0^t \|u_t\|^2 \,\mathrm{d}s \le |v|_1^2 + \int_0^t \|f\|^2 \,\mathrm{d}s,$$

which is (8.32). □

It follows in the standard way from (8.31) that if u_1 and u_2 are solutions corresponding to the right-hand sides f_1 and f_2 and initial values v_1 and v_2, then we have

$$\|u_1(t)-u_2(t)\|^2+\int_0^t |u_1-u_2|_1^2 \,\mathrm{d}s \le \|v_1-v_2\|^2+C\int_0^t \|f_1-f_2\|^2 \,\mathrm{d}s, \quad \text{for } t \ge 0,$$

and a similar bound is obtained from (8.32). Note that these estimates also bound the error in H_0^1 and uses the L_2-norm in time rather than the L_1-norm employed in (8.28).

8.4 A Maximum Principle

We now consider the generalization of the mixed initial-boundary value problem of Sect. 8.2 which allows a source term and inhomogeneous boundary conditions, i.e., to find u on $\bar{\Omega} \times \bar{I}$ such that

$$\begin{aligned} u_t - \Delta u &= f, && \text{in } \Omega \times I, \\ u &= g, && \text{on } \Gamma \times I, \\ u(\cdot,0) &= v, && \text{in } \Omega, \end{aligned} \tag{8.34}$$

where Ω is a bounded domain in $\mathbf{R}^d$ and $I = (0,T)$ is a finite interval in time. In order to show a maximum principle for this problem it is convenient to introduce the *parabolic boundary* of $\Omega \times I$ as the set $\Gamma_p = (\Gamma \times \bar{I}) \cup (\Omega \times \{t = 0\})$, i.e., the boundary of $\Omega \times I$ minus the interior of the top part of this boundary, $\Omega \times \{t = T\}$.

Theorem 8.6. *Let u be smooth and assume that $u_t - \Delta u \le 0$ in $\Omega \times I$. Then u attains its maximum on the parabolic boundary Γ_p.*

Proof. If this were not true, then the maximum would be attained either at an interior point of $\Omega \times I$ or at a point of $\Omega \times \{t = T\}$, i.e., at a point $(\bar{x}, \bar{t}) \in \Omega \times (0,T]$, and we would have

$$u(\bar{x},\bar{t}) = \max_{\bar{\Omega}\times\bar{I}} u = M > m = \max_{\Gamma_p} u.$$

In such a case, for $\epsilon > 0$ sufficiently small, the function

$$w(x,t) = u(x,t) + \epsilon |x|^2$$

would also take its maximum at a point in $\Omega \times (0,T]$, since, for ϵ small,

$$\max_{\Gamma_p} w \le m + \epsilon \max_{\Gamma_p} |x|^2 < M \le \max_{\bar{\Omega}\times\bar{I}} w.$$

By our assumption we have since $\Delta(|x|^2) = 2d$ that

$$w_t - \Delta w = u_t - \Delta u - 2d\epsilon < 0, \quad \text{in } \Omega \times I. \tag{8.35}$$

On the other hand, at the point $(\tilde{x}, \tilde{t})$, where w takes its maximum, we have

$$-\Delta w(\tilde{x}, \tilde{t}) = -\sum_{i=1}^{d} w_{x_i x_i}(\tilde{x}, \tilde{t}) \geq 0,$$

and

$$w_t(\tilde{x}, \tilde{t}) = 0, \quad \text{if } \tilde{t} < T, \quad \text{or} \quad w_t(\tilde{x}, \tilde{t}) \geq 0, \quad \text{if } \tilde{t} = T,$$

so that in both cases

$$w_t(\tilde{x}, \tilde{t}) - \Delta w(\tilde{x}, \tilde{t}) \geq 0.$$

This is a contradiction to (8.35) and thus shows our claim. □

By considering the functions $\pm u$, it follows, in particular, that a solution of the homogeneous heat equation ($f = 0$) attains both its maximum and its minimum on Γ_p, so that in this case, with $\|w\|_{\mathcal{C}(\bar{M})} = \max_{x \in \bar{M}} |w(x)|$,

$$\|u\|_{\mathcal{C}(\bar{\Omega} \times \bar{I})} \leq \max \left\{ \|g\|_{\mathcal{C}(\Gamma \times \bar{I})}, \|v\|_{\mathcal{C}(\bar{\Omega})} \right\}.$$

For the inhomogeneous equation one may show the following inequality, the proof of which we leave as an exercise, see Problem 8.7.

Theorem 8.7. *The solution of* (8.34) *satisfies*

$$\|u\|_{\mathcal{C}(\bar{\Omega} \times \bar{I})} \leq \max \left\{ \|g\|_{\mathcal{C}(\Gamma \times \bar{I})}, \|v\|_{\mathcal{C}(\bar{\Omega})} \right\} + \frac{r^2}{2d} \|f\|_{\mathcal{C}(\bar{\Omega} \times \bar{I})},$$

where r is the radius of a ball containing Ω.

As usual such a result shows uniqueness and stability for the initial-boundary value problem.

We close this section by proving the uniqueness of a bounded solution to the pure initial value problem considered in Sect. 8.1.

Theorem 8.8. *The initial value problem* (8.1) *has at most one solution which is bounded in $\mathbf{R}^d \times [0, T]$, where T is arbitrary.*

Proof. If there were two solutions of (8.1), then their difference would be a solution with initial data zero. It suffices therefore to show that the only bounded solution u of

$$\begin{aligned} u_t &= \Delta u, && \text{in } \mathbf{R}^d \times I, \quad \text{where } I = (0, T), \\ u(\cdot, 0) &= 0, && \text{in } \mathbf{R}^d, \end{aligned}$$

is $u = 0$, or that and if (x_0, t_0) is an arbitrary point in $\mathbf{R}^d \times I$, and $\epsilon > 0$ is arbitrary, then $|u(x_0, t_0)| \leq \epsilon$. We introduce the auxiliary function

$$w(x,t) = \frac{|x|^2 + 2d\,t}{|x_0|^2 + 2d\,t_0},$$

and note that $w_t = \Delta w$. Let now

$$h_\pm(x,t) = -\epsilon w(x,t) \pm u(x,t).$$

Then

$$(h_\pm)_t - \Delta h_\pm = 0, \quad \text{in } \mathbf{R}^d \times I.$$

Since u is bounded we have $|u(x,t)| \le M$ on $\mathbf{R}^d \times I$ for some M. Defining R by $R^2 = \max\big(|x_0|^2, M(|x_0|^2 + 2dt_0)/\epsilon\big)$, we have

$$h_\pm(x,t) \le -\epsilon \frac{R^2}{|x_0|^2 + 2d\,t_0} + M \le 0, \quad \text{if } |x| = R,$$

and

$$h_\pm(x,0) = -\epsilon|x|^2/(|x_0|^2 + 2d\,t_0) \le 0, \quad \text{for } x \in \mathbf{R}^d.$$

Hence we may apply Theorem 8.6 with $\Omega = \{|x| < R\}$ and conclude that $h_\pm(x,t) \le 0$ for $(x,t) \in \Omega \times I$. In particular, at (x_0,t_0) we have $\pm u(x_0,t_0) = h_\pm(x_0,t_0) + \epsilon \le \epsilon$, which completes the proof of the theorem. □

The assumption of Theorem 8.8 that the solutions are bounded in $\mathbf{R}^d \times [0,T]$ may be relaxed to the requirement that $|u(x,t)| \le M\mathrm{e}^{c|x|^2}$ for all $x \in \mathbf{R}^d$, $0 \le t \le T$, and for some $M, c > 0$, but without some such restriction on the growth of the solution for large $|x|$, uniqueness is not guaranteed. For instance, the following function is a solution of the homogeneous heat equation which has initial values zero but does not vanish identically for $t > 0$:

$$u(x,t) = \sum_{n=0}^{\infty} f^{(n)}(t)\frac{x^{2n}}{(2n)!}, \quad \text{where } f(t) = \mathrm{e}^{-1/t^2} \text{ for } t > 0,\ f(0) = 0.$$

The technical part of the proof is to show that the series converges so rapidly that it may be differentiated termwise. Then it is obvious that $u_t = u_{xx}$ and that $u(x,0) = 0$.

8.5 Problems

Problem 8.1. Show that if u is a solution of (8.1) with

$$\int_{\mathbf{R}^d} |v(x)|\,\mathrm{d}x < \infty,$$

then

$$\int_{\mathbf{R}^d} u(x,t)\,\mathrm{d}x = \text{constant} = \int_{\mathbf{R}^d} v(x)\,\mathrm{d}x, \quad \text{for } t \ge 0.$$

Give a physical interpretation of this result.

Problem 8.2. Find the solution of the initial-boundary value problem (8.19) with

(a) $v(x) = 1$, for $0 < x < \pi$;

(b) $v(x) = x(\pi - x)$, for $0 < x < \pi$.

Sketch the solutions $u(x,t)$ at various time levels t.

Problem 8.3. Consider the function

$$u(x,t) = \begin{cases} xt^{-3/2}\mathrm{e}^{-x^2/4t}, & \text{for } t > 0, \\ 0, & \text{for } t = 0. \end{cases}$$

Show that u is a solution of

$$u_t - u_{xx} = 0, \quad \text{in } \mathbf{R} \times \mathbf{R}_+,$$

and that, for each x,

$$u(x,t) \to 0, \quad \text{as } t \to 0.$$

Why is this not a counter-example to the uniqueness result of Theorem 8.8? Hint: set $x = t$.

Problem 8.4. Let $u = E(t)v$ be the solution of (8.10). Show that E has the semigroup property (7.1). Show the estimates

(a) $\|u(t)\| \le \mathrm{e}^{-\lambda_1 t}\|v\|$, for $t \ge 0$,

(b) $\|\Delta^k D_t^j u(t)\| \le Ct^{-(j+k)}\mathrm{e}^{-\lambda_1 t/2}\|v\|$, for $t > 0$.

Problem 8.5. Prove by the energy method that there is a constant $C = C(T)$ such that if u satisfies (8.29) and (8.30), then

(a) $$\int_0^t s\|u_t(s)\|^2 \,\mathrm{d}s \le C\Big(\|v\|^2 + \int_0^t \|f(s)\|^2 \,\mathrm{d}s\Big), \qquad \text{for } 0 \le t \le T,$$

(b) $$|u(t)|_1^2 \le Ct^{-1}\Big(\|v\|^2 + \int_0^t \|f(s)\|^2 \,\mathrm{d}s\Big), \qquad \text{for } 0 < t \le T.$$

Problem 8.6. Let u be the solution of (8.29) and (8.30) with $v = 0$. Show that

$$\int_0^t \big(\|u_t(s)\|^2 + \|\Delta u(s)\|^2\big)\,\mathrm{d}s \le C\int_0^t \|f(s)\|^2 \,\mathrm{d}s, \quad \text{for } t \ge 0.$$

Problem 8.7. Prove Theorem 8.7. Hint: See the proof of Theorem 3.2.

Problem 8.8. Show estimates analogous to those of Theorem 8.5 when the term $-\Delta u$ in (8.21) is replaced by $\mathcal{A}u = -\nabla \cdot (a\nabla u) + b \cdot \nabla u + cu$ as in Sect. 3.5.

Problem 8.9. Prove (8.4).

Problem 8.10. Show that if u satisfies (8.10), then there is a constant C such that

$$\|u(t)\|_2^2 + \int_0^t |u_t(s)|_1^2 \, \mathrm{d}s \le C\|v\|_2^2, \quad \forall v \in H^2 \cap H_0^1, \ t \ge 0.$$

You can use either the spectral method or the energy method. You also need the elliptic regularity estimate (3.36).

Problem 8.11. Let u be the solution of

$$\begin{aligned} &u_t - \Delta u = 0, && \text{in } \Omega \times \mathbf{R}_+, \\ &u(x,t) = 0, && \text{on } \Gamma \times \mathbf{R}_+, \\ &u(\cdot,0) = v, && \text{in } \Omega, \end{aligned}$$

where $\Omega = \{x \in \mathbf{R}^2 : 0 < x_i < 1, \ i = 1,2\}$. Let $\varphi(x) = A\sin(\pi x_1)\sin(\pi x_2)$ with $A > 0$. Show that if $0 \le v(x) \le \varphi(x)$ for $x \in \Omega$, then $0 \le u(x,t) \le \mathrm{e}^{-2\pi^2 t}\varphi(x)$ for $x \in \Omega$, $t > 0$. Hint: Use the maximum principle.

Problem 8.12. Prove the L_2 version of Theorem 8.1: If $v \in L_2(\mathbf{R}^d)$ then $\|E(t)v\|_{L_2} \le \|v\|_{L_2}$ for $t \ge 0$ and $\|E(t)v - v\|_{L_2} \to 0$ as $t \to 0$. Hint: Parseval's formula (A.32).

Problem 8.13. Consider the heat equation with Neumann's boundary condition:

$$\begin{aligned} &u_t - \Delta u = 0, && \text{in } \Omega \times \mathbf{R}_+, \\ &\frac{\partial u}{\partial n} = 0, && \text{on } \Gamma \times \mathbf{R}_+, \\ &u(\cdot,0) = v, && \text{in } \Omega, \end{aligned}$$

where $\partial u/\partial n$ is the outward normal derivative.

(a) Show that $\overline{u(t)} = \overline{v}$ for $t \ge 0$, where $\overline{v} = \frac{1}{|\Omega|}\int_\Omega v(x)\,\mathrm{d}x$ denotes the average value of v.
(b) Show that $\|u(t) - \overline{v}\| \to 0$ as $t \to \infty$.

Problem 8.14. Suppose that u satisfies the initial-boundary value problem

$$\begin{aligned} &u_t - \Delta u = f, && \text{in } \Omega \times \mathbf{R}_+, \\ &\frac{\partial u}{\partial n} = g, && \text{on } \Gamma \times \mathbf{R}_+, \\ &u(\cdot,0) = v, && \text{in } \Omega, \end{aligned}$$

where $\Omega \subset \mathbf{R}^d$ is a bounded domain in $\mathbf{R}^d$ with a smooth boundary Γ and $\partial u/\partial n$ is the exterior normal derivative. Assume in addition that $f(x,t) \ge 0$, $v(x) \ge 0$ for $x \in \Omega$, $t \ge 0$, and $g(x,t) > 0$ for $x \in \Gamma$, $t \ge 0$. Show that $u(x,t) \ge 0$ for $x \in \Omega$, $t \ge 0$. (In fact, it is sufficient to assume that $g(x,t) \ge 0$.)

Problem 8.15. Consider the Stokes equations describing the 2-dimensional motion of a viscous and incompressible fluid at small Reynolds number R:

$$\begin{aligned} &\frac{\partial u}{\partial t} - R^{-1}\Delta u + \nabla p = 0, && \text{in } \mathbf{R}^2 \times \mathbf{R}_+, \\ &\nabla \cdot u = 0, && \text{in } \mathbf{R}^2 \times \mathbf{R}_+, \\ &u(\cdot, 0) = v, && \text{in } \mathbf{R}^2, \end{aligned} \tag{8.36}$$

where $u(x,t) \in \mathbf{R}^2$ is the dimensionless velocity and $p(x,t) \in \mathbf{R}$ the dimensionless pressure. In this form R^{-1} represents the viscosity. Let us define the vorticity ω by

$$\omega = \nabla \times u = \partial u_2/\partial x_1 - \partial u_1/\partial x_2.$$

Show that (8.36) can be rewritten in the vorticity variable as

$$\begin{aligned} &\frac{\partial \omega}{\partial t} - R^{-1}\Delta \omega = 0, && \text{in } \mathbf{R}^2 \times \mathbf{R}_+, \\ &\omega(\cdot, 0) = \nabla \times v, && \text{in } \mathbf{R}^2. \end{aligned}$$

Problem 8.16. Let $u(x,t) = (E(t)v)(x)$ be the solution of (8.10) and let $\{\lambda_j\}_{j=1}^{\infty}$ and $\{\varphi_j\}_{j=1}^{\infty}$ be the eigenvalues and normalized eigenfunctions of (6.5) as in Theorem 6.4. Show that

$$u(x,t) = (E(t)v)(x) = \int_\Omega G(x,y,t)v(y)\,\mathrm{d}y,$$

where the Green's function is

$$G(x,y,t) = \sum_{j=1}^{\infty} \mathrm{e}^{-\lambda_j t}\varphi_j(x)\varphi_j(y).$$

Hint: see Problem 6.7.

9 Finite Difference Methods for Parabolic Problems

In this chapter we give an introduction to the numerical solution of parabolic equations by finite differences, and consider the application of such methods to the homogeneous heat equation in one space dimension. We first discuss, in Sect. 9.1, the pure initial value problem, with data given on the unrestricted real axis, and then in Sect. 9.2 the mixed initial-boundary value problem on a finite interval in space, under Dirichlet boundary conditions. We discuss stability and error bounds for various choices of finite difference approximations, in maximum-norm by maximum principle type arguments and in L_2-norm by Fourier analysis. For the unrestricted problem we consider explicit schemes, and on a finite interval also implicit ones, such as the Crank-Nicolson scheme.

9.1 The Pure Initial Value Problem

Consider first the pure initial value problem to find $u = u(x,t)$ such that

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial^2 u}{\partial x^2}, \qquad \text{in } \mathbf{R}\times\mathbf{R}_+, \\ u(\cdot,0) &= v, \qquad \text{in } \mathbf{R}, \end{aligned} \tag{9.1}$$

where v is a given smooth bounded function. We recall from Sect. 8.1 that this problem has a unique solution, many properties of which may be deduced, for instance, from the representation

$$u(x,t) = \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} \mathrm{e}^{-y^2/4t} v(x-y)\,\mathrm{d}y = \big(E(t)v\big)(x), \tag{9.2}$$

where $E(t)$ denotes the solution operator of (9.1). In particular, we note that the solution is bounded with respect to the maximum-norm, or, more precisely,

$$\|u(\cdot,t)\|_{\mathcal{C}} = \|E(t)v\|_{\mathcal{C}} \le \|v\|_{\mathcal{C}} = \sup_{x\in\mathbf{R}} |v(x)|, \qquad \text{for } t \ge 0. \tag{9.3}$$

For the numerical solution of this problem by finite differences one introduces a grid of mesh-points $(x,t) = (x_j, t_n)$. Here $x_j = jh,\ t_n = nk$, where j

and n are integers, $n \geq 0$, h the mesh-width in x, and k the time step, with both h and k small. One then seeks an approximate solution U_j^n at these mesh-points, determined by an equation obtained by replacing the derivatives in (9.1) by difference quotients. For functions defined on the grid we introduce thus the forward and backward difference quotients with respect to x,

$$\partial_x U_j^n = h^{-1}(U_{j+1}^n - U_j^n) \quad \text{and} \quad \bar{\partial}_x U_j^n = h^{-1}(U_j^n - U_{j-1}^n),$$

and similarly with respect to t, for instance,

$$\partial_t U_j^n = k^{-1}(U_j^{n+1} - U_j^n).$$

The simplest finite difference scheme corresponding to (9.1) is then the *forward Euler method*

$$\begin{aligned} &\partial_t U_j^n = \partial_x \bar{\partial}_x U_j^n, && \text{for } j,\ n \in \mathbf{Z},\ n \geq 0, \\ &U_j^0 = v_j := v(x_j), && \text{for } j \in \mathbf{Z}, \end{aligned}$$

where $\mathbf{Z}$ denotes the integers. The difference equation may also be written

$$\frac{U_j^{n+1} - U_j^n}{k} = \frac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{h^2},$$

or, if we introduce the mesh-ratio $\lambda = k/h^2$,

$$U_j^{n+1} = (E_k U^n)_j = \lambda U_{j-1}^n + (1 - 2\lambda) U_j^n + \lambda U_{j+1}^n, \tag{9.4}$$

which defines the local discrete solution operator E_k. We shall consider h and k related by $k/h^2 = \lambda = \text{constant}$, and may therefore omit the dependence on h in the notation. The scheme (9.4) is called *explicit*, since it expresses the solution at $t = t_{n+1}$ explicitly in terms of the values at $t = t_n$. Iterating the operator we find that the solution of the discrete problem is

$$U_j^n = (E_k^n U^0)_j = (E_k^n v)_j, \quad \text{for } j,\ n \in \mathbf{Z},\ n \geq 0.$$

Assume now that $\lambda \leq \frac{1}{2}$. All the coefficients of the operator E_k in (9.4) are then non-negative, and since their sum is 1, we find

$$|U_j^{n+1}| \leq \lambda |U_{j-1}^n| + (1 - 2\lambda)|U_j^n| + \lambda |U_{j+1}^n| \leq \sup_{j \in \mathbf{Z}} |U_j^n|,$$

so that

$$\sup_{j \in \mathbf{Z}} |U_j^{n+1}| \leq \sup_{j \in \mathbf{Z}} |U_j^n|.$$

Defining for mesh-functions $v = (v_j)$, a discrete maximum-norm by

$$\|v\|_{\infty,h} = \sup_{j \in \mathbf{Z}} |v_j|, \tag{9.5}$$

we thus have

$$\|U^{n+1}\|_{\infty,h} = \|E_k U^n\|_{\infty,h} \le \|U^n\|_{\infty,h},$$

and hence by repeated application

$$\|U^n\|_{\infty,h} = \|E_k^n v\|_{\infty,h} \le \|v\|_{\infty,h}, \tag{9.6}$$

which is a discrete analogue of the estimate (9.3) for the continuous problem.

The boundedness of the discrete solution operator is referred to as the *stability* of this operator. We shall now see that if λ is chosen as a constant bigger than $\frac{1}{2}$, then the method is unstable. To see this, we choose $v_j = (-1)^j\epsilon$, where ϵ is a small positive number, so that $\|v\|_{\infty,h} = \epsilon$. Then

$$U_j^1 = \big(\lambda(-1)^{j-1} + (1-2\lambda)(-1)^j + \lambda(-1)^{j+1}\big)\epsilon = (1-4\lambda)(-1)^j\epsilon,$$

or, more generally,

$$U_j^n = (1-4\lambda)^n(-1)^j\epsilon,$$

whence

$$\|U^n\|_{\infty,h} = (4\lambda-1)^n\epsilon \to \infty, \quad \text{as } n \to \infty.$$

We thus find that in this case, even though the initial data are very small, the norm of the discrete solution tends to infinity as $n \to \infty$ when $k = t/n \to 0$, even if $t = t_n$ is bounded. This may be interpreted to mean that very small perturbations of the initial data (for instance by round-off errors) may cause so big changes in the discrete solution at later times that it becomes useless.

We now restrict the considerations to the stable case, $\lambda \le \frac{1}{2}$, and we shall show that the discrete solution converges to the exact solution as the mesh-parameters tend to zero, provided the initial data, and thus the exact solution of (9.1), are smooth enough. In order to demonstrate this, we need to use that the exact solution satisfies the difference equation except for a small error, which tends to zero with h and k. More precisely, setting $u_j^n = u(x_j, t_n)$ we have by Taylor's formula for the solution of (9.1), with appropriate $\bar{x}_j \in (x_{j-1}, x_{j+1})$, $\bar{t}_n \in (t_n, t_{n+1})$,

$$\begin{aligned}\tau_j^n &= \partial_t u_j^n - \partial_x\bar{\partial}_x u_j^n = \Big(\partial_t u_j^n - u_t(x_j,t_n)\Big) - \Big(\partial_x\bar{\partial}_x u_j^n - u_{xx}(x_j,t_n)\Big)\\ &= \tfrac{1}{2}k u_{tt}(x_j,\bar{t}_n) - \tfrac{1}{12}h^2 u_{xxxx}(\bar{x}_j,t_n).\end{aligned}$$

Since $u_{tt} = u_{xxxx}$ and since one easily sees from (9.2) that $|u(\cdot,t)|_{\mathcal{C}^4} \le |v|_{\mathcal{C}^4}$ for the solution of (9.1), we obtain

$$\begin{aligned}\|\tau^n\|_{\infty,h} &\le Ck\max_{t\in I_n}|u_{tt}(\cdot,t)|_{\mathcal{C}} + Ch^2|u(\cdot,t_n)|_{\mathcal{C}^4}\\ &\le Ch^2\max_{t\in I_n}|u(\cdot,t)|_{\mathcal{C}^4} \le Ch^2|v|_{\mathcal{C}^4}, \quad \text{for } \lambda \le \frac{1}{2}.\end{aligned} \tag{9.7}$$

The expression τ_j^n is referred to as the *truncation error* (or *local discretization error*). We now have the following error estimate.

Theorem 9.1. *Let U^n and u be the solutions of* (9.4) *and* (9.1), *and let $k/h^2 = \lambda \le \frac{1}{2}$. Then there is constant C such that*

$$\|U^n - u^n\|_{\infty,h} \le C t_n h^2 |v|_{\mathcal{C}^4} \quad \text{for } t_n \ge 0.$$

Proof. Set $z^n = U^n - u^n$. Then

$$\partial_t z_j^n - \partial_x \bar{\partial}_x z_j^n = -\tau_j^n,$$

and hence

$$z_j^{n+1} = (E_k z^n)_j - k\tau_j^n.$$

By repeated application this yields

$$z_j^n = (E_k^n z^0)_j - k \sum_{l=0}^{n-1} (E_k^{n-1-l} \tau^l)_j.$$

Since $z_j^0 = U_j^0 - u_j^0 = v_j - v_j = 0$ we find, using the stability estimate (9.6) and the truncation error estimate (9.7),

$$\|z^n\|_{\infty,h} \le k \sum_{l=0}^{n-1} \|\tau^l\|_{\infty,h} \le C\, nk\, h^2 |v|_{\mathcal{C}^4},$$

which is the desired result. □

The method described has first order *accuracy* in time and second order in space, but since k and h are tied by $k/h^2 = \lambda \le \frac{1}{2}$, the total effect is second order accuracy with respect to the mesh-width h.

More generally we may consider finite difference operators of the form

$$U_j^{n+1} = (E_k U^n)_j := \sum_p a_p U_{j-p}^n, \quad \text{for } j,\, n \in \mathbf{Z},\ n \ge 0, \tag{9.8}$$

where $a_p = a_p(\lambda)$, $\lambda = k/h^2$, and where the sum is finite. One may associate with this operator the trigonometric polynomial

$$\tilde{E}(\xi) = \sum_p a_p \mathrm{e}^{-\mathrm{i}p\xi}. \tag{9.9}$$

This polynomial is relevant to the stability analysis and is called the *symbol* or the *characteristic polynomial* of E_k. We find at once the following result.

Theorem 9.2. *A necessary condition for stability of the operator E_k in* (9.8) *with respect to the discrete maximum-norm defined in* (9.5) *is that*

$$|\tilde{E}(\xi)| \le 1, \quad \text{for } \xi \in \mathbf{R}. \tag{9.10}$$

Proof. Assume that E_k is stable and that $|\tilde{E}(\xi_0)| > 1$ for some $\xi_0 \in \mathbf{R}$. Then, for $v_j = \mathrm{e}^{\mathrm{i}j\xi_0}\epsilon$,

$$U_j^1 = \epsilon \sum_p a_p \mathrm{e}^{\mathrm{i}(j-p)\xi_0} = \tilde{E}(\xi_0) v_j,$$

and by repeated application this yields

$$\|U^n\|_{\infty,h} = |\tilde{E}(\xi_0)|^n \epsilon \to \infty, \quad \text{as } n \to \infty.$$

Since $\|v\|_{\infty,h} = \epsilon$ this contradicts the stability and proves the theorem. □

For the finite difference operator defined in (9.4) we have $\tilde{E}(\xi) = 1 - 2\lambda + 2\lambda\cos\xi$, and since $\cos\xi$ varies in $[-1, 1]$ the condition (9.10) is equivalent to $1 - 4\lambda \geq -1$, or $\lambda \leq \frac{1}{2}$, which agrees with our previous stability condition.

The condition (9.10) is a special case of *von Neumann's stability condition.* We shall see that in a slightly different setting this condition is also sufficient for stability.

By its definition the symbol of a discrete solution operator is particularly suited for investigating finite difference methods in the framework of Fourier analysis. It is then convenient to use the l_2-norm to measure the mesh-functions. Let thus $V = \{V_j\}_{-\infty}^{\infty}$ be a mesh-function in the space variable and set

$$\|V\|_{2,h} = \Big(h \sum_{j=-\infty}^{\infty} V_j^2\Big)^{1/2}.$$

The set of mesh-functions normed in this way and with finite norm will be denoted by $l_{2,h}$. Let us also define for such a mesh-function its discrete Fourier transform

$$\hat{V}(\xi) = h \sum_{j=-\infty}^{\infty} V_j \mathrm{e}^{-\mathrm{i}j\xi},$$

where we assume that the sum is absolutely convergent. The function $\hat{V}(\xi)$ is 2π-periodic and V can be retrieved from $\hat{V}(\xi)$ by

$$V_j = \frac{1}{2\pi h} \int_{-\pi}^{\pi} \hat{V}(\xi) \mathrm{e}^{\mathrm{i}j\xi} \,\mathrm{d}\xi.$$

We recall Parseval's relation

$$\|V\|_{2,h}^2 = \frac{1}{2\pi h} \int_{-\pi}^{\pi} |\hat{V}(\xi)|^2 \,\mathrm{d}\xi = \frac{1}{2\pi} \int_{-\pi/h}^{\pi/h} |\hat{V}(h\xi)|^2 \,\mathrm{d}\xi. \tag{9.11}$$

We may now define stability with respect to the norm $\|\cdot\|_{2,h}$, or *stability in* $l_{2,h}$, to mean, analogously to (9.6), but allowing a constant factor C on the right,

$$\|E_k^n V\|_{2,h} \leq C \|V\|_{2,h}, \quad \text{for } n \geq 0,\ h \in (0, 1), \tag{9.12}$$

and find the following:

Theorem 9.3. *Von Neumann's condition* (9.10) *is a necessary and sufficient condition for stability of the operator* E_k^n *in* $l_{2,h}$.

Proof. We note that

$$(E_k V)\hat{}(\xi) = h\sum_j \sum_p a_p V_{j-p} \mathrm{e}^{-\mathrm{i}j\xi}$$
$$= \sum_p a_p \mathrm{e}^{-\mathrm{i}p\xi} h \sum_j V_{j-p} \mathrm{e}^{-\mathrm{i}(j-p)\xi} = \tilde{E}(\xi)\hat{V}(\xi).$$

Hence

$$(E_k^n V)\hat{}(\xi) = \tilde{E}(\xi)^n \hat{V}(\xi),$$

and using Parseval's relation (9.11), the stability of E_k in $l_{2,h}$ is equivalent to

$$\int_{-\pi}^{\pi} |\tilde{E}(\xi)|^{2n} |\hat{V}(\xi)|^2 \,\mathrm{d}\xi \le C^2 \int_{-\pi}^{\pi} |\hat{V}(\xi)|^2 \,\mathrm{d}\xi, \quad \text{for } n \ge 0,$$

for all admissible $\hat{V}$. But this is easily seen to hold if and only if

$$|\tilde{E}(\xi)|^n \le C, \quad \text{for } n \ge 0, \quad \xi \in \mathbf{R},$$

which is equivalent to (9.10) (and we thus have $C = 1$ in (9.12)). □

In the discussion of an explicit finite difference method of the form (9.8) it is sometimes convenient to consider the functions of the space variable x to be defined not just at the mesh-points, but for all $x \in \mathbf{R}$, so that we are given an initial function $U^0(x) = v(x)$ and seek an approximate solution $U^n(x)$ at $t = t_n$, $n = 1, 2, \ldots$, from

$$U^{n+1}(x) = (E_k U^n)(x) = \sum_p a_p U^n(x - x_p), \quad a_p = a_p(\lambda),\ \lambda = k/h^2. \tag{9.13}$$

One advantage of this point of view is that all U^n then lie in the same function space, independently of h, for instance in $L_2(\mathbf{R})$ or $\mathcal{C}(\mathbf{R})$.

We consider briefly the situation in which the analysis takes place in $L_2 = L_2(\mathbf{R})$ and set, allowing now also complex-valued functions,

$$\|u\| = \Big(\int_{-\infty}^{\infty} |u(x)|^2 \,\mathrm{d}x \Big)^{1/2}.$$

We shall then use the Fourier transform defined by (cf. Appendix A.3)

$$(\mathcal{F}v)(\xi) = \hat{v}(\xi) = \int_{-\infty}^{\infty} v(x) \mathrm{e}^{-\mathrm{i}x\xi} \,\mathrm{d}x, \tag{9.14}$$

and note that here, with $\tilde{E}(\xi)$ defined by (9.9),

$$(E_k v)\hat{}(\xi) = \sum_p a_p (\mathcal{F} v(\cdot - ph))(\xi) = \Big(\sum_p a_p \mathrm{e}^{-\mathrm{i}ph\xi} \Big) \hat{v}(\xi) = \tilde{E}(h\xi)\hat{v}(\xi).$$

Recalling Parseval's relation for (9.14),

$$\|v\|^2 = (2\pi)^{-1} \|\hat{v}\|^2,$$

we thus find

$$\|U^n\| = (2\pi)^{-1/2} \|\tilde{E}(h\xi)^n \hat{v}\| \le \sup_{\xi \in \mathbf{R}} |\tilde{E}(h\xi)|^n \|v\|,$$

and therefore that stability with respect to L_2 holds if and only if

$$\sup_{\xi \in \mathbf{R}} |\tilde{E}(h\xi)|^n \le C, \quad n \ge 0,$$

which is again equivalent to von Neumann's condition (9.10).

Also the convergence analysis may be expressed in L_2. We say that the finite difference operator E_k defined in (9.13) is *accurate of order* r if

$$\tilde{E}(\xi) = \mathrm{e}^{-\lambda\xi^2} + O(|\xi|^{r+2}), \quad \text{as } \xi \to 0. \tag{9.15}$$

For instance, for the operator defined in (9.4) we have

$$\begin{aligned} \tilde{E}(\xi) &= 1 - 2\lambda + 2\lambda \cos\xi = 1 - \lambda\xi^2 + \tfrac{1}{12}\lambda\xi^4 + O(\xi^6) \\ &= \mathrm{e}^{-\lambda\xi^2} + \big(\tfrac{1}{12}\lambda - \tfrac{1}{2}\lambda^2\big)\xi^4 + O(\xi^6), \end{aligned}$$

so that (9.4) is accurate of order 2, or, for the special choice $\lambda = \frac{1}{6}$, of order 4.

By comparing the coefficients in the Taylor expansion of $\tilde{E}(\xi) - \mathrm{e}^{-\lambda\xi^2}$ around $\xi = 0$ with those in the expansion of $E_k u(x,t) - u(x,t+k)$ around (x,t), with $k = \lambda h^2$, it is easy to see that the definition (9.15) is equivalent to saying that, for the exact solution of (9.1),

$$u^{n+1}(x) - E_k u^n(x) = kO(h^r), \quad \text{as } h \to 0, \quad \lambda = k/h^2 = \text{ constant}, \tag{9.16}$$

i.e., that the one step discrete solution operator approximates the exact solution operator to order $kO(h^r)$, see Problem 9.1.

We have then the following result, where we recall that $|\cdot|_s = |\cdot|_{H^s}$.

Theorem 9.4. *Assume that E_k is defined by (9.13) with $\lambda = k/h^2 =$ constant and is accurate of order r and stable in L_2. Then*

$$\|U^n - u^n\| \le C t_n h^r |v|_{r+2}, \quad \textit{for } t_n \ge 0.$$

Proof. Since $\tilde{E}(\xi)$ is bounded on $\mathbf{R}$, we have by (9.15)

$$|\tilde{E}(\xi) - \mathrm{e}^{-\lambda\xi^2}| \le C|\xi|^{r+2}, \quad \text{for } \xi \in \mathbf{R}.$$

By stability it follows that

$$|\tilde{E}(\xi)^n - \mathrm{e}^{-n\lambda\xi^2}| = |(\tilde{E}(\xi) - \mathrm{e}^{-\lambda\xi^2}) \sum_{j=0}^{n-1} \tilde{E}(\xi)^{n-1-j} \mathrm{e}^{-j\lambda\xi^2}| \le Cn|\xi|^{r+2}. \tag{9.17}$$

Now, by Fourier transformation of (9.1) with respect to x, we have as in Sect. 8.1, that

$$\frac{d\hat{u}}{dt}(\xi, t) = -\xi^2 \hat{u}(\xi, t), \quad \text{for } t > 0, \quad \text{with } \hat{u}(\xi, 0) = \hat{v}(\xi),$$

and hence

$$\hat{u}(\xi, t) = \mathrm{e}^{-\xi^2 t} \hat{v}(\xi).$$

We conclude that

$$(U^n - u^n)\hat{}(\xi) = \big(\tilde{E}(h\xi)^n - \mathrm{e}^{-nk\xi^2}\big)\hat{v}(\xi),$$

and hence

$$\|U^n - u^n\|^2 = (2\pi)^{-1} \|(\tilde{E}(h\xi)^n - \mathrm{e}^{-nk\xi^2})\hat{v}(\xi)\|^2.$$

Now, by (9.17)

$$|\tilde{E}(h\xi)^n - \mathrm{e}^{-nk\xi^2}| \le Cnh^{r+2}|\xi|^{r+2},$$

so that, using the facts that $(dv/dx)\hat{}(\xi) = -\mathrm{i}\xi\hat{v}(\xi)$ and $\lambda = k/h^2$,

$$\|U^n - u^n\| \le (2\pi)^{-1/2} Cnh^{r+2} \|\xi^{r+2}\hat{v}(\xi)\| \le C\, nk\, h^r \|v^{(r+2)}\|.$$

This shows the conclusion of the theorem under the assumption that the initial data are such that $v^{(r+2)}$ belongs to L_2. In fact, by a more precise argument, using the smoothing property of the solution operator $E(t)$. one may reduce this regularity requirement by almost two derivatives. □

In the above discussion we have only considered finite difference schemes of one-step (or two-level) type, that is, schemes that use the values at time $t = t_n$ to compute the approximate solution at $t = t_{n+1}$. It would also be natural to replace the derivatives in the model heat equation (9.1) by difference quotients in a symmetric fashion around (x, t_n), which would result in the equation

$$\frac{U^{n+1}(x) - U^{n-1}(x)}{2k} = \partial_x \bar{\partial}_x U^n(x). \tag{9.18}$$

In this case, in addition to $U^0 = v$, we also need to prescribe U^1 (presumably by some approximation of $u(\cdot, k)$) in order to be able to use (9.18) to find U^n

for $n \geq 0$. This two-step, or three-level, scheme would formally be accurate of second order in both x and t. Although the particular scheme (9.18) turns out to be unstable for any combination of h and k (cf. Problem 9.6), other multi-step schemes are useful in applications. For instance, one may show that the scheme (9.18) may be stabilized, for any constant λ, by replacing $U^n(x)$ on the right by the average $\frac{1}{2}(U^{n+1}(x)+U^{n-1}(x))$, so that the scheme becomes the Dufort-Frankel scheme

$$\frac{U^{n+1}(x) - U^{n-1}(x)}{2k} = \frac{U^n(x+h) - U^{n+1}(x) - U^{n-1}(x) + U^n(x-h)}{h^2}.$$

We shall end this discussion by making an observation concerning the accuracy of the Dufort-Frankel scheme. Let thus u be a smooth function and replace U by u above. With $\partial_x\bar{\partial}_x$ as before and correspondingly for $\partial_t\bar{\partial}_t$ and with $\hat{\partial}_t$ denoting the symmetric difference quotient

$$\hat{\partial}_t u(x,t) = \frac{u(x,t+k) - u(x,t-k)}{2k} = \tfrac{1}{2}(\partial_t + \bar{\partial}_t)u(x,t),$$

we have for the truncation error

$$\begin{aligned}
&\tau_{h,k,n}(x) \\
&= \frac{u^{n+1}(x) - u^{n-1}(x)}{2k} - \frac{u^n(x+h) - u^{n+1}(x) - u^{n-1}(x) + u^n(x-h)}{h^2} \\
&= \hat{\partial}_t u(x,t_n) - \partial_x\bar{\partial}_x u(x,t_n) + \frac{k^2}{h^2}\partial_t\bar{\partial}_t u(x,t_n) \\
&= (u_t - u_{xx})(x,t_n) + O(k^2) + O(h^2) + \frac{k^2}{h^2}u_{tt}(x,t_n) + O\Big(\frac{k^4}{h^2}\Big).
\end{aligned}$$

Consistency with the heat equation therefore requires that k/h tends to zero, which is the case, for instance, if $k/h^2 = \lambda = \text{constant}$. However, if instead $k/h = \lambda = \text{constant}$, we obtain

$$\tau_{h,k,n}(x) = (u_t - u_{xx} + \lambda^2 u_{tt})(x,t_n) + O(h^2), \qquad \text{as } h \to 0,$$

which shows that the scheme is then consistent, not with the heat equation, but with the second order hyperbolic equation

$$\lambda^2 u_{tt} + u_t - u_{xx} = 0.$$

Much of the analysis of this section generalizes to the initial-value problem for the inhomogeneous equation,

$$\begin{aligned}
u_t &= u_{xx} + f(x,t), && \text{in } \mathbf{R} \times \mathbf{R}_+, \\
u(\cdot,0) &= v, && \text{in } \mathbf{R}.
\end{aligned}$$

For instance, we may apply the forward Euler scheme

$$\begin{aligned} &\partial_t U_j^n = \partial_x \bar{\partial}_x U_j^n + f_j^n, && \text{for } j,\ n \in \mathbf{Z},\ n \geq 0, \\ &U_j^0 = v_j := v(x_j), && \text{for } j \in \mathbf{Z}, \end{aligned}$$

or, with E_k defined as in (9.4),

$$U_j^{n+1} = (E_k U^n)_j + k\, f_j^n.$$

One may conclude at once by the stability of E_k in maximum-norm that

$$\|U^n\|_{\infty,h} \leq \|v\|_{\infty,h} + k \sum_{l=0}^{n-1} \|f^l\|_{\infty,h}.$$

Moreover, in the same way as in the proof of Theorem 9.1 one may easily prove the error estimate

$$\|U^n - u^n\|_{\infty,h} \leq C t_n h^2 \max_{s \leq t_n} \big(|u_{tt}(\cdot, s)|_{\mathcal{C}} + |u(\cdot, s)|_{\mathcal{C}^4} \big).$$

9.2 The Mixed Initial-Boundary Value Problem

In many physical situations our above pure initial value model problem (9.1) is inadequate, and instead it is required to solve the heat equation on a finite interval with boundary values given at the end points of the interval for positive time. We thus have reason to consider the following model problem

$$\begin{aligned} &u_t = u_{xx}, && \text{in } \Omega = (0,1),\ t > 0, \\ &u(0,t) = u(1,t) = 0, && \text{for } t > 0, \\ &u(\cdot, 0) = v, && \text{in } \Omega. \end{aligned} \tag{9.19}$$

For the approximate solution we may again cover the domain with a grid of mesh-points, this time by dividing Ω into subintervals of equal length $h = 1/M$, where M is a positive integer, and setting $(x_j, t_n) = (jh, nk)$ with $j = 0, \ldots, M$ and $n = 0, 1, \ldots$. With U_j^n denoting the approximation of $u(x_j, t_n)$, the explicit *forward Euler scheme* is now

$$\begin{aligned} &\partial_t U_j^n = \partial_x \bar{\partial}_x U_j^n, && \text{for } j = 1, \ldots, M-1,\ n \geq 0, \\ &U_0^n = U_M^n = 0, && \text{for } n > 0, \\ &U_j^0 = V_j = v(x_j), && \text{for } j = 0, \ldots, M, \end{aligned} \tag{9.20}$$

or, for U_j^n, $j = 0, \ldots, M$, given,

$$\begin{aligned} &U_j^{n+1} = \lambda(U_{j-1}^n + U_{j+1}^n) + (1 - 2\lambda) U_j^n, \quad j = 1, \ldots, M-1, \\ &U_0^{n+1} = U_M^{n+1} = 0. \end{aligned}$$

In this case we are thus looking for a sequence of $(M+1)$-vectors $U^n = (U_0^n, \ldots, U_M^n)^T$ with $U_0^n = U_M^n = 0$ satisfying these equations. In the analysis we shall first use the discrete maximum-norm

$$\|U^n\|_{\infty,h} = \max_{0 \le j \le M} |U_j^n|.$$

When $\lambda = k/h^2 \le \frac{1}{2}$ we conclude, as for the pure initial value problem, that

$$\|U^{n+1}\|_{\infty,h} \le \|U^n\|_{\infty,h},$$

or, defining the local solution operator E_k in the obvious way,

$$\|E_k^n V\|_{\infty,h} \le \|V\|_{\infty,h}, \quad \text{for } n \ge 0.$$

The scheme is thus maximum-norm stable for $\lambda \le \frac{1}{2}$.

In order to see that this condition is necessary for stability also in the present case, we modify our counter-example from Sect. 9.1 so as to incorporate the boundary conditions and set

$$U_j^0 = V_j = (-1)^j \sin(\pi j h), \quad \text{for } j = 0, \ldots, M.$$

By a simple calculation analogous to that of the proof of Theorem 9.2 we then have

$$U_j^n = (1 - 2\lambda - 2\lambda\cos(\pi h))^n V_j, \quad \text{for } j = 0, \ldots, M.$$

If $\lambda > \frac{1}{2}$ we have for h sufficiently small

$$|1 - 2\lambda - 2\lambda\cos(\pi h)| \ge \gamma > 1,$$

and hence, if $t_n = 1$, say,

$$\|U^n\|_{\infty,h} \ge \gamma^n \|V\|_{\infty,h} \to \infty, \quad \text{as } h \to 0.$$

In the presence of stability we may derive an error estimate in the same way as for the pure initial value problem. The estimate in (9.7) now shows for the truncation error $\tau_j^n = \partial_t u_j^n - \partial_x \bar{\partial}_x u_j^n$,

$$|\tau_j^n| \le Ch^2 \max_{t \in I_n} |u(\cdot, t)|_{\mathcal{C}^4}, \quad \text{where } I_n = (t_n, t_{n+1}),$$

and we obtain the following error estimate.

Theorem 9.5. *Let U^n and u be the solutions of* (9.20)*, with $\lambda \le \frac{1}{2}$, and* (9.19)*. Then*

$$\|U^n - u^n\|_{\infty,h} \le Ct_n h^2 \max_{t \le t_n} |u(\cdot, t)|_{\mathcal{C}^4}, \quad \textit{for } t_n \ge 0.$$

We remark that in this case, in order for u to be sufficiently regular to guarantee that the right hand side of (9.7) is bounded by $Ch^2|v|_{\mathcal{C}^4}$, we need to require certain compatibility conditions for v with the boundary conditions, namely $v(x) = v''(x) = v^{(iv)}(x) = 0$ for $x = 0, 1$.

We note that a method of the form

$$U_j^{n+1} = \sum_p a_p U_{j-p}^n, \quad \text{for } j = 1, \dots, M-1,$$

is not suitable here if $a_p \neq 0$ for some $|p| > 1$, since then for some interior mesh-point of Ω the equation uses mesh-points outside this interval. In such a case the equation has to be modified near the endpoints, which significantly complicates the analysis.

The stability requirement $k \leq \frac{1}{2}h^2$ used for the forward Euler method is quite restrictive in practice, and it would be desirable to relax it to be able to use h and k of the same order of magnitude. For this purpose one may define an *implicit method*, instead of the explicit method considered above, by the *backward Euler scheme*

$$\begin{aligned} &\bar{\partial}_t U_j^{n+1} = \partial_x \bar{\partial}_x U_j^{n+1}, && \text{for } j = 1, \dots, M-1,\ n \geq 0, \\ &U_0^{n+1} = U_M^{n+1} = 0, && \text{for } n \geq 0, \\ &U_j^0 = V_j = v(x_j), && \text{for } j = 0, \dots, M. \end{aligned} \tag{9.21}$$

For U^n given this may be put in the form

$$\begin{aligned} &(1+2\lambda)U_j^{n+1} - \lambda(U_{j-1}^{n+1} + U_{j+1}^{n+1}) = U_j^n, \quad j = 1, \dots, M-1, \\ &U_0^{n+1} = U_M^{n+1} = 0, \end{aligned}$$

which is a linear system of equations for the determination of U^{n+1}. In matrix notation it may be written as

$$B\bar{U}^{n+1} = \bar{U}^n, \tag{9.22}$$

where $\bar{U}^{n+1}$ and $\bar{U}^n$ are now thought of as vectors with $M-1$ components corresponding to the interior mesh-points and B is the diagonally dominant, symmetric, tridiagonal matrix

$$B = \begin{bmatrix} 1+2\lambda & -\lambda & 0 & \dots & 0 \\ -\lambda & 1+2\lambda & -\lambda & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -\lambda & 1+2\lambda & -\lambda \\ 0 & \dots & 0 & -\lambda & 1+2\lambda \end{bmatrix}.$$

Clearly the system (9.22) may easily be solved for $\bar{U}^{n+1}$.

Introducing the finite dimensional space l_h^0 of $(M+1)$-vectors $\{V_j\}_{j=0}^M$ with $V_0 = V_M = 0$, and the operator B_{kh} on l_h^0 defined by

$$(B_{kh}V)_j = (1+2\lambda)V_j - \lambda(V_{j-1}+V_{j+1}) = V_j - k\partial_x\bar{\partial}_x V_j, \quad j = 1,\dots,M-1,$$

we may write the above system may as

$$B_{kh}U^{n+1} = U^n,$$

or, again with E_k denoting the local solution operator,

$$U^{n+1} = B_{kh}^{-1}U^n = E_k U^n.$$

We shall now show that this method is stable in maximum-norm without any restrictions on k and h, or, more precisely,

$$\|U^{n+1}\|_{\infty,h} \le \|U^n\|_{\infty,h}, \quad \text{for } n \ge 0. \tag{9.23}$$

In fact, with suitable j_0,

$$\begin{aligned}\|U^{n+1}\|_{\infty,h} = |U_{j_0}^{n+1}| &\le \frac{1}{1+2\lambda}\Big(\lambda\big(|U_{j_0-1}^{n+1}| + |U_{j_0+1}^{n+1}|\big) + |U_{j_0}^n|\Big)\\ &\le \frac{2\lambda}{1+2\lambda}\,\|U^{n+1}\|_{\infty,h} + \frac{1}{1+2\lambda}\,\|U^n\|_{\infty,h},\end{aligned}$$

from which (9.23) follows at once. This implies the stability estimate

$$\|U^n\|_{\infty,h} = \|E_k^n V\|_{\infty,h} \le \|V\|_{\infty,h}. \tag{9.24}$$

The solution operator E_k^n is thus stable in maximum-norm and convergence of U^n to $u(t_n)$ may also be proved. This time we have for the truncation error

$$\tau_j^n = \bar{\partial}_t u_j^{n+1} - \partial_x\bar{\partial}_x u_j^{n+1} = O(k+h^2), \quad \text{as } k,h \to 0, \text{ for } j = 1,\dots,M-1,$$

where, since h and k are unrelated, the latter expression does not reduce to $O(h^2)$. As a consequence the convergence result now reads as follows.

Theorem 9.6. *Let U^n and u be the solutions of* (9.19) *and* (9.21). *Then*

$$\|U^n - u^n\|_{\infty,h} \le C\,t_n(h^2+k)\max_{t\le t_n}|u(\cdot,t)|_{\mathcal{C}^4}, \quad \textit{for } t_n \ge 0.$$

Proof. Defining the error $z^n = U^n - u^n$ we may write

$$B_{kh}z^{n+1} = B_{kh}U^{n+1} - B_{kh}u^{n+1} = U^n - (u^{n+1} - k\partial_x\bar{\partial}_x u^{n+1}) = z^n - k\,\tau^n,$$

where we consider τ^n to be an element of l_h^0. Thus

$$z^{n+1} = E_k z^n - kE_k\tau^n,$$

and hence

$$z^n = -k \sum_{l=0}^{n-1} E_k^{n-l} \tau^l.$$

The estimate (9.7) is now replaced by

$$\|\tau^n\|_{\infty,h} \le C(h^2 + k) \max_{t \in I_{n-1}} |u(\cdot, t)|_{\mathcal{C}^4}.$$

Using (9.24) we then obtain

$$\|z^n\|_{\infty,h} \le k \sum_{l=0}^{n-1} \|\tau^l\|_{\infty,h} \le C\, t_n (h^2 + k) \max_{t \le t_n} |u(\cdot, t)|_{\mathcal{C}^4},$$

which concludes the proof. □

The above convergence result for the backward Euler method is satisfactory in that it requires no restriction on the mesh-ratio $\lambda = k/h^2$. On the other hand, since it is only first order accurate in time, the error in the time discretization will dominate unless k is chosen much smaller than h. It would thus be desirable to find a stable method which is second order accurate also with respect to time. Such a method is provided by the *Crank-Nicolson scheme*, which was introduced for a system of ordinary differential equations in Sect. 7.2. This uses symmetry around the point $(x_j, t_{n+1/2})$ and is defined by

$$\begin{aligned} &\bar\delta_t U_j^{n+1} = \tfrac{1}{2} \partial_x \bar\delta_x \big(U_j^n + U_j^{n+1}\big), && \text{for } j = 1, \dots, M-1,\ n \ge 0, \\ &U_0^{n+1} = U_M^{n+1} = 0, && \text{for } n \ge 0, \\ &U_j^0 = V_j := v(jh), && \text{for } j = 0, \dots, M. \end{aligned} \tag{9.25}$$

The first equation may also be written

$$(I - \tfrac{1}{2} k \partial_x \bar\delta_x) U_j^{n+1} = (I + \tfrac{1}{2} k \partial_x \bar\delta_x) U_j^n,$$

or

$$(1+\lambda) U_j^{n+1} - \tfrac{1}{2}\lambda (U_{j-1}^{n+1} + U_{j+1}^{n+1}) = (1-\lambda) U_j^n + \tfrac{1}{2}\lambda (U_{j-1}^n + U_{j+1}^n),$$

and, in matrix form, with $\bar U^n$ again denoting the $(M-1)$-vector associated with U^n,

$$B \bar U^{n+1} = A \bar U^n,$$

where now both A and B are symmetric tridiagonal matrices, with B diagonally dominant:

$$B = \begin{bmatrix} 1+\lambda & -\frac{1}{2}\lambda & 0 & \dots & 0 \\ -\frac{1}{2}\lambda & 1+\lambda & -\frac{1}{2}\lambda & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -\frac{1}{2}\lambda & 1+\lambda & -\frac{1}{2}\lambda \\ 0 & \dots & 0 & -\frac{1}{2}\lambda & 1+\lambda \end{bmatrix},$$

and

$$A = \begin{bmatrix} 1-\lambda & \frac{1}{2}\lambda & 0 & \dots & 0 \\ \frac{1}{2}\lambda & 1-\lambda & \frac{1}{2}\lambda & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \frac{1}{2}\lambda & 1-\lambda & \frac{1}{2}\lambda \\ 0 & \dots & 0 & \frac{1}{2}\lambda & 1-\lambda \end{bmatrix}.$$

With obvious notation we also have

$$B_{kh}U^{n+1} = A_{kh}U^n,$$

or

$$U^{n+1} = B_{kh}^{-1}A_{kh}U^n = E_kU^n,$$

where, similarly to the above,

$$\|B_{kh}^{-1}V\|_{\infty,h} \le \|V\|_{\infty,h}.$$

The same approach to stability as for the backward Euler method gives, for $\lambda \le 1$, since the coefficients on the right are then non-negative,

$$(1+\lambda)\|U^{n+1}\|_{\infty,h} \le \lambda\|U^{n+1}\|_{\infty,h} + \|U^n\|_{\infty,h},$$

or

$$\|U^{n+1}\|_{\infty,h} \le \|U^n\|_{\infty,h},$$

which shows stability. However, if $\lambda > 1$, which is the interesting case if we want to be able to take h and k of the same order, one obtains instead

$$(1+\lambda)\|U^{n+1}\|_{\infty,h} \le \lambda\|U^{n+1}\|_{\infty,h} + (2\lambda-1)\|U^n\|_{\infty,h},$$

which does not yield maximum-norm stability, since $2\lambda - 1 > 1$. For $\lambda \le 1$ we have immediately as before a $O(k^2+h^2) = O(h^2)$ convergence estimate.

In order to be able to deal with $\lambda > 1$, we now instead turn to an analysis in an l_2 type norm. We introduce thus for vectors $V = (V_0, \dots, V_M)^T$ the inner product

$$(V,W)_h = h\sum_{j=0}^{M} V_jW_j,$$

and the corresponding norm

$$\|V\|_{2,h} = (V,V)_h^{1/2} = \Big(h\sum_{j=0}^{M} V_j^2\Big)^{1/2}.$$

We denote by $l_{2,h}^0$ the space l_h^0 equipped with this inner product and norm, and note that this space is spanned by the $M-1$ vectors φ_p, $p = 1, \dots, M-1$, with components

$$\varphi_{p,j} = \sqrt{2}\sin(\pi pjh), \quad \text{for } j = 0, \dots, M,$$

and that these form an orthonormal basis with respect to the above inner product (cf. Problem 9.7), i.e.,

$$(\varphi_p, \varphi_q)_h = \delta_{pq} = \begin{cases} 1, & \text{if } p = q, \\ 0, & \text{if } p \neq q. \end{cases}$$

We also observe that the φ_p are eigenfunctions of the finite difference operators $-\partial_x\bar{\partial}_x$,

$$-\partial_x\bar{\partial}_x\varphi_{p,j} = \frac{2}{h^2}(1 - \cos(\pi ph))\varphi_{p,j}, \quad \text{for } j = 1, \dots, M-1.$$

We shall now to discuss the stability within this framework of the three difference schemes considered above. Let V be given initial data in $l^0_{2,h}$. Then

$$V = \sum_{p=1}^{M-1} \hat{V}_p\varphi_p, \quad \text{where } \hat{V}_p = (V, \varphi_p)_h.$$

The forward Euler method then gives

$$U^1_j = V_j + k\partial_x\bar{\partial}_x V_j = \sum_{p=1}^{M-1} \hat{V}_p\big(1 - 2\lambda(1 - \cos(\pi ph))\big)\varphi_{p,j}, \ j = 1, \dots, M-1,$$

with $U^1_0 = U^1_M = 0$, or, more generally,

$$U^n_j = \sum_{p=1}^{M-1} \hat{V}_p\tilde{E}(\pi ph)^n\varphi_{p,j}, \quad j = 0, \dots, M, \tag{9.26}$$

where $\tilde{E}(\xi)$ is the symbol of the local discrete solution operator E_k,

$$\tilde{E}(\xi) = 1 - 2\lambda + 2\lambda\cos\xi.$$

By Parseval's relation we have thus

$$\|U^n\|_{2,h} = \Big(\sum_{p=1}^{M-1} \hat{V}_p^2\tilde{E}(\pi ph)^{2n}\Big)^{1/2} \le \max_p |\tilde{E}(\pi ph)^n| \, \|V\|_{2,h},$$

with equality for the appropriate V. Now for $1 \le p \le M-1$ we have

$$\begin{aligned} |\tilde{E}(\pi ph)| &= \max\big\{|1 - 2\lambda(1 - \cos(\pi h))|, |1 - 2\lambda(1 - \cos(\pi(M-1)h))|\big\} \\ &= \max\big\{|1 - 2\lambda(1 - \cos(\pi h))|, |1 - 2\lambda(1 + \cos(\pi h))|\big\}. \end{aligned}$$

We thus have $\max_p |\tilde{E}(ph\pi)| \le 1$ for small h if and only if $4\lambda - 1 \le 1$, or $\lambda \le \frac{1}{2}$, and it follows in this case that

$$\|U^n\|_{2,h} \le \|V\|_{2,h}. \tag{9.27}$$

Consequently, the forward Euler scheme is stable in $l^0_{2,h}$ if and only if $\lambda \le \frac{1}{2}$, i.e., under the same conditions as for the maximum-norm.

The corresponding analysis for the backward Euler scheme gives (9.26), where now

$$\tilde{E}(\xi) = \frac{1}{1 + 2\lambda(1 - \cos\xi)}.$$

In this case $0 \le \tilde{E}(\pi ph) \le 1$ for all p and λ and hence (9.27) is valid for any value of λ.

Similarly, for the Crank-Nicolson scheme, (9.26) holds with

$$\tilde{E}(\xi) = \frac{1 - \lambda(1 - \cos\xi)}{1 + \lambda(1 - \cos\xi)},$$

and we now note that $|\tilde{E}(\xi)| \le 1$ and all ξ for any $\lambda > 0$. Thus, the Fourier analysis method shows stability in $l^0_{2,h}$ for any λ. The convergence follows again by the standard method and gives the following.

Theorem 9.7. *Let U^n and u be the solutions of* (9.25) *and* (9.19). *Then*

$$\|U^n - u^n\|_{2,h} \le C\, t_n (h^2 + k^2) \max_{t \le t_n} |u(\cdot, t)|_{\mathcal{C}^6}, \quad \textit{for } t_n \ge 0,$$

Proof. We write the truncation error

$$\tau_j^n = \bar{\partial}_t u_j^{n+1} - \partial_x \bar{\partial}_x \frac{u_j^n + u_j^{n+1}}{2} = \Big(\bar{\partial}_t u_j^{n+1} - u_t(x_j, t_{n+1/2})\Big)$$
$$+ \partial_x \bar{\partial}_x \Big(\frac{u_j^n + u_j^{n+1}}{2} - u_j^{n+1/2}\Big) + \Big(\partial_x \bar{\partial}_x u_j^{n+1/2} - u_{xx}(x_j, t_{n+1/2})\Big),$$

and hence, using Taylor expansions as earlier,

$$\|\tau^n\|_{2,h} \le C(h^2 + k^2) \max_{t \in I_n} |u(\cdot, t)|_{\mathcal{C}^6}.$$

In the same way as before the error $z_j^n = U_j^n - u_j^n$ satisfies

$$z^{n+1} = U^{n+1} - u^{n+1} = E_k U^n - B_{kh}^{-1} B_{kh} u^{n+1} = E_k z^n - k\, B_{kh}^{-1} \tau^n,$$

or

$$z^n = -k \sum_{l=0}^{n-1} E_k^{n-1-l} B_{kh}^{-1} \tau^l,$$

from which the result follows by using the stability of the Crank-Nicolson operator E_k^n and the boundedness of B_{kh}^{-1}. □

The forward and backward Euler methods and the Crank-Nicolson method may all be considered to be special cases of the *θ-method* defined by

$$\bar{\partial}_t U_j^{n+1} = \theta \partial_x \bar{\partial}_x U_j^{n+1} + (1-\theta)\partial_x \bar{\partial}_x U_j^n, \quad j = 1, \dots, M-1, \tag{9.28}$$

with $\theta = 0$ for the forward Euler, $\theta = 1$ for the backward Euler, and $\theta = 1/2$ for the Crank-Nicolson method. The equation (9.28) may be written as

$$(I - \theta k \partial_x \bar{\partial}_x) U^{n+1} = (I + (1-\theta) k \partial_x \bar{\partial}_x) U^n,$$

and we find this time for the symbol

$$\tilde{E}(\xi) = \frac{1 - 2(1-\theta)\lambda(1-\cos\xi)}{1 + 2\theta\lambda(1-\cos\xi)}.$$

Assuming $0 \le \theta \le 1$ we have $\tilde{E}(\xi) \le 1$ for $\xi \in \mathbf{R}$, and the stability requirement reduces to

$$\min_\xi \tilde{E}(\xi) = \frac{1 - 4(1-\theta)\lambda}{1 + 4\theta\lambda} \ge -1,$$

or

$$(1 - 2\theta)\lambda \le \tfrac{1}{2}.$$

Hence the θ method is unconditionally stable in $l^0_{2,h}$, i.e., stable in $l^0_{2,h}$ for all λ, if $\theta \ge 1/2$, whereas for $\theta < 1/2$ stability holds if and only if

$$\lambda \le \frac{1}{2(1-2\theta)}.$$

9.3 Problems

Problem 9.1. Show the equivalence of definitions (9.15) and (9.16) of accuracy of order r. Use the alternate definition (9.16) to show that the accuracy of (9.4) is of order 4, if $\lambda = 1/6$.

Problem 9.2. Formulate and prove an analogue of Theorem 9.2 in two space dimensions.

Problem 9.3. Let (a_{jk}) be a symmetric, positive definite 2×2 matrix. For the solution of the initial value problem

$$\begin{aligned} \frac{\partial u}{\partial t} &= \sum_{j,k=1}^{2} a_{jk} \frac{\partial^2 u}{\partial x_j \partial x_k}, && \text{in } \mathbf{R}^2 \times \mathbf{R}_+, \\ u(\cdot, 0) &= v, && \text{in } \mathbf{R}^2, \end{aligned}$$

we wish to apply the finite difference method

$$\partial_t U_{ij}^n = \sum_{k,l=1}^{2} a_{kl}\, \partial_{x_k} \bar{\partial}_{x_l} U_{ij}^n.$$

(a) Give sufficient conditions on the coefficients for the method to be stable in the maximum-norm.
(b) Is the method stable in $l_{2,h}$?

Problem 9.4. Find an explicit 5-point finite difference operator for (9.1) of the form (9.8) (i.e., with five terms on the right-hand side of (9.8)) of order of accuracy 4. Discuss the stability of this operator.

Problem 9.5. Formulate a finite difference method for

$$\begin{aligned} u_t &= \Delta u, && \text{in } \mathbf{R}^2 \times \mathbf{R}_+, \\ u(\cdot, 0) &= v, && \text{in } \mathbf{R}^2, \end{aligned}$$

such that

$$\|U^n - u^n\|_{\infty,h} = O(h^4), \quad \text{as } h \to 0.$$

Problem 9.6. Consider the three-level finite difference method (9.18), and let $U^0 = V$ and $U^1 = W$, with V, $W \in L_2(\mathbf{R})$. Show that

$$\hat{U}^n(\xi) = c_1(\xi)\tau_1(\xi)^n + c_2(\xi)\tau_2(\xi)^n,$$

where $\tau_{1,2}(\xi)$ are the roots of the equation

$$\tau^2 + 4\lambda(1 - \cos\xi)\tau - 1 = 0, \quad \text{with } \lambda = k/h^2,$$

and $c_1(\xi)$ and $c_2(\xi)$ are determined from

$$c_1(\xi) + c_2(\xi) = \hat{V}(\xi), \quad c_1(\xi)\tau_1(\xi) + c_2(\xi)\tau_2(\xi) = \hat{W}(\xi).$$

Use this to show that $\|U^n\| \to \infty$ as $n \to \infty$ for any $\lambda > 0$, and thus that (9.18) is unstable.

Problem 9.7. Let $\{\varphi_p\}_{p=1}^{M-1}$ be defined by (9.2). Show that they form an orthonormal basis for $l_{2,h}^0$ and that they are eigenfunctions of the difference operator $-\partial_x \bar{\partial}_x$ with eigenvalues $2h^{-2}(1 - \cos(\pi p h))$. Compare with the eigenfunctions and eigenvalues of $-d^2/dx^2$. Note that one of the φ_p gives the counter-example to stability in the beginning of Sect. 9.2.

Problem 9.8. (A discrete maximum principle.) Let $\Omega \subset \mathbf{R}$ be a bounded interval and $I = (0, T]$. Show that if $\lambda = kh^{-2} \leq \frac{1}{2}$ and

$$\partial_t U_j^n - \partial_x \bar{\partial}_x U_j^n \leq 0, \quad \text{for } (x_j, t_n) \in \Omega \times I,$$

then U_j^n attains its maximum on the parabolic boundary Γ_p, cf. Theorem 8.6. Hint: Use the argument leading to (9.6). Prove a similar result for the backward Euler method.

Problem 9.9. We know that all norms on the finite dimensional space l_h^0 are equivalent. For example, show that

$$\|V\|_{2,h} \le \|V\|_{\infty,h} \le h^{-1/2}\|V\|_{2,h}, \qquad \text{for } V \in l_h^0,$$

and that these inequalities are sharp. Note that the equivalence is not uniform in h and is lost when $h \to 0$, that is, when the dimension of l_h^0 tends to infinity. The second inequality above has the same character as the inverse inequality (6.37), relating a stronger norm ($\|\cdot\|_{\infty,h}$) to a weaker norm ($\|\cdot\|_{2,h}$).

Problem 9.10. Show that the function $\varphi(x) = \mathrm{e}^{\mathrm{i}x\xi}$ is an eigenfunction of the differential and difference operators $\partial/\partial x$, ∂_x, and $\bar{\partial}_x$.

Problem 9.11. (Computer exercise.) Consider the initial-boundary value problem (9.19) with $v(x) = \sin(\pi x) - \sin(3\pi x)$. Apply the forward Euler method with $h = 1/10$ and $k = 1/600, 1/300, 1/100$. Apply also the Crank-Nicolson method with $h = k = 1/10$. Calculate the error at $(1/2, 1)$.

10 The Finite Element Method for a Parabolic Problem

In this chapter we consider the approximation of solutions of the model heat equation in two space dimensions by means of Galerkin's method, using piecewise linear trial functions. In Sect. 10.1 we consider the discretization with respect to the space variables only, and in the following Sect. 10.2 we study some completely discrete schemes.

10.1 The Semidiscrete Galerkin Finite Element Method

Let $\Omega \subset \mathbf{R}^2$ be a bounded convex domain with smooth boundary Γ, and consider the initial-boundary value problem,

$$
\begin{aligned}
&u_t - \Delta u = f, && \text{in } \Omega \times \mathbf{R}_+,\\
&u = 0, && \text{on } \Gamma \times \mathbf{R}_+,\\
&u(\cdot, 0) = v, && \text{in } \Omega,
\end{aligned}
\tag{10.1}
$$

where u_t denotes $\partial u/\partial t$ and Δ the Laplacian $\partial^2/\partial x_1^2 + \partial^2/\partial x_2^2$. In the first step we shall approximate the solution $u(x,t)$ by means of a function $u_h(x,t)$ which, for each fixed t, is a piecewise linear function of x over a triangulation $\mathcal{T}_h$ of Ω, thus depending on a finite number of parameters.

Thus, let $\mathcal{T}_h = \{K\}$ denote a triangulation of Ω of the type considered in Sect. 5.2 and let $\{P_j\}_{j=1}^{M_h}$ be the interior nodes of $\mathcal{T}_h$. Further, let S_h denote the continuous piecewise linear functions on $\mathcal{T}_h$ which vanish on $\partial\Omega$ and let $\{\Phi_j\}_{j=1}^{M_h}$ be the standard basis of S_h corresponding to the nodes $\{P_j\}_{j=1}^{M_h}$. Recall the definition (5.28) of the interpolant $I_h : \mathcal{C}_0(\bar{\Omega}) \to S_h$, and the error bounds (5.34) with $r = 2$.

For the purpose of defining thus an approximate solution to the initial boundary value problem (10.1) we first write this in weak form as in Sect. 8.3, i.e., with the definitions there,

$$
(u_t, \varphi) + a(u, \varphi) = (f, \varphi), \quad \forall \varphi \in H_0^1,\ t > 0. \tag{10.2}
$$

We then pose the approximate problem to find $u_h(t) = u_h(\cdot, t)$, belonging to S_h for each t, such that

$$\begin{aligned}(u_{h,t},\chi)+a(u_h,\chi)&=(f,\chi),\quad \forall\chi\in S_h,\ t>0,\\ u_h(0)&=v_h,\end{aligned}\tag{10.3}$$

where $v_h \in S_h$ is some approximation of v. Since we have discretized only in the space variables, this is referred to as a *spatially semidiscrete* problem. In the next section, we shall discretize also in the time variable to produce completely discrete schemes.

In terms of the basis $\{\Phi_j\}_{j=1}^{M_h}$ our semidiscrete problem may be stated: Find the coefficients $\alpha_j(t)$ in

$$u_h(x,t)=\sum_{j=1}^{M_h}\alpha_j(t)\Phi_j(x),$$

such that

$$\sum_{j=1}^{M_h}\alpha_j'(t)(\Phi_j,\Phi_k)+\sum_{j=1}^{M_h}\alpha_j(t)a(\Phi_j,\Phi_k)=(f(t),\Phi_k),\quad k=1,\dots,M_h,$$

and, with γ_j denoting the nodal values of the given initial approximation v_h,

$$\alpha_j(0)=\gamma_j,\quad j=1,\dots,M_h.$$

In matrix notation this may be expressed as

$$B\alpha'(t)+A\alpha(t)=b(t),\quad \text{for } t>0,\quad \text{with } \alpha(0)=\gamma,\tag{10.4}$$

where $B=(b_{kj})$ is the mass matrix with elements $b_{kj}=(\Phi_j,\Phi_k)$, $A=(a_{kj})$ the stiffness matrix with $a_{kj}=a(\Phi_j,\Phi_k)$, $b=(b_k)$ the vector with entries $b_k=(f,\Phi_k)$, $\alpha(t)$ the vector of unknowns $\alpha_j(t)$, and $\gamma=(\gamma_j)$. The dimension of all these items equals M_h, the number of interior nodes of $\mathcal{T}_h$.

We recall from Sect. 5.2 that the stiffness matrix A is symmetric positive definite, and this holds also for the mass matrix B since

$$\sum_{k,j=1}^{M_h}\xi_j\xi_k(\Phi_j,\Phi_k)=\Big\|\sum_{j=1}^{M_h}\xi_j\Phi_j\Big\|^2\ge 0,$$

and since equality can only occur if the vector $\xi=0$. In particular, B is invertible, and therefore the above system of ordinary differential equations may be written

$$\alpha'(t)+B^{-1}A\alpha(t)=B^{-1}b(t),\quad \text{for } t>0,\quad \text{with } \alpha(0)=\gamma,$$

and hence obviously has a unique solution for t positive.

We begin our analysis by considering the stability of the semidiscrete method. Since $u_h(t)\in S_h$ we may choose $\chi=u_h(t)$ in (10.3) to obtain

$$(u_{h,t}, u_h) + a(u_h, u_h) = (f, u_h), \quad \text{for } t > 0,$$

or, since the first term equals $\frac{1}{2}\frac{d}{dt}\|u_h\|^2$ and the second is non-negative,

$$\tfrac{1}{2}\frac{d}{dt}\|u_h\|^2 = \|u_h\|\frac{d}{dt}\|u_h\| \le \|f\|\,\|u_h\|.$$

This yields

$$\frac{d}{dt}\|u_h\| \le \|f\|,$$

which after integration shows the stability estimate

$$\|u_h(t)\| \le \|v_h\| + \int_0^t \|f\|\,\mathrm{d}s. \tag{10.5}$$

For the purpose of writing equation in (10.3) in operator form, we introduce a *discrete Laplacian* Δ_h, which we think of as an operator from S_h into itself, defined by

$$(-\Delta_h\psi, \chi) = a(\psi, \chi), \quad \forall \psi, \chi \in S_h. \tag{10.6}$$

This discrete analogue of Green's formula clearly defines $\Delta_h\psi = \sum_{j=1}^{M_h} d_j\Phi_j$ from

$$\sum_{j=1}^{M_h} d_j(\Phi_j, \Phi_k) = -a(\psi, \Phi_k), \quad k = 1, \dots, M_h,$$

since the matrix of this system is the positive definite mass matrix encountered above. The operator Δ_h is easily seen to be selfadjoint and $-\Delta_h$ is positive definite in S_h with respect to the L_2-inner product, see Problem 10.3. With P_h denoting the L_2-projection onto S_h, the equation in (10.3) may now be written

$$(u_{h,t} - \Delta_h u_h - P_h f, \chi) = 0, \quad \forall \chi \in S_h,$$

or, noting that the first factor is in S_h, so that χ may be chosen equal to it, it follows that

$$u_{h,t} - \Delta_h u_h = P_h f, \quad \text{for } t > 0, \quad \text{with } u_h(0) = v_h, \tag{10.7}$$

We denote by $E_h(t)$ the solution operator of the homogeneous case of the semidiscrete equation in (10.7), with $f = 0$. Hence $E_h(t)$ is the operator which takes the initial data $u_h(0) = v_h$ into the solution $u_h(t)$ at time t, so that $u_h(t) = E_h(t)v_h$. It is then easy to show (cf. Duhamel's principle (8.22)) that the solution of the initial value problem (10.7) is

$$u_h(t) = E_h(t)v_h + \int_0^t E_h(t-s)P_h f(s)\,\mathrm{d}s. \tag{10.8}$$

We now note that it follows from (10.5) that $E_h(t)$ is stable in L_2, or

$$\|E_h(t)v_h\| \le \|v_h\|, \quad \forall v_h \in S_h. \tag{10.9}$$

Since also P_h has unit norm in L_2 this, together with (10.8), re-establishes the stability estimate (10.5) for the inhomogeneous equation, so that, in fact, it suffices to show stability for the homogeneous equation.

We shall prove the following estimate for the error between the solutions of the semidiscrete and continuous problems.

Theorem 10.1. *Let u_h and u be the solutions of* (10.3) *and* (10.1). *Then*

$$\|u_h(t) - u(t)\| \le \|v_h - v\| + Ch^2\Big(\|v\|_2 + \int_0^t \|u_t\|_2 \,\mathrm{d}s\Big), \quad \textit{for } t \ge 0.$$

Here we require, as usual, that the solution of the continuous problem has the regularity implicitly assumed by the presence of the norms on the right. Note also that for $v_h = I_h v$, (5.31) shows that

$$\|v_h - v\| \le Ch^2\|v\|_2, \tag{10.10}$$

in which case the first term on the right is dominated by the second. The same holds true if $v_h = P_h v$, where P_h denotes the orthogonal projection of L_2 onto S_h, since this choice is the best approximation of v in S_h with respect to the L_2-norm, see (5.39). Another choice of optimal order is $v_h = R_h v$, where R_h is the elliptic (or Ritz) projection onto S_h defined in (5.49) by

$$a(R_h v, \chi) = a(v, \chi), \quad \forall \chi \in S_h. \tag{10.11}$$

Thus $R_h v$ is the finite element approximation of the solution of the elliptic problem whose exact solution is v. We recall the error estimates of Theorem 5.5,

$$\|R_h v - v\| + h|R_h v - v|_1 \le Ch^s\|v\|_s, \quad \text{for } s = 1, 2. \tag{10.12}$$

We now turn to the

Proof of Theorem 10.1. In the main step of the proof we shall compare the solution of the semidiscrete problem to the elliptic projection of the exact solution. We write

$$u_h - u = (u_h - R_h u) + (R_h u - u) = \theta + \rho. \tag{10.13}$$

The second term is easily bounded using (10.12) and obvious estimates by

$$\|\rho(t)\| \le Ch^2\|u(t)\|_2 = Ch^2\Big\|v + \int_0^t u_t \,\mathrm{d}s\Big\|_2 \le Ch^2\Big(\|v\|_2 + \int_0^t \|u_t\|_2 \,\mathrm{d}s\Big).$$

In order to bound θ, we note that

$$(\theta_t,\chi)+a(\theta,\chi)=(u_{h,t},\chi)+a(u_h,\chi)-(R_hu_t,\chi)-a(R_hu,\chi)$$
$$=(f,\chi)-(R_hu_t,\chi)-a(u,\chi)=(u_t-R_hu_t,\chi), \tag{10.14}$$

or

$$(\theta_t,\chi)+a(\theta,\chi)=-(\rho_t,\chi),\quad \forall\chi\in S_h. \tag{10.15}$$

In this derivation we have used (10.3), (10.2), the definition of R_h in (10.11), and the easily established fact that this operator commutes with time differentiation, i.e., $R_hu_t=(R_hu)_t$. We may now apply the stability estimate (10.5) to (10.15) to obtain

$$\|\theta(t)\|\le\|\theta(0)\|+\int_0^t\|\rho_t\|\,\mathrm{d}s.$$

Here

$$\|\theta(0)\|=\|v_h-R_hv\|\le\|v_h-v\|+\|R_hv-v\|\le\|v_h-v\|+Ch^2\|v\|_2,$$

and further

$$\|\rho_t\|=\|R_hu_t-u_t\|\le Ch^2\|u_t\|_2.$$

Together these estimates prove the theorem. □

We see from the proof of Theorem 10.1 that the error estimate for the semidiscrete parabolic problem is thus a consequence of the stability for this problem combined with the error estimate for the elliptic problem, expressed in terms of $\rho=(R_h-I)u$.

Recalling the maximum principle for parabolic equations, Theorem 8.7, we find at once that, for the solution operator $E(t)$ of the homogeneous case of the initial boundary value problem (10.1), we have $\|E(t)v\|_{\mathcal{C}}\le\|v\|_{\mathcal{C}}$ for $t\ge0$. The corresponding maximum principle does not hold for the finite element problem, but it may be shown that, if the family $\{\mathcal{T}_h\}$ of triangulations is quasi-uniform, cf. (5.52), then for some $C>1$,

$$\|E_h(t)v_h\|_{\mathcal{C}}\le C\|v_h\|_{\mathcal{C}},\quad \text{for } t\ge0.$$

This may be combined with the error estimate (5.53) for the stationary problem to show a maximum-norm error estimate for the parabolic problem.

In this regard we mention a variant of the semidiscrete problem (10.2) for which a maximum principle sometimes holds, namely the *lumped mass method*. To define this we replace the matrix B in (10.4) by a diagonal matrix $\bar{B}$, in which the diagonal elements are the row sums of B. One can show that this method can also be defined by

$$(u_{h,t},\chi)_h+a(u_h,\chi)=(f,\chi),\quad \forall\chi\in S_h,\quad \text{for } t>0, \tag{10.16}$$

where the inner product in the first term has been obtained by computing the first term in (10.2) by using the nodal quadrature rule (5.64). For this

method one may derive a $O(h^2)$ error estimate similar to that of Theorem 10.1. If we now assume that all angles of the triangulations are $\leq \pi/2$, then the off-diagonal elements of the stiffness matrix A are nonpositive, and as a result of this one may show that, if $\bar{E}_h(t)$ denotes the solution operator of the modified problem, then

$$\|\bar{E}_h(t)v_h\|_{\mathcal{C}} \leq \|v_h\|_{\mathcal{C}}, \quad \text{for } t \geq 0.$$

This is a discrete maximum principle, which is not true for the standard finite element method.

Returning to the standard Galerkin method (10.3) we now prove the following estimate for the error in the gradient.

Theorem 10.2. *Under the assumptions of Theorem 10.1, we have for* $t \geq 0$,

$$|u_h(t) - u(t)|_1 \leq |v_h - v|_1 + Ch\Big\{\|v\|_2 + \|u(t)\|_2 + \Big(\int_0^t \|u_t\|_1^2 \,\mathrm{d}s\Big)^{1/2}\Big\}.$$

Proof. As before we write the error in the form (10.13). Here by (10.12),

$$|\rho(t)|_1 = |R_h u(t) - u(t)|_1 \leq Ch\|u(t)\|_2.$$

In order to estimate $\nabla\theta$ we use again (10.15), now with $\chi = \theta_t$. We obtain

$$\|\theta_t\|^2 + \tfrac{1}{2}\frac{d}{dt}|\theta|_1^2 = -(\rho_t, \theta_t) \leq \tfrac{1}{2}(\|\rho_t\|^2 + \|\theta_t\|^2),$$

so that

$$\frac{d}{dt}|\theta|_1^2 \leq \|\rho_t\|^2,$$

or

$$|\theta(t)|_1^2 \leq |\theta(0)|_1^2 + \int_0^t \|\rho_t\|^2 \,\mathrm{d}s \leq \big(|v_h - v|_1 + |R_h v - v|_1\big)^2 + \int_0^t \|\rho_t\|^2 \,\mathrm{d}s.$$

Hence, since $a^2 + b^2 \leq (|a| + |b|)^2$ and in view of (10.12), we conclude

$$|\theta(t)|_1 \leq |v_h - v|_1 + Ch\Big\{\|v\|_2 + \Big(\int_0^t \|u_t\|_1^2 \,\mathrm{d}s\Big)^{1/2}\Big\}, \tag{10.17}$$

which completes the proof. □

Note that if $v_h = I_h v$ or $R_h v$, then

$$|v_h - v|_1 \leq Ch\|v\|_2,$$

so that the first term on the right in Theorem 10.2 is dominated by the second.

We make the following observation concerning $\theta = u_h - R_h u$: Assume that we choose $v_h = R_h v$, so that $\theta(0) = 0$. Then in addition to (10.17) we have

$$|\theta(t)|_1 \le \Big(\int_0^t \|\rho_t\|_2 \, \mathrm{d}s \Big)^{1/2} \le Ch^2 \Big(\int_0^t \|u_t\|_2^2 \, \mathrm{d}s \Big)^{1/2}.$$

Hence the gradient of θ is of second order $O(h^2)$, whereas the gradient of the total error is only of order $O(h)$ as $h \to 0$. Thus ∇u_h is a better approximation to $\nabla R_h u$ than is possible to ∇u. This is an example of a phenomenon which is sometimes referred to as *superconvergence*.

The discrete solution operator $E_h(t)$ introduced above also has smoothing properties analogous to the corresponding results in Sect. 8.2 for the continuous problem, such as, for instance

$$|E_h(t)v_h|_1 \le Ct^{-1/2}\|v_h\|, \quad \text{for } t > 0, \quad v_h \in S_h,$$

and

$$\Big\| D_t^k E_h(t) v_h \Big\| = \|\Delta_h^k E_h(t) v_h\| \le C_k t^{-k} \|v_h\|, \quad \text{for } t > 0, \ v_h \in S_h. \tag{10.18}$$

Such results may be used to show, e.g., the following non-smooth data error estimate for the homogeneous equation.

Theorem 10.3. *Assume that $f = 0$ and let u_h and u be the solutions of* (10.3) *and* (10.1)*, respectively, where now the initial data for* (10.3) *are chosen as $v_h = P_h v$. Then*

$$\|u_h(t) - u(t)\| \le Ch^2 t^{-1} \|v\|, \quad \text{for } t > 0.$$

The proof is left as an exercise (Problem 10.4). This result shows that the convergence rate is $O(h^2)$ for t bounded away from zero, even when v is only assumed to belong to L_2.

The above theory easily extends to finite elements of higher order, under the appropriate regularity assumptions on the solution. Thus, if the finite element subspace is such that

$$\|R_h w - w\| \le Ch^r \|w\|_r, \quad \forall w \in H^r \cap H_0^1, \tag{10.19}$$

then we may show the following theorem.

Theorem 10.4. *Let u_h and u be the solutions of* (10.3) *and* (10.1)*, respectively, and assume that* (10.19) *holds. Then, for v_h suitably chosen, we have*

$$\|u_h(t) - u(t)\| \le Ch^r \Big(\|v\|_r + \int_0^t \|u_t\|_r \, \mathrm{d}s \Big), \quad \text{for } t \ge 0.$$

Recall from (5.50) that for $r > 2$ the estimate (10.19) holds for piecewise polynomials of degree $r-1$, but that the regularity assumption $w \in H^r \cap H_0^1$ is then somewhat unrealistic for a polygonal domain Ω. For a domain Ω with a smooth boundary Γ, special considerations are needed in the boundary layer $\Omega \setminus \Omega_h$.

10.2 Some Completely Discrete Schemes

We shall now turn our attention to some simple schemes for discretization also with respect to the time variable, and let S_h be the space of piecewise linear finite element functions as before. We begin with the *backward Euler-Galerkin method*. With k the time step and $U^n \in S_h$ the approximation of $u(t)$ at $t = t_n = nk$, this method is defined by replacing the time derivative in (10.3) by a backward difference quotient, or with $\bar{\partial}_t U^n = k^{-1}(U^n - U^{n-1})$,

$$\begin{aligned} &(\bar{\partial}_t U^n, \chi) + a(U^n, \chi) = (f(t_n), \chi), \quad \forall \chi \in S_h,\ n \geq 1, \\ &U^0 = v_h. \end{aligned} \tag{10.20}$$

Given U^{n-1} this defines U^n implicitly from the discrete elliptic problem

$$(U^n, \chi) + ka(U^n, \chi) = (U^{n-1} + kf(t_n), \chi), \quad \forall \chi \in S_h.$$

Expressing U^n in terms of the basis $\{\Phi_j\}_{j=1}^{M_h}$ as $U^n(x) = \sum_{j=1}^{M_h} \alpha_j^n \Phi_j(x)$, we may write this equation in the matrix notation introduced in Sect. 10.1 as

$$B\alpha^n + kA\alpha^n = B\alpha^{n-1} + kb^n, \quad \text{for } n \geq 1,$$

where α^n is the vector with components α_j^n, or

$$\alpha^n = (B + kA)^{-1} B\alpha^{n-1} + k(B + kA)^{-1} b^n, \quad \text{for } n \geq 1, \quad \text{with } \alpha^0 = \gamma.$$

We begin our analysis of the backward Euler method by showing that it is unconditionally stable, i.e., that it is stable independently of the relation between h and k. Choosing $\chi = U^n$ in (10.20) we have, since $a(U^n, U^n) \geq 0$,

$$(\bar{\partial}_t U^n, U^n) \leq \|f^n\| \, \|U^n\|, \quad \text{where } f^n = f(t_n),$$

or

$$\|U^n\|^2 - (U^{n-1}, U^n) \leq k\|f^n\| \, \|U^n\|.$$

Since $(U^{n-1}, U^n) \leq \|U^{n-1}\| \, \|U^n\|$, this shows

$$\|U^n\| \leq \|U^{n-1}\| + k\|f^n\|, \quad \text{for } n \geq 1,$$

and hence, by repeated application,

$$\|U^n\| \leq \|U^0\| + k \sum_{j=1}^{n} \|f^j\|. \tag{10.21}$$

We shall now prove the following error estimate.

Theorem 10.5. *With U^n and u the solutions of* (10.20) *and* (10.1)*, respectively, and with v_h chosen so that* (10.10) *holds, we have, for $n \geq 0$,*

$$\|U^n - u(t_n)\| \leq Ch^2 \Big(\|v\|_2 + \int_0^{t_n} \|u_t\|_2 \, \mathrm{d}s \Big) + Ck \int_0^{t_n} \|u_{tt}\| \, \mathrm{d}s.$$

Proof. In analogy with (10.13) we write

$$U^n - u(t_n) = \big(U^n - R_h u(t_n)\big) + \big(R_h u(t_n) - u(t_n)\big) = \theta^n + \rho^n.$$

As before, by (10.12),

$$\|\rho^n\| \le Ch^2 \|u(t_n)\|_2 \le Ch^2 \Big(\|v\|_2 + \int_0^{t_n} \|u_t\|_2 \, \mathrm{d}s \Big).$$

This time, a calculation corresponding to (10.14) yields

$$(\bar\partial_t \theta^n, \chi) + a(\theta^n, \chi) = -(\omega^n, \chi), \tag{10.22}$$

where

$$\omega^n = R_h \bar\partial_t u(t_n) - u_t(t_n) = (R_h - I)\bar\partial_t u(t_n) + (\bar\partial_t u(t_n) - u_t(t_n)) = \omega_1^n + \omega_2^n.$$

By application of the stability estimate (10.21) to (10.22) we obtain

$$\|\theta^n\| \le \|\theta^0\| + k \sum_{j=1}^n \|\omega_1^j\| + k \sum_{j=1}^n \|\omega_2^j\|.$$

Here, as before, by (10.10) and (10.12),

$$\|\theta^0\| = \|v_h - R_h v\| \le \|v_h - v\| + \|v - R_h v\| \le Ch^2 \|v\|_2.$$

Note now that

$$\omega_1^j = (R_h - I) k^{-1} \int_{t_{j-1}}^{t_j} u_t \, \mathrm{d}s = k^{-1} \int_{t_{j-1}}^{t_j} (R_h - I) u_t \, \mathrm{d}s,$$

whence

$$k \sum_{j=1}^n \|\omega_1^j\| \le \sum_{j=1}^n \int_{t_{j-1}}^{t_j} Ch^2 \|u_t\|_2 \, \mathrm{d}s = Ch^2 \int_0^{t_n} \|u_t\|_2 \, \mathrm{d}s.$$

Further, by Taylor's formula,

$$\omega_2^j = k^{-1}(u(t_j) - u(t_{j-1})) - u_t(t_j) = -k^{-1} \int_{t_{j-1}}^{t_j} (s - t_{j-1}) u_{tt}(s) \, \mathrm{d}s,$$

so that

$$k \sum_{j=1}^n \|\omega_2^j\| \le \sum_{j=1}^n \Big\| \int_{t_{j-1}}^{t_j} (s - t_{j-1}) u_{tt}(s) \, \mathrm{d}s \Big\| \le k \int_0^{t_n} \|u_{tt}\| \, \mathrm{d}s.$$

Together our estimates complete the proof of the theorem. □

Replacing the backward difference quotient with respect to time in (10.20) by a forward difference quotient we arrive at the *forward Euler-Galerkin method*, or with $\partial_t U^n = (U^{n+1} - U^n)/k$,

$$\begin{aligned} &(\partial_t U^n, \chi) + a(U^n, \chi) = (f(t_n), \chi), \quad \forall \chi \in S_h,\ n \geq 1, \\ &U^0 = v_h. \end{aligned}$$

In matrix form this may be expressed as

$$B\alpha^{n+1} = (B - kA)\alpha^n + kb^n, \quad \text{for } n \geq 0,$$

Since B is not a diagonal matrix this method is not explicit. However, if this time discretization method is applied to the lumped mass semidiscrete equation (10.16), and thus B replaced by the diagonal matrix $\bar{B}$, then the corresponding forward Euler method becomes an explicit one.

Using the discrete Laplacian defined in (10.6), the forward Euler method may also be defined by

$$U^{n+1} = (I + k\Delta_h)U^n + kP_h f(t_n), \quad \text{for } n \geq 0, \quad \text{with } U^0 = v_h. \tag{10.23}$$

This method is not unconditionally stable as the backward Euler method, but considering for simplicity only the homogeneous equation, we shall show stability under the condition that the family $\{S_h\}$ is such that

$$\lambda_{M_h,h}\, k \leq 2, \tag{10.24}$$

where $\lambda_{M_h,h}$ is the largest eigenvalue of $-\Delta_h$. Recalling (6.38), we note that this holds, e.g., if the S_h satisfy the inverse inequality (6.37) and if $k \leq 2C^{-1}h^2$, where C is the constant in (6.38), which thus shows conditional stability.

It is clear that (10.23) is stable if and only if $\|(I + k\Delta_h)\chi\| \leq \|\chi\|$ for all $\chi \in S_h$, and since $-\Delta_h$ is symmetric positive definite, this holds if and only if all eigenvalues of $I + k\Delta_h$ belong to $[-1, 1]$. By the positivity of $-\Delta_h$ this is the same as requiring the smallest eigenvalue of $I + k\Delta_h$ to be ≥ -1, or that the largest eigenvalue of $-\Delta_h$ is $\leq 2/k$, which is (10.24). See also Problem 10.3.

Note that because of the non-symmetric choice of the discretization in time, the backward Euler-Galerkin method is only first order accurate in time. We therefore now turn to the *Crank-Nicolson-Galerkin method*, in which the semidiscrete equation is discretized in a symmetric fashion around the point $t_{n-1/2} = (n - \frac{1}{2})k$, which yields a method which is second order accurate in time. More precisely, we define $U^n \in S_h$ recursively for $n \geq 1$ by

$$\begin{aligned} &(\bar{\partial}_t U^n, \chi) + a(\tfrac{1}{2}(U^n + U^{n-1}), \chi) = (f(t_{n-1/2}), \chi), \quad \forall \chi \in S_h, \\ &U^0 = v_h. \end{aligned} \tag{10.25}$$

In matrix notation this takes the form,

$$B\alpha^n + \tfrac{1}{2}kA\alpha^n = B\alpha^{n-1} - \tfrac{1}{2}kA\alpha^{n-1} + kb^{n-1/2}, \quad \text{for } n \geq 1,$$

or, with $\alpha^0 = \gamma$,

$$\alpha^n = (B + \tfrac{1}{2}kA)^{-1}(B - \tfrac{1}{2}kA)\alpha^{n-1} + k(B + \tfrac{1}{2}kA)^{-1}b^{n-1/2}, \quad n \geq 1.$$

This method is also unconditionally stable which may be shown by choosing $\chi = U^n + U^{n-1}$ in (10.25) and using the Cauchy-Schwarz inequality on the right. Then

$$k(\bar{\partial}_t U^n, U^n + U^{n+1}) = \|U^n\|^2 - \|U^{n-1}\|^2 = (\|U^n\| - \|U^{n-1}\|)(\|U^n\| + \|U^{n-1}\|).$$

Using the positivity of $a(U^n, U^n)$ and cancelling $\|U^n\| + \|U^{n-1}\|$ we find

$$\|U^n\| \leq \|U^{n-1}\| + k\|f^{n-1/2}\|, \quad \text{where } f^{n-1/2} = f(t_{n-1/2}),$$

or after summation

$$\|U^n\| \leq \|v_h\| + k\sum_{j=1}^{n} \|f^{j-1/2}\|.$$

This time the error estimate reads as follows. Its proof is similar to that of Theorem 10.5 and is left to Problem 10.7.

Theorem 10.6. *With U^n and u the solutions of* (10.25) *and* (10.1)*, respectively, and with v_h chosen so that* (10.10) *holds, we have for $n \geq 0$,*

$$\|U^n - u(t_n)\| \leq Ch^2\Big(\|v\|_2 + \int_0^{t_n} \|u_t\|_2 \,\mathrm{d}s\Big) + Ck^2 \int_0^{t_n} \big(\|u_{ttt}\| + \|\Delta u_{tt}\|\big)\,\mathrm{d}s.$$

10.3 Problems

Problem 10.1. Consider the problem (10.1) in the case of one space dimension with $\Omega = (0, 1)$. For the numerical solution, we use the piecewise linear functions based on the partition

$$0 < x_1 < x_2 < \ldots < x_M < 1, \quad x_j = jh, \ h = 1/(M+1).$$

Determine the mass matrix B and the stiffness matrix A and write down the semidiscrete problem, the backward Euler equations, and the Crank-Nicolson equations.

Problem 10.2. (Computer exercise.) Consider the initial boundary value problem (10.1) with $\Omega = (-\pi, \pi)$ and $v = \operatorname{sign} x$.

(a) Determine the exact solution by eigenfunction expansion.
(b) Apply the backward Euler method (10.20) based on piecewise linear finite elements with $v_h = P_h v$ and $(h, k) = (\pi/5, 1/10),\ (\pi/10, 1/40)$. Determine the maximal error at the mesh-points for $t = 0.1,\ 0.5,\ 1.0$.

Problem 10.3. (a) Show that the operator $-\Delta_h : S_h \to S_h$ defined in (10.6) is selfadjoint positive definite with respect to $(\cdot, \cdot)$.
(b) Show that, with the notation of Theorem 6.7,

$$-\Delta_h v_h = \sum_{i=1}^{M_h} \lambda_{i,h}(v_h, \varphi_{i,h})\varphi_{i,h} \quad \text{and} \quad \|\Delta_h\| = \lambda_{M_h,h}.$$

Hint: The left side of the second identity is the operator norm of Δ_h, see (A.7). Thus, you must show that $\|\Delta_h \chi\| \le \lambda_{M_h,h}\|\chi\|$ for all $\chi \in S_h$ with equality for some χ.
(c) Assume that the family of finite element spaces $\{S_h\}$ satisfies the inverse inequality (6.37). Show that

$$\|\Delta_h\| \le Ch^{-2}.$$

Hint: See (6.38).

Problem 10.4. Assume that $f = 0$ and let u_h and u be the solutions of (10.3) and (10.1), respectively, with $v_h = P_h v$.
(a) Assume that $v \in H^2 \cap H_0^1$. Show that

$$\|u_h(t) - u(t)\| \le Ch^2 \|v\|_2, \quad \text{for } t \ge 0.$$

(b) Assume that $v \in L_2$. Show that

$$\|u_h(t) - u(t)\| \le Ch^2 t^{-1} \|v\|, \quad \text{for } t > 0.$$

Hint: For (a) deduce from (10.15) that

$$\theta(t) = E_h(t)\theta(0) - \int_0^t E_h(t-s)P_h\rho_t(s)\,\mathrm{d}s. \tag{10.26}$$

Split the integral as $\int_0^t = \int_0^{t/2} + \int_{t/2}^t$ and integrate by parts in the first term to get, with $e = u_h - u$,

$$\begin{aligned}\theta(t) = {} & E_h(t)P_h e(0) - E_h(t/2)P_h\rho(t/2) \\ & + \int_0^{t/2} D_s E_h(t-s)P_h\rho(s)\,\mathrm{d}s - \int_{t/2}^t E_h(t-s)P_h\rho_t(s)\,\mathrm{d}s.\end{aligned}$$

Then use (10.18), (10.12), (8.18), and Problem 8.10. Note also $P_h e(0) = 0$, since $v_h = P_h v$.

For (b) integrate by parts once more to get the additional terms

$$D_t E_h(t/2) P_h \tilde{\rho}(t/2) - \int_0^{t/2} D_s^2 E_h(t-s) P_h \tilde{\rho}(s)\,\mathrm{d}s,$$

where $\tilde{\rho}(t) = \int_0^t \rho(s)\,\mathrm{d}s$, $\|\tilde{\rho}\| \le Ch^2 \|\tilde{u}\|_2$, $\|\tilde{u}\|_2 \le C\|\Delta\tilde{u}\|$, and $\Delta\tilde{u}(t) = \int_0^t u_t(s)\,\mathrm{d}s = u(t) - v$.

Problem 10.5. Assume that the family of finite element spaces $\{S_h\}$ is such that $\|\Delta_h\| \le Ch^{-2}$, cf. Problem 10.3. Let u_h and u be the solutions of (10.3) and (10.1), respectively. Assume that $\|v_h - v\| \le Ch^2\|v\|_2$. Show that

$$\|u_h(t) - u(t)\| \le C\big(1 + \log(t/h^2)\big)h^2 \max_{0\le s\le t} \|u(s)\|_2, \qquad \text{for } t \ge h^2.$$

Hint: Integrate by parts in (10.26) to get

$$\theta(t) = E_h(t)P_h e(0) - P_h\rho(t) + \int_0^t D_s E_h(t-s) P_h \rho(s)\,\mathrm{d}s.$$

Split the integral as $\int_0^t = \int_0^{t-h^2} + \int_{t-h^2}^t$ and treat the first part as in Problem 10.4 (a). For the second part use $\|D_s E_h(t-s)\| = \|\Delta_h E_h(t-s)\| \le \|\Delta_h\|\|E_h(t-s)\| \le Ch^{-2}$, see Problem 10.3 (b).

Problem 10.6. Show error estimates analogous to those of Theorem 10.1 when the term $-\Delta u$ in (10.1) is replaced by $\mathcal{A}u = -\nabla\cdot(a\nabla u) + b\cdot\nabla u + cu$ as in Sect. 3.5. Hint: See Problems 5.7 and 8.8.

Problem 10.7. Prove Theorem 10.6.

11 Hyperbolic Equations

In this chapter we present basic concepts and results for hyperbolic equations. We begin in Sect. 11.1 with a short discussion of characteristic directions, curves, and surfaces. In Sect. 11.2 we study the model wave equation. We use the method of eigenfunction expansions to solve the standard initial boundary value problem, and apply the energy method to study uniqueness and domains of dependence. In Sect. 11.3 we reduce the solution of first order scalar first order partial differential equations to integration along characteristic curves, and in Sect. 11.4 we extend this approach to symmetric first order system, and consider finally symmetric hyperbolic systems in more than one space variable by energy arguments.

11.1 Characteristic Directions and Surfaces

Consider the scalar linear partial differential equation

$$\mathcal{L}u = \mathcal{L}(x,D)u := \sum_{|\alpha|\le m} a_\alpha(x) D^\alpha u = f(x), \quad \text{in } \Omega, \tag{11.1}$$

where Ω is a domain in $\mathbf{R}^d$. We say that the direction $\xi \in \mathbf{R}^d$, $\xi \neq 0$, is a *characteristic direction* for the operator $\mathcal{L}(x,D)$ at x if

$$\Lambda(\xi) = \Lambda(x,\xi) := \sum_{|\alpha|=m} a_\alpha(x)\xi^\alpha = 0. \tag{11.2}$$

The polynomial $\Lambda(\xi) = \Lambda(x,\xi)$ is called the *characteristic polynomial* of $\mathcal{L}$ at x. Note that the summation in (11.2) is only over $|\alpha| = m$, i.e., it corresponds to the *principal part* of $\mathcal{L}$, the terms of order exactly m.

Sometimes we shall consider also systems of linear partial differential equations. These may be included in (11.1) if we interpret the coefficients $a_\alpha(x)$ as matrices. In the case that these matrices are square matrices of order N with $N \ge 2$, we say that $\xi \in \mathbf{R}^d$ is a characteristic direction at x if

$$\det \Lambda(x,\xi) = 0. \tag{11.3}$$

A $(d-1)$-dimensional surface in $\mathbf{R}^d$ is said to be a *characteristic surface* if its normal at each point x is a characteristic direction at x. In the case of the plane, $d = 2$, we call this a *characteristic curve* or simply a *characteristic*.

Example 11.1. For the first order scalar equation

$$\sum_{j=1}^{d} a_j(x)\frac{\partial u}{\partial x_j} + a_0(x)u = f(x), \tag{11.4}$$

the characteristic directions are given by the equation

$$\Lambda(x,\xi) = \sum_{j=1}^{d} a_j(x)\xi_j = 0,$$

and hence any direction orthogonal to the vector $a(x) = (a_1(x), \dots, a_d(x))$ is characteristic.

The hyperplane $x_1 = 0$ has the normal $(1, 0, \dots, 0)$ and hence it is a characteristic surface if $a_1(x) = 0$ for all $x = (0, x_2, \dots, x_d)$. It is non-characteristic if $a_1(x) \neq 0$ at each point $x = (0, x_2, \dots, x_d)$, which is equivalent to saying that the equation (11.4) may be solved for $\partial u/\partial x_1$. In such a case the equation may be written

$$\frac{\partial u}{\partial x_1} = \sum_{j=2}^{d} \tilde{a}_j(x)\frac{\partial u}{\partial x_j} + \tilde{a}_0(x)u + \tilde{f}(x)$$

near the hyperplane.

Example 11.2. Poisson's equation,

$$-\Delta u = f,$$

has no characteristic directions, since $\Lambda(\xi) = -(\xi_1^2 + \cdots + \xi_d^2) = -|\xi|^2$ vanishes only for $\xi = 0$.

Example 11.3. The heat equation,

$$\frac{\partial u}{\partial t} - \Delta u = f,$$

now considered in $\mathbf{R}^{d+1}$ with points (x,t), $x \in \mathbf{R}^d$, $t \in \mathbf{R}$, has the characteristic equation $\Lambda(\xi,\tau) = -|\xi|^2 = 0$. In this case, the variable is $(\xi,\tau) \in \mathbf{R}^{d+1}$, $\xi \in \mathbf{R}^d$, $\tau \in \mathbf{R}$, which means that $(0, \dots, 0, 1)$ is a characteristic direction and the hyperplane $t = 0$ a characteristic surface.

Example 11.4. The wave equation,

$$\frac{\partial^2 u}{\partial t^2} - \Delta u = f,$$

similarly corresponds to $\Lambda(\xi,\tau) = \tau^2 - |\xi|^2 = 0$, so that $(\xi, \pm|\xi|)$ is a characteristic direction for any choice of $\xi \neq 0$. For instance, the circular cone with vertex $(\bar{x}, \bar{t})$, defined by the equation

$$F(x,t) := |x-\bar{x}|^2 - (t-\bar{t})^2 = 0,$$

has for its normal at a point (x,t) on the cone

$$\Big(\frac{\partial F}{\partial x_1},\dots,\frac{\partial F}{\partial x_d},\frac{\partial F}{\partial t}\Big) = 2(x-\bar{x},-(t-\bar{t})) = 2(x-\bar{x},\mp|x-\bar{x}|).$$

This is thus a characteristic direction and the cone itself a characteristic surface. Now $t=0$ is non-characteristic.

The characteristic polynomial may be used to classify partial differential equations into different types. For example, $\mathcal{L}$ is said to be *elliptic* if it has no characteristic directions. For a second order equation with constant coefficients, $\Lambda(\xi)$ is a homogeneous quadratic polynomial, so that

$$\Lambda(\xi) = \sum_{j,k=1}^{d} a_{jk}\xi_j\xi_k, \quad \text{where } a_{jk} = a_{kj}.$$

After an orthogonal transformation of variables, $\xi = P\eta$, such a polynomial may be written in the form

$$\Lambda(P\eta) = \sum_{j=1}^{d} \lambda_j \eta_j^2,$$

where $\{\lambda_j\}_{j=1}^d$ are the eigenvalues of the matrix $A = (a_{jk})$. The differential equation is said to be *elliptic* if all the λ_j are of the same sign, as in Example 11.2, which is equivalent to the above definition that it has no characteristics. The equation is said to be *hyperbolic*, if all but one of the λ_j have the same sign and the remaining λ_j has the opposite sign, as in Example 11.4. In Example 11.3, all but one λ_j have the same sign and the remaining λ_j is zero, and we then have a *parabolic* equation.

Example 11.5. Let now A be an $N \times N$ diagonal matrix with diagonal elements $\{\lambda_j\}_{j=1}^N$, and consider the system

$$\frac{\partial u}{\partial t} - A\frac{\partial u}{\partial x} = f.$$

The characteristic directions (ξ,τ) are then determined by the equation

$$\det(\tau I - \xi A) = 0.$$

The matrix $\tau I - \xi A$ is a diagonal matrix with elements $\tau - \lambda_j\xi$, $j = 1,\dots,N$, and thus (ξ,τ) is a characteristic direction exactly when one of these elements vanishes. This gives the characteristic directions $(1,\ \lambda_j)$, $j = 1,\dots,N$. Thus the straight lines

$$x + \lambda_j t = \text{constant}, \quad j = 1,\dots,N,$$

are characteristic curves and $t=0$ is non-characteristic.

11.2 The Wave Equation

In this section we first consider the initial-boundary value problem for the wave equation,

$$\begin{aligned} &u_{tt} - \Delta u = 0, && \text{in } \Omega \times \mathbf{R}_+, \\ &u = 0, && \text{on } \Gamma \times \mathbf{R}_+, \\ &u(\cdot, 0) = v,\ u_t(\cdot, 0) = w, && \text{in } \Omega, \end{aligned} \tag{11.5}$$

where Ω is a bounded domain in $\mathbf{R}^d$ with boundary Γ, and v and w are given functions of x in Ω.

The existence of a solution of (11.5) may be shown by eigenfunction expansion, in a way analogous to the case of the heat equation in Sect. 8.2. For the purpose of demonstrating this, we introduce the eigenfunctions $\{\varphi_j\}_{j=1}^{\infty}$ and corresponding eigenvalues $\{\lambda_j\}_{j=1}^{\infty}$ of the elliptic operator $-\Delta$ and assume that (11.5) has a solution the form

$$u(x,t) = \sum_{j=1}^{\infty} \hat{u}_j(t)\varphi_j(x).$$

Inserting this into the differential equation we find

$$\sum_{j=1}^{\infty} \big(\hat{u}_j''(t) + \lambda_j \hat{u}_j(t)\big)\varphi_j(x) = 0.$$

Correspondingly, we have for the initial conditions

$$\sum_{j=1}^{\infty} \hat{u}_j(0)\varphi_j(x) = v(x), \qquad \sum_{j=1}^{\infty} \hat{u}_j'(0)\varphi_j(x) = w(x).$$

Since the φ_j form an orthonormal basis of $L_2 = L_2(\Omega)$ we have, for $j \geq 1$,

$$\begin{aligned} &\hat{u}_j'' + \lambda_j \hat{u}_j = 0, \quad \text{for } t > 0, \\ &\hat{u}_j(0) = \hat{v}_j = (v, \varphi_j), \qquad \hat{u}_j'(0) = \hat{w}_j = (w, \varphi_j), \end{aligned}$$

and by solving this initial-value problem we conclude

$$\hat{u}_j(t) = \hat{v}_j \cos(\sqrt{\lambda_j}t) + \hat{w}_j \frac{1}{\sqrt{\lambda_j}} \sin(\sqrt{\lambda_j}t), \quad \text{for } j \geq 1,$$

and hence

$$u(x,t) = \sum_{j=1}^{\infty} \big(\hat{v}_j \cos(\sqrt{\lambda_j}t) + \hat{w}_j \lambda_j^{-1/2} \sin(\sqrt{\lambda_j}t)\big)\varphi_j(x). \tag{11.6}$$

It is clear that if v and w are sufficiently regular for the series to converge also after differentiation, then this represents a solution of (11.5), see Problem 11.4. We have thus arrived at the following.

Theorem 11.1. *Assume that $v \in H^2 \cap H_0^1$, $w \in H_0^1$. Then the series* (11.6) *is a solution of* (11.5).

We shall now prove an energy estimate for the solution u of (11.5). By this estimate we easily obtain, in the standard way, the uniqueness and stability of the solution of the problem. The energy method described here is useful also in situations when the eigenfunction expansion approach does not apply.

Theorem 11.2. *Let $u = u(x,t)$ be a sufficiently smooth solution of* (11.5). *Then the total energy $\mathcal{E}(t)$ of u is constant in time, i.e.,*

$$\mathcal{E}(t) := \tfrac{1}{2}\int_\Omega \left(u_t^2 + |\nabla u|^2\right) \mathrm{d}x = \mathcal{E}(0). \tag{11.7}$$

Proof. Multiplying the differential equation in (11.5) by u_t and integrating with respect to x over Ω, using also Green's formula, we find

$$\int_\Omega u_{tt} u_t \,\mathrm{d}x + \int_\Omega \nabla u \cdot \nabla u_t \,\mathrm{d}x = 0,$$

or, with the notation of Sect. 8.3,

$$(u_{tt}, u_t) + a(u, u_t) = 0,$$

Hence,

$$\tfrac{1}{2}\frac{d}{dt}\|u_t\|^2 + \tfrac{1}{2}\frac{d}{dt}\|\nabla u\|^2 = 0,$$

or

$$\frac{d}{dt}\mathcal{E}(t) = 0, \quad \text{for } t > 0.$$

This immediately implies the statement of the theorem. □

We shall now prove an energy estimate for the pure initial value problem for the wave equation, from which we infer that the solution at a given point (x,t) with $t > 0$ only depends on the initial data in a certain sphere in the initial plane $t = 0$. The problem considered is then

$$\begin{aligned} &u_{tt} - \Delta u = 0, && \text{in } \mathbf{R}^d \times \mathbf{R}_+, \\ &u(\cdot,0) = v, \quad u_t(\cdot,0) = w, && \text{in } \mathbf{R}^d. \end{aligned} \tag{11.8}$$

Theorem 11.3. *Let u be a solution of the wave equation in* (11.8). *For $(\bar{x},\bar{t})$ a given point in $\mathbf{R}^d \times \mathbf{R}_+$, let K denote the circular cone, cf. Fig. 11.1,*

$$K = \left\{(x,t) \in \mathbf{R}^d \times \mathbf{R}_+ : |x - \bar{x}| \le \bar{t} - t,\ t \le \bar{t}\right\}, \tag{11.9}$$

and set

$$\mathcal{E}_K(t) = \tfrac{1}{2}\int_{B_t} \left(u_t(x,t)^2 + |\nabla u(x,t)|^2\right) \mathrm{d}x,$$

where $B_t = \left\{x \in \mathbf{R}^d : (x,t) \in K\right\}$. Then

$$\mathcal{E}_K(t) \le \mathcal{E}_K(0), \quad \textit{for } 0 \le t \le \bar{t}.$$

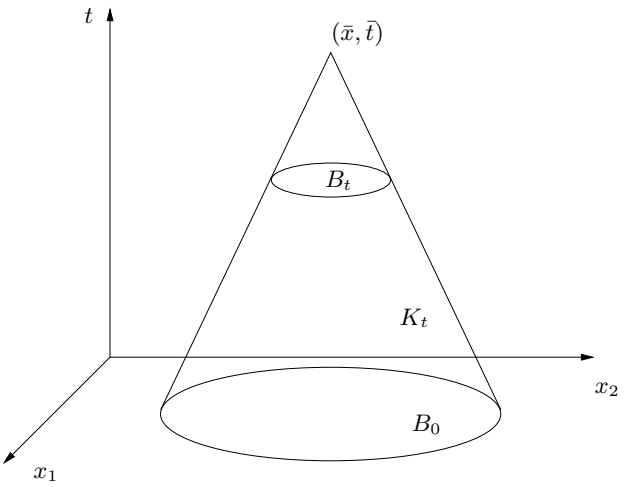

Fig. 11.1. The light cone.

Proof. We introduce the mantle surface of K, $M = \{(x,t) : |x - \bar{x}| = \bar{t} - t\}$, and set $M_t = \{(x,\tau) \in M : \tau \le t\}$. By multiplication of the differential equation by $2u_t$ we find

$$\begin{aligned} 0 &= 2\big(u_{tt} - \nabla \cdot \nabla u\big)\, u_t = 2u_{tt}u_t + 2\nabla u \cdot \nabla u_t - 2\nabla \cdot \big(\nabla u u_t\big) \\ &= D_t\big(u_t^2 + |\nabla u|^2\big) - 2\nabla \cdot \big(\nabla u u_t\big). \end{aligned}$$

Integrating over $K_t = \{(x,\tau) \in K : 0 \le \tau \le t\}$, cf. Fig. 11.1, and using the divergence theorem, we obtain, with $n = (n_x, n_t) = (n_{x_1}, \ldots, n_{x_d}, n_t)$ the exterior normal of ∂K_t,

$$\begin{aligned} 0 &= \int_{\partial K_t} \Big(n_t\big(u_t^2 + |\nabla u|^2\big) - 2n_x \cdot \nabla u\, u_t\Big)\, \mathrm{d}s \\ &= \int_{B_t} \big(u_t^2 + |\nabla u|^2\big)\, \mathrm{d}x - \int_{B_0} \big(u_t^2 + |\nabla u|^2\big)\, \mathrm{d}x \\ &\quad + \int_{M_t} \Big(n_t\big(u_t^2 + |\nabla u|^2\big) - 2u_t\, n_x \cdot \nabla u\Big)\, \mathrm{d}s. \end{aligned}$$

To complete the proof we now show that the integrand of the last term is nonnegative. We have $n_t^2 = |n_x|^2$ on M, and because $n_t = 1/\sqrt{2}$ this yields, by the Cauchy-Schwarz inequality

$$|n_x \cdot \nabla u| \le |n_x|\, |\nabla u| = n_t\, |\nabla u|.$$

Using also the inequality $2\,ab \le a^2 + b^2$, we obtain

$$2\,|u_t\; n_x \cdot \nabla u| = 2\,|u_t|\, |n_x \cdot \nabla u| \le 2\, n_t\, |u_t|\, |\nabla u| \le n_t\, \big(u_t^2 + |\nabla u|^2\big),$$

which completes the proof. □

It follows from Theorem 11.3 that, if $v = w = 0$ in B_0, then $u = 0$ in K, and thus, in particular, at $(\bar{x}, \bar{t})$. This shows that the value of the solution of (11.8) at $(\bar{x}, \bar{t})$ depends only on the values of v and w in the ball B_0 defined by the circular cone K with vertex $(\bar{x}, \bar{t})$, and not on the values of v and w outside this ball.

The existence of a solution of the pure initial value problem (11.8) may be shown in different ways. For the particular equation considered here, an explicit solution may be written down in the form of an integral representation, which takes different forms depending on the number d of space dimensions. For instance, for $d = 1$, it is easy to verify, cf. Problem 11.5, that

$$u(x,t) = \tfrac{1}{2}\big(v(x+t) + v(x-t)\big) + \tfrac{1}{2}\int_{x-t}^{x+t} w(y)\,\mathrm{d}y, \tag{11.10}$$

which is called d'Alembert's formula, and for $d = 3$ one may show

$$u(x,t) = \frac{\partial}{\partial t}\Big\{\frac{1}{4\pi t}\int_{|y-x|=t} v(y)\,\mathrm{d}s_y\Big\} + \frac{1}{4\pi t}\int_{|y-x|=t} w(y)\,\mathrm{d}s_y.$$

In this case the solution at (x,t) depends only on the values of the data on the sphere cut out by the characteristic cone at time zero. More generally, this holds when d is an odd integer. When d is even the solution at (x,t) depends on the initial data in the "ball" $|y - x| \le t$.

11.3 First Order Scalar Equations

We now turn to the first order scalar differential equation

$$\sum_{j=1}^{d} a_j(x)\frac{\partial u}{\partial x_j} + a_0(x)u = f(x), \quad x \in \Omega, \tag{11.11}$$

where $\Omega \subset \mathbf{R}^d$ is a bounded domain with boundary Γ, the vector field $a = a(x) = (a_1(x), \ldots, a_d(x))$ is smooth and does not vanish at any point, and a_0, f are given smooth functions.

We say that $x = x(s) = (x_1(s), \ldots, x_d(s))$, with s a real parameter, is a *characteristic curve*, or simply a *characteristic*, for (11.11) if

$$\frac{d}{ds}x(s) = a(x(s)), \tag{11.12}$$

that is, if the curve in $\mathbf{R}^d$ defined by $x = x(s)$ has the vector $a(x)$ as a tangent at each of its points. Note that a characteristic direction is a normal to the characteristic curve. In particular, in the special case $d = 2$, a characteristic is a characteristic curve in the sense described in Sect. 11.1.

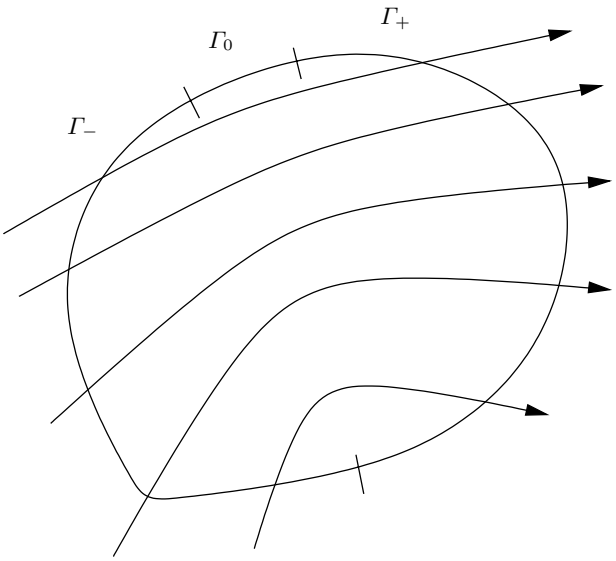

Fig. 11.2. Inflow and outflow boundaries.

In coordinate form (11.12) may be written as the system of ordinary differential equations

$$\frac{dx_j}{ds} = a_j(x), \quad \text{for } j = 1, \ldots, d,$$

and, since the vector field does not vanish, it is clear from the theory of such equations that for each $x_0 \in \Omega$ there exists a unique such curve in some neighborhood of x_0 such that $x(0) = x_0$.

Let Γ be the boundary of Ω and denote by Γ_- the inflow boundary defined by

$$\Gamma_- = \big\{x \in \Gamma : n(x) \cdot a(x) < 0\big\},$$

where $n(x)$ is the exterior normal to Γ at x. Through each point of Γ_- there is a unique characteristic which enters Ω, and we prescribe for the solution of (11.11) the boundary condition

$$u = v, \quad \text{on } \Gamma_-, \tag{11.13}$$

where v is a given smooth function on Γ_-. We also introduce the outflow and the characteristic boundaries,

$$\Gamma_+ = \big\{x \in \Gamma : n(x) \cdot a(x) > 0\big\}, \quad \Gamma_0 = \big\{x \in \Gamma : n(x) \cdot a(x) = 0\big\}.$$

Consider now a solution u of (11.11), (11.13) along a characteristic $x = x(s)$, i.e., consider the function $w(s) = u(x(s))$. We have by the chain rule

$$\frac{dw}{ds} = \nabla u \cdot \frac{dx}{ds} = a(x) \cdot \nabla u,$$

so that, by (11.11), w satisfies

$$\begin{aligned} &\frac{dw}{ds} + a_0(x(s))w = f(x(s)), \quad \text{for } s > 0, \\ &w(0) = v(x_0), \quad \text{with } x(0) = x_0 \in \Gamma_-. \end{aligned} \tag{11.14}$$

This is an initial value problem for a linear ordinary differential equation, which may be solved for the value of w at the points along the characteristic. To find the solution of (11.11), (11.13) at a point $\bar{x} \in \Omega$ we thus determine the characteristic through $\bar{x}$, find its intersection x_0 with Γ_-, and then solve the equation (11.14) with $x(0) = x_0$. The solution at $\bar{x}$ thus only depends on $v(x_0)$ and the values of f on the characteristic.

In the special case that $a_0 = f = 0$ in Ω, the equation (11.14) reduces to

$$\frac{dw}{ds} = 0, \quad \text{for } s > 0, \quad \text{with } w(0) = v(x_0),\ x(0) = x_0 \in \Gamma_-.$$

Thus in this case $u(x(s))$ is constant along the characteristic and the value of the solution at $\bar{x}$ is the same as at $x(0)$, i.e., $u(x(s)) = u(x(0)) = v(x(0))$.

This procedure is often referred to as the *method of characteristics*.

Equations of the form (11.11) are often obtained in the limit from the stationary heat or diffusion equation with convection when the heat conduction or diffusion coefficient vanishes, see (1.18). Such equations can be written in the form (11.11) also in the time-dependent case, if one of the independent variables is interpreted as time. Writing the time variable explicitly in (11.11), we have

$$\begin{aligned} &u_t + a \cdot \nabla u + a_0 u = f, && \text{in } \Omega \times \mathbf{R}_+, \\ &u = g, && \text{in } \Gamma_{-,x}, \\ &u(\cdot, 0) = v, && \text{in } \Omega. \end{aligned}$$

Now $\Omega \subset \mathbf{R}^d$ denotes a spatial domain with boundary Γ, and the inflow boundary of $\Omega \times \mathbf{R}_+$ is split into its spatial part $\Gamma_{-,x} = \big\{(x,t) \in \Gamma \times \mathbf{R}_+ : a(x,t) \cdot n < 0\big\}$ and its temporal part $\Gamma_{-,t} = \Omega \times \{0\}$ corresponding to $t = 0$. We may then use the time variable to parametrize the characteristic curves, $x = x(t)$, which are often called streamlines in this situation.

Example 11.6. Consider the problem

$$\begin{aligned} &u_t + \lambda u_x = 0, && \text{in } \mathbf{R} \times \mathbf{R}_+, \\ &u(\cdot, 0) = v, && \text{in } \mathbf{R}. \end{aligned}$$

Here the characteristics $(x(s), t(s))$ are determined by

$$\frac{dx}{ds} = \lambda, \quad \frac{dt}{ds} = 1.$$

We may thus take t as the parameter along the characteristic and obtain

$$x = \lambda t + C.$$

The characteristic through $(\bar{x}, \bar{t})$ is

$$x - \bar{x} = \lambda(t - \bar{t}), \tag{11.15}$$

and, since the solution is constant on this line,

$$u(\bar{x}, \bar{t}) = v(\bar{x} - \lambda\bar{t}).$$

Example 11.7. With $\Omega = (0, 1)$, we now ask for a solution of

$$\begin{aligned} &u_t + \lambda u_x + u = 1, && \text{in } \Omega \times \mathbf{R}_+, \\ &u = 0, && \text{on } \Gamma_-, \end{aligned}$$

where $\lambda = \text{constant} > 0$. Here

$$\Gamma_- = \big(\{0\} \times \mathbf{R}_+\big) \cup \big(\bar{\Omega} \times \{0\}\big) = \Gamma_{-,x} \cup \Gamma_{-,t},$$

and the characteristic through $(\bar{x}, \bar{t})$ is again defined by (11.15).

We consider first the case $\bar{x} \geq \lambda\bar{t}$ (see Fig. 11.3). Then the characteristic through $(\bar{x}, \bar{t})$ starts at $(\bar{x} - \lambda\bar{t}, 0) \in \Gamma_{-,x}$. With $s = t$ as a parameter we introduce $w(s) = u(\bar{x} + \lambda(s - \bar{t}), s)$ and find that the equation for w is

$$w' + w = 1, \quad \text{for } s > 0, \quad \text{with } w(0) = 0.$$

Hence

$$w(s) = 1 - \mathrm{e}^{-s}, \tag{11.16}$$

and

$$u(\bar{x}, \bar{t}) = 1 - \mathrm{e}^{-\bar{t}}.$$

In the case $\bar{x} < \lambda\bar{t}$ the characteristic through $(\bar{x}, \bar{t})$ starts at $(0, \bar{t} - \bar{x}/\lambda) \in \Gamma_{-,t}$ and with $s = t - (\bar{t} - \bar{x}/\lambda)$ as a parameter we find again (11.16) and thus

$$u(\bar{x}, \bar{t}) = 1 - \mathrm{e}^{-\bar{x}/\lambda}.$$

Altogether, we thus have

$$u(x, t) = \begin{cases} 1 - \mathrm{e}^{-t}, & \text{if } x \geq \lambda t, \\ 1 - \mathrm{e}^{-x/\lambda}, & \text{if } x < \lambda t. \end{cases}$$

Note that the solution is continuous at $x = \lambda t$, but that the derivatives u_t and u_x are not.

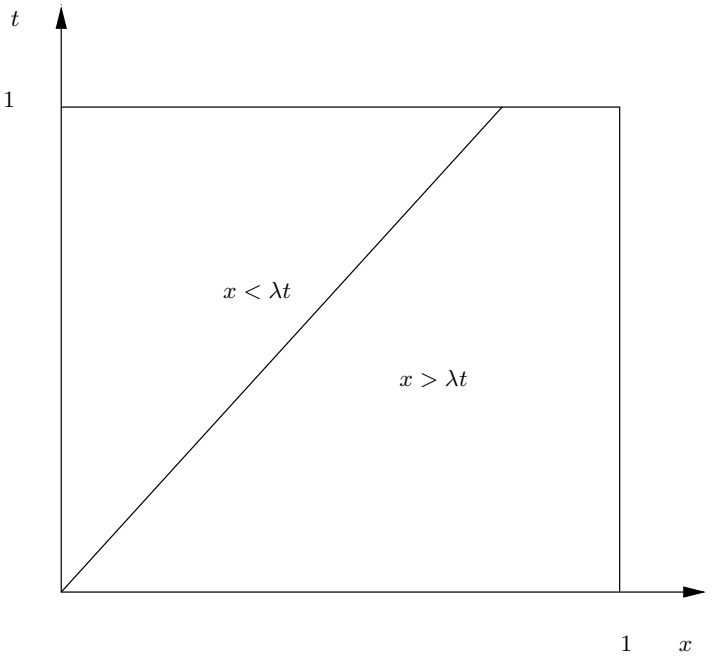

Fig. 11.3. Example 11.7.

Example 11.8. With the same domain as in Example 11.7 we now consider the problem

$$\begin{aligned} &u_t + (1+t)u_x = 0, && \text{in } \Omega \times \mathbf{R}_+, \\ &u = x^2, && \text{on } \Gamma_-. \end{aligned}$$

The characteristic through $(\bar{x}, \bar{t})$ is now (see Fig. 11.4)

$$x = t + \tfrac{1}{2}t^2 + \bar{x} - \bar{t} - \tfrac{1}{2}\bar{t}^2,$$

and starts at $(\bar{x} - \bar{t} - \frac{1}{2}\bar{t}^2, 0) \in \Gamma_{-,x}$, if $\bar{x} \geq \bar{t} + \frac{1}{2}\bar{t}^2$, and somewhere on $\Gamma_{-,t}$, if $\bar{x} < \bar{t} + \frac{1}{2}\bar{t}^2$. The solution is therefore

$$u(x,t) = \begin{cases} (x - t - \frac{1}{2}t^2)^2, & \text{if } x \geq t + \frac{1}{2}t^2, \\ 0, & \text{if } x < t + \frac{1}{2}t^2. \end{cases}$$

11.4 Symmetric Hyperbolic Systems

We first consider an initial value problem in one space dimension of the form

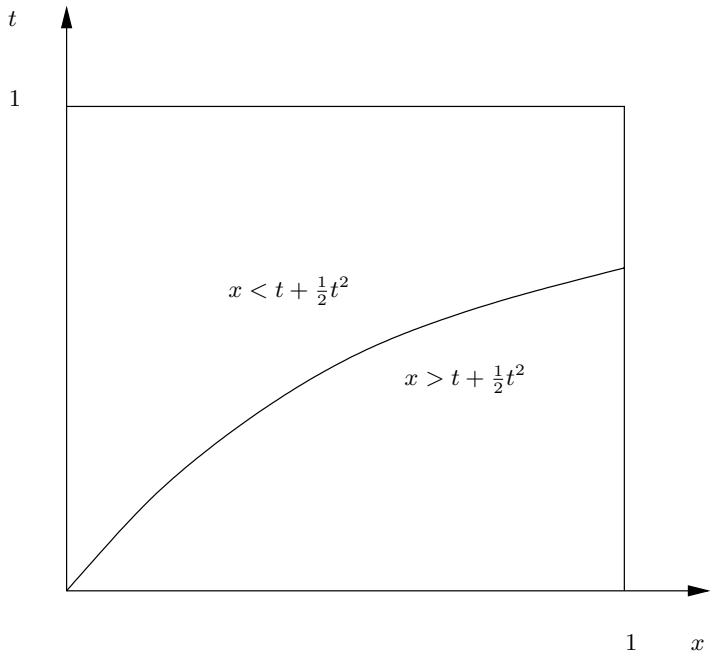

Fig. 11.4. Example 11.8.

$$\begin{aligned} &\frac{\partial u}{\partial t} + A(x,t)\frac{\partial u}{\partial x} + B(x,t)u = f(x,t), && \text{for } x \in \mathbf{R},\ t > 0, \\ &u(x,0) = v(x), && \text{for } x \in \mathbf{R}, \end{aligned} \tag{11.17}$$

where $u = u(x,t)$ and $f = f(x,t)$ are N-vector valued functions and $A = A(x,t)$ and $B = B(x,t)$ are smooth $N \times N$ matrices, with A symmetric. The matrix A then has real eigenvalues $\{\lambda_j\}_{j=1}^N$, with $\lambda_j = \lambda_j(x,t)$, and we make the additional assumption that these are distinct. The system (11.17) is then called *strictly hyperbolic*. Under this assumption one may find a smooth orthogonal matrix $P = P(x,t)$, which diagonalizes A, so that

$$P^{\mathrm{T}} A P = \Lambda = \operatorname{diag}(\lambda_j)_{j=1}^N,$$

see Problem 11.17. Introducing a new dependent variable w by $u = Pw$ we find

$$\frac{\partial u}{\partial t} + A\frac{\partial u}{\partial x} + Bu = P\frac{\partial w}{\partial t} + AP\frac{\partial w}{\partial x} + \Big(\frac{\partial P}{\partial t} + A\frac{\partial P}{\partial x} + BP\Big)w = f,$$

or,

$$\frac{\partial w}{\partial t} + \Lambda\frac{\partial w}{\partial x} + \tilde{B}w = P^{\mathrm{T}} f, \quad \text{where } \tilde{B} = P^{\mathrm{T}}\Big(\frac{\partial P}{\partial t} + A\frac{\partial P}{\partial x} + BP\Big),$$

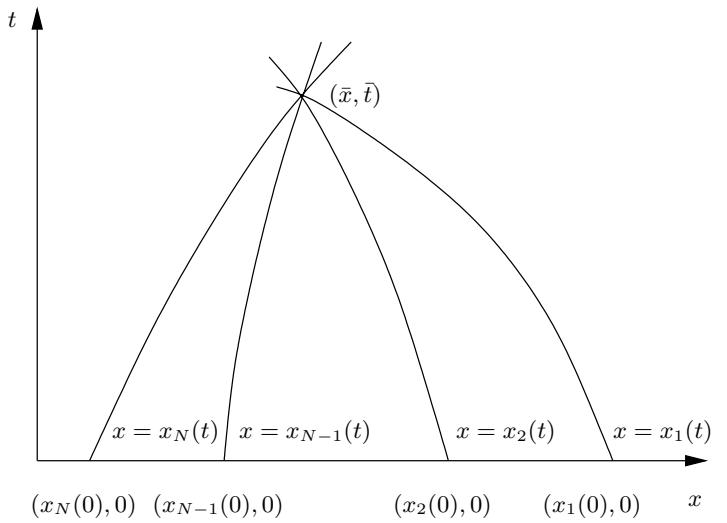

Fig. 11.5. Characteristic curves. Domain of dependence.

which is a system of the form (11.17), but with A diagonal.

We now suppose, thus without restricting the generality, that A in (11.17) is itself a diagonal matrix, and that the λ_j are arranged in increasing order, $\lambda_1 < \lambda_2 < \cdots < \lambda_N$.

Consider first the case that $B = 0$. The system then consists of N uncoupled equations

$$\frac{\partial u_j}{\partial t} + \lambda_j(x,t)\frac{\partial u_j}{\partial x} = f_j(x,t), \quad \text{with } u_j(x,0) = v_j(x), \quad \text{for } j = 1, \ldots, N,$$

each of which is a scalar problem of the kind considered in Sect. 11.3 above. Corresponding to each j there exists a characteristic through $(\bar{x}, \bar{t})$ determined by

$$\frac{dx}{dt} = \lambda_j(x,t), \quad \text{with } x(\bar{t}) = \bar{x}.$$

Denoting the solution of this initial value problem by $x_j(t)$, so that the characteristic through $(\bar{x}, \bar{t})$ is $x = x_j(t)$, we have

$$u_j(\bar{x}, \bar{t}) = v_j(x_j(0)) + \int_0^{\bar{t}} f_j(x_j(s), s)\, \mathrm{d}s, \tag{11.18}$$

and thus $u_j(\bar{x}, \bar{t})$ depends on v_j at only one point and on f_j along the characteristic through $(\bar{x}, \bar{t})$, see Fig. 11.5.

Consider now the case that $B \neq 0$. We may then use an iterative scheme for the solution of (11.17) by setting

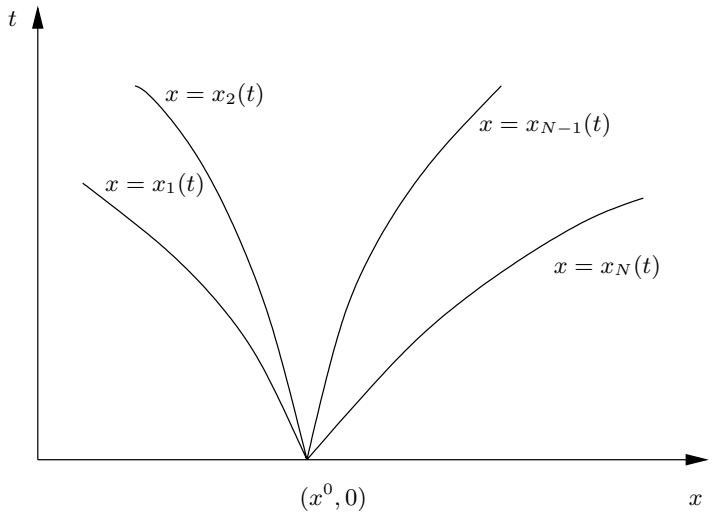

Fig. 11.6. Domain of influence.

$$u^0 = 0, \quad \text{in } \mathbf{R} \times \mathbf{R}_+,$$

and with u^{k+1} definied from u^k for $k \geq 0$ by

$$\begin{aligned} &\frac{\partial u^{k+1}}{\partial t} + A\frac{\partial u^{k+1}}{\partial x} = f - Bu^k, && \text{in } \mathbf{R} \times \mathbf{R}_+, \\ &u^{k+1}(\cdot, 0) = v, && \text{in } \mathbf{R}, \end{aligned}$$

or, in view of (11.18)

$$\begin{aligned} &u^0 = 0, \\ &u_j^{k+1}(\bar{x}, \bar{t}) = v_j(x_j(0)) + \int_0^{\bar{t}} (f - Bu^k)_j(x_j(s), s)\, \mathrm{d}s, \qquad k \geq 0. \end{aligned} \tag{11.19}$$

It is not difficult to show that the u^k converge to a solution of (11.17) as $k \to \infty$, so that the following holds (cf. Problem 7.4).

Theorem 11.4. *The strictly hyperbolic system* (11.17) *has a solution if* A, B, f, *and* v *are appropriately smooth. When* A *is diagonal this solution may be obtained from the iterative scheme* (11.19).

The uniqueness of the solution will follow from Theorem 11.5 below.

We note from (11.19) and Fig. 11.5 that only the values of v in the interval $(x_N(0),\, x_1(0))$ enter in the successive definitions of the u^k, and that f and B are only evaluated in the curvilinear triangle determined by the extreme characteristics $x = x_N(t)$ and $x = x_1(t)$. This thus determines the domain of

dependence of the solution at $(\bar{x}, \bar{t})$ upon the data. Similarly, the initial values at a point $(x^0, 0)$ only influence the solution for $t > 0$ in a wedge between the characteristics corresponding to λ_1 and λ_N, and originating at $(x^0, 0)$, see Fig. 11.6.

Example 11.9. Consider the initial value problem for the wave equation

$$\begin{aligned} &\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}, && \text{in } \mathbf{R} \times \mathbf{R}_+, \\ &u(\cdot, 0) = v, \ \frac{\partial u}{\partial t}(\cdot, 0) = w, && \text{in } \mathbf{R}. \end{aligned} \tag{11.20}$$

We introduce new variables $U_1 = \partial u/\partial t$, $U_2 = \partial u/\partial x$ and find for $U = (U_1, U_2)^{\mathrm{T}}$ the system

$$\begin{aligned} &\frac{\partial U_1}{\partial t} - \frac{\partial U_2}{\partial x} = 0, \\ &\frac{\partial U_2}{\partial t} - \frac{\partial U_1}{\partial x} = 0, && \text{in } \mathbf{R} \times \mathbf{R}_+, \\ &U_1(\cdot, 0) = w, \ U_2(\cdot, 0) = v', && \text{in } \mathbf{R}, \end{aligned}$$

or

$$\frac{\partial U}{\partial t} + A \frac{\partial U}{\partial x} = 0, \quad \text{with } U(x, 0) = \begin{bmatrix} w(x) \\ v'(x) \end{bmatrix},$$

where $A = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$. The eigenvalues of A are $\lambda_1 = -1$, $\lambda_2 = 1$. Setting

$$P = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad U = PV,$$

we find for the new dependent variable $V = (V_1, V_2)^{\mathrm{T}}$ the system

$$\frac{\partial V}{\partial t} + \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \frac{\partial V}{\partial x} = 0, \quad \text{in } \mathbf{R} \times \mathbf{R}_+,$$

or

$$\begin{aligned} &\frac{\partial V_1}{\partial t} - \frac{\partial V_1}{\partial x} = 0, \\ &\frac{\partial V_2}{\partial t} + \frac{\partial V_2}{\partial x} = 0. \end{aligned}$$

Hence

$$V_1(x, t) = V_1(x + t, 0), \quad V_2(x, t) = V_2(x - t, 0).$$

Going back to the original variables U, this may be used to derive d'Alembert's formula (11.10) for the solution of (11.20) (cf. Problem 11.5).

Consider now the generalization of the system (11.17) to d space dimensions,

$$\begin{aligned} &\frac{\partial u}{\partial t} + \sum_{j=1}^{d} A_j \frac{\partial u}{\partial x_j} + Bu = f, && \text{in } \mathbf{R}^d \times \mathbf{R}_+, \\ &u(\cdot, 0) = v, && \text{in } \mathbf{R}^d, \end{aligned} \tag{11.21}$$

where $u = u(x,t)$ is an N-vector valued function, $A_j = A_j(x,t)$ are symmetric $N \times N$ matrices, $B = B(x,t)$ an $N \times N$ matrix, and $f = f(x,t)$ and $v = v(x)$ N-vectors, all of which depending smoothly and boundedly on their variables. We also assume that solutions are small for large $|x|$ in such a way that the following analysis is valid.

A system such as in (11.17) is called a *symmetric hyperbolic system* or a *Friedrichs system*. A special case is Maxwell's equations in electro-dynamics, see Problem 11.15. The classical wave equation $u_{tt} = \Delta u$ may be transformed into a symmetric hyperbolic system by introduction of the first order derivatives as new dependent variables. We leave the verification to Problem 11.11. More generally, many other important equations of mathematical physics can be written as symmetric hyperbolic systems, sometimes after a transformation of the dependent variables.

According to (11.3) the characteristic directions $(\xi, \tau) = (\xi_1, \dots, \xi_d, \tau)$ are given by

$$\det A(\xi, \tau) = \det \Big(\tau I + \sum_{j=1}^{d} \xi_j A_j \Big) = 0.$$

It is clear that for any given ξ this equation has N real roots $\tau_j(\xi)$, $j = 1, \dots, N$, namely the eigenvalues of the symmetric $N \times N$ matrix $-\sum_{j=1}^{d} \xi_j A_j$.

In general, if $d > 1$, it is not possible to simultaneously diagonalize the matrices A_j, and thus the problem may not be treated as above. We shall therefore restrict ourselves here to applying the energy method to show a stability estimate for this problem with respect to $\|\cdot\| = \|\cdot\|_{L_2(\mathbf{R}^d)}$.

Theorem 11.5. *We have for the solution of* (11.21)*, with* $C = C(T)$*,*

$$\|u(t)\| \le C\Big(\|v\| + \Big(\int_0^T \|f\|^2 \, \mathrm{d}s \Big)^{1/2} \Big), \quad \text{for } 0 \le t \le T.$$

Proof. We multiply the equation by u and integrate over $\mathbf{R}^d$ to obtain

$$(\frac{\partial u}{\partial t}, u) + \sum_{j=1}^{d} (A_j \frac{\partial u}{\partial x_j}, u) + (Bu, u) = (f, u).$$

Here

$$(\frac{\partial u}{\partial t}, u) = \int_\Omega \langle \frac{\partial u}{\partial t}, u\rangle \,\mathrm{d}x = \tfrac{1}{2}\frac{d}{dt}\|u\|^2,$$

and

$$(A_j \frac{\partial u}{\partial x_j}, u) = \int_{\mathbf{R}^d} \langle A_j \frac{\partial u}{\partial x_j}, u\rangle \,\mathrm{d}x,$$

where $\langle \cdot, \cdot \rangle$ is the standard inner product in $\mathbf{R}^N$. We have

$$\frac{\partial}{\partial x_j}\langle A_j u, u\rangle = \langle \frac{\partial A_j}{\partial x_j} u, u\rangle + \langle A_j \frac{\partial u}{\partial x_j}, u\rangle + \langle A_j u, \frac{\partial u}{\partial x_j}\rangle,$$

and, since A_j is symmetric, the last two terms are equal. Further, assuming that u is small for large $|x|$,

$$\int_{\mathbf{R}^d} \frac{\partial}{\partial x_j}\langle A_j u, u\rangle \,\mathrm{d}x = 0,$$

and hence

$$(A_j \frac{\partial u}{\partial x_j}, u) = -\tfrac{1}{2}(\frac{\partial A_j}{\partial x_j} u, u).$$

We conclude that

$$\tfrac{1}{2}\frac{d}{dt}\|u\|^2 + (\tilde{B}u, u) \le \|f\|\,\|u\|,$$

where

$$\tilde{B} = B - \tfrac{1}{2}\sum_{j=1}^{d} \frac{\partial A_j}{\partial x_j},$$

and hence

$$\frac{d}{dt}\|u\|^2 \le 2\|\tilde{B}\|_{\mathcal{C}}\|u\|^2 + 2\|f\|\|u\| \le C_0\|u\|^2 + \|f\|^2$$

with $C_0 = 2\|\tilde{B}\|_{\mathcal{C}} + 1$. This implies

$$\|u(t)\|^2 \le \|v\|^2 + \int_0^T \|f\|^2 \,\mathrm{d}s + C_0 \int_0^t \|u\|^2 \,\mathrm{d}s, \quad \text{for } 0 \le t \le T,$$

so that by Gronwall's lemma (cf. Problem 7.6),

$$\|u(t)\|^2 \le \mathrm{e}^{C_0 T}\Big(\|v\|^2 + \int_0^T \|f\|^2 \,\mathrm{d}s\Big), \quad \text{for } 0 \le t \le T.$$

□

In the usual way, this inequality implies uniqueness and stability for the problem (11.21). Existence of a solution may be shown, for instance, by constructing a finite difference approximation on a mesh with mesh-width h and then showing convergence as $h \to 0$.

It is also possible to show here that, as in the case of one space dimension treated above, the value of the solution of (11.21) at a point $(\bar{x}, \bar{t})$ with $\bar{t} > 0$ only depends on data in a finite domain. To do so we consider for simplicity the case of a homogeneous equation with constant coefficients and with no lower order term, so that the problem is

$$\begin{aligned} &\frac{\partial u}{\partial t} + \sum_{j=1}^{d} A_j \frac{\partial u}{\partial x_j} = 0, \qquad && \text{in } \mathbf{R} \times \mathbf{R}_+, \\ &u(\cdot, 0) = v, && \text{in } \mathbf{R}^d. \end{aligned}$$

Then the characteristic polynomial is the symmetric matrix

$$\Lambda(\xi, \tau) = \tau I + \sum_{j=1}^{d} \xi_j A_j.$$

Consider now a circular cone K with vertex $(\bar{x}, \bar{t})$, restricted to $t \leq \bar{t}$, cf. (11.9), and with such an opening angle that the exterior unit normal (n_x, n_t) on the mantle M of the cone makes $\Lambda(n_x, n_t)$ positive definite. That it is possible to find such a cone follows from the fact that for the direction (0,1) we have $\Lambda(0, 1) = I$ which is positive definite, and hence $\Lambda(\xi, 1)$ is also positive definite for small $|\xi|$.

Let B_0 be the domain in the plane $t = 0$ cut out by the cone. We claim that if $v = 0$ in B_0, then $u(\bar{x}, \bar{t}) = 0$.

To prove this we use again the energy method. We multiply the equation by u and integrate over K, using the assumption that the A_j are symmetric and constant, to obtain

$$\begin{aligned} 0 &= \int_K \Big(\langle \frac{\partial u}{\partial t}, u \rangle + \sum_{j=1}^{d} \langle A_j \frac{\partial u}{\partial x_j}, u \rangle \Big) \, \mathrm{d}x \, \mathrm{d}t \\ &= \tfrac{1}{2} \int_K \Big(\frac{\partial}{\partial t} \langle u, u \rangle + \sum_{j=1}^{d} \frac{\partial}{\partial x_j} \langle A_j u, u \rangle \Big) \, \mathrm{d}x \, \mathrm{d}t. \end{aligned}$$

By the divergence theorem we have

$$\int_M \Big(\langle u, u \rangle \, n_t + \sum_{j=1}^{d} \langle A_j u, u \rangle \, n_{x_j} \Big) \, \mathrm{d}s = \int_{B_0} \langle u, u \rangle \, \mathrm{d}x,$$

or, since $u = 0$ in B_0,

$$\int_M \langle \Lambda(n_x, n_t) u, u \rangle \, \mathrm{d}s = 0,$$

which implies $u = 0$ on M, since $\Lambda(n_x, n_t)$ is positive definite. In particular, $u(\bar{x}, \bar{t}) = 0$, which is our claim.

11.5 Problems

Problem 11.1. Determine the characteristics for the Tricomi equation

$$\frac{\partial^2 u}{\partial x_1^2} + x_1 \frac{\partial^2 u}{\partial x_2^2} = f, \quad \text{for } x = (x_1, x_2) \in \mathbf{R}^2.$$

Problem 11.2. Find the characteristic directions of the Cauchy-Riemann equations

$$\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} = 0, \quad \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} = 0.$$

Problem 11.3. Show (11.7) directly from (11.6).

Problem 11.4. Let u be as in (11.6) and assume that $v \in H^2 \cap H_0^1$, $w \in H_0^1$. Show that

$$\begin{aligned} \|u(t)\| &\le C\big(\|v\| + \|w\|\big), \\ \|\nabla u(t)\| &\le C\big(\|\nabla v\| + \|w\|\big), \\ \|u_{tt}(t)\| = \|\Delta u(t)\| &\le C\big(\|\Delta v\| + \|\nabla w\|\big), \\ \|u(t) - v\| &\le Ct\big(\|\nabla v\| + \|w\|\big), \\ \|u_t(t) - w\| &\le Ct\big(\|\Delta v\| + \|\nabla w\|\big). \end{aligned}$$

Hence u is a solution of (11.5) at least in the L_2 sense. Hint: Recall Theorem 6.4 and Problem 6.3. Show that

$$\|u(t) - v\|^2 = t^2 \sum_{j=1}^{\infty} \Big(\sqrt{\lambda_j}\hat{v}_j \frac{\cos(\sqrt{\lambda_j}t) - 1}{\sqrt{\lambda_j}t} + \hat{w}_j \frac{\sin(\sqrt{\lambda_j}t)}{\sqrt{\lambda_j}t} \Big)^2.$$

Problem 11.5. Prove d'Alembert's solution formula (11.10) for the Cauchy problem for the one-dimensional wave equation, i.e.,

$$\begin{aligned} &u_{tt} - u_{xx} = 0 && \text{in } \mathbf{R} \times \mathbf{R}_+, \\ &u(\cdot, 0) = v, \ u_t(\cdot, 0) = w, && \text{in } \mathbf{R}. \end{aligned}$$

Problem 11.6. (a) Solve the initial-value problem

$$\begin{aligned} &\frac{\partial u}{\partial t} + \begin{bmatrix} 0 & x \\ x & 0 \end{bmatrix} \frac{\partial u}{\partial x} = 0, && x \in \mathbf{R}, \ t > 0, \\ &u(x, 0) = v(x), && x \in \mathbf{R}, \end{aligned}$$

by the method of characteristics.
(b) Prove a stability estimate by the energy method.

Problem 11.7. (a) Solve the initial value problem

$$u_t + (x + t)u_x = 0 \text{ for } (x, t) \in \mathbf{R} \times \mathbf{R}_+, \quad \text{with } u(x, 0) = v(x) \text{ for } x \in \mathbf{R},$$

by means of the method of characteristics.
(b) Show that

$$\|u(\cdot,t)\| = \mathrm{e}^{t/2}\|v\| \quad \text{and} \quad \|u_x(\cdot,t)\| = \mathrm{e}^{-t/2}\|v_x\|, \quad \text{for } t \geq 0,$$

by the energy method. Check these results by a direct calculation using the solution formula from (a).

Problem 11.8. Solve the problem

$$x_1 \frac{\partial u}{\partial x_1} - x_2 \frac{\partial u}{\partial x_2} = 0 \quad \text{for } x \in \mathbf{R}^2, \quad \text{with } u(x) = \varphi(x) \quad \text{for } x \in S,$$

where S is a non-characteristic curve.

Problem 11.9. Solve the problem

$$\begin{aligned} x_1 \frac{\partial u}{\partial x_1} + 2x_2 \frac{\partial u}{\partial x_2} + \frac{\partial u}{\partial x_3} &= 3u \qquad && \text{for } x \in \mathbf{R}^3, \\ u(x_1, x_2, 0) &= \varphi(x_1, x_2) && \text{for } (x_1, x_2) \in \mathbf{R}^2. \end{aligned}$$

Problem 11.10. Prove the following stability estimate for the problem (11.11), (11.13) under a suitable condition on the coefficients 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. Show that the wave equation $u_{tt} - \Delta u = 0$ can be written as a symmetric hyperbolic system by introduction of the first order derivatives $u_t, u_{x_1}, \ldots, u_{x_d}$ as new dependent variables.

Problem 11.12. Modify the proof of Theorem 11.5 to show the slightly stronger result

$$\|u(t)\| \leq C(T)\Big(\|v\| + \int_0^T \|f(s)\| \,\mathrm{d}s \Big), \quad \text{for } 0 \leq t \leq T.$$

Problem 11.13. In addition to the assumptions of Theorem 11.5 assume that A_j are constant and B symmetric positive semidefinite. Prove

$$\|u(t)\| \leq \|v\| + \int_0^t \|f\| \,\mathrm{d}s, \quad \text{for } t \geq 0.$$

Problem 11.14. Generalize Theorem 11.5 to symmetric hyperbolic systems of the form

$$M \frac{\partial u}{\partial t} + \sum_{j=1}^{d} A_j \frac{\partial u}{\partial x_j} + Bu = f, \quad \text{in } \mathbf{R}^d \times \mathbf{R}_+,$$

where A_j and B are as before and $M = M(x,t)$ is symmetric positive definite uniformly with respect to x, t, so that $\langle M(x,t)\xi, \xi \rangle \geq \alpha\, |\xi|^2$ for all $\xi \in \mathbf{R}^N$, $(x,t) \in \mathbf{R}^d \times \mathbf{R}_+$, with $\alpha > 0$.

Problem 11.15. The evolution of the electric field $E(x,t) \in \mathbf{R}^3$ and magnetic field $H(x,t) \in \mathbf{R}^3$ in a homogeneous and isotropic space can be described by the following two of Maxwell's equations (Ampère's law and Faraday's law)

$$\begin{aligned} &\frac{1}{c}\frac{\partial E}{\partial t} - \nabla \times H + \frac{4\pi}{c} J = 0, && \text{in } \mathbf{R}^3 \times \mathbf{R}_+, \\ &\frac{1}{c}\frac{\partial H}{\partial t} + \nabla \times E = 0, && \text{in } \mathbf{R}^3 \times \mathbf{R}_+, \end{aligned} \tag{11.22}$$

where c is a positive constant and

$$\nabla \times H = \operatorname{curl} H = \Big(\frac{\partial H_3}{\partial x_2} - \frac{\partial H_2}{\partial x_3}, \frac{\partial H_1}{\partial x_3} - \frac{\partial H_3}{\partial x_1}, \frac{\partial H_2}{\partial x_1} - \frac{\partial H_1}{\partial x_2}\Big).$$

Let us also assume that the density of current J satisfies Ohm's law $J = \sigma E$ with σ a nonnegative constant. Show that (11.22) with E and H given at $t = 0$ constitute a well posed problem by showing that (11.22) is a Friedrichs system. What can be said about the stability of the energy density $e = \frac{1}{2}(E \cdot E + H \cdot H)$? Hint: Problem 11.13.

Problem 11.16. Recall the equation

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

for the longitudinal motion of an elastic bar from Problem 1.2.
(a) Assume for simplicity that ρ and E are constant and show that it can be written as a symmetric hyperbolic system

$$\begin{bmatrix} \rho & 0 \\ 0 & E \end{bmatrix} U_t - \begin{bmatrix} 0 & E \\ E & 0 \end{bmatrix} U_x = 0$$

in the variables $U_1 = u_t$, $U_2 = u_x$, cf. Problem 11.14.
(b) Assume, e.g., the boundary conditions $u(0) = 0$, $u_x(L) = 0$. Show that the mechanical energy is conserved, i.e., with $e = \frac{1}{2}(\rho u_t^2 + E u_x^2)$,

$$\int_0^L e(x,t)\,\mathrm{d}x = \int_0^L e(x,0)\,\mathrm{d}x.$$

Problem 11.17. Compute the eigenvalues and normalized eigenvectors of the symmetric matrix $A(x,t) = \begin{bmatrix} x & t \\ t & -x \end{bmatrix}$. Show that the eigenvector matrix $P(x,t)$ is discontinuous at $x = 0$, $t = 0$, where the eigenvalues are multiple.

12 Finite Difference Methods for Hyperbolic Equations

Solution of hyperbolic equations is perhaps the area in which finite difference methods have most successfully continued to play an important role. This is particularly true for nonlinear conservation laws, which, however, are beyond the scope of this elementary presentation. Here we begin in Sect. 12.1 with the pure initial-value problem for a first order scalar equation in one space variable and study stability and error estimates for the basic upwind scheme, the Friedrichs scheme, and the Lax-Wendroff scheme. In Sect. 12.2 we extend these considerations to symmetric hyperbolic systems and also to higher space dimension, and in Sect. 12.3 we treat the Wendroff box scheme for a mixed initial-boundary value problem in one space dimension.

12.1 First Order Scalar Equations

In this first section we consider the simple model initial value problem

$$\begin{aligned} &\frac{\partial u}{\partial t} = a\frac{\partial u}{\partial x}, && \text{in } \mathbf{R}\times\mathbf{R}_+,\\ &u(\cdot,0) = v, && \text{in } \mathbf{R}, \end{aligned} \tag{12.1}$$

where a is a constant. We recall that for $v \in \mathcal{C}^1$ this problem admits the unique classical solution

$$u(x,t) = (E(t)v)(x) = v(x+at), \tag{12.2}$$

which may thus be found by following the characteristic $x + at =$ constant through (x,t) backwards to $t = 0$ and taking the value of v at that point. A similar statement holds for variable coefficient $a = a(x)$, in which case the characteristic is curved. Since the solution operator just affects a shift of the argument both the maximum-norm and the L_2-norm are constant in time,

$$\|E(t)v\|_{\mathcal{C}} = \|v\|_{\mathcal{C}} \quad \text{and} \quad \|E(t)v\| = \|v\|, \quad \text{for } t \geq 0, \tag{12.3}$$

and thus, in particular, $E(t)$ is stable in both norms.

For the purpose of solving the model problem approximately by the finite difference method we introduce, as earlier for parabolic equations in Sect. 9.1,

a mesh size h in space and a time step k and denote the approximation of $u(x,t)$ at $(x_j,t_n) = (jh,nk)$ by U_j^n, for $j, n \in \mathbf{Z},\ n \geq 0$. Here $\mathbf{Z} = \{\dots,-1,-2,0,1,2,\dots\}$ is the set of all integers. Assuming that $a > 0$ we replace (12.1) by

$$\begin{aligned} &\partial_t U_j^n = a\partial_x U_j^n, && \text{for } j,n \in \mathbf{Z},\ n \geq 0, \\ &U_j^0 = V_j = v(x_j), && \text{for } j \in \mathbf{Z}, \end{aligned} \tag{12.4}$$

where as earlier ∂_t and ∂_x denote forward difference quotients, so that the difference equation reads

$$\frac{U_j^{n+1} - U_j^n}{k} = a\,\frac{U_{j+1}^n - U_j^n}{h}.$$

Introducing this time the mesh ratio $\lambda = k/h$, which we assume is kept constant as h and k tend to zero, we see that (12.4) is an explicit scheme, which defines the approximation at $t = t_{n+1}$ by

$$U_j^{n+1} = (E_k U^n)_j = a\lambda U_{j+1}^n + (1-a\lambda)U_j^n, \quad \text{for } j,n \in \mathbf{Z},\ n \geq 0. \tag{12.5}$$

If we think of U^n as being defined for all x in $\mathbf{R}$ and not only at the mesh points $x = x_j$, we may write

$$U^{n+1}(x) = (E_k U^n)(x) = a\lambda U^n(x+h) + (1-a\lambda)U^n(x),\ x \in \mathbf{R}. \tag{12.6}$$

By iteration we find for the approximate solution at $t = t_n$

$$U^n(x) = (E_k^n v)(x), \quad \text{for } x \in \mathbf{R}.$$

Similarly to the situation for the heat equation we find that E_k is stable in maximum-norm if $a\lambda \leq 1$, since the coefficients of E_k are then positive and add up to 1, so that

$$\|E_k v\|_{\mathcal{C}} \leq \|v\|_{\mathcal{C}},$$

and hence also

$$\|U^n\|_{\mathcal{C}} = \|E_k^n v\|_{\mathcal{C}} \leq \|v\|_{\mathcal{C}}.$$

It is also easy to see that the condition $a\lambda \leq 1$ is necessary for stability. As earlier stability implies convergence:

Theorem 12.1. *Let U^n and u be defined by* (12.6) *and* (12.1)*, and assume that $0 < a\lambda \leq 1$. Then*

$$\|U^n - u^n\|_{\mathcal{C}} \leq Ct_n h|v|_{\mathcal{C}^2}, \quad \textit{for } t_n \geq 0.$$

Proof. We introduce the truncation error

$$\tau^n(x) := \partial_t u^n(x) - a\partial_x u^n(x), \tag{12.7}$$

and find by Taylor expansion for an exact solution u of the differential equation, with $I_n = (t_n, t_{n+1})$,

$$
\begin{aligned}
|\tau^n(x)| &\le |\partial_t u^n(x) - u_t(x,t_n)| + a|\partial_x u^n(x) - a u_x(x,t_n)| \\
&\le C(h+k)\max_{t\in I_n}\big(|u_{tt}(\cdot,t)| + |u_{xx}(\cdot,t)|\big) \le Ch|v|_{\mathcal{C}^2},
\end{aligned}
\tag{12.8}
$$

where we have used that $k \le \lambda h$, that $u_{tt} = a u_{xx}$, and that $|u_{xx}(\cdot,t)|_{\mathcal{C}} \le |v|_{\mathcal{C}^2}$.

We may also write (12.7) in the form

$$
u^{n+1}(x) = E_k u^n(x) + k\tau^n(x), \quad \text{for } x \in \mathbf{R}.
$$

Setting $z^n = U^n - u^n$ we therefore have

$$
z^{n+1} = E_k z^n - k\tau^n,
$$

or, by repeated application,

$$
z^n = E_k^n z^0 - k\sum_{j=0}^{n-1} E_k^{n-1-j}\tau^j.
$$

Since $z^0 = U^0 - u^0 = v - v = 0$, we conclude by stability and (12.8),

$$
\|z^n\|_{\mathcal{C}} \le k\sum_{j=0}^{n-1}\|\tau^j\|_{\mathcal{C}} \le C\,nk\,h|v|_{\mathcal{C}^2}, \quad \text{for } t_n \ge 0,
$$

which completes the proof of the theorem. □

Note that if $a < 0$, the natural choice of finite difference approximation is, instead of (12.4),

$$
\partial_t U_j^n = a\bar{\partial}_x U_j^n, \quad \text{for } n \ge 0,
\tag{12.9}
$$

or, cf. (12.5),

$$
U_j^{n+1} = (E_k U^n)_j = -a\lambda U_{j-1}^n + (1+a\lambda)U_j^n, \quad \text{for } j \in \mathbf{Z},\ n \ge 0.
$$

The stability condition is now $0 < -a\lambda \le 1$. Since both (12.4) and (12.9) use points in the direction of the flow, these difference schemes are referred to as *upwind schemes*.

Let us consider more generally an explicit finite difference scheme

$$
U_j^{n+1} = (E_k U^n)_j = \sum_p a_p U_{j-p}^n, \quad \text{for } j, n \in \mathbf{Z},\ n \ge 0,
\tag{12.10}
$$

where $a_p = a_p(\lambda)$ with $\lambda = k/h = \text{const.}$, or, with x allowed to vary over $\mathbf{R}$,

$$
\begin{aligned}
&U^{n+1}(x) = (E_k U^n)(x) = \sum_p a_p U^n(x - ph), \quad \text{for } x \in \mathbf{R},\ n \geq 0,\\
&U^0(x) = v(x), \quad \text{for } x \in \mathbf{R}.
\end{aligned}
\tag{12.11}
$$

We say that such a method is *accurate of order* r if

$$\tau^n = k^{-1}(u^{n+1} - E_k u^n) = O(h^r), \quad \text{as } h \to 0,$$

where u is the exact solution of (12.1) and $k/h = \lambda = \text{constant}$.

We note that as in Sect. 9.1 we have for the Fourier transform of $E_k v$

$$(E_k v)\hat{}\,(\xi) = \tilde{E}(h\xi)\hat{v}(\xi), \quad \text{where } \tilde{E}(\xi) = \sum_p a_p \mathrm{e}^{-\mathrm{i}p\xi},$$

and, in exactly the same way as in the parabolic case, a necessary and sufficient condition for stability in L_2 is the *von Neumann condition*

$$|\tilde{E}(\xi)| \leq 1, \quad \text{for } \xi \in \mathbf{R}. \tag{12.12}$$

For our above scheme (12.5) we have

$$\tilde{E}(\xi) = a\lambda \mathrm{e}^{\mathrm{i}\xi} + 1 - a\lambda,$$

and as ξ varies, $\tilde{E}(\xi)$ belongs to a circle in the complex plane with center at $1 - a\lambda$ and radius $a\lambda$. In order for (12.12) to hold it is therefore necessary and sufficient that $a\lambda \leq 1$, which is our old stability condition.

In the same way as for the parabolic problem in Sect. 9.1, the definition of the accuracy of the method may also be expressed in terms of the trigonometric polynomial $\tilde{E}(\xi)$: The method is accurate of order r if and only if

$$\tilde{E}(\xi) = \mathrm{e}^{\mathrm{i}a\lambda\xi} + O(\xi^{r+1}), \quad \text{as } \xi \to 0, \tag{12.13}$$

where in the proof we use that for the exact solution we have

$$(E(t)v)\hat{}\,(\xi) = \int_{\mathbf{R}} v(x + at)\mathrm{e}^{-\mathrm{i}x\xi}\,\mathrm{d}x = \mathrm{e}^{\mathrm{i}at\xi}\hat{v}(\xi).$$

As in Sect. 9.1 one may then prove the following error estimate.

Theorem 12.2. *Let U^n and u be defined by* (12.11) *and* (12.1)*, and assume E_k is accurate of order r and stable in L_2. Then*

$$\|U^n - u^n\| \leq C t_n h^r |v|_{r+1}, \quad \textit{for } t_n \geq 0.$$

Another natural choice for a difference approximation to (12.1) is obtained by replacing the derivative with respect to x by the symmetric difference quotient

$$\hat{\partial}_x U^n(x) = \frac{U^n(x + h) - U^n(x - h)}{2h},$$

which results in the finite difference equation

$$\frac{U^{n+1}(x) - U^n(x)}{k} = a\,\frac{U^n(x+h) - U^n(x-h)}{2h}, \tag{12.14}$$

and thus in the difference scheme

$$U^{n+1}(x) = (E_k U^n)(x) = U^n(x) + \tfrac{1}{2}a\lambda(U^n(x+h) - U^n(x-h)),\ n \geq 0,$$
$$U^0(x) = v(x), \quad x \in \mathbf{R}.$$

In this case the symbol of E_k is

$$\tilde{E}(\xi) = 1 + \tfrac{1}{2}a\lambda(\mathrm{e}^{\mathrm{i}\xi} - \mathrm{e}^{-\mathrm{i}\xi}) = 1 + a\lambda \mathrm{i}\sin\xi.$$

Since

$$|\tilde{E}(\xi)|^2 = 1 + a^2\lambda^2\sin^2\xi > 1, \quad \text{except at } \xi = m\pi,$$

we conclude that this method is unstable for any choice of λ.

The latter scheme may be stabilized by replacing $U^n(x)$ on the left in (12.14) by the average $\frac{1}{2}(U^n(x+h) + U^n(x-h))$, which results in

$$\frac{U^{n+1}(x) - \frac{1}{2}(U^n(x+h) + U^n(x-h))}{k} = a\,\frac{U^n(x+h) - U^n(x-h)}{2h},$$

or

$$U^{n+1}(x) = (E_k U^n)(x) = \tfrac{1}{2}(1 + a\lambda)U^n(x+h) + \tfrac{1}{2}(1 - a\lambda)U^n(x-h).$$

This is a special case of the *Friedrichs scheme* which we shall study in more generality below. Here

$$\tilde{E}(\xi) = \cos\xi + \mathrm{i}a\lambda\sin\xi,$$

and we find

$$|\tilde{E}(\xi)|^2 = \cos^2\xi + a^2\lambda^2\sin^2\xi \leq 1, \quad \text{for } \xi \in \mathbf{R},$$

if and only if $|a\lambda| \leq 1$.

This case of the Friedrichs scheme may also be written in the form

$$\frac{U^{n+1}(x) - U^n(x)}{k} = a\,\frac{U^n(x+h) - U^n(x-h)}{2h} + \frac{1}{2k}\Big(U^n(x+h) - 2U^n(x) + U^n(x-h)\Big),$$

or

$$\partial_t U^n = a\hat{\partial}_x U^n + \frac{1}{2}\frac{h}{\lambda}\partial_x\bar{\partial}_x U^n. \tag{12.15}$$

This equation may be thought of as an approximation to a parabolic equation with the small diffusion coefficient $\frac{1}{2}h/\lambda$. The stability of this scheme may

be interpreted as the result of introducing *artificial diffusion* in the original scheme (12.14). (This is also referred to as artificial viscosity in computational fluid dynamics.)

Let us note that for the Friedrichs scheme $U^n(x)$ may be expressed in terms of the initial data in the form

$$U^n(x) = \sum_{j=-n}^{n} a_{nj} v(x - jh),$$

and thus uses the values of $v(x)$ in the interval $[x - nh, x + nh] = [x - t_n/\lambda, x + t_n/\lambda]$. The exact solution at $t = t_n$ is given by (12.2) as the value of v at $x + t_n a$. It is clear that if the domain of dependence of the difference scheme, i.e., the interval $[x - t_n/\lambda, x + t_n/\lambda]$, does not contain the domain of dependence of the exact solution, namely the point $x + t_n a$, then the difference method could not possibly be successful. This condition reduces to $-1 \le a\lambda \le 1$, which is our old stability criterion.

For a general scheme of the form (12.10) we may thus formulate the *Courant-Friedrichs-Lewy condition* (or the *CFL condition*) for stability: In order for the scheme to be stable it is necessary that the domain of dependence of the finite difference scheme at (x, t) contains the domain of dependence of the continuous problem.

In our first example (12.4) we find that the finite difference scheme has the interval of dependence $[x, x + t_n/\lambda]$ and thus that the CFL condition requires $0 \le a\lambda \le 1$. In particular, we recover our old stability condition $a\lambda \le 1$, and also note that the scheme (12.4) cannot be used for $a < 0$. In the same way we find that for $a > 0$ the forward difference quotient in (12.4) could not successfully be replaced by a backward difference quotient. For $a < 0$, however, as we have learned, the scheme (12.4) with ∂_x replaced by $\bar{\partial}_x$ is stable if $a\lambda \ge -1$.

That the CFL condition is not sufficient for stability is shown by the scheme (12.14), which has the same domain of dependence as the Friedrichs scheme but which is unstable for all λ.

Like our first example (12.4) the Friedrichs scheme is also first order accurate: If the exact solution u of (12.1) is sufficiently regular, then we have by the representation (12.15) that

$$\begin{aligned}
\partial_t u^n - a\hat{\partial}_x u^n - \frac{1}{2}\frac{h}{\lambda}\partial_x\bar{\partial}_x u^n &= u_t^n + \frac{1}{2}k u_{tt}^n - a u_x^n - \frac{1}{2}\frac{h}{\lambda}u_{xx}^n + O(h^2) \\
&= \frac{1}{2}\frac{h}{\lambda}(\lambda^2 u_{tt}^n - u_{xx}^n) + O(h^2) \\
&= \frac{1}{2}\frac{h}{\lambda}(a^2\lambda^2 - 1)u_{xx}^n + O(h^2), \quad \text{as } h \to 0.
\end{aligned}$$

Thus the error is first order except for the special choice $\lambda = 1/|a|$, in which case the approximate solution is equal to the exact solution (cf. (12.13)).

Next we propose to determine a second order accurate scheme of the form

$$U^{n+1}(x) = (E_k U^n)(x) = a_1 U^n(x-h) + a_0 U^n(x) + a_{-1} U^n(x+h).$$

The formula (12.13) shows that the condition for this is

$$a_1 \mathrm{e}^{-\mathrm{i}\xi} + a_0 + a_{-1}\mathrm{e}^{\mathrm{i}\xi} = \mathrm{e}^{\mathrm{i}a\lambda\xi} + O(\xi^3), \quad \text{as } \xi \to 0,$$

or, by Taylor expansion,

$$\begin{aligned}(a_1 + a_0 + a_{-1}) - \mathrm{i}(a_1 - a_{-1})\xi - \tfrac{1}{2}(a_1 + a_{-1})\xi^2 \\ = 1 + \mathrm{i}a\lambda\xi - \tfrac{1}{2}a^2\lambda^2\xi^2 + O(\xi^3), \quad \text{as } \xi \to 0,\end{aligned}$$

that is,

$$\begin{aligned}a_1 + a_0 + a_{-1} &= 1,\\ a_1 - a_{-1} &= -a\lambda,\\ a_1 + a_{-1} &= a^2\lambda^2.\end{aligned}$$

This results in

$$a_{-1} = \tfrac{1}{2}(a^2\lambda^2 + a\lambda), \quad a_0 = 1 - a^2\lambda^2, \quad a_1 = \tfrac{1}{2}(a^2\lambda^2 - a\lambda),$$

and thus

$$\begin{aligned}(E_k U^n)(x) = \tfrac{1}{2}(a^2\lambda^2 + a\lambda)U^n(x+h) + (1 - a^2\lambda^2)U^n(x)\\ + \tfrac{1}{2}(a^2\lambda^2 - a\lambda)U^n(x-h),\end{aligned}$$

which gives

$$\tilde{E}(\xi) = 1 - a^2\lambda^2 + a^2\lambda^2\cos\xi + \mathrm{i}a\lambda\sin\xi.$$

We find by a simple calculation

$$|\tilde{E}(\xi)|^2 = 1 - a^2\lambda^2(1 - a^2\lambda^2)(1 - \cos\xi)^2, \tag{12.16}$$

and hence the method is stable in L_2 exactly if $a^2\lambda^2 \le 1$, see Problem 12.2. Again this agrees with the CFL necessary condition for stability.

This latter method is referred to as the *Lax-Wendroff method*. We remark that this is an example of a method which is not stable in maximum-norm even though it is L_2-stable. In fact, it can be shown that, if $0 < a^2\lambda^2 < 1$, then

$$\|E_k^n v\|_{\mathcal{C}} \le C n^{1/12} \|v\|_{\mathcal{C}},$$

and that this estimate is sharp in terms of the power of n. However, this power is small and the effect of the instability is in general not noticeable.

12.2 Symmetric Hyperbolic Systems

Much of what has been said in Sect. 12.1 generalizes to systems in one space dimension,

$$\frac{\partial u}{\partial t} = A\frac{\partial u}{\partial x}, \quad \text{in } \mathbf{R}\times\mathbf{R}_+,$$

where $u = (u_1,\dots,u_N)^{\mathrm{T}}$ is a vector with N components and A is a symmetric $N\times N$ matrix.

For instance, the Friedrichs scheme now takes the form

$$U^{n+1}(x) = (E_kU^n)(x) = \tfrac{1}{2}(I+\lambda A)U^n(x+h)+\tfrac{1}{2}(I-\lambda A)U^n(x-h) \tag{12.17}$$

and the Lax-Wendroff scheme is

$$\begin{aligned} U^{n+1}(x) = (E_kU^n)(x) &= \tfrac{1}{2}(\lambda^2A^2+\lambda A)U^n(x+h)\\ &\quad + (I-\lambda^2A^2)U^n(x) + \tfrac{1}{2}(\lambda^2A^2-\lambda A)U^n(x-h). \end{aligned} \tag{12.18}$$

The symbols of these operators now become the matrix-valued periodic functions

$$\tilde{E}(\xi) = I\cos\xi + \mathrm{i}\lambda A\sin\xi,$$

and

$$\tilde{E}(\xi) = I-\lambda^2A^2+\lambda^2A^2\cos\xi + \mathrm{i}\lambda A\sin\xi,$$

respectively. These may be diagonalized by the same orthogonal transformation as A and one finds easily that the stability requirement in both cases is that

$$|A|\lambda \le 1,$$

where $|A|$ is the matrix norm subordinate to the Euclidean norm on $\mathbf{R}^N$, i.e.,

$$|A| = \sup_{v\neq 0}\frac{|Av|}{|v|} = \max_{j=1,\dots,N}|\mu_j|,$$

where μ_j are the eigenvalues of A.

For an example of such a system we consider the initial value problem

$$\begin{aligned} &\frac{\partial^2 w}{\partial t^2} = a^2\,\frac{\partial^2 w}{\partial x^2}, \quad \text{in } \mathbf{R}\times\mathbf{R}_+,\\ &w(\cdot,0) = w_0, \quad \frac{\partial w}{\partial t}(\cdot,0) = w_1, \quad \text{in } \mathbf{R}. \end{aligned} \tag{12.19}$$

The second order hyperbolic equation may be reduced to a system by setting

$$u_1 = a\,\frac{\partial w}{\partial x},\ u_2 = \frac{\partial w}{\partial t}.$$

These functions then satisfy

$$\frac{\partial u_1}{\partial t} = a\,\frac{\partial u_2}{\partial x}, \quad \frac{\partial u_2}{\partial t} = a\,\frac{\partial u_1}{\partial x},$$

i.e., we have for $u = (u_1,\, u_2)^{\mathrm{T}}$,

$$\begin{aligned} \frac{\partial u}{\partial t} &= \begin{bmatrix} 0 & a \\ a & 0 \end{bmatrix} \frac{\partial u}{\partial x}, \quad \text{in } \mathbf{R} \times \mathbf{R}_+, \\ u(\cdot, 0) &= \begin{bmatrix} a w_0' \\ w_1 \end{bmatrix}, \quad \text{in } \mathbf{R}. \end{aligned} \tag{12.20}$$

Conversely, the solution of (12.20) determines the solution of (12.19).

Either of the two schemes (12.17) and (12.18) may now be applied to the present system. Since the matrix in (12.20) has eigenvalues $\pm a$ we rediscover our standard stability criterion $|a|\lambda \le 1$.

We briefly turn to the case of more than one space dimension and consider a symmetric hyperbolic system (or Friedrichs system),

$$\begin{aligned} \frac{\partial u}{\partial t} &= \sum_{j=1}^{d} A_j \frac{\partial u}{\partial x_j}, \qquad \text{in } \mathbf{R} \times \mathbf{R}_+, \\ u(\cdot, 0) &= v, \qquad\qquad\quad \text{in } \mathbf{R}^d, \end{aligned} \tag{12.21}$$

where u is an N-vector valued function and the A_j are symmetric $N \times N$ matrices. We recall from Sect. 11.4 that the corresponding initial value problem is correctly posed in L_2.

We consider now an associated finite difference operator

$$U^{n+1}(x) = (E_k U^n)(x) = \sum_{\beta} a_\beta U^n(x - \beta h), \tag{12.22}$$

where $\beta = (\beta_1, \ldots, \beta_d)$ has integer components and the a_β are finitely many constant $N \times N$ matrices. With $\xi = (\xi_1, \ldots, \xi_d) \in \mathbf{R}^d$ the symbol is now the matrix

$$\tilde{E}(\xi) = \sum_{\beta} a_\beta \mathrm{e}^{-\mathrm{i}\beta\cdot\xi}, \quad \text{where } \beta \cdot \xi = \beta_1\xi_1 + \cdots + \beta_d\xi_d.$$

After Fourier transformation we have now

$$(U^{n+1})\hat{}\,(\xi) = \tilde{E}(h\xi)(U^n)\hat{}\,(\xi), \quad \text{for } n \ge 0,$$

and hence

$$(U^n)\hat{}\,(\xi) = \tilde{E}(h\xi)^n \hat{v}(\xi).$$

It follows easily that a necessary and sufficient condition for stability in L_2 is that the matrix norm of the symbol satisfies

$$|\tilde{E}(\xi)^n| \le C, \quad \text{for } n \ge 0,\ \xi \in \mathbf{R}^d. \tag{12.23}$$

In contrast to the scalar case it is not true for matrices that $|A^n| \le C$ for $n \ge 0$ implies $|A| \le 1$, as the example

$$\begin{bmatrix} a & 1 \\ 0 & a \end{bmatrix}^n = \begin{bmatrix} a^n & na^{n-1} \\ 0 & a^n \end{bmatrix}, \tag{12.24}$$

with $|a| < 1$ shows. However, if (12.23) holds, it follows that for each eigenvalue $\lambda_j(\xi)$ of $\tilde{E}(\xi)$ we have $|\lambda_j(\xi)^n| \le C$, for $\xi \in \mathbf{R}^d$, $n \ge 0$, and hence

$$|\lambda_j(\xi)| \le 1, \quad \text{for } \xi \in \mathbf{R}^d.$$

This is referred to as *von Neumann's stability condition*, and is thus a *necessary* condition for stability in L_2. It is not a sufficient condition as illustrated by the example (12.24) with $a = 1$.

A sufficient condition for stability in L_2 is obviously

$$|\tilde{E}(\xi)| \le 1, \quad \text{for } \xi \in \mathbf{R}^d.$$

In order to be able to construct a stable difference scheme of the above form we shall need the following result.

Lemma 12.1. *Assume that a_β are symmetric, positive semidefinite matrices with $\sum_\beta a_\beta = I$. Then*

$$|\tilde{E}(\xi)| = \Big| \sum_\beta a_\beta \mathrm{e}^{-\mathrm{i}\beta\xi} \Big| \le 1, \quad \textit{for } \xi \in \mathbf{R}^d.$$

Proof. Let

$$\langle u, v \rangle = \sum_{j=1}^N u_j \overline{v_j}, \quad \text{for } u, v \in \mathbf{C}^N.$$

Since a_β is real, symmetric and positive semidefinite we see that the corresponding bilinear form $\langle a_\beta u, v \rangle$ satisfies

$$\langle a_\beta u, v \rangle = \overline{\langle a_\beta v, u \rangle}, \quad \langle a_\beta u, u \rangle \ge 0, \quad \text{for } u, v \in \mathbf{C}^N.$$

Using these properties it is easy to prove the generalized Cauchy-Schwarz inequality

$$|\langle a_\beta u, v \rangle| \le \langle a_\beta u, u \rangle^{1/2} \langle a_\beta v, v \rangle^{1/2}.$$

(The proof of the standard Cauchy-Schwarz inequality works; this inequality is a generalization because $\langle a_\beta u, u \rangle^{1/2}$ is only a seminorm.) Hence, using also the inequality $2ab \le a^2 + b^2$, we have

$$|\langle a_\beta u, v \rangle| \le \tfrac{1}{2} \langle a_\beta u, u \rangle + \tfrac{1}{2} \langle a_\beta v, v \rangle.$$

Therefore,

$$|\langle \tilde{E}(\xi)v, w\rangle| \le \sum_\beta |\langle a_\beta e^{-i\beta\cdot\xi} v, w\rangle|$$
$$\le \tfrac{1}{2}\sum_\beta \langle a_\beta v, v\rangle + \tfrac{1}{2}\sum_\beta \langle a_\beta w, w\rangle = \tfrac{1}{2}|v|^2 + \tfrac{1}{2}|w|^2.$$

Taking $w = \tilde{E}(\xi)v$, we conclude

$$|w|^2 \le \tfrac{1}{2}|v|^2 + \tfrac{1}{2}|w|^2,$$

which completes the proof. □

As an application, we take the Friedrichs scheme (of which we have seen particular cases above)

$$U^{n+1}(x) = (E_k U^n)(x)$$
$$= \frac{1}{2}\sum_{j=1}^{d} \Big\{ \Big(\frac{1}{d} I + \lambda A_j\Big) U^n(x + he_j) + \Big(\frac{1}{d} I - \lambda A_j\Big) U^n(x - he_j) \Big\}, \tag{12.25}$$

where e_j is the unit vector in the direction of x_j. This may also be written, similarly to (12.15),

$$\partial_t U^n = \sum_{j=1}^{d} A_j \hat{\partial}_{x_j} U^n + \frac{1}{2}\frac{h}{\lambda d}\sum_{j=1}^{d} \partial_{x_j}\bar{\partial}_{x_j} U^n,$$

and is thus, in particular, consistent with (12.21).

Now, if λ is such that

$$0 < \lambda \le \min_{1\le j\le d} (d|A_j|)^{-1}, \tag{12.26}$$

then the coefficients of (12.25) are positive semidefinite and hence the lemma shows stability in L_2.

For instance, for the system

$$\frac{\partial u}{\partial t} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \frac{\partial u}{\partial x_1} + \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \frac{\partial u}{\partial x_2},$$

the condition (12.26) reduces to $0 < \lambda \le 1/2$. In this particular case

$$\tilde{E}(\xi) = \tfrac{1}{2} I(\cos\xi_1 + \cos\xi_2) + \lambda \mathrm{i} \begin{bmatrix} \sin\xi_1 & \sin\xi_2 \\ \sin\xi_2 & -\sin\xi_1 \end{bmatrix},$$

which is a normal matrix (one which commutes with its adjoint). For such a matrix its norm equals the maximum modulus of its eigenvalues, and using this one may prove that $|\tilde{E}(\xi)| \le 1$ for $\lambda \le 1/\sqrt{2}$, which is less restrictive a condition in this case than the above condition based on the lemma.

The Friedrichs scheme is again only first order accurate. In fact, it may be shown that schemes of the form (12.22) with a_β positive semidefinite may in general only be accurate of first order.

12.3 The Wendroff Box Scheme

In this final section we shall describe the second order Wendroff box scheme, which is suitable for mixed initial-boundary value problems and also for systems in the case of one space dimension.

We consider thus the initial boundary value problem

$$\begin{aligned} &\frac{\partial u}{\partial t} + a\,\frac{\partial u}{\partial x} + bu = f, && \text{in } \Omega \times J, \quad \text{where } \Omega = (0,1),\ J = (0,T),\\ &u(0,\cdot) = g, && \text{on } J,\\ &u(\cdot,0) = v, && \text{in } \Omega, \end{aligned}$$

where a, b, f, g, and v are smooth functions with a positive and where $g(0) = v(0)$ for compatibility at $(x,t) = (0,0)$. Note that since a is positive the boundary values have been prescribed on the left boundary.

With U_j^n the value of the mesh function at $(x_j, t_n) = (jh, nk)$, $0 \le j \le M$, $0 \le n \le N$, where $Mh = 1$, $Nk = T$, we define also

$$U_{j+1/2} = \tfrac{1}{2}(U_j + U_{j+1}) \quad \text{and} \quad U^{n+1/2} = \tfrac{1}{2}(U^n + U^{n+1}),$$

and

$$U_{j+1/2}^{n+1/2} = \tfrac{1}{4}(U_j^n + U_j^{n+1} + U_{j+1}^n + U_{j+1}^{n+1}).$$

The *Wendroff box scheme* is then

$$\begin{aligned} &\partial_t U_{j+1/2}^n + a\,\partial_x U_j^{n+1/2} + bU_{j+1/2}^{n+1/2} = f, && 0 \le j < M,\ 0 \le n < N,\\ &U_0^n = G^n = g(t_n), && 0 \le n \le N,\\ &U_j^0 = V_j = v(x_j), && 0 \le j \le M, \end{aligned} \tag{12.27}$$

where a, b, and f are evaluated at $(x_{j+1/2}, t_{n+1/2})$. The difference equation may also be written

$$\begin{aligned} \frac{U_j^{n+1} + U_{j+1}^{n+1} - U_j^n - U_{j+1}^n}{2k} + a\,\frac{U_{j+1}^{n+1} + U_{j+1}^n - U_j^{n+1} - U_j^n}{2h}\\ + b\,\frac{U_j^{n+1} + U_{j+1}^{n+1} + U_j^n + U_{j+1}^n}{4} = f; \end{aligned} \tag{12.28}$$

and we see by symmetry that it is second order accurate. This equation may also be expressed as

$$\begin{aligned} (1 + a\lambda + \tfrac{1}{2}bk)U_{j+1}^{n+1} = (1 + a\lambda - \tfrac{1}{2}bk)U_j^n + (1 - a\lambda - \tfrac{1}{2}bk)U_{j+1}^n\\ - (1 - a\lambda + \tfrac{1}{2}bk)U_j^{n+1} + 2kf, \quad \text{where } \lambda = k/h. \end{aligned} \tag{12.29}$$

This defines the solution at (x_{j+1}, t_{n+1}) in terms of the values at (x_j, t_n), (x_{j+1}, t_n) and (x_j, t_{n+1}). Given U^n one may therefore find U^{n+1} explicitly in the order $U_0^{n+1} = G^{n+1}$, U_1^{n+1}, $U_2^{n+1}, \ldots, U_M^{n+1}$.

We shall show the stability of this method in the discrete L_2-norm

$$\|V\|_h = \Big(h \sum_{j=1}^{M} V_j^2\Big)^{1/2},$$

and restrict ourselves, for simplicity only, to the case that a is constant and $b = f = g = 0$. In this case (12.28) reduces to

$$U_{j+1}^{n+1} = U_j^n + \frac{1-a\lambda}{1+a\lambda}(U_{j+1}^n - U_j^{n+1}),$$

and we shall show

$$\|U^n\|_h \le C\|V\|_h, \quad \text{for } 0 \le t_n \le T. \tag{12.30}$$

For this purpose we multiply (12.27) by $U_{j+1/2}^{n+1/2}$ and note that

$$\partial_t U_{j+1/2}^n \, U_{j+1/2}^{n+1/2} = \tfrac{1}{2}\partial_t (U_{j+1/2}^n)^2,$$

and

$$\partial_x U_j^{n+1/2} \, U_{j+1/2}^{n+1/2} = \tfrac{1}{2}\partial_x (U_j^{n+1/2})^2,$$

so that we obtain

$$\partial_t (U_{j+1/2}^m)^2 + a\partial_x (U_j^{m+1/2})^2 = 0.$$

After summation over $j = 0, \dots, M-1$, $m = 0, \dots, n-1$ and multiplication by hk this yields, for $n \le N$,

$$h \sum_{j=0}^{M-1} (U_{j+1/2}^n)^2 + ak \sum_{m=0}^{n-1} (U_M^{m+1/2})^2 = h \sum_{j=0}^{M-1} (U_{j+1/2}^0)^2 + ak \sum_{m=0}^{n-1} (U_0^{m+1/2})^2,$$

which implies, since we have assumed $U_0^{m+1/2} = 0$,

$$h \sum_{j=0}^{M-1} (U_j^n + U_{j+1}^n)^2 \le C\|V\|_h^2. \tag{12.31}$$

Similarly, multiplying (12.27) instead by

$$hk\partial_x \partial_t U_j^n = U_{j+1}^{n+1} - U_j^{n+1} - U_{j+1}^n + U_j^n,$$

we obtain after a simple calculation

$$h \sum_{j=0}^{M-1} (U_{j+1}^n - U_j^n)^2 \le C\|V\|_h^2. \tag{12.32}$$

Together, (12.31) and (12.32) complete the proof of (12.30).

The stability and the second order of accuracy imply second order convergence, provided u is smooth enough, i.e.,

$$\|U^n - u^n\|_h \le C(u)h^2, \quad \text{for } k/h = \lambda = \text{constant}.$$

12.4 Problems

Problem 12.1. Prove that the Friedrichs scheme for (12.1) is stable in the maximum norm if and only if $|a|\lambda \leq 1$.

Problem 12.2. Show (12.16) and hence that the Lax-Wendroff scheme is stable in L_2 if and only if $|a|\lambda \leq 1$.

Problem 12.3. Let E_k be an explicit finite difference operator defined by

$$(E_k V)_j = \sum_p a_p V_{j-p}.$$

(a) Show that

$$\|E_k V\|_{\infty,h} \leq C\|v\|_{\infty,h}, \quad \text{with } C = \sum_p |a_p|,$$

and that this inequality does not hold with any smaller constant.
(b) Show that $(E_k^n V)_j = \sum_p a_{np} V_{j-p}$, where

$$a_{np} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \tilde{E}(\xi)^n \mathrm{e}^{\mathrm{i}p\xi}\, \mathrm{d}\xi,$$

with $\tilde{E}(\xi) = \sum_j a_j \mathrm{e}^{-\mathrm{i}j\xi}$ the symbol of E_k.
(c) Show that E_k is maximum-norm stable if and only if

$$\sum_p |a_{np}| \leq C, \quad \forall n \geq 0.$$

Problem 12.4. Prove that the Lax-Wendroff scheme is unstable in the maximum norm if $|a|\lambda > 1$. Hint: Problem 12.3.

Problem 12.5. Recall from Sect. 7.1 that for an $m \times m$ matrix M one can define $\exp(M) = \sum_{j=0}^{\infty} \frac{1}{j!} M^j$. Consider the symmetric hyperbolic system $\partial u/\partial t = A\partial u/\partial x$.
(a) Show that a finite difference scheme for this system of the form (cf. (12.17) and (12.18))

$$(E_k V)_j = \sum_p a_p(\lambda A) V_{j-p}$$

is accurate of order r, $r = 1, 2$, if

$$\tilde{E}(\xi) = \exp(\mathrm{i}\lambda A\xi) + O(\xi^{r+1}), \quad \text{as } \xi \to 0.$$

(b) Check this condition for the Friedrichs and Lax-Wendroff schemes (12.17) and (12.18).

Problem 12.6. Discuss the meaning of the CFL condition for the initial-boundary value problem in Sect. 12.3 and show that it is satisfied for the Wendroff box scheme defined in (12.27).

Problem 12.7. (Computer exercise.) Apply the Wendroff box scheme to the problem in Example 11.8 with $h = k = 1/10$ and $h = k = 1/20$. Calculate the errors at $(1, 1/2)$.

13 The Finite Element Method for Hyperbolic Equations

In this chapter we apply the finite element method to hyperbolic equations. In Sect. 13.1 we study an initial-boundary value problem for the wave equation, and discuss semidiscrete and completely discrete schemes based on the standard finite element discretization in the spatial variables. In Sect. 13.2 we consider a scalar partial differential equation of first order in two independent variables. We begin by treating the equation as an evolution equation and show a nonoptimal order $O(h)$ error estimate for the standard Galerkin method. Looking instead of the associated boundary value problem as a two-dimensional problem of the type treated in Sect. 11.3, we introduce the streamline diffusion modification and demonstrate a $O(h^{3/2})$ convergence estimate. We finally return to the evolution aspect and combine streamline diffusion with the so-called discontinuous Galerkin method to design a time stepping scheme by using two-dimensional approximating functions which may be discontinuous at the time levels.

13.1 The Wave Equation

In this section we briefly discuss some results concerning semidiscrete and completely discrete schemes for the following initial-boundary value problem for the wave equation,

$$\begin{aligned} &u_{tt} - \Delta u = f, && \text{in } \Omega \times \mathbf{R}_+, \\ &u = 0, && \text{on } \Gamma \times \mathbf{R}_+, \\ &u(\cdot, 0) = v,\ u_t(\cdot, 0) = w, && \text{in } \Omega. \end{aligned} \tag{13.1}$$

As often earlier we assume that $\Omega \subset \mathbf{R}^2$ is a bounded convex domain whose boundary Γ is a polygon, and denote by $S_h \subset H_0^1$ a family of spaces of piecewise linear finite element functions in the spatial variables.

The semidiscrete analogue of (13.1) is then to find $u_h(t) \in S_h$ such that, with our previous notation, in particular with $a(v, w) = (\nabla v, \nabla w)$,

$$\begin{aligned} &(u_{h,tt}, \chi) + a(u_h, \chi) = (f, \chi), \quad \forall \chi \in S_h, \quad \text{for } t > 0, \\ &u_h(0) = v_h,\ u_{h,t}(0) = w_h. \end{aligned} \tag{13.2}$$

This is an initial value problem for a system of ordinary differential equations of second order for the coefficients of u_h with respect to the standard basis $\{\Phi_j\}_{j=1}^{M_h}$ of S_h. If

$$u_h(x,t) = \sum_{j=1}^{M_h} \alpha_j(t)\Phi_j(x),$$

then (13.2) is equivalent to

$$B\alpha''(t) + A\alpha(t) = b(t), \quad \text{for } t > 0,$$

where the elements of B, A, and b are $b_{kj} = (\Phi_j, \Phi_k)$, $a_{kj} = a(\Phi_j, \Phi_k)$, and $b_k = (f, \Phi_k)$, respectively. The initial conditions are

$$\alpha(0) = \beta, \quad \alpha'(0) = \gamma,$$

where

$$v_h = \sum_{j=1}^{M_h} \beta_j \Phi_j, \quad w_h = \sum_{j=1}^{M_h} \gamma_j \Phi_j.$$

We begin by showing a discrete version of the energy conservation result in Theorem 11.2.

Lemma 13.1. *Let u_h be the solution of* (13.2) *with $f = 0$. Then*

$$\|u_{h,t}(t)\|^2 + |u_h(t)|_1^2 = \|w_h\|^2 + |v_h|_1^2, \quad \textit{for } t \geq 0.$$

Proof. Choosing $\chi = u_{h,t}$ in (13.2) we have

$$\tfrac{1}{2}\frac{d}{dt}\big(\|u_{h,t}\|^2 + |u_h|_1^2\big) = 0,$$

from which the result immediately follows. □

We now show the following error estimate, where R_h denotes the elliptic projection defined in (5.49).

Theorem 13.1. *Let u_h and u be the solutions of* (13.2) *and* (13.1). *Then we have, for $t \geq 0$,*

$$\begin{aligned}
\|u_{h,t}(t) - u_t(t)\| &\leq C\Big(|v_h - R_h v|_1 + \|w_h - R_h w\|\Big) \\
&\quad + Ch^2\Big(\|u_t(t)\|_2 + \int_0^t \|u_{tt}\|_2 \,\mathrm{d}s\Big), \\
\|u_h(t) - u(t)\| &\leq C\Big(|v_h - R_h v|_1 + \|w_h - R_h w\|\Big) \\
&\quad + Ch^2\Big(\|u(t)\|_2 + \int_0^t \|u_{tt}\|_2 \,\mathrm{d}s\Big), \\
|u_h(t) - u(t)|_1 &\leq C\Big(|v_h - R_h v|_1 + \|w_h - R_h w\|\Big) \\
&\quad + Ch\Big(\|u(t)\|_2 + \int_0^t \|u_{tt}\|_1 \,\mathrm{d}s\Big).
\end{aligned}$$

Proof. Writing as usual

$$u_h - u = (u_h - R_h u) + (R_h u - u) = \theta + \rho,$$

we may bound ρ and ρ_t as in the proof of Theorem 10.1 by

$$\|\rho(t)\| + h|\rho(t)|_1 \le Ch^2\|u(t)\|_2, \quad \|\rho_t(t)\| \le Ch^2\|u_t(t)\|_2. \tag{13.3}$$

For $\theta(t)$ we have, after a calculation analogous to that in (10.14),

$$(\theta_{tt}, \chi) + a(\theta, \chi) = -(\rho_{tt}, \chi), \quad \forall \chi \in S_h, \quad \text{for } t > 0. \tag{13.4}$$

Imitating the proof of Lemma 13.1, we choose $\chi = \theta_t$:

$$\tfrac{1}{2}\frac{d}{dt}\big(\|\theta_t\|^2 + |\theta|_1^2\big) \le \|\rho_{tt}\|\,\|\theta_t\|.$$

After integration in t we obtain

$$\begin{aligned}
\|\theta_t(t)\|^2 + |\theta(t)|_1^2 &\le \|\theta_t(0)\|^2 + |\theta(0)|_1^2 + 2\int_0^t \|\rho_{tt}\|\,\|\theta_t\|\,\mathrm{d}s \\
&\le \|\theta_t(0)\|^2 + |\theta(0)|_1^2 + 2\int_0^t \|\rho_{tt}\|\,\mathrm{d}s \max_{s\in[0,t]}\|\theta_t\| \\
&\le \|\theta_t(0)\|^2 + |\theta(0)|_1^2 + 2\Big(\int_0^T \|\rho_{tt}\|\,\mathrm{d}s\Big)^2 + \tfrac{1}{2}\Big(\max_{s\in[0,T]}\|\theta_t\|\Big)^2,
\end{aligned}$$

for $t \in [0, T]$. This implies

$$\tfrac{1}{2}\Big(\max_{s\in[0,T]}\|\theta_t\|\Big)^2 \le \|\theta_t(0)\|^2 + |\theta(0)|_1^2 + 2\Big(\int_0^T \|\rho_{tt}\|\,\mathrm{d}s\Big)^2$$

and hence

$$\|\theta_t(t)\|^2 + |\theta(t)|_1^2 \le 2\|\theta_t(0)\|^2 + 2|\theta(0)|_1^2 + 4\Big(\int_0^T \|\rho_{tt}\|\,\mathrm{d}s\Big)^2,$$

for $t \in [0, T]$. In particular this holds with $t = T$ where T is arbitrary. Using also bounds for ρ_{tt} similar to (13.3), we obtain

$$\begin{aligned}
\|\theta_t(t)\| + \|\theta(t)\| &\le C\Big(\|\theta_t(t)\| + |\theta(t)|_1\Big) \\
&\le C\Big(\|w_h - R_h w\| + |v_h - R_h v|_1\Big) + Ch^2\int_0^t \|u_{tt}\|_2\,\mathrm{d}s,
\end{aligned}$$

and

$$|\theta(t)|_1 \le C\Big(\|w_h - R_h w\| + |v_h - R_h v|_1\Big) + Ch\int_0^t \|u_{tt}\|_1\,\mathrm{d}s.$$

Together with the bounds in (13.3) this completes the proof. □

We remark that the choices $v_h = R_h v$ and $w_h = R_h w$ in Theorem 13.1 give optimal order error estimates for all the three quantities considered, but that other optimal choices of v_h could cause a loss of one power of h, because of the gradient in the first term on the right. This can be avoided by a more refined argument. The regularity requirement on the exact solution can also be reduced.

We shall now briefly discuss the discretization also in time, and let $U^n \in S_h$ denote the approximation at time $t_n = nk$, where k is the time step. One possible method is then to determine U^n for $n \geq 2$ by posing for $n \geq 1$ the equations

$$(13.5)\quad (\partial_t \bar{\partial}_t U^n, \chi) + a(\tfrac{1}{4}(U^{n+1} + 2U^n + U^{n-1}), \chi) = (f(t_n), \chi), \quad \forall \chi \in S_h,$$

where U^0 and U^1 are given approximations of $u(0) = v$ and $u(t_1)$, respectively. The choice of the average in the second term is motivated by a combination of stability and accuracy considerations. As regards stability we show the following fully discrete analogue of the semidiscrete energy conservation law of Lemma 13.1, where we define $U^{n+1/2} = (U^n + U^{n+1})/2$.

Lemma 13.2. *We have for the solution of* (13.5)*, with* $f = 0$*,*

$$\|\partial_t U^n\|^2 + |U^{n+1/2}|_1^2 = \|\partial_t U^0\|^2 + |U^{1/2}|_1^2, \quad \text{for } n \geq 0.$$

Proof. We apply (13.5) with

$$\chi = \frac{1}{2k}(U^{n+1} - U^{n-1}) = \frac{1}{2}(\partial_t U^n + \partial_t U^{n-1}) = \frac{1}{k}(U^{n+1/2} - U^{n-1/2}).$$

With this χ we have

$$(\partial_t \bar{\partial}_t U^n, \chi) = \frac{1}{2k}(\partial_t U^n - \partial_t U^{n-1}, \partial_t U^n + \partial_t U^{n-1}) = \frac{1}{2}\bar{\partial}_t \|\partial_t U^n\|^2$$

and

$$\begin{aligned} a(\tfrac{1}{4}U^{n+1} + \tfrac{1}{2}U^n + \tfrac{1}{4}U^{n-1}, \chi) &= \frac{1}{2k} a(U^{n+1/2} + U^{n-1/2}, U^{n+1/2} - U^{n-1/2}) \\ &= \frac{1}{2}\bar{\partial}_t |U^{n+1/2}|_1^2. \end{aligned}$$

Hence

$$\bar{\partial}_t \big(\|\partial_t U^n\|^2 + |U^{n+1/2}|_1^2 \big) = 0,$$

from which the result follows. □

Using this stability result together with direct analogues of the arguments in the proof of Theorem 13.1 one may show the following, where we use our usual notation $\theta^n = U^n - R_h u(t_n)$. We leave the details to Problem 13.4.

Theorem 13.2. *Let U^n and u be the solutions of* (13.5) *and* (13.1)*, and assume that the initial values U^0 and U^1 are chosen in such a way that*

$$\|\partial_t \theta^0\| + |\theta^0|_1 + |\theta^1|_1 \le C(h^2 + k^2).$$

Then, under the appropriate regularity conditions for u, we have, with $C(u,t)$ nondecreasing in t,

$$\|U^{n+1/2} - u(t_n + \tfrac{1}{2}k)\| + \|\partial_t U^n - u_t(t_n + \tfrac{1}{2}k)\| \le C(u,t_n)(h^2 + k^2),$$

and

$$|U^{n+1/2} - u(t_n + \tfrac{1}{2}k)|_1 \le C(u,t_n)(h + k^2), \quad \text{for } n \ge 0.$$

The conditions for the initial values may be satisfied by taking $U^0 = R_h v$ and $U^1 = R_h(v + kw + \frac{1}{2}k^2 u_{tt}(0))$, where $u_{tt}(0) = \Delta v + f(0)$.

Although Theorem 13.2 estimates the error at the points $t_n + \frac{1}{2}k$ it is clear that optimal order approximations at the points t_n may also be obtained, since, e.g.,

$$\begin{aligned}\|\tfrac{1}{4}(U^{n+1} + 2U^n + U^{n-1}) - u(t_n)\| &\le \tfrac{1}{2}\|U^{n+1/2} - u(t_n + \tfrac{1}{2}k)\| + \tfrac{1}{2}\|U^{n-1/2} - u(t_n - \tfrac{1}{2}k)\| \\ &\quad + \|\tfrac{1}{2}(u(t_n + \tfrac{1}{2}k) + u(t_n - \tfrac{1}{2}k)) - u(t_n)\| \le C(u,t_n)(h^2 + k^2).\end{aligned}$$

13.2 First Order Hyperbolic Equations

We begin by considering the initial-boundary value problem, cf. Example 11.7,

$$\begin{aligned} u_t + u_x &= f, && \text{in } \Omega = (0,1), \quad \text{for } t > 0, \\ u(0,t) &= 0, && \text{for } t > 0, \\ u(\cdot,0) &= v, && \text{in } \Omega. \end{aligned} \tag{13.6}$$

With $0 = x_0 < x_1 < \cdots < x_M = 1$ and $K_j = [x_{j-1}, x_j]$, we shall seek an approximate solution in the space

$$S_h^- = \{\chi \in \mathcal{C}(\bar{\Omega}) : \chi \text{ linear in } K_j,\ j = 1, \ldots, M,\ \chi(0) = 0\}. \tag{13.7}$$

Note that we require the functions in S_h^- to vanish at $x = 0$, i.e., on the spatial part $\Gamma_{-,x}$ of the inflow boundary, but not at $x = 1$, which is part of the outflow boundary.

The spatially discrete standard Galerkin method is then to find $u_h(t) \in S_h^-$ for $t \ge 0$ such that

$$\begin{aligned} (u_{h,t} + u_{h,x}, \chi) &= (f, \chi), \quad \forall \chi \in S_h^-,\ t > 0, \\ u_h(0) &= v_h \approx v. \end{aligned} \tag{13.8}$$

In terms of the standard basis $\{\Phi_j\}_{j=1}^M$ of hat functions this may be written in the form

$$B\alpha'(t) + A\alpha(t) = f, \quad \text{for } t > 0, \quad \text{with } \alpha(0) = \gamma,$$

where as usual B is the symmetric positive definite matrix with elements $b_{kj} = (\Phi_j, \Phi_k)$, so that, in particular, the problem has a well defined solution for $t \geq 0$, but where the matrix A with elements $a_{kj} = (\Phi_j', \Phi_k) = -(\Phi_k', \Phi_j) = -a_{jk}$ is now skew-symmetric.

We begin to show the stability of this method, and choose $\chi = u_h$ in (13.8). This gives

$$\tfrac{1}{2}\frac{d}{dt}\|u_h\|^2 + (u_{h,x}, u_h) = (f, u_h) \leq \|f\|\,\|u_h\|.$$

Here

$$(u_{h,x}, u_h) = \tfrac{1}{2}\Big[u_h^2\Big]_0^1 = \tfrac{1}{2}u_h(1)^2 \geq 0,$$

and thus

$$\frac{d}{dt}\|u_h\| \leq \|f\|,$$

so that after integration

$$\|u_h(t)\| \leq \|v_h\| + \int_0^t \|f\|\,\mathrm{d}s, \quad \text{for } t \geq 0. \tag{13.9}$$

We now show an error estimate.

Theorem 13.3. *Let u_h and u be the solutions of* (13.8) *and* (13.6). *Then, with v_h suitably chosen, we have*

$$\|u_h(t) - u(t)\| \leq Ch\Big(\|v\|_1 + \int_0^t \big(\|u\|_2 + \|u_t\|_1\big)\,\mathrm{d}s\Big), \quad \textit{for } t \geq 0.$$

Proof. With I_h the standard interpolation operator into S_h we write

$$u_h - u = (u_h - I_h u) + (I_h u - u) = \theta + \rho.$$

Here by Theorem 5.5

$$\|\rho(t)\| \leq Ch\|u(t)\|_1 \leq Ch\Big(\|v\|_1 + \int_0^t \|u_t\|_1\,\mathrm{d}s\Big),$$

which is bounded as desired. By our definitions we have $\theta \in S_h^-$ and

$$(\theta_t, \chi) + (\theta_x, \chi) = -(\omega, \chi), \quad \forall \chi \in S_h^-, \quad \text{with } \omega = \rho_t + \rho_x.$$

From the stability estimate (13.9) and Theorem 5.5 we conclude, if $v_h = I_h v$ so that $\theta(0) = 0$,

$$\|\theta(t)\| \leq \int_0^t \Big(\|\rho_t\| + \|\rho_x\|\Big)\,\mathrm{d}s \leq Ch\int_0^t \|u_t\|_1\,\mathrm{d}s + Ch\int_0^t \|u\|_2\,\mathrm{d}s.$$

This completes the proof. □

We note that the error bound is not of optimal order $O(h^2)$ because the bound for $\theta(t)$ contains the derivative of the interpolation error.

This analysis of the spatially semidiscrete problem may be carried over to fully discrete methods. We exemplify this by the backward Euler method, i.e., with our standard notation,

$$\begin{aligned}(\bar\partial_t U^n\chi) + (U_x^n,\chi) &= (f^n,\chi), \quad \forall\chi\in S_h^-,\ n>0,\\ U^0 &= v_h.\end{aligned} \tag{13.10}$$

Now the stability bound is (Problem 13.5)

$$\|U^n\| \le \|v_h\| + k\sum_{j=1}^{n}\|f^j\|, \quad \text{for } n\ge 0, \tag{13.11}$$

and the error estimate reads as follows.

Theorem 13.4. *Let U^n and u be the solutions of* (13.10) *and* (13.6)*. Then, with v_h suitably chosen, we have for $n\ge 0$,*

$$\|U^n - u(t_n)\| \le Ch\Big(\|v\|_1 + \int_0^{t_n}(\|u\|_2 + \|u_t\|_1)\,\mathrm{d}s\Big) + Ck\int_0^{t_n}\|u_{tt}\|\,\mathrm{d}s.$$

Proof. This time $\theta^n = U^n - I_h u^n$ satisfies, with $u^n = u(t_n)$,

$$\begin{aligned}(\bar\partial_t\theta^n,\chi) + (\theta_x^n,\chi) &= -(\omega^n,\chi), \quad \forall\chi\in S_h^-,\\ &\text{where } \omega^n = \bar\partial_t\rho^n + \rho_x^n + (u_t^n - \bar\partial_t u^n).\end{aligned}$$

The only essentially new term in ω^n is the last one, which is bounded by

$$\|u_t^n - \bar\partial_t u^n\| = \Big\|\int_{t_{n-1}}^{t_n}(s-t_{n-1})\,u_{tt}(s)\,\mathrm{d}s\Big\| \le k\int_{t_{n-1}}^{t_n}\|u_{tt}\|\,\mathrm{d}s.$$

Using the stability estimate (13.11) completes the proof. □

In order to proceed further with finite element methods for equations of first order we temporarily abandon the evolution aspect and consider the two-dimensional problem, which we discussed in Sect. 11.3,

$$\begin{aligned}a\cdot\nabla u + a_0 u &= f, && \text{in } \Omega,\\ u &= g, && \text{on } \Gamma_-,\end{aligned} \tag{13.12}$$

where, for brevity of presentation we assume that the velocity field $a = (a_1,\dots,a_d)$ and the coefficient a_0 are constant with $a_0 > 0$, and where we recall that we have defined the inflow and outflow boundaries by

$$\Gamma_- = \big\{x\in\Gamma : a\cdot n < 0\big\}, \quad \Gamma_+ = \big\{x\in\Gamma : a\cdot n > 0\big\}.$$

We shall keep track of the dependence on the constant a_0 in our estimates below but assume that it is bounded above.

We now discretize this by means of a standard two-dimensional finite element method. As in Sect. 5.2 we assume that $\Omega \subset \mathbf{R}^2$ is a bounded convex domain whose boundary Γ is a polygon, and we let S_h be a family of spaces of piecewise linear finite element functions with respect to a family of triangulations of Ω, without imposing any boundary conditions on the functions in S_h. Thus, instead of $S_h \subset H_0^1$, we now have $S_h \subset H^1$. We will use the interpolation operator I_h defined in Sect. 5.3 and recall the interpolation error estimates

$$\|I_h v - v\| \le Ch^2 \|v\|_2, \quad |I_h v - v|_1 \le Ch\|v\|_2. \tag{13.13}$$

Finally, we assume that the triangulations are matched to the boundary so that the inflow boundary is exactly a union of triangle edges and set

$$S_h^- = \{\chi \in S_h : \chi = 0 \text{ on } \Gamma_-\}.$$

We emphasize that the norms in (13.13) are taken over the two-dimensional domain Ω.

The standard Galerkin finite element method for the present problem is then to find $u_h \in S_h$ such that

$$\begin{aligned} &(a \cdot \nabla u_h, \chi) + a_0(u_h, \chi) = (f, \chi), && \forall \chi \in S_h^-, \\ &u_h = g_h = I_h g, && \text{on } \Gamma_-, \end{aligned} \tag{13.14}$$

where the inner products are now over the two-dimensional domain Ω.

Using Green's formula we have the identity (recall that a is constant)

$$(a \cdot \nabla v, v) = \tfrac{1}{2}(a \cdot n\, v, v)_\Gamma = \tfrac{1}{2}|v|_{\Gamma_+}^2 - \tfrac{1}{2}|v|_{\Gamma_-}^2, \tag{13.15}$$

where we have introduced the weighted norms

$$|v|_{\Gamma_\pm}^2 = \pm(a \cdot n\, v, v)_{\Gamma_\pm} = \int_{\Gamma_\pm} |a \cdot n|\, v^2 \, \mathrm{d}s.$$

We consider now a solution $w_h \in S_h$ of (13.14), which satisfies the homogeneous boundary condition $w_h = 0$ on Γ_-. Since $w_h \in S_h^-$ we may then choose $\chi = w_h$ to obtain, in view of (13.15),

$$\tfrac{1}{2}|w_h|_{\Gamma_+}^2 + a_0\|w_h\|^2 = (f, w_h).$$

For $f = 0$ this immediately shows that $w_h = 0$ and hence the uniqueness of the solution of (13.14), and therefore also the existence. We also easily conclude the stability estimate

$$|w_h|_{\Gamma_+}^2 + a_0\|w_h\|^2 \le C\|f\|^2, \quad \text{with } C = 1/a_0. \tag{13.16}$$

We continue our discussion by proving the following simple error estimate.

Theorem 13.5. *Let u_h and u be the solutions of* (13.14) *and* (13.12). *Then we have*

$$\|u_h - u\| \le Ch\|u\|_2.$$

Proof. We write $u_h - u = (u_h - I_h u) + (I_h u - u) = \theta + \rho$. Then, in view of (13.13), we have

$$\|\rho\| + h\|\rho\|_1 \le Ch^2\|u\|_2. \tag{13.17}$$

In order to estimate θ we note that $\theta \in S_h^-$ and, by (13.14) and (13.12),

$$(a \cdot \nabla\theta, \chi) + a_0(\theta, \chi) = -(a \cdot \nabla\rho + a_0\rho, \chi), \quad \forall \chi \in S_h^-. \tag{13.18}$$

Since $\theta \in S_h^-$ the stability estimate (13.16) together with (13.17) shows

$$|\theta|_{\Gamma_+}^2 + a_0\|\theta\|^2 \le C(\|\nabla\rho\|^2 + \|\rho\|^2) \le Ch^2\|u\|_2^2,$$

which completes the proof. □

We observe that, as in Theorems 13.3 and 13.4, the error estimate of Theorem 13.5 is of non-optimal order $O(h)$, as a result of the fact that the gradient of the interpolation error occurs on the right hand side of (13.18), and it is known that this error bound cannot be improved. Nevertheless, this means that the standard Galerkin method works adequately when the solution is smooth. However, solutions of (13.14) need not be smooth and experience shows that the method then performs less well, and may, for example, produce oscillations near layers where the solution changes rapidly.

In order to reduce such oscillations one may add artificial diffusion, as was done to obtain the Friedrichs scheme from the unstable scheme (12.15). The standard Galerkin method with artificial diffusion is to find $u_h \in S_h$ such that

$$\begin{aligned} &(a \cdot \nabla u_h, \chi) + a_0(u_h, \chi) + h(\nabla u_h, \nabla\chi) = (f, \chi), && \forall \chi \in S_h^-. \\ &u_h = g_h = I_h g, && \text{on } \Gamma_-. \end{aligned} \tag{13.19}$$

This method is consistent with the elliptic equation $a \cdot \nabla u + a_0 u - h\Delta u = f$, and the error is therefore still expected to be $O(h)$ for smooth solutions, see Problem 13.6, and for non-smooth solutions the method has been observed to smoothen discontinuities more than desirable.

More elaborate ways of adding diffusion have been developed. We now describe one such method, the so-called *streamline diffusion method*, which is to find $u_h \in S_h$ such that

$$\begin{aligned} &(a \cdot \nabla u_h + a_0 u_h, \chi + h\, a \cdot \nabla\chi) = (f, \chi + h\, a \cdot \nabla\chi), && \forall \chi \in S_h^-. \\ &u_h = g_h = I_h g, && \text{on } \Gamma_-. \end{aligned} \tag{13.20}$$

We note that the exact solution of (13.12) satisfies

$$(a\cdot\nabla u + a_0 u, \chi + h\,a\cdot\nabla\chi) = (f, \chi + h\,a\cdot\nabla\chi), \quad \forall\chi\in S_h^-, \tag{13.21}$$

which means that (13.20) is consistent with (13.12). This method is an example of a *Petrov-Galerkin method* in that we have chosen to multiply the equation by test functions other than those in S_h.

The rest of this section will perhaps require a little harder work to get through than has been the case with what we have presented earlier, but we include this material because we think that it illustrates the difficulties in applying the finite element method to first order hyperbolic equations.

We begin by discussing the stability and restrict ourselves again to a solution $w_h\in S_h^-$ of (13.20), thus vanishing on Γ_-. We then choose $\chi = w_h$, and use (13.15) and $ab\le a^2+\frac14 b^2$ to obtain

$$\tfrac12(1+ha_0)|w_h|^2_{\Gamma_+} + a_0\|w_h\|^2 + h\,\|a\cdot\nabla w_h\|^2 = (f,w_h) + h(f, a\cdot\nabla w_h). \tag{13.22}$$

As before this implies uniqueness and existence of solutions to (13.20) by setting $f=0$. Using the Cauchy-Schwarz inequality in the obvious way and $ha_0>0$ this yields the stability estimate

$$|w_h|^2_{\Gamma_+} + a_0\|w_h\|^2 + h\|a\cdot\nabla w_h\|^2 \le (a_0^{-1}+h)\|f\|^2. \tag{13.23}$$

Note the extra stability given by the presence of the term $h\|a\cdot\nabla w_h\|^2$. We may interpret this by saying that the method adds artificial diffusion, but only along the characteristic curves (streamlines).

For the error analysis we shall also need a somewhat stronger stability estimate for the case that f has the form $f = a\cdot\nabla F$, which reads

$$|w_h|^2_{\Gamma_+} + a_0\|w_h\|^2 + h\|a\cdot\nabla w_h\|^2 \le C(h\|F\|_1^2 + h^{-1}\|F\|^2), \tag{13.24}$$

where C is independent of a_0. Starting again with (13.22) we have now

$$h|(f, a\cdot\nabla w_h)| = h|(a\cdot\nabla F, a\cdot\nabla w_h)| \le \tfrac14 h\|a\cdot\nabla w_h\|^2 + Ch\|F\|_1^2.$$

Moreover, by Green's formula,

$$(f,w_h) = (a\cdot\nabla F, w_h) = (a\cdot n\,F, w_h)_{\Gamma_+} - (F, a\cdot\nabla w_h),$$

and hence

$$|(f,w_h)| \le C|F|^2_{\Gamma_+} + \tfrac14|w_h|^2_{\Gamma_+} + h^{-1}\|F\|^2 + \tfrac14 h\|a\cdot\nabla w_h\|^2.$$

Using the trace inequality

$$|F|^2_{\Gamma_+} \le C\|F\|\,\|F\|_1 \le Ch\|F\|_1^2 + Ch^{-1}\|F\|^2, \tag{13.25}$$

cf. Problem A.16, the proof of (13.24) is completed as in (13.22).

We are now ready for the following error estimate which shows an improvement of half a power of h compared to the standard Galerkin method. We also remark that the error in the flux is of optimal order, $\|a\cdot\nabla e\| = O(h)$.

Theorem 13.6. *Let u_h and u be the solutions of* (13.20) *and* (13.12). *Then we have for $e = u_h - u$,*

$$|e|_{\Gamma_+} + a_0^{1/2}\|e\| + h^{1/2}\|a \cdot \nabla e\| \le Ch^{3/2}\|u\|_2.$$

Proof. We write again $u_h - u = (u_h - I_h u) + (I_h u - u) = \theta + \rho$ and have using (13.17) and (13.25)

$$|\rho|^2_{\Gamma_+} + a_0\|\rho\|^2 + h\|a \cdot \nabla\rho\|^2 \le C\big(h\|\rho\|_1^2 + h^{-1}\|\rho\|^2\big) \le Ch^3\|u\|_2^2.$$

This time we have by (13.20) and (13.21),

$$(a \cdot \nabla\theta + a_0\theta, \chi + h\, a \cdot \nabla\chi) = -(a \cdot \nabla\rho + a_0\rho, \chi + h\, a \cdot \nabla\chi), \quad \forall \chi \in S_h^-.$$

Choosing $\chi = \theta \in S_h^-$ we conclude from (13.23) with $f = a_0\rho$ and (13.24) with $F = \rho$, together with (13.17), with the last C depending only on an upper bound for a_0,

$$|\theta|^2_{\Gamma_+} + a_0\|\theta\|^2 + h\|a \cdot \nabla\theta\|^2 \le C\big(a_0\|\rho\|^2 + h\|\rho\|_1^2 + h^{-1}\|\rho\|^2\big) \le Ch^3\|u\|_2^2,$$

which completes the proof. □

The error estimate of the previous theorem thus shows that streamline diffusion performs slightly better than the standard Galerkin method for smooth solutions, but the main reason for its use is that it performs better for non-smooth solutions, due to the fact that artificial diffusion is added only in the characteristic direction so that internal layers are not smeared out, while the added diffusion removes oscillations near boundary layers. We shall not go into the details.

The result of Theorem 13.6 is valid also when $a_0 = 0$. In this case the problem (13.12) still admits a unique solution, since the uniqueness, and hence the existence, again follows directly from (13.22) with $f = 0$. We have assumed $a_0 > 0$ in order to be able to bound the error to order $O(h^{3/2})$ in the L_2-norm. One may easily show the Poincaré type inequality $\|w\| \le C\|a \cdot \nabla w\|$ for $w = 0$ on Γ_-, and hence, in the absence of an estimate for $\|e\|$ we then have to be content with a $O(h)$ error bound in L_2-norm. We still have a $O(h^{3/2})$ error bound on Γ_+, and in our next result we will solve the problem in a sequence of domains in such a way that the bounds on the corresponding Γ_+ will result in a global $O(h^{3/2})$ error bound.

The above approach thus treats the first order hyperbolic problem as a two-dimensional one, and, in particular, solves the discrete equations simultaneously for all the nodal values of the solution. Applied to the initial-boundary value problem (13.6) the evolution aspect is therefore lost. We shall now turn to a modification that retains the advantages of the streamline diffusion method, but recovers the time stepping character. This will be done by dividing the domain $\Omega \times \mathbf{R}_+$ into strips parallel to the x-axis, and then using

approximating functions which are allowed to be discontinuous when passing from one strip to the next. This method is referred to as the *discontinuous Galerkin method*.

Considering thus the initial-boundary value problem (13.6), we use as earlier the partition of Ω defined by $0 = x_0 < x_1 < \cdots < x_M = 1$ and introduce now also a partition $0 = t_0 < t_1 < \dots$ of $\mathbf{R}_+$. We assume, for simplicity, that both partitions are quasi-uniform, and that the increments in space and time are of the same magnitude. Setting $h_j = x_j - x_{j-1}$, $k_j = t_j - t_{j-1}$, and $h = \max h_j$, $k = \max k_j$, this means that $ch \le h_j \le h$, $ck \le k_j \le k$, for all j, with $c > 0$, and that $ch \le k \le Ch$. These partitions in space and time define a partition of $Q = \Omega \times \mathbf{R}_+$ into rectangles. These could be further subdivided into triangles by means of diagonals with positive slopes, say, to form a triangulation which would permit application of our above discussion of the streamline diffusion method on any finite interval in time. We shall use the undivided rectangles in our analysis, and define, with $M_n = (t_{n-1}, t_n)$,

$$S_{h,k} = \Big\{ V(x,t) = \alpha^n(x)\frac{t - t_{n-1}}{k_n} + \beta^n(x)\frac{t_n - t}{k_n}, \\ \text{for } t \in M_n, \quad \text{with } \alpha^n, \beta^n \in S_h^- \Big\},$$

where S_h^- denotes the piecewise linear space defined in (13.7). Note that $V \in S_{h,k}$ may be discontinuous at t_n and that $V_{n-1}^+ = V(t_{n-1}^+) = \beta^n$, $V_n^- = V(t_n^-) = \alpha^n$, and $V_t(t) = (\alpha^n - \beta^n)/k_n$ for $t \in M_n$.

The discontinuous Galerkin method with streamline diffusion for the solution of (13.6) is to find $U \in S_{h,k}$ such that $U_0^- = v_h$ and then, for $n = 1, 2, \dots$,

$$\begin{aligned} &\int_{M_n} (U_t + U_x, \chi + h(\chi_t + \chi_x))\,\mathrm{d}t + (U_{n-1}^+, \chi_{n-1}^+) \\ &\qquad = \int_{M_n} (f, \chi + h(\chi_t + \chi_x))\,\mathrm{d}t + (U_{n-1}^-, \chi_{n-1}^+), \quad \forall \chi \in S_{h,k}, \end{aligned} \tag{13.26}$$

where the inner products are over the one-dimensional interval Ω. We note that if f and U_{n-1}^- vanish, then we may choose $\chi = U$ to conclude easily that $U = 0$ on $\Omega \times M_n$. Hence this equation can be solved for U_n^- and U_{n-1}^+, if U_{n-1}^- is given together with f on $\Omega \times M_n$, and the method is therefore a time stepping procedure.

We remark that the local equation (13.26) is of the form (13.20) on the domain $\Omega \times M_n$, with the boundary condition $U = 0$ on $\Gamma_{-,x}$, but with the boundary condition $U_{n-1}^+ = U_{n-1}^-$ on $\Gamma_{-,t}$ only weakly imposed, see Problem 13.7.

By adding the equations in (13.26) and the initial condition $(U_0^- - v_h, \chi_0^+) = 0$ we can write the equations on Q in weak form as

$$B_n(U, \chi) = L_n(v_h, f; \chi), \quad \forall \chi \in S_{h,k}, \quad \text{for } n \ge 1,$$

where, with $[v]_j = v_j^+ - v_j^-$,

$$B_n(v,w) = \sum_{j=1}^{n} \int_{M_j} (v_t + v_x, w + h(w_t + w_x))\,\mathrm{d}t + \sum_{j=1}^{n-1} ([v]_j, w_j^+) + (v_0^+, w_0^+)$$

and

$$L_n(v,f;w) = (v, w_0^+) + \sum_{j=1}^{n} \int_{M_j} (f, w + h(w_t + w_x))\,\mathrm{d}t. \tag{13.27}$$

We note that since the exact solution is continuous in time, and thus the jump terms vanish, it satisfies

$$B_n(u,\chi) = L_n(v,f;\chi), \quad \forall \chi \in S_{h,k}, \quad \text{for } n \geq 1. \tag{13.28}$$

By integration by parts we can write $B_n(\cdot,\cdot)$ as

$$\begin{aligned} B_n(v,w) = &\sum_{j=1}^{n} \int_{M_j} \Big((v, -w_t - w_x) + h(v_t + v_x, w_t + w_x) \Big)\,\mathrm{d}t \\ &+ \sum_{j=1}^{n-1} (v_j^-, -[w]_j) + (v_n^-, w_n^-) + \int_0^{t_n} v(1,t)w(1,t)\,\mathrm{d}t. \end{aligned} \tag{13.29}$$

By adding the two forms of $B_n(\cdot,\cdot)$, using $v_j^- = v_j^+ - [v]_j$, we obtain for $w = v$

$$\begin{aligned} B_n(v,v) = &\tfrac{1}{2}\|v_n^-\|^2 + \tfrac{1}{2}\sum_{j=1}^{n-1} \|[v]_j\|^2 + h\sum_{j=1}^{n} \int_{M_j} \|v_t + v_x\|^2\,\mathrm{d}t \\ &+ \tfrac{1}{2}\|v_0^+\|^2 + \tfrac{1}{2}\int_0^{t_n} v(1,t)^2\,\mathrm{d}t. \end{aligned} \tag{13.30}$$

We now turn to the error analysis.

Theorem 13.7. *Let U and u be the solutions of* (13.26) *and* (13.6)*. Then, for v_h is suitably chosen, we have for $e = U - u$,*

$$\begin{aligned} &\|e_n^-\|^2 + \sum_{j=1}^{n-1} \|[e]_j\|^2 + h\sum_{j=1}^{n} \int_{M_j} \|e_t + e_x\|^2\,\mathrm{d}t \\ &\qquad \leq Ch^3 \int_0^{t_n} \left(\|u\|_2^2 + \|u_t\|_1^2 + \|u_{tt}\|^2 \right) \mathrm{d}t, \quad \textit{for } n \geq 0. \end{aligned} \tag{13.31}$$

We remark that the first term on the left shows a $O(h^{3/2})$ estimate for the error to the left at t_n. Since the jumps at the time levels are bounded in the second term, the error to the right at t_n, $e_n^+ = e_n^- + [e]_n$, is also of order $O(h^{3/2})$, and we may conclude that this holds everywhere on M_n, since $e(t) = k^{-1}(t - t_{n-1})e_n^- + k^{-1}(t_n - t)e_{n-1}^+ + (\bar{u}(t) - u(t))$, where $\bar{u}$ denotes the linear interpolant of u so that the last term is $O(h^2)$.

Proof of Theorem 13.7. The proof proceeds in a way similar to that of Theorem 13.6 and all terms that occur below have their counterparts there. Let I_h denote the interpolation operator into S_h^- and J_k the piecewise linear interpolation operator in time. We write

$$U - u = (U - \widetilde{u}) + (\widetilde{u} - u) = \theta + \rho, \quad \text{where } \widetilde{u} = J_k I_h u.$$

Note that $\widetilde{u}(\cdot, t)$ is continuous in time. In view of (13.30) it suffices to bound $B_n(e, e)$, and we note that

$$B_n(e, e) \le 2B_n(\theta, \theta) + 2B_n(\rho, \rho).$$

We begin by bounding the first term on the right. We find in view of (13.28) and (13.26), for any $\chi \in S_{h,k}$,

$$\begin{aligned} B_n(\theta, \chi) &= B_n(U, \chi) - B_n(\widetilde{u}, \chi) = L_n(v_h, f; \chi) - B_n(\widetilde{u}, \chi) \\ &= L_n(v_h, f; \chi) + \big(B_n(u, \chi) - L_n(v, f; \chi)\big) - B_n(\widetilde{u}, \chi) \\ &= (v_h - v, \chi_0^+) - B_n(\rho, \chi) = (e_0, \chi_0^+) - B_n(\rho, \chi) = -B_n(\rho, \chi), \end{aligned}$$

where we have now chosen the initial value $v_h = P_h v$ so that $(e_0, \chi_0^+) = 0$. Setting $\chi = \theta$ and using (13.29) for $B_n(\cdot, \cdot)$, we conclude

$$\begin{aligned} B_n(\theta, \theta) = |B_n(\rho, \theta)| &\le \sum_{j=1}^{n} \int_{M_j} \big(\|\rho\| + h\|\rho_t + \rho_x\|\big)\, \|\theta_t + \theta_x\| \, \mathrm{d}t \\ &\quad + \sum_{j=1}^{n-1} \|\rho_j\|\, \|[\theta]_j\| + \|\rho_n\|\, \|\theta_n^-\| + \int_0^{t_n} |\rho(1,t)|\, |\theta(1,t)| \, \mathrm{d}t \\ &\le \tfrac{1}{2} B_n(\theta, \theta) + Ch^{-1} \int_0^{t_n} \|\rho\|^2 \, \mathrm{d}t + \sum_{j=1}^{n} \|\rho_j\|^2 + CB_n(\rho, \rho). \end{aligned}$$

Here we noted that ρ is continuous in time so that $\rho_n^- = \rho_n$. To complete the proof we kick back $B_n(\theta, \theta)$ and then bound the last three terms appropriately. Note first that, by (13.30),

$$B_n(\rho, \rho) = \tfrac{1}{2}\|\rho_n\|^2 + h \int_0^{t_n} \|\rho_t + \rho_x\|^2 \, \mathrm{d}t + \tfrac{1}{2}\|\rho_0\|^2 + \tfrac{1}{2} \int_0^{t_n} \rho(1,t)^2 \, \mathrm{d}t.$$

We write

$$\rho = J_k I_h u - u = J_k(I_h u - u) + (J_k u - u) = J_k \eta + \omega.$$

Using the standard estimates for J_k, with $\|v\|_{M_j}^2 = \int_{M_j} \|v\|^2 \, \mathrm{d}t$,

$$\|J_k v - v\|_{M_j} + k_j \|D_t(J_k v - v)\|_{M_j} \le C k_j^s \|D_t^s v\|_{M_j}, \quad \text{for } s = 1, 2,$$

and since the partitions are quasi-uniform, with h and k of the same order of magnitude, we get

$$\begin{aligned}\sum_{j=1}^{n}\int_{M_j}\big(h^{-1}\|\omega\|^2+h\|\omega_t\|^2+h\|\omega_x\|^2\big)\,\mathrm{d}t&\\ \le C(h^{-1}k^4+hk^2)\int_0^{t_n}\|u_{tt}\|^2\,\mathrm{d}t+Chk^2\int_0^{t_n}\|u_t\|_1^2\,\mathrm{d}t&\\ \le Ch^3\int_0^{t_n}\big(\|u_{tt}\|^2+\|u_t\|_1^2\big)\,\mathrm{d}t.&\end{aligned}$$

Next we note that $\|J_k\eta(t)\| \le \max_{s\in M_j}\|\eta(s)\|$ for $t\in M_j$, and use a trace inequality from Problem A.12, to see that

$$\begin{aligned}\|\eta(t)\|^2 &\le Ck_j^{-1}\int_{M_j}\|\eta\|^2\,\mathrm{d}t+Ck_j\int_{M_j}\|\eta_t\|^2\,\mathrm{d}t\\ &\le Ch^3\int_{M_j}\big(\|u\|_2^2+\|u_t\|_1^2\big)\,\mathrm{d}t,\quad \text{for } t\in M_j,\end{aligned}\tag{13.32}$$

so that

$$h^{-1}\int_0^{t_n}\|J_k\eta\|^2\,\mathrm{d}t\le Ch^3\int_0^{t_n}\big(\|u\|_2^2+\|u_t\|_1^2\big)\,\mathrm{d}t.$$

In a similar way, using

$$\|(J_k\eta)_t(t)\|=k_j^{-1}\|\eta_j-\eta_{j-1}\|\le 2k_j^{-1}\max_{s\in M_j}\|\eta(s)\|,\quad \text{for } t\in M_j,$$

we get

$$h\sum_{j=1}^{n}\int_{M_j}\big(\|(J_k\eta)_x\|^2+\|(J_k\eta)_t\|^2\big)\,\mathrm{d}t\le Ch^3\int_0^{t_n}(\|u\|_2^2+\|u_t\|_1^2)\,\mathrm{d}t.$$

Further, from (13.32),

$$\sum_{j=0}^{n}\|\rho_j\|^2=\sum_{j=0}^{n}\|\eta_j\|^2\le Ch^3\int_0^{t_n}\big(\|u\|_2^2+\|u_t\|_1^2\big)\,\mathrm{d}t.$$

Finally, using again a trace inequality from Problem A.12, we get

$$\begin{aligned}\int_0^{t_n}|\rho(1,t)|^2\,\mathrm{d}t=\int_0^{t_n}|\omega(1,t)|^2\,\mathrm{d}t&\le C\int_0^{t_n}\|\omega(\cdot,t)\|\,\|\omega(\cdot,t)\|_1\,\mathrm{d}t\\ &\le Ck^3\Big(\int_0^{t_n}\|u_{tt}\|^2\,\mathrm{d}t\int_0^{t_n}\|u_t\|_1^2\,\mathrm{d}t\Big)^{1/2}\\ &\le Ch^3\int_{M_j}\big(\|u_{tt}\|^2+\|u_t\|_1^2\big)\,\mathrm{d}t.\end{aligned}$$

This completes the proof. □

13.3 Problems

Problem 13.1. Consider the initial-boundary value problem

$$\begin{aligned} &u_{tt} = u_{xx}, && x \in (0,1),\ t > 0,\\ &u(0,t) = u(1,t) = 0, && t > 0,\\ &u(x,0) = v(x),\ u_t(x,0) = w(x), && x \in (0,1). \end{aligned}$$

For the numerical solution by the Galerkin finite element method, consider the piecewise linear continuous functions based on the partition of $[0,1]$ into M intervals of equal lengths $h = 1/M$. Find the matrix forms of the semidiscrete and completely discrete methods analogous to those described above in (13.2) and (13.5).

Problem 13.2. (Computer exercise.) Solve the initial-boundary value problem in Problem 13.1 with $v(x) = 0,\ w(x) = \sin(2\pi x)$ using $M = 10,\ 20$ and the time stepping method in (13.5) with $k = 1/10,\ 1/20$, and compare with the exact solution given in Sect. 11.2 at time $t = 3/4$.

Problem 13.3. Write the wave equation (13.1) as a system of two equations of first order in time by setting $w_1 = u,\ w_2 = u_t$. Discretize this system by means of the standard finite element method in the spatial variables and by means of the Crank-Nicolson method in the time variable. Show, by elimination of W_2^n, that the resulting scheme is essentially the same as (13.5). Prove stability in the case $f = 0$. Compare with Lemma 13.2. Hint: Multiply the system by $(W_1^{n-\frac{1}{2}}, -\Delta_h W_2^{n-\frac{1}{2}})$.

Problem 13.4. Prove Theorem 13.2.

Problem 13.5. Show the stability bound (13.11).

Problem 13.6. Prove stability and error estimates for the standard Galerkin method with artificial diffusion (13.19).

Problem 13.7. (Weakly imposed boundary condition.) The boundary condition $u = g$ is imposed strongly in (13.14) and (13.20). It is also possible to impose the boundary condition weakly in the standard Galerkin method: Find $u_h \in S_h$ such that

$$(a \cdot \nabla u_h, \chi) + (a_0 u_h, \chi) - (a \cdot n\, u_h, \chi)_{\Gamma_-} = (f, \chi) - (a \cdot n\, g, \chi)_{\Gamma_-}, \quad \forall \chi \in S_h,$$

and in its streamline diffusion modification: Find $u_h \in S_h$ such that

$$\begin{aligned} (a \cdot \nabla u_h, \chi + ha \cdot \nabla \chi) + (a_0 u_h, \chi + ha \cdot \nabla \chi) - (a \cdot n\, u_h, \chi)_{\Gamma_-} \\ = (f, \chi + ha \cdot \nabla \chi) - (a \cdot n\, g, \chi)_{\Gamma_-}, \quad \forall \chi \in S_h. \end{aligned}$$

Prove stability and error estimates for these methods.

14 Some Other Classes of Numerical Methods

Numerical methods other than finite difference and finite element methods, but often closely related to these, have been developed and are also of interest. In this chapter we review briefly four such classes of methods, namely *collocation methods, spectral methods, finite volume methods*, and *boundary element methods*.

14.1 Collocation methods

In a *collocation method* one seeks an approximate solution of a differential equation in a finite dimensional space of sufficiently regular functions by requiring that the equation is satisfied exactly at a finite number of points. We describe one such method for the parabolic model problem

$$\begin{aligned} &u_t = u_{xx} && \text{in } \Omega = (0,1), \quad \text{for } t > 0, \\ &u(0,t) = u(1,t) = 0 && \text{for } t > 0, \\ &u(\cdot,0) = v && \text{in } \Omega. \end{aligned}$$

Setting $h = 1/M,\ x_j = jh$ for $0 \le j \le M$, and $K_j = [x_{j-1}, x_j]$, we introduce the piecewise polynomial space

$$S_h = \big\{ v \in \mathcal{C}^1(\bar{\Omega}) : v\big|_{K_j} \in \Pi_{r-1},\ v(0) = v(1) = 0 \big\}, \quad \text{with } r \ge 4.$$

Letting $\xi_i,\ i = 1, \ldots, r-2$, be the Gauss points in $(0,1)$, i.e., the zeros of the Legendre polynomial $\tilde{P}_{r-2}(x) = P_{r-2}(2x-1)$, re-scaled from the interval $(-1,1)$ to $(0,1)$, we define the collocation points $x_{j,i} = x_{j-1} + h\xi_i$ in K_j, and pose the spatially semidiscrete problem to find $u_h(\cdot,t) \in S_h$ for $t > 0$ such that

$$u_{h,t}(x_{j,i},t) = u_{h,xx}(x_{j,i},t), \quad \text{for } 1 \le j \le M,\ 1 \le i \le r-2,\ t > 0, \tag{14.1}$$

with $u_h(\cdot,0) = v_h \in S_h$ an approximation of v. This method may be considered as a Galerkin method using a discrete inner product based on the Gauss quadrature rule. In fact, letting ω_i be the weights in the Gauss formula

$$\sum_{i=1}^{r-2} \omega_i\, \varphi(\xi_i) \approx \int_0^1 \varphi(x)\, \mathrm{d}x,$$

which is exact for polynomials of degree at most $2r-5$, we set

$$(\psi,\chi)_h = h\sum_{j=1}^{M}\sum_{i=1}^{r-2}\omega_i\,\psi(x_{j,i})\,\chi(x_{j,i}) \approx (\psi,\chi), \tag{14.2}$$

and we may then write (14.1) as

$$(u_{h,t},\chi)_h - (u_{h,xx},\chi)_h = 0, \quad \forall\chi\in S_h,\ t>0.$$

For v_h appropriately chosen one may show the global error estimate

$$\|u_h(t)-u(t)\|_{\mathcal{C}} \le Ch^r\Big\{\max_{s\le t}\|u(s)\|_{r+2} + \Big(\int_0^t \|u_t(s)\|_{r+2}^2\,\mathrm{d}s\Big)^{1/2}\Big\}.$$

Further, for $r>4$, and with a more refined choice of initial approximation v_h, superconvergence takes place at the nodes, so that

$$|u_h(x_j,t)-u(x_j,t)| \le C_T h^{2r-4}\sup_{s\le t}\sum_{p+2q\le 2r-1}\|u^{(q)}(s)\|_p, \quad \text{for } t\le T.$$

We note the more stringent regularity requirements than for the finite difference and finite element methods discussed in Chaps. 9–10. These results carry over to fully discrete schemes using both finite difference approximations and collocation in time.

14.2 Spectral Methods

Spectral methods are in many ways similar to finite element and collocation methods. The main difference is in the choice of finite dimensional approximating spaces.

Consider the initial value problem

$$\begin{aligned} &u_t - u_{xx} = f && \text{in } \Omega=(0,1), \quad \text{for } t>0,\\ &u(0,t)=u(1,t)=0 && \text{for } t>0,\\ &u(\cdot,0)=v && \text{in } \Omega. \end{aligned} \tag{14.3}$$

Let now $\{\varphi_j\}_{j=1}^{\infty}$ be a sequence of linearly independent functions in $H^2\cap H_0^1$, which span L_2, and set $S_N = \operatorname{span}\{\varphi_j\}_{j=1}^N$. Using Galerkin's method we define a spatially semidiscrete approximation $u_N = u_N(t)\in S_N$ of (14.3) by

$$(u_{N,t},\chi) - (u_{N,xx},\chi) = (f,\chi), \quad \forall\chi\in S_N,\ t>0, \tag{14.4}$$

with $u_N(0)=v_N\in S_N$ suitably chosen. Introducing the orthogonal projection $P_N : L_2\to S_N$, we may write (14.4) as

$$u_{N,t} + \mathcal{A}_N u_N = P_N f, \quad \text{for } t>0, \quad \text{where } \mathcal{A}_N = P_N\mathcal{A}P_N,\ \mathcal{A}=\partial^2/\partial x^2.$$

Here $(\mathcal{A}_N\chi,\chi) = (\mathcal{A}P_N\chi, P_N\chi) \geq 0$. With $u_N(x,t) = \sum_{j=1}^N \alpha_j(t)\varphi_j(x)$, this equation may be written $B\alpha'(t) + A\alpha(t) = b(t)$ for $t > 0$, where the elements of the N by N matrices A and B are $(\mathcal{A}\varphi_i,\varphi_j)$ and (φ_i,φ_j), respectively. It is easy to see that B is positive definite.

We note that the error $e_N = u_N - u$ satisfies

$$e_{N,t} + \mathcal{A}_N e_N = (P_N - I)f - (\mathcal{A}_N - \mathcal{A})u \quad \text{for } t > 0, \quad \text{with } e_N(0) = v_N - v,$$

and hence, since the corresponding solution operator $E_N(t) = e^{-\mathcal{A}_N t}$ is easily seen to be bounded by 1 in L_2 operator norm,

$$\|u_N(t) - u(t)\| \leq \|v_N - v\| + \int_0^t \big(\|(P_N - I)f\| + \|(\mathcal{A}_N - \mathcal{A})u\|\big)\,\mathrm{d}s. \tag{14.5}$$

It follows that the error is small with $v_N - v$, $(P_N - I)f$, and $(\mathcal{A}_N - \mathcal{A})u$.

As a simple example, let the $\varphi_j(x) = c\sin(j\pi x)$ be the normalized eigenfunctions of $\mathcal{A}$, with homogeneous Dirichlet boundary conditions. Then $B = I$, A is positive definite and P_N is simply the truncation of the Fourier series, $P_N v = \sum_{j=1}^N (v,\varphi_j)\varphi_j$, so that $\mathcal{A}_N v = \sum_{j=1}^N (j\pi)^2(v,\varphi_j)\varphi_j = P_N\mathcal{A}v$. Thus, if $v_N = P_N v$ and if the Fourier series of v, f, and u_{xx} converge, then the error is small. In particular, the convergence is of order $O(N^{-r})$ for any r, provided that the solution is sufficiently regular.

Another way to define a semidiscrete numerical method employing the space S_N of our example is to make S_N a Hilbert space with the inner product $(v,w)_N = h\sum_{j=0}^{N-1} v(x_j)w(x_j)$, where $x_j = j/(N-1)$. This gives rise to a projection P_N defined by $P_N u(x_j) = u(x_j)$, $j = 0,\ldots,N-1$, and the semidiscrete equation (14.4) now becomes the collocation equation

$$u_{N,t}(x_j,t) - u_{N,xx}(x_j,t) = f(x_j,t), \quad \text{for } j = 0,\ldots,N-1,\ t > 0.$$

This is also referred to as a *pseudospectral method* and the error estimate (14.5) is valid in the discrete norm corresponding to $(\cdot,\cdot)_N$.

Spectral and pseudospectral methods using the above sinusoidal basis functions are particularly useful for periodic problems. For initial-boundary value problems for hyperbolic equations basis functions related to Chebyshev and Legendre polynomials are sometimes useful, e.g., in connection with fluid dynamics calculations.

14.3 Finite Volume Methods

We illustrate the use of the *finite volume methods* in the case of the model problem

$$-\Delta u = f \quad \text{in } \Omega, \quad \text{with } u = 0 \quad \text{on } \Gamma, \tag{14.6}$$

where Ω is a convex polygonal domain in $\mathbf{R}^2$ with boundary Γ. The basis for this approach is the observation that, by Green's formula, for any $V \subset \Omega$ we have

$$\int_{\partial V} \frac{\partial u}{\partial n}\, \mathrm{d}s = \int_V f\, \mathrm{d}x. \tag{14.7}$$

We begin by describing the *cell centered* finite volume difference method. Let $\mathcal{T}_h = \{K_j\}$ be a triangulation of Ω of the type considered in Sect. 5 in which all angles of the K_j are $< \pi/2$, and consider (14.7) with $V = K_j \in \mathcal{T}_h$. Then $\partial V = \partial K_j$ is the union of the edges γ_{ji} common with three other triangles K_i, and we want to approximate $\partial u/\partial n$ on each of these edges. With Q_j the center of the circumscribed circle of K_j (which then lies in the interior of K_j), the vector Q_jQ_j is orthogonal to γ_{ji}, and $\partial u/\partial n$ in (14.7) may be approximated by the difference quotient $(U(Q_i) - U(Q_j))/|Q_iQ_j|$. Using the boundary values in (14.6) for the Q_j associated with the boundary triangles, this produces a finite difference scheme on the nonuniform mesh $\{Q_j\}$. Writing the discrete problem in matrix form as $AU = b$, one may show that the matrix A is symmetric positive definite and diagonally dominant. When the $\mathcal{T}_h$ are quasi-uniform one may show the error estimate

$$\|U - u\|_{1,h} \leq Ch\|u\|_2$$

in a certain discrete H^1-norm.

An alternative approach is the following *vertex centered method*, also referred to as the *finite volume element method*: Let $S_h \subset H_0^1$ be the piecewise linear finite element space defined by $\mathcal{T}_h$. For $K \in \mathcal{T}_h$ the straight lines connecting a vertex with the midpoint of the opposite edge intersect at the barycenter of K and divide K into six triangles. Let $B_{j,K}$ be the union of the two of these which have P_j as a vertex. For each interior vertex P_j we consider the union B_j of the corresponding $B_{j,K}$, and let $\bar{S}_h$ denote the associated piecewise constant functions. Using (14.7) for each of the B_j we are lead to the Petrov-Galerkin method to find $u_h \in S_h$ such that

$$\bar{a}(u_h, \psi) := \sum_j \psi_j \int_{\partial B_j} \frac{\partial u_h}{\partial n}\, \mathrm{d}s = (f, \psi) \quad \forall \psi \in \bar{S}_h, \tag{14.8}$$

which may also be thought of as a finite difference scheme on the irregular mesh $\{P_j\}$. The B_j are referred to as control volumes. Associating with $\chi \in S_h$ the function $\bar{\chi} \in \bar{S}_h$, which agrees with χ at the vertices of $\mathcal{T}_h$, one may show that (cf. Problem 14.3)

$$\bar{a}(\psi, \bar{\chi}) = a(\psi, \chi), \quad \forall \psi, \chi \in S_h, \tag{14.9}$$

so that (14.8) may be written

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

(This does not hold exactly for elliptic operators with variable coefficients.) It may be shown that the standard error estimate

$$\|u_h - u\|_1 \leq Ch\|u\|_2$$

holds for this method, and also, under slightly more stringent regularity assumptions, that $\|u_h - u\| = O(h^2)$.

Finite volume methods are useful for operators in divergence form, particularly for time dependent conservation laws.

14.4 Boundary Element Methods

In a *boundary integral method* a boundary value problem for a homogeneous partial differential equation in a domain Ω with the solution u given on the boundary Γ is reformulated as an integral equation over Γ. This equation may then be used as a basis for numerical approximation. We shall illustrate this approach for the model problem

$$\Delta u = 0 \quad \text{in } \Omega \subset \mathbf{R}^2, \quad \text{with } u = g \quad \text{on } \Gamma, \tag{14.10}$$

where we assume Γ smooth. To pose the boundary integral equation, let $U(x) = -(2\pi)^{-1} \log|x|$ be the fundamental solution of the Laplacian in $\mathbf{R}^2$, see Theorem 3.5. For any u with $\Delta u = 0$ on Γ we have by Green's formula

$$u(x) = \int_\Gamma U(x-y) \frac{\partial u}{\partial n_y}(y)\, \mathrm{d}s_y - \int_\Gamma \frac{\partial U}{\partial n_y}(x-y) u(y)\, \mathrm{d}s_y, \quad x \in \Omega. \tag{14.11}$$

With x on Γ the integrals on the right define the single and double layer potentials $V\partial u/\partial n$ and Wu. We note that although $\nabla U(x-y)$ has a singularity of order $O(|x-y|^{-1})$, the kernel $(\partial U/\partial n_y)(x-y)$ is bounded for $x, y \in \Gamma$, so that the operator W is well defined. For $x \in \Omega$ approaching Γ the two integrals tend to $V\partial u/\partial n$ and $\frac{1}{2}u + Wu$, respectively, so that (14.11) yields

$$\tfrac{1}{2}u = V\partial u/\partial n - Wu \quad \text{on } \Gamma.$$

With $u = g$ on Γ this is a Fredholm integral equation of the first kind for the determination of $\partial u/\partial n$ on Γ, which inserted into (14.11) together with $u = g$ on Γ gives the solution of (14.10).

Instead of this direct method one may use the indirect method of assuming that the solution of (14.11) may be represented as a potential of a function on Γ, so that

$$u(x) = \int_\Gamma \Phi(x-y) v(y)\, \mathrm{d}s_y \text{ or } u(x) = \int_\Gamma \frac{\partial \Phi}{\partial n_y}(x-y) w(y)\, \mathrm{d}s_y, \quad x \in \Omega.$$

With V and W as above, if such functions v and w exist, they satisfy the first and second kind Fredholm integral equations

$$Vv = g \quad \text{and} \quad \tfrac{1}{2}w + Ww = g \quad \text{on } \Gamma. \tag{14.12}$$

Writing $H^s = H^s(\Gamma)$, V and W are so-called pseudodifferential operators of order -1, i.e., bounded linear operators $H^s \to H^{s+1}$, in particular compact on H^s. The first kind equation is uniquely solvable provided a certain measure, the transfinite diameter δ_Γ of Γ, is such that $\delta_\Gamma \neq 1$, and the second kind equation in (14.12) always has a unique solution. Similar reformulations may be used also for Neumann boundary conditions, for a large number of other problems involving elliptic type equations, and for exterior problems; in fact, this approach to the numerical solution is particularly useful in the latter case.

In the Boundary Element Method (BEM) one determines the approximate solution in a piecewise polynomial finite element type space of a boundary integral formulation such as the above, using the Galerkin or the collocation method.

For the second kind equation in (14.12), using Galerkin's method and a finite dimensional subspace S_h of $L_2(\Gamma)$, we determine the discrete approximation $w_h \in S_h$ to w from

$$\tfrac{1}{2}\langle w_h, \chi\rangle + \langle Ww_h, \chi\rangle = \langle g, \chi\rangle, \quad \forall \chi \in S_h, \quad \text{where } \langle\cdot,\cdot\rangle = (\cdot,\cdot)_{L_2(\Gamma)}.$$

Writing $|\cdot|_s$ for the norm in $H^s(\Gamma)$, one has $|w_h - w|_0 \leq C_r(w)h^r$ if S_h is accurate of order $O(h^r)$, and by a duality argument one may show the superconvergent order negative norm estimate $|w_h - w|_{-r} \leq C_r(w)h^{2r}$; using an iteration argument this may be used to define an approximate solution $\tilde{w}_h$ with $|\tilde{w}_h - w|_0 = O(h^{2r})$.

Consider for example the numerical solution of the first kind equation in (14.12) in the finite dimensional space S_h of periodic smoothest splines of order r, i.e., $S_h \subset \mathcal{C}^{r-2}$ consists of piecewise polynomials in Π_{r-1}. Our discrete problem is then to find $v_h \in S_h$ such that

$$\langle Vv_h, \chi\rangle = \langle g, \chi\rangle, \quad \forall \chi \in S_h.$$

It can be shown that the bilinear form $\langle Vv, w\rangle$ associated with V is symmetric, bounded, and coercive with respect to the norm $|\cdot|_{-1/2}$ in a certain Sobolev space $H^{-1/2}(\Gamma)$, so that

$$\langle Vv, w\rangle = \langle v, Vw\rangle \leq C|v|_{-1/2}|w|_{-1/2} \quad \text{and} \quad \langle Vv, v\rangle \geq c|v|^2_{-1/2}, \quad \text{with } c > 0.$$

One may then show that

$$|v_h - v|_{-1/2} \leq C \inf_{\chi \in S_h} |\chi - v|_{-1/2} \leq Ch^{r+1/2}|v|_r,$$

and a duality argument implies $|v_h - v|_{-r-1} \leq Ch^{2r+1}|v|_r$, where we use the norm in $H^{-r-1}(\Gamma)$. For x an interior point of Ω we therefore find for $u_h = Vv_h$ that $|u_h(x) - u(x)| \leq C_x|v_h - v|_{-r-1} \leq Ch^{2r+1}$, since $\Phi(x-y)$ is smooth when $y \neq x$.

Expressed in terms of a basis $\{\phi_j\}$ of S_h this problem may be written in matrix form as $A\alpha = \tilde{g}$, where A is symmetric positive definite. However, although the dimension of A has been reduced by the reduction of the original two-dimensional problem to a one-dimensional one, in contrast to the finite element method for a differential equation problem, the matrix A is now not sparse. We also note that the elements $\langle V\Phi_i, \Phi_j\rangle$ require two integrations, one in forming $V\Phi_i$ and one in forming the inner product.

In order to reduce this work one may apply the collocation method and determine v_h from $Vv_h(x(s_j)) = g(x(s_j))$ at M_h quadrature points s_j in $[0, l]$, where $x = x(s)$ is a parametrization of Γ and $M_h = \dim(S_h)$.

In the vast literature on the numerical boundary integral methods much attention has been paid to the complications arising when our above regularity assumptions fail to be satisfied, such as for domains with corners in which case V and W are not compact.

14.5 Problems

Problem 14.1. Let $r = 4$ and let $(\cdot,\cdot)_h$ be defined by the corresponding case of (14.2).

(a) Show that $\|\chi\|_h := (\chi,\chi)_h^{1/2}$ is a norm on S_h.
(b) Show that

$$-(\chi'',\chi)_h \geq -(\chi'',\chi) = \|\chi'\|^2, \quad \text{for } \chi \in S_h.$$

(c) Show the stability of the solution of (14.1) with respect to $\|\cdot\|_h$.

Hint for (b): Let $\tilde{P}_2(x) = P_2(2x-1) = x^2 - x + \frac{1}{6}$ be the Legendre polynomial corresponding to $(0,1)$ with zeros $\xi_{1,2} = \frac{1}{2} \pm \frac{\sqrt{3}}{6}$. Recall that Gauss quadrature with two Gauss points is exact for cubic polynomials. Restrict the consideration to one interval $(0,1)$ and let $\chi \in \Pi_3$ with coefficient 1 for x^3. Then $\chi''\chi - 6\tilde{P}_2^2 \in \Pi_3$ and hence, since $\tilde{P}(\xi_i) = 0,\ i = 1, 2,$

$$-\tfrac{1}{2}\sum_{i=1}^{2} \chi''(\xi_i)\chi(\xi_i) = -\int_0^1 \chi''\chi \,\mathrm{d}x + 6\int_0^1 \tilde{P}_2^2 \,\mathrm{d}x \geq -\int_0^1 \chi''\chi \,\mathrm{d}x.$$

Problem 14.2. Consider the first order initial value problem with periodic boundary conditions

$$\begin{aligned} &u_t + u_x = 0, && \text{in } \Omega = (-\pi,\pi), \quad \text{for } t > 0,\\ &u(-\pi,t) = u(\pi,t), && \text{for } t > 0,\\ &u(\cdot,0) = v, && \text{in } \Omega. \end{aligned}$$

Formulate the spectral method based on

$$S_N = \{1, \cos x, \sin x, \cos 2x, \sin 2x, \ldots, \cos Nx, \sin Nx\}.$$

Let $\mathcal{A} = \partial/\partial x$, determine $\mathcal{A}_N$ and show that $\mathcal{A}_N^* = -\mathcal{A}_N$, where $*$ denotes the adjoint. Show also that $\|u_N(t)\| \leq \|v\| = \|v\|_{L_2(\Omega)}$, and hence that $\|E_N(t)\| = 1$, where $E_N(t) = \mathrm{e}^{-t\mathcal{A}_N}$.

Problem 14.3. Show (14.9). Hint: Write Ω as unions of the B_j and of the K, and write these in turn as unions of the sets $B_{j,K}$. Note that

$$\int_e \bar{\chi}\,\mathrm{d}s = \int_e \chi\,\mathrm{d}s, \quad \text{for } \chi \in S_h,$$

for any edge e of the triangulation $\mathcal{T}_h$.

A Some Tools from Mathematical Analysis

In this appendix we give a short survey of results, essentially without proofs, from mathematical, particularly functional, analysis which are needed in our treatment of partial differential equations. We begin in Sect. A.1 with a simple account of abstract linear spaces with emphasis on Hilbert space, including the Riesz representation theorem and its generalization to bilinear forms of Lax and Milgram. We continue in Sect. A.2 with function spaces, where after a discussion of the spaces $\mathcal{C}^k$, integrability, and the L_p-spaces, we turn to L_2-based Sobolev spaces, with the trace theorem and Poincaré's inequality. The final Sect. A.3 is concerned with the Fourier transform.

A.1 Abstract Linear Spaces

Let V be a linear space (or vector space) with real scalars, i.e., a set such that if $u, v \in V$ and $\lambda, \mu \in \mathbf{R}$, then $\lambda u + \mu v \in V$. A *linear functional* (or *linear form*) L on V is a function $L : V \to \mathbf{R}$ such that

$$L(\lambda u + \mu v) = \lambda L(u) + \mu L(v), \quad \forall u, v \in V,\ \lambda, \mu \in \mathbf{R}.$$

A *bilinear form* $a(\cdot, \cdot)$ on V is a function $a : V \times V \to \mathbf{R}$, which is linear in each argument separately, i.e., such that, for all $u, v, w \in V$ and $\lambda, \mu \in \mathbf{R}$,

$$\begin{aligned} a(\lambda u + \mu v, w) &= \lambda a(u, w) + \mu a(v, w), \\ a(w, \lambda u + \mu v) &= \lambda a(w, u) + \mu a(w, v). \end{aligned}$$

The bilinear form $a(\cdot, \cdot)$ is said to be *symmetric* if

$$a(w, v) = a(v, w), \quad \forall v, w \in V,$$

and *positive definite* if

$$a(v, v) > 0, \quad \forall v \in V,\ v \neq 0.$$

A positive definite, symmetric, bilinear form on V is also called an *inner product* (or *scalar product*) on V. A linear space V with an inner product is called an *inner product space*.

If V is an inner product space and $(\cdot,\cdot)$ is an inner product on V, then we define the corresponding *norm* by

$$\|v\| = (v,v)^{1/2}, \quad \text{for } v \in V. \tag{A.1}$$

We recall the *Cauchy-Schwarz inequality*,

$$|(w,v)| \le \|w\|\,\|v\|, \quad \forall v, w \in V, \tag{A.2}$$

with equality if and only if $w = \lambda v$ or $v = \lambda w$ for some $\lambda \in \mathbf{R}$, and the *triangle inequality*,

$$\|w+v\| \le \|w\| + \|v\|, \quad \forall v, w \in V. \tag{A.3}$$

Two elements $v, w \in V$ for which $(v,w) = 0$ are said to be *orthogonal*.

An infinite sequence $\{v_i\}_{i=1}^{\infty}$ in V is said to converge to $v \in V$, also written $v_i \to v$ as $i \to \infty$ or $v = \lim_{i\to\infty} v_i$, if

$$\|v_i - v\| \to 0 \quad \text{as } i \to \infty.$$

The sequence $\{v_i\}_{i=1}^{\infty}$ is called a *Cauchy sequence* in V if

$$\|v_i - v_j\| \to 0 \quad \text{as } i, j \to \infty.$$

The inner product space V is said to be *complete* if every Cauchy sequence in V is convergent, i.e., if every Cauchy sequence $\{v_i\}_{i=1}^{\infty}$ has a limit $v = \lim v_i \in V$. A complete inner product space is called a *Hilbert space*.

When we want to emphasize that an inner product or a norm is associated to a specific space V, we write $(\cdot,\cdot)_V$ and $\|\cdot\|_V$.

It is sometimes important to permit the scalars in a linear space V to be complex numbers. Such a space is then an inner product space if there is a functional (v,w) defined on $V \times V$, which is linear in the first variable and hermitian, i.e., $(w,v) = \overline{(v,w)}$. The norm is then again defined by (A.1) and V is a complex Hilbert space if completeness holds with respect to this norm. For brevity we generally consider the case of real-valued scalars in the sequel.

More generally, a *norm* in a linear space V is a function $\|\cdot\| : V \to \mathbf{R}_+$ such that

$$\begin{aligned} \|v\| &> 0, && \forall v \in V,\ v \neq 0, \\ \|\lambda v\| &= |\lambda|\,\|v\|, && \forall \lambda \in \mathbf{R} \text{ (or } \mathbf{C}),\ v \in V, \\ \|v+w\| &\le \|v\| + \|w\|, && \forall v, w \in V. \end{aligned}$$

A function $|\cdot|$ is called a *seminorm* if these conditions hold with the exception that the first one is replaced by $|v| \ge 0,\ \forall v \in V$, i.e., if it is only positive semidefinite, and thus can vanish for some $v \neq 0$. A linear space with a norm is called a *normed linear space*. As we have seen, an inner product space is a normed linear space, but not all normed linear spaces are inner product spaces. A complete normed space is called a *Banach space*.

Let V be a Hilbert space and let $V_0 \subset V$ be a linear subspace. Such a subspace V_0 is said to be *closed* if it contains all limits of sequences in V_0, i.e., if $\{v_j\}_{j=1}^{\infty} \subset V_0$ and $v_j \to v$ as $j \to \infty$ implies $v \in V_0$. Such a V_0 is itself a Hilbert space, with the same inner product as V.

Let V_0 be a closed subspace of V. Then any $v \in V$ may be written uniquely as $v = v_0 + w$, where $v_0 \in V_0$ and w is orthogonal to V_0. The element v_0 may be characterized as the unique element in V_0 which is closest to v, i.e.,

$$\|v - v_0\| = \min_{u \in V_0} \|v - u\|. \tag{A.4}$$

This is called the *projection theorem* and is a basic result in Hilbert space theory. The element v_0 is called the *orthogonal projection* of v onto V_0 and is also denoted $P_{V_0} v$. One useful consequence of the projection theorem is that if the closed linear subspace V_0 is not equal to the whole space V, then it has a normal vector, i.e., there is a nonzero vector $w \in V$ which is orthogonal to V_0.

Two norms $\|\cdot\|_a$ and $\|\cdot\|_b$ are said to be *equivalent* in V if there are positive constants c and C such that

$$c\|v\|_b \le \|v\|_a \le C\|v\|_b, \quad \forall v \in V. \tag{A.5}$$

Let V, W be two Hilbert spaces. A linear operator $B : V \to W$ is said to be *bounded*, if there is a constant C such that

$$\|Bv\|_W \le C\|v\|_V, \quad \forall v \in V. \tag{A.6}$$

The norm of a bounded linear operator B is

$$\|B\| = \sup_{v \in V \setminus \{0\}} \frac{\|Bv\|_W}{\|v\|_V}. \tag{A.7}$$

Thus

$$\|Bv\|_W \le \|B\| \, \|v\|_V, \quad \forall v \in V,$$

and, by definition, $\|B\|$ is the smallest constant C such that (A.6) holds.

Note that a bounded linear operator $B : V \to W$ is continuous. In fact, if $v_j \to v$ in V, then $Bv_j \to Bv$ in W as $j \to \infty$, because

$$\|Bv_j - Bv\|_W = \|B(v_j - v)\|_W \le \|B\| \, \|v_j - v\| \to 0, \quad \text{as } j \to \infty.$$

One can show that, conversely, a continuous linear operator is bounded.

In the special case that $W = \mathbf{R}$ the definition of an operator reduces to that of a linear functional. The set of all bounded linear functionals on V is called the *dual space* of V, denoted V^*. By (A.7) the norm in V^* is

$$\|L\|_{V^*} = \sup_{v \in V \setminus \{0\}} \frac{|L(v)|}{\|v\|_V}. \tag{A.8}$$

Note that V^* is itself a linear space if we define $(\lambda L + \mu M)(v) = \lambda L(v) + \mu M(v)$ for $L, M \in V^*$, $\lambda, \mu \in \mathbf{R}$. With the norm defined by (A.8), V^* is a normed linear space, and one may show that V^* is complete, and thus itself also a Banach space.

Similarly, we say that the bilinear form $a(\cdot, \cdot)$ on V is *bounded* if there is a constant M such that

$$|a(w, v)| \leq M\|w\| \, \|v\|, \quad \forall w, v \in V. \tag{A.9}$$

The next theorem states an important property of Hilbert spaces.

Theorem A.1. (Riesz' representation theorem.) *Let V be a Hilbert space with scalar product $(\cdot, \cdot)$. For each bounded linear functional L on V there is a unique $u \in V$ such that*

$$L(v) = (v, u), \quad \forall v \in V.$$

Moreover,

$$\|L\|_{V^*} = \|u\|_V. \tag{A.10}$$

Proof. The uniqueness is clear since $(v, u_1) = (v, u_2)$ with $v = u_1 - u_2$ implies $\|u_1 - u_2\|^2 = (u_1 - u_2, u_1 - u_2) = 0$. If $L(v) = 0$ for all $v \in V$, then we may take $u = 0$. Assume now that $L(\bar{v}) \neq 0$ for some $\bar{v} \in V$. We will construct u as a suitably normalized "normal vector" to the "hyperplane" $V_0 = \{v \in V : L(v) = 0\}$, which is easily seen to be a closed subspace of V, see Problem A.2. Then $\bar{v} = v_0 + w$ with $v_0 \in V_0$ and w orthogonal to V_0 and $L(w) = L(\bar{v}) \neq 0$. But then $L(v - w\, L(v)/L(w)) = 0$, so that $(v - w\, L(v)/L(w), w) = 0$ and hence $L(v) = (v, u)$, $\forall v \in V$, where $u = w\, L(w)/\|w\|^2$. □

This result makes it natural to identify the linear functionals $L \in V^*$ with the associated $u \in V$, and thus V^* is equivalent to V, in the case of a Hilbert space.

We sometimes want to solve equations of the form: Find $u \in V$ such that

$$a(u, v) = L(v), \quad \forall v \in V, \tag{A.11}$$

where V is a Hilbert space, L is a bounded linear functional on V, and $a(\cdot, \cdot)$ is a symmetric bilinear form, which is *coercive* in V, i.e.,

$$a(v, v) \geq \alpha \|v\|_V^2, \quad \forall v \in V, \quad \text{with } \alpha > 0. \tag{A.12}$$

This implies that $a(\cdot, \cdot)$ is symmetric, positive definite, i.e., an inner product on V, and the Riesz representation theorem immediately gives the existence of a unique solution $u \in V$ for each $L \in V^*$.

Moreover, by taking $v = u$ in (A.11) we get

$$\alpha \|u\|_V^2 \leq a(u, u) = L(u) \leq \|L\|_{V^*} \, \|u\|_V,$$

so that, after cancelling one factor $\|u\|_V$,

$$\|u\|_V \le C\|L\|_{V^*}, \quad \text{where } C = 1/\alpha. \tag{A.13}$$

This is an example of an *energy estimate*.

If $a(\cdot,\cdot)$ is a symmetric bilinear form, which is coercive and bounded in V, so that (A.12) and (A.9) hold, then we may define a norm $\|\cdot\|_a$, the *energy norm*, by

$$\|v\|_a = a(v,v)^{1/2}, \quad \text{for } v \in V,$$

By (A.12) and (A.9) we then have

$$\sqrt{\alpha}\|v\|_V \le \|v\|_a \le \sqrt{M}\|v\|_V, \quad \forall v \in V, \tag{A.14}$$

and thus the norm $\|\cdot\|_a$ on V is equivalent to $\|\cdot\|_V$. Clearly, V is then also a Hilbert space with respect to the scalar product $a(\cdot,\cdot)$ and norm $\|\cdot\|_a$.

The solution of (A.11) may also be characterized in terms of a minimization problem.

Theorem A.2. *Assume that $a(\cdot,\cdot)$ is a symmetric, positive definite bilinear form and that L is a bounded linear form on the Hilbert space V. Then $u \in V$ satisfies* (A.11) *if and only if*

$$F(u) \le F(v), \quad \forall v \in V, \quad \textit{where } F(v) = \tfrac{1}{2}a(v,v) - L(v). \tag{A.15}$$

Proof. Suppose first that u satisfies (A.11). Let $v \in V$ be arbitrary and define $w = v - u \in V$. Then $v = u + w$ and

$$\begin{aligned} F(v) &= \tfrac{1}{2}a(u+w,u+w) - L(u+w) \\ &= \tfrac{1}{2}a(u,u) - L(u) + a(u,w) - L(w) + \tfrac{1}{2}a(w,w) \\ &= F(u) + \tfrac{1}{2}a(w,w), \end{aligned}$$

where we have used (A.11) and the symmetry of $a(\cdot,\cdot)$. Since a is positive definite, this proves (A.15).

Conversely, if (A.15) holds, then for $v \in V$ given we have

$$g(t) := F(u+tv) \ge F(u) = g(0), \quad \forall\, t \in \mathbf{R},$$

so that $g(t)$ has a minimum at $t = 0$. But $g(t)$ is the quadratic polynomial

$$\begin{aligned} g(t) &= \tfrac{1}{2}a(u+tv,u+tv) - L(u+tv) \\ &= \tfrac{1}{2}a(u,u) - L(u) + t\big(a(u,v) - L(v)\big) + \tfrac{1}{2}t^2 a(v,v), \end{aligned}$$

and thus $0 = g'(0) = a(u,v) - L(v)$, which is (A.11). □

Thus, $u \in V$ satisfies (A.11) if and only if u minimizes the energy functional F. This method of studying the minimization problem by varying the argument of the functional F around the given vector u is called a variational method, and the equation (A.11) is called the *variational equation* of F.

The following theorem, which is known as the *Lax-Milgram lemma*, extends the Riesz representation theorem to nonsymmetric bilinear forms.

Theorem A.3. *If the bilinear form $a(\cdot,\cdot)$ is bounded and coercive in the Hilbert space V, and L is a bounded linear form in V, then there exists a unique vector $u \in V$ such that* (A.11) *is satisfied. Moreover, the energy estimate* (A.13) *holds.*

Proof. With $(\cdot,\cdot)$ the inner product in V we have by Riesz' representation theorem that there exists a unique $b \in V$ such that

$$L(v) = (b, v), \quad \forall v \in V.$$

Moreover, for each $u \in V$, $a(u,\cdot)$ is clearly also a bounded linear functional on V, so that there exists a unique $A(u) \in V$ such that

$$a(u, v) = (A(u), v), \quad \forall v \in V.$$

It is easy to check that $A(u)$ depends linearly and boundedly on u, so that $Au = A(u)$ defines $A : V \to V$ as a bounded linear operator. The equation (A.11) is therefore equivalent to $Au = b$, and to complete the proof of the theorem we shall show that this equation has a unique solution $u = A^{-1}b$ for each b.

Using the coercivity we have

$$\alpha\|v\|_V^2 \le a(v, v) = (Av, v) \le \|Av\|_V \|v\|_V,$$

so that

$$\alpha\|v\|_V \le \|Av\|_V, \quad \forall v \in V. \tag{A.16}$$

This shows uniqueness, since $Av = 0$ implies $v = 0$. This may also be expressed by saying that the null space $N(A) = \{v \in V : Av = 0\} = 0$, or that A is *injective*.

To show that there exists a solution u for each $b \in V$ means to show that each $b \in V$ belongs to the range $R(A) = \{w \in V : w = Av \text{ for some } v \in V\}$, i.e., $R(A) = V$, or A is *surjective*. To see this we first note that $R(A)$ is a closed linear subspace of V. To show that $R(A)$ is closed, assume that $Av_j \to w$ in V as $j \to \infty$. Then by (A.16) we have $\|v_j - v_i\|_V \le \alpha^{-1}\|Av_j - Av_i\|_V \to 0$ as $i, j \to \infty$. Hence $v_j \to v \in V$ as $j \to \infty$, and by the continuity of A, also $Av_j \to Av = w$. Therefore, $w \in R(A)$ and $R(A)$ is closed.

Assume now that $R(A) \neq V$. Then, by the projection theorem, there exists $w \neq 0$, which is orthogonal to $R(A)$. But, by the orthogonality,

$$\alpha\|w\|_V^2 \le a(w, w) = (Aw, w) = 0,$$

so that $w = 0$, which is a contradiction. Hence $R(A) = V$. This completes the proof that there is a unique solution for each $b \in V$. The energy estimate is proved in the same way as before. □

In the unsymmetric case there is no characterization of the solution in terms of energy minimization.

We finally make a remark about linear equations in finite-dimensional spaces. Let $V = \mathbf{R}^N$ and consider a linear equation in V, which may be written in matrix form as

$$Au = b,$$

where A is a $N \times N$ matrix and u, b are N-vectors. It is well-known that this equation has a unique solution $u = A^{-1}b$ for each $b \in V$, if the matrix A is nonsingular, i.e., if its determinant $\det(A) \neq 0$. If $\det(A) = 0$, then the homogeneous equation $Au = 0$ has nontrivial solutions $u \neq 0$, and $R(A) \neq V$ so that the inhomogeneous equation is not always solvable. Thus we have neither uniqueness nor existence for all $b \in V$. In particular, uniqueness only holds when $\det(A) \neq 0$, and we then also have existence. It is sometimes easy to prove uniqueness, and we then also obtain the existence of the solution at the same time.

A.2 Function Spaces

The Spaces $\mathcal{C}^k$

For $M \subset \mathbf{R}^d$ we denote by $\mathcal{C}(M)$ the linear space of continuous functions on M. The subspace $\mathcal{C}_\mathrm{b}(M)$ of all bounded functions is made into a normed linear space by setting (with a slight abuse of notation)

$$\|v\|_{\mathcal{C}(M)} = \sup_{x \in M} |v(x)|. \tag{A.17}$$

For example, this defines $\|v\|_{\mathcal{C}(\mathbf{R}^d)}$, which we use frequently. When M is a bounded and closed set, i.e., a compact set, the supremum in (A.17) is attained in M and we may write

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

The norm (A.17) is therefore called the *maximum-norm*. Note that convergence in $\mathcal{C}(M)$,

$$\|v_i - v\|_{\mathcal{C}(M)} = \sup_{x \in M} |v_i(x) - v(x)| \to 0, \quad \text{as } i \to \infty,$$

is the same as uniform convergence in M. Recall that if a sequence of continuous functions is uniformly convergent in M, then the limit function is continuous. Using this fact it is not difficult to prove that $\mathcal{C}(M)$ is a complete normed space, i.e., a Banach space. $\mathcal{C}(M)$ is not a Hilbert space, because the maximum-norm is not associated with a scalar product as in (A.1).

Let now $\Omega \subset \mathbf{R}^d$ be a *domain*, i.e., a connected open set. For any integer $k \geq 0$, we denote by $\mathcal{C}^k(\Omega)$ the linear space of all functions v that are k times continuously differentiable in Ω, and by $\mathcal{C}^k(\bar{\Omega})$ the functions in $\mathcal{C}^k(\Omega)$, for which $D^\alpha v \in \mathcal{C}(\bar{\Omega})$ for all $|\alpha| \leq k$, where $D^\alpha v$ denotes the partial derivative of v defined in (1.8). If Ω is bounded, then the latter space is a Banach space with respect to the norm

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

For functions in $\mathcal{C}^k(\bar{\Omega})$, $k \geq 1$, we sometimes also use the seminorm containing only the derivatives of highest order,

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

A function has compact support in Ω if it vanishes outside some compact subset of Ω. We write $\mathcal{C}_0^k(\Omega)$ for the space of functions in $\mathcal{C}^k(\Omega)$ with compact support in Ω. In particular, such functions vanish near the boundary Γ, and for very large x if Ω is unbounded.

We say that a function is *smooth* if, depending on the context, it has sufficiently many continuous derivatives for the purpose at hand.

When there is no risk for confusion, we omit the domain of the functions from the notation of the spaces and write, e.g., $\mathcal{C}$ for $\mathcal{C}(\bar{\Omega})$ and $\|\cdot\|_{\mathcal{C}^k}$ for $\|\cdot\|_{\mathcal{C}^k(\bar{\Omega})}$, and similarly for other spaces that we introduce below.

Integrability, the Spaces L_p

Let Ω be a domain in $\mathbf{R}^d$. We shall need to work with integrals of functions $v = v(x)$ in Ω which are more general than those in $\mathcal{C}(\bar{\Omega})$. For a nonnegative function one may define the so-called *Lebesgue integral*

$$I_\Omega(v) = \int_\Omega v(x)\,\mathrm{d}x,$$

which may be either finite or infinite, and which agrees with the standard Riemann integral for $v \in \mathcal{C}(\bar{\Omega})$. The functions we consider are assumed measurable; we shall not go into details about this concept but just note that all functions that we encounter in this text will satisfy this requirement. A nonnegative function v is said to be integrable if $I_\Omega(v) < \infty$, and a general real or complex-valued function v is similarly integrable if $|v|$ is integrable. A subset Ω_0 of Ω is said to be a nullset, or a set of measure 0, if its volume $|\Omega|$ equals 0. Two functions which are equal except on a nullset are said to be equal almost everywhere (a.e.), and they then have the same integral. Thus if $v_1(x) = 1$ in a bounded domain Ω and if $v_2(x) = 1$ in Ω except at $x_0 \in \Omega$ where $v_2(x_0) = 2$, then $I_\Omega(v_1) = I_\Omega(v_2) = |\Omega|$. In particular, from the fact that a function is integrable we cannot draw any conclusion about its value

at a point $x_0 \in \Omega$, i.e., the point values are not well defined. Also, since the boundary Γ of Ω is a nullset, $I_{\bar{\Omega}}(v) = I_\Omega(v)$ for any v.

We now define

$$\|v\|_{L_p} = \|v\|_{L_p(\Omega)} = \begin{cases} \left(\int_\Omega |v(x)|^p \, \mathrm{d}x \right)^{1/p}, & \text{for } 1 \le p < \infty, \\ \operatorname*{ess\,sup}_\Omega |v(x)|, & \text{for } p = \infty, \end{cases}$$

and say that $v \in L_p = L_p(\Omega)$ if $\|v\|_{L_p} < \infty$. Here the ess sup means the *essential supremum*, disregarding values on nullsets, so that, e.g., $\|v_2\|_{L_\infty} = 1$ for the function v_2 above, even though $\sup_\Omega v_2 = 2$. One may show that L_p is a complete normed space, i.e., a Banach space; the triangle inequality in L_p is called Minkowski's inequality. Clearly, any $v \in \mathcal{C}$ belongs to L_p for $1 \le p \le \infty$ if Ω is bounded, and

$$\|v\|_{L_p} \le C\|v\|_{\mathcal{C}}, \quad \text{with } C = |\Omega|^{1/p}, \quad \text{for } 1 \le p < \infty, \quad \text{and} \quad \|v\|_{L_\infty} = \|v\|_{\mathcal{C}},$$

but L_p also contains functions that are not continuous. Moreover, it is not difficult to show that $\mathcal{C}(\bar{\Omega})$ is incomplete with respect to the L_p-norm for $1 \le p < \infty$. To see this one constructs a sequence $\{v_i\}_{i=1}^\infty \subset \mathcal{C}(\bar{\Omega})$, which is a Cauchy sequence with respect to the L_p-norm, i.e., such that $\|v_i - v_j\|_{L_p} \to 0$, but whose limit $v = \lim_{i\to\infty} v_i$ is discontinuous. However, $\mathcal{C}(\bar{\Omega})$ is a *dense subspace* of $L_p(\Omega)$ for $1 \le p < \infty$, if Γ is sufficiently smooth. By this we mean that for any $v \in L_p$ there is a sequence $\{v_i\}_{i=1}^\infty \subset \mathcal{C}$ such that $\|v_i - v\|_{L_p} \to 0$ as $i \to \infty$. In other words, any function $v \in L_p$ can be approximated arbitrarily well in the L_p-norm by functions in $\mathcal{C}$ (in fact, for any k by functions in $\mathcal{C}_0^k$). In contrast, $\mathcal{C}$ is not dense in L_∞ since a discontinuous function cannot be well approximated uniformly by a continuous function.

The case L_2 is of particular interest to us, and this space is an inner product space, and hence a Hilbert space, with respect to the inner product

$$(v, w) = \int_\Omega v(x) w(x) \, \mathrm{d}x. \tag{A.18}$$

In the case of complex-valued functions one takes the complex conjugate of $w(x)$ in the integrand.

Sobolev Spaces

We shall now introduce some particular Hilbert spaces which are natural to use in the study of partial differential equations. These spaces consist of functions which are square integrable together with their partial derivatives up to a certain order. To define them we first need to generalize the concept of a partial derivative.

Let Ω be a domain in $\mathbf{R}^d$ and let first $v \in \mathcal{C}^1(\bar{\Omega})$. Integration by parts yields

$$\int_\Omega \frac{\partial v}{\partial x_i}\phi \,\mathrm{d}x = -\int_\Omega v\frac{\partial \phi}{\partial x_i}\,\mathrm{d}x, \quad \forall \phi \in \mathcal{C}_0^1 = \mathcal{C}_0^1(\Omega).$$

If $v \in L_2 = L_2(\Omega)$, then $\partial v/\partial x_i$ does not necessarily exist in the classical sense, but we may define $\partial v/\partial x_i$ to be the linear functional

$$L(\phi) = \frac{\partial v}{\partial x_i}(\phi) = -\int_\Omega v\frac{\partial \phi}{\partial x_i}\,\mathrm{d}x, \quad \forall \phi \in \mathcal{C}_0^1. \tag{A.19}$$

This functional is said to be a *generalized* or *weak derivative* of v. When L is bounded in L_2, i.e., $|L(\phi)| \le C\|\phi\|$, it follows from Riesz' representation theorem that there exists a unique function $w \in L_2$, such that $L(\phi) = (w, \phi)$ for all $\phi \in L_2$, and in particular

$$-\int_\Omega v\frac{\partial \phi}{\partial x_i}\,\mathrm{d}x = \int_\Omega w\phi \,\mathrm{d}x, \quad \forall \phi \in \mathcal{C}_0^1.$$

We then say that the weak derivative belongs to L_2 and write $\partial v/\partial x_i = w$. In this case we thus have

$$\int_\Omega \frac{\partial v}{\partial x_i}\phi \,\mathrm{d}x = -\int_\Omega v\frac{\partial \phi}{\partial x_i}\,\mathrm{d}x, \quad \forall \phi \in \mathcal{C}_0^1. \tag{A.20}$$

In particular, if $v \in \mathcal{C}^1(\bar{\Omega})$, then the generalized derivative $\partial v/\partial x_i$ coincides with the classical derivative $\partial v/\partial x_i$.

In a similar way, with $D^\alpha v$ denoting the partial derivative of v defined in (1.8), we define the weak partial derivative $D^\alpha v$ as the linear functional

$$D^\alpha v(\phi) = (-1)^{|\alpha|}\int_\Omega v\, D^\alpha \phi \,\mathrm{d}x, \quad \forall \phi \in \mathcal{C}_0^{|\alpha|}. \tag{A.21}$$

When this functional is bounded in L_2, Riesz' representation theorem shows that there exists a unique function in L_2, which we denote by $D^\alpha v$, such that

$$(D^\alpha v, \phi) = (-1)^{|\alpha|}(v, D^\alpha \phi), \quad \forall \phi \in \mathcal{C}_0^{|\alpha|}.$$

We refer to Problem A.9 for further discussion of generalized functions.

We now define $H^k = H^k(\Omega)$, for $k \ge 0$, to be the space of all functions whose weak partial derivatives of order $\le k$ belong to L_2, i.e.,

$$H^k = H^k(\Omega) = \big\{v \in L_2 : D^\alpha v \in L_2 \text{ for } |\alpha| \le k\big\},$$

and we equip this space with the inner product

$$(v, w)_k = (v, w)_{H^k} = \sum_{|\alpha|\le k}\int_\Omega D^\alpha v D^\alpha w \,\mathrm{d}x,$$

and the corresponding norm

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

In particular, $\|v\|_0 = \|v\|_{L_2}$, and in this case we normally omit the subscript 0 and write $\|v\|$. Also

$$\|v\|_1 = \Big(\int_\Omega \Big\{ v^2 + \sum_{j=1}^d \big(\frac{\partial v}{\partial x_j}\big)^2 \Big\} \,\mathrm{d}x \Big)^{1/2} = \Big(\|v\|^2 + \|\nabla v\|^2 \Big)^{1/2}$$

and

$$\|v\|_2 = \Big(\int_\Omega \Big\{ v^2 + \sum_{j=1}^d \big(\frac{\partial v}{\partial x_j}\big)^2 + \sum_{i,j=1}^d \big(\frac{\partial^2 v}{\partial x_i \partial x_j}\big)^2 \Big\} \,\mathrm{d}x \Big)^{1/2}.$$

We sometimes also use the seminorm, for $k \ge 1$,

$$|v|_k = |v|_{H^k} = \Big(\sum_{|\alpha|=k} \int_\Omega (D^\alpha v)^2 \,\mathrm{d}x \Big)^{1/2}. \tag{A.22}$$

Note that the seminorm vanishes for constant functions. Using the fact that L_2 is complete, one may show that H^k is complete and thus a Hilbert space, see Problem A.4. The space H^k is an example of a more general class of function spaces called Sobolev spaces.

It may be shown that $\mathcal{C}^l = \mathcal{C}^l(\bar\Omega)$ is dense in $H^k = H^k(\Omega)$ for any $l \ge k$, if Γ is sufficiently smooth. This is useful because it allows us to obtain certain results for H^k by carrying out the proof for functions in $\mathcal{C}^k$, which may be technically easier, and then extend the result to all $v \in H^k$ by using the density, cf. the proof of Theorem A.4 below.

Similarly, we denote by $W_p^k = W_p^k(\Omega)$ the normed space defined by the norm

$$\|v\|_{W_p^k} = \Big(\int_\Omega \sum_{|\alpha|\le k} |D^\alpha v|^p \,\mathrm{d}x \Big)^{1/p}, \quad \text{for } 1 \le p < \infty,$$

with the obvious modification for $p = \infty$. This space is in fact complete and hence a Banach space. For $p = 2$ we have $W_2^k = H^k$. Again, for $v \in \mathcal{C}^k$ we have $\|v\|_{W_\infty^k} = \|v\|_{\mathcal{C}^k}$.

Trace Theorems

If $v \in \mathcal{C}(\bar\Omega)$, then $v(x)$ is well defined for $x \in \Gamma$, the boundary of Ω. The *trace* γv of such a v on Γ is the restriction of v to Γ, i.e.,

$$(\gamma v)(x) = v(x), \quad \text{for } x \in \Gamma. \tag{A.23}$$

Recall that since Γ is a nullset, the trace of $v \in L_2(\Omega)$ is not well defined.

Suppose now that $v \in H^1(\Omega)$. Is it then possible to define v uniquely on Γ, i.e., to define its trace γv on Γ? (One may show that functions in

$H^1(\Omega)$ are not necessarily continuous, cf. Theorem A.5 and Problems A.6, A.7 below.) This question can be made more precise by asking if it possible to find a norm $\|\cdot\|_{(\Gamma)}$ for functions on Γ and some constant C that

$$\|\gamma v\|_{(\Gamma)} \le C\|v\|_1, \quad \forall v \in \mathcal{C}^1(\bar{\Omega}). \tag{A.24}$$

An inequality of this form is called a *trace inequality*. If (A.24) holds, then by a density argument (see below) it is possible to extend the domain of definition of the trace operator γ from $\mathcal{C}^1(\bar{\Omega})$ to $H^1(\Omega)$, and (A.24) will also hold for all $v \in H^1(\Omega)$. The function space to which γv will belong will be defined by the norm $\|\cdot\|_{(\Gamma)}$ in (A.24).

We remark that in the above discussion the boundary Γ could be replaced by some other subset of Ω of dimension smaller than d.

In order to proceed with the trace theorems, we first consider a one-dimensional case, with Γ corresponding to a single point.

Lemma A.1. *Let $\Omega = (0,1)$. Then there is a constant C such that*

$$|v(x)| \le C\|v\|_1, \quad \forall x \in \bar{\Omega},\ \forall v \in \mathcal{C}^1(\bar{\Omega}).$$

Proof. For $x, y \in \Omega$ we have $v(x) = v(y) + \int_y^x v'(s)\,\mathrm{d}s$, and hence by the Cauchy-Schwarz inequality

$$|v(x)| \le |v(y)| + \int_0^1 |v'(s)|\,\mathrm{d}s \le |v(y)| + \|v'\|.$$

Squaring both sides and integrating with respect to y, we obtain,

$$v(x)^2 \le 2\big(\|v\|^2 + \|v'\|^2\big) = 2\|v\|_1^2. \tag{A.25}$$

which shows the desired estimate. □

We now show a simple trace theorem. By $L_2(\Gamma)$ we denote the Hilbert space of all functions that are square integrable on Γ with norm

$$\|w\|_{L_2(\Gamma)} = \Big(\int_\Gamma w^2\,\mathrm{d}s\Big)^{1/2}.$$

Theorem A.4. (Trace theorem.) *Let Ω be a bounded domain in $\mathbf{R}^d$ ($d \ge 2$) with smooth or polygonal boundary Γ. Then the trace operator $\gamma : \mathcal{C}^1(\bar{\Omega}) \to \mathcal{C}(\Gamma)$ may be extended to $\gamma : H^1(\Omega) \to L_2(\Gamma)$, which defines the trace $\gamma v \in L_2(\Gamma)$ for $v \in H^1(\Omega)$. Moreover, there is a constant $C = C(\Omega)$ such that*

$$\|\gamma v\|_{L_2(\Gamma)} \le C\|v\|_1, \quad \forall v \in H^1(\Omega). \tag{A.26}$$

Proof. We first show the trace inequality (A.26) for functions $v \in \mathcal{C}^1(\bar{\Omega})$. For simplicity we consider only the case when $\Omega = (0,1) \times (0,1)$, the unit square

in $\mathbf{R}^2$. The proof in the general case is similar. For $x = (x_1, x_2) \in \Omega$ we have by (A.25)

$$v(0,x_2)^2 \le 2\Big(\int_0^1 v(x_1,x_2)^2\,\mathrm{d}x_1 + \int_0^1 \Big(\frac{\partial v}{\partial x_1}(x_1,x_2)\Big)^2\,\mathrm{d}x_1\Big),$$

and hence by integration with respect to x_2,

$$\int_0^1 v(0,x_2)^2\,\mathrm{d}x_2 \le 2\big(\|v\|^2 + \|\nabla v\|^2\big) = 2\|v\|_1^2.$$

The analogous estimates for the remaining parts of Γ complete the proof of (A.26) for $v \in \mathcal{C}^1$.

Let now $v \in H^1(\Omega)$. Since $\mathcal{C}^1$ is dense in H^1 there is a sequence $\{v_i\}_{i=1}^\infty \subset \mathcal{C}^1$ such that $\|v - v_i\|_1 \to 0$. This sequence is then a Cauchy sequence in H^1, i.e., $\|v_i - v_j\|_1 \to 0$ as $i,j \to \infty$. Applying (A.26) to $v_i - v_j \in \mathcal{C}^1$, we find

$$\|\gamma v_i - \gamma v_j\|_{L_2(\Gamma)} \le C\|v_i - v_j\|_1 \to 0, \quad \text{as } i,j \to \infty,$$

i.e., $\{\gamma v_i\}_{i=1}^\infty$ is a Cauchy sequence in $L_2(\Gamma)$, and thus there exists $w \in L_2(\Gamma)$ such that $\gamma v_i \to w$ in $L_2(\Gamma)$ as $i \to \infty$. We define $\gamma v = w$. It is easy to show that (A.26) then holds for $v \in H^1$. This extends γ to a bounded linear operator $\gamma : H^1(\Omega) \to L_2(\Gamma)$. Since $\mathcal{C}^1$ is dense in H^1, there is only one such extension (prove this!). In particular, γ is independent of the choice of the sequence $\{v_i\}$. □

The constant in Theorem A.4 depends on the size and shape of the domain Ω. It is sometimes important to have more detailed information about this dependence, and in Problem A.15 we assume that the shape is fixed (a square) and investigate the dependence of the constant on the size of Ω.

The following result, of a somewhat similar nature, is a special case of the well-known and important Sobolev inequality.

Theorem A.5. *Let Ω be a bounded domain in $\mathbf{R}^d$ with smooth or polygonal boundary and let $k > d/2$. Then $H^k(\Omega) \subset \mathcal{C}(\bar\Omega)$, and there exists a constant $C = C(\Omega)$ such that*

$$\|v\|_{\mathcal{C}} \le C\|v\|_k, \quad \forall v \in H^k(\Omega). \tag{A.27}$$

In the same way as for the trace theorem it suffices to show (A.27) for smooth v, see Problem A.20. The particular case when $d = k = 1$ is given in Lemma A.1, and Problem A.13 considers the case $\Omega = (0,1)\times(0,1)$. The general case is more complicated. As shown in Problems A.6, A.7, a function in $H^1(\Omega)$ with $\Omega \subset \mathbf{R}^d$ is not necessarily continuous when $d \ge 2$.

If we apply Sobolev's inequality to derivatives of v, we get

$$\|v\|_{\mathcal{C}^\ell} \le C\|v\|_k, \quad \forall v \in H^k(\Omega), \text{ if } k > \ell + d/2, \tag{A.28}$$

and we may similarly conclude that $H^k(\Omega) \subset \mathcal{C}^\ell(\bar\Omega)$ if $k > \ell + d/2$.

The Space $H_0^1(\Omega)$. Poincaré's Inequality

Theorem A.4 shows that the trace operator $\gamma : H^1(\Omega) \to L_2(\Gamma)$ is a bounded linear operator. This implies that its null space,

$$H_0^1(\Omega) = \big\{v \in H^1(\Omega) : \gamma v = 0\big\},$$

is a closed subspace of $H^1(\Omega)$, and hence a Hilbert space with the norm $\|\cdot\|_1$. It is the set of functions in H^1 that vanish on Γ in the sense of trace. We note that the seminorm $|v|_1 = \|\nabla v\|$ defined in (A.22) is in fact a norm on $H_0^1(\Omega)$, equivalent to $\|\cdot\|_1$, as follows from the following result.

Theorem A.6. (Poincaré's inequality.) *If Ω is a bounded domain in $\mathbf{R}^d$, then there exists a constant $C = C(\Omega)$ such that*

$$\|v\| \le C\|\nabla v\|, \quad \forall v \in H_0^1(\Omega). \tag{A.29}$$

Proof. As an example we show the result for $\Omega = (0,1) \times (0,1)$. The proof in the general case is similar.

Since $\mathcal{C}_0^1$ is dense in H_0^1, it suffices to show (A.29) for $v \in \mathcal{C}_0^1$. For such a v we write

$$v(x) = \int_0^{x_1} \frac{\partial v}{\partial x_1}(s, x_2)\,\mathrm{d}s, \quad \text{for } x = (x_1, x_2) \in \Omega,$$

and hence by the Cauchy-Schwarz inequality

$$|v(x)|^2 \le \int_0^1 \mathrm{d}s \int_0^1 \Big(\frac{\partial v}{\partial x_1}(s, x_2)\Big)^2 \mathrm{d}s.$$

The result now follows by integration with respect to x_2 and x_1, with $C = 1$ in this case. □

The equivalence of the norms $|\cdot|_1$ and $\|\cdot\|_1$ on $H_0^1(\Omega)$ now follows from

$$\|\nabla v\|^2 \le \|v\|_1^2 = \|v\|^2 + \|\nabla v\|^2 \le (C+1)\|\nabla v\|^2, \quad \forall v \in H_0^1(\Omega).$$

The dual space of $H_0^1(\Omega)$ is denoted $H^{-1}(\Omega)$. Thus $H^{-1} = (H_0^1)^*$ is the space of all bounded linear functionals on H_0^1. The norm in H^{-1} is (cf. (A.8))

$$\|L\|_{(H_0^1)^*} = \|L\|_{-1} = \sup_{v \in H_0^1} \frac{|L(v)|}{|v|_1}. \tag{A.30}$$

A.3 The Fourier Transform

Let v be a real or complex function in $L_1(\mathbf{R}^d)$. We define its Fourier transform for $\xi = (\xi_1, \ldots, \xi_d) \in \mathbf{R}^d$ by

$$\mathcal{F}v(\xi) = \hat{v}(\xi) = \int_{\mathbf{R}^d} v(x)\mathrm{e}^{-\mathrm{i}x\cdot\xi}\,\mathrm{d}x, \quad \text{where } x\cdot\xi = \sum_{j=1}^{d} x_j\xi_j.$$

The inverse Fourier transform is

$$\mathcal{F}^{-1}v(x) = \check{v}(x) = (2\pi)^{-d}\int_{\mathbf{R}^d} v(\xi)\mathrm{e}^{\mathrm{i}x\cdot\xi}\,\mathrm{d}\xi = (2\pi)^{-d}\hat{v}(-x), \quad \text{for } x\in\mathbf{R}^d.$$

If v and $\hat{v}$ are both in $L_1(\mathbf{R}^d)$, then Fourier's inversion formula holds, i.e.,

$$\mathcal{F}^{-1}(\mathcal{F}\ v) = (\hat{v})\check{} = v.$$

The inner product in $L_2(\mathbf{R}^d)$ of two functions can be expressed in terms of their Fourier transforms according to Parseval's formula,

$$\int_{\mathbf{R}^d} v(x)\overline{w(x)}\,\mathrm{d}x = (2\pi)^{-d}\int_{\mathbf{R}^d} \hat{v}(\xi)\overline{\hat{w}(\xi)}\,\mathrm{d}\xi, \tag{A.31}$$

or

$$(v,w) = (2\pi)^{-d}(\hat{v},\hat{w}), \quad \text{where } (v,w) = (v,w)_{L_2(\mathbf{R}^d)}.$$

In particular, we have for the corresponding norms

$$\|v\| = (2\pi)^{-d/2}\|\hat{v}\|. \tag{A.32}$$

Let $D^\alpha v$ be a partial derivative of v as defined in (1.8). We then have, assuming that v and its derivatives are sufficiently small for $|x|$ large,

$$\mathcal{F}(D^\alpha v)(\xi) = (i\xi)^\alpha\hat{v}(\xi) = i^{|\alpha|}\xi^\alpha\hat{v}(\xi), \quad \text{where } \xi^\alpha = \xi_1^{\alpha_1}\cdots\xi_d^{\alpha_d}.$$

In fact, by integration by parts,

$$\int_{\mathbf{R}^d} D^\alpha v(x)\mathrm{e}^{-\mathrm{i}x\cdot\xi}\,\mathrm{d}x = (-1)^{|\alpha|}\int_{\mathbf{R}^d} v(x)D^\alpha(\mathrm{e}^{-\mathrm{i}x\cdot\xi})\,\mathrm{d}x = (i\xi)^\alpha\hat{v}(\xi).$$

Further, translation of the argument of the function corresponds to multiplication of its Fourier transform by an exponential,

$$\mathcal{F}v(\cdot+y)(\xi) = \mathrm{e}^{\mathrm{i}y\cdot\xi}\hat{v}(\xi), \quad \text{for } y\in\mathbf{R}^d, \tag{A.33}$$

and for scaling of the argument we have

$$\mathcal{F}v(a\cdot)(\xi) = a^{-d}\hat{v}(a^{-1}\xi), \quad \text{for } a>0. \tag{A.34}$$

The convolution of two functions v and w is defined by

$$(v*w)(x) = \int_{\mathbf{R}^d} v(x-y)w(y)\,\mathrm{d}y = \int_{\mathbf{R}^d} v(y)w(x-y)\,\mathrm{d}y,$$

and we have

$$\mathcal{F}(v * w)(\xi) = \hat{v}(\xi)\hat{w}(\xi),$$

because

$$\begin{aligned}\int_{\mathbf{R}^d} \Big(\int_{\mathbf{R}^d} v(x-y)w(y)\,\mathrm{d}y \Big) \mathrm{e}^{-\mathrm{i}x\cdot\xi}\,\mathrm{d}x \\ = \int_{\mathbf{R}^d}\int_{\mathbf{R}^d} v(x-y)w(y)\mathrm{e}^{-\mathrm{i}(x-y)\cdot\xi}\mathrm{e}^{-\mathrm{i}y\cdot\xi}\,\mathrm{d}x\,\mathrm{d}y \\ = \int_{\mathbf{R}^d}\int_{\mathbf{R}^d} v(z)w(y)\mathrm{e}^{-\mathrm{i}z\cdot\xi}\mathrm{e}^{-\mathrm{i}y\cdot\xi}\,\mathrm{d}z\,\mathrm{d}y.\end{aligned}$$

It follows, which can also easily be shown directly, that differentiation of a convolution can be carried out on either factor,

$$D^\alpha(v * w) = D^\alpha v * w = v * D^\alpha w.$$

A.4 Problems

Problem A.1. Let V be a Hilbert space with scalar product $(\cdot,\cdot)$ and let $u \in V$ be given. Define $L : V \to \mathbf{R}$ by $L(v) = (u, v)\ \forall v \in V$. Prove that L is a bounded linear functional on V. Determine $\|L\|$.

Problem A.2. Prove that if $L : V \to \mathbf{R}$ is a bounded linear functional and $\{v_i\}$ is a sequence with $L(v_i) = 0$ that converges to $v \in V$, then $L(v) = 0$. This proves that the subspace V_0 in the proof of Theorem A.1 is closed.

Problem A.3. Prove the energy estimate (A.13) by using (A.10) and (A.14). Hint: Recall (A.8) and note that (A.10) means

$$\sup_{v\in V\setminus\{0\}} \frac{|L(v)|}{\|v\|_a} = \|u\|_a.$$

Problem A.4. Given that $L_2(\Omega)$ is complete, prove that $H^1(\Omega)$ is complete. Hint: Assume that $\|v_j - v_i\|_1 \to 0$ as $i, j \to \infty$. Show that there are v, w_k such that $\|v_j - v\| \to 0$, $\|\partial v_j/\partial x_k - w_k\| \to 0$, and that $w_k = \partial v/\partial x_k$ in the sense of weak derivative.

Problem A.5. Let $\Omega = (-1, 1)$ and let $v : \Omega \to \mathbf{R}$ be defined by $v(x) = 1$ if $x \in (-1, 0)$ and $v(x) = 0$ if $x \in (0, 1)$. Prove that $v \in L_2(\Omega)$ and that v can be approximated arbitrarily well in L_2-norm by $\mathcal{C}^0$-functions.

Problem A.6. Let Ω be the unit ball in $\mathbf{R}^d$, $d = 1, 2, 3$, i.e., $\Omega = \{x \in \mathbf{R}^d : |x| < 1\}$. For which values of $\lambda \in \mathbf{R}$ does the function $v(x) = |x|^\lambda$ belong to (a) $L_2(\Omega)$, (b) $H^1(\Omega)$?

Problem A.7. Check if the function $v(x) = \log(-\log|x|^2)$ belongs to $H^1(\Omega)$ if $\Omega = \{x \in \mathbf{R}^2 : |x| < \frac{1}{2}\}$. Are functions in $H^1(\Omega)$ necessarily bounded and continuous?

Problem A.8. It is known that $\mathcal{C}_0^1(\Omega)$ is dense in $L_2(\Omega)$ and $H_0^1(\Omega)$. Explain why $\mathcal{C}_0^1(\Omega)$ is not dense in $H^1(\Omega)$.

Problem A.9. The generalized (or weak) derivative defined in (A.19) is a special case of the so-called *generalized functions* or *distributions*. Another important example is *Dirac's delta*, which is defined as a linear functional acting on continuous test functions, for $\Omega \subset \mathbf{R}^d$,

$$\delta(\phi) = \phi(0), \quad \forall \phi \in \mathcal{C}_0^0(\Omega).$$

Let now $d = 1$, $\Omega = (-1, 1)$ and

$$f(x) = \begin{cases} x, & x \geq 0, \\ 0, & x \leq 0, \end{cases} \qquad g(x) = \begin{cases} 1, & x > 0, \\ 0, & x < 0. \end{cases}$$

Show that $f' = g$, $g' = \delta$ in the sense of generalized derivative, i.e.,

$$f'(\phi) = -\int_\Omega f\phi' \,\mathrm{d}x = \int_\Omega g\phi \,\mathrm{d}x, \qquad \forall \phi \in \mathcal{C}_0^1(\Omega),$$

$$g'(\phi) = -\int_\Omega g\phi' \,\mathrm{d}x = \phi(0), \qquad \forall \phi \in \mathcal{C}_0^1(\Omega).$$

Conclude that the generalized derivative $f' = g$ belongs to L_2, but that $g' = \delta$ does not. For the latter statement, you must show that δ is not bounded with respect to the L_2-norm, i.e., you need to find a sequence of test functions such that $\|\phi_i\|_{L_2} \to 0$, but $\phi_i(0) = 1$ as $i \to \infty$. Thus, $f \in H^1(\Omega)$ and $g \notin H^1(\Omega)$.

Problem A.10. For $f \in L_2(\Omega)$ we define the linear functional $f(v) = (f, v)$ $\forall v \in L_2(\Omega)$. Show the inequality, cf. (A.30),

$$\|f\|_{-1} \leq C\|f\|, \quad \forall f \in L_2(\Omega).$$

Conclude that $L_2(\Omega) \subset H^{-1}(\Omega)$.

Problem A.11. Let $\Omega = (0, 1)$ and $f(x) = 1/x$. Show that $f \notin L_2(\Omega)$. Show that $f \in H^{-1}(\Omega)$ by defining the linear functional $f(v) = (f, v)$ $\forall v \in H_0^1(\Omega)$, and proving the inequality

$$|(f, v)| \leq C\|v'\|, \quad \forall v \in H_0^1(\Omega).$$

Conclude that $H^{-1}(\Omega) \not\subset L_2(\Omega)$.

Problem A.12. Prove that if $\Omega = (0, L)$ is a finite interval, then there is a constant $C = C(L)$ such that, for all $x \in \bar{\Omega}$ and $v \in \mathcal{C}^1(\bar{\Omega})$,

(a) $$|v(x)| \leq L^{-1} \int_\Omega |v| \,\mathrm{d}y + \int_\Omega |v'| \,\mathrm{d}y \leq C\|v\|_{W_1^1(\Omega)},$$

(b) $$|v(x)|^2 \leq L^{-1} \int_\Omega |v|^2 \,\mathrm{d}y + L \int_\Omega |v'|^2 \,\mathrm{d}y \leq C\|v\|_1^2,$$

(c) $$|v(x)|^2 \leq L^{-1}\|v\|^2 + 2\|v\| \, \|v'\| \leq C\|v\| \, \|v\|_1.$$

Problem A.13. Prove that if Ω is the unit square in $\mathbf{R}^2$, then there exists a constant C such that

(a) $$\|v\|_{L_1(\Gamma)} \leq C\|v\|_{W_1^1(\Omega)}, \qquad \forall v \in \mathcal{C}^1(\bar{\Omega}),$$

(b) $$\|v\|_{\mathcal{C}} \leq C\|v\|_{W_1^2}, \qquad \forall v \in \mathcal{C}^2(\bar{\Omega}).$$

Since $\|v\|_{W_1^2} \leq 3^{1/2}|\Omega|^{1/2}\|v\|_{H^2}$, part (b) implies the special case of Theorem A.5 with $k = d = 2$ and Ω a square domain. Part (b) directly generalizes to $\|v\|_{\mathcal{C}} \leq C\|v\|_{W_1^d}$ for $\Omega \subset \mathbf{R}^d$. Hint: Proof of Theorem A.4.

Problem A.14. (Scaling of Sobolev norms.) Let L be a positive number and consider the coordinate transformation $x = L\hat{x}$, which maps the bounded domain $\Omega \subset \mathbf{R}^d$ onto $\hat{\Omega}$. A function v defined on Ω is transformed to a function $\hat{v}$ on $\hat{\Omega}$ according to $\hat{v}(\hat{x}) = v(L\hat{x})$. Prove the scaling identities

(a) $$\|v\|_{L_2(\Omega)} = L^{d/2}\|\hat{v}\|_{L_2(\hat{\Omega})},$$

(b) $$\|\nabla v\|_{L_2(\Omega)} = L^{d/2-1}\|\hat{\nabla}\hat{v}\|_{L_2(\hat{\Omega})},$$

(c) $$\|v\|_{L_2(\Gamma)} = L^{d/2-1/2}\|\hat{v}\|_{L_2(\hat{\Gamma})}.$$

Problem A.15. (Scaled trace inequality.) Let $\Omega = (0, L)\times(0, L)$ be a square domain of side L. Prove the scaled trace inequality

$$\|v\|_{L_2(\Gamma)} \leq C\Big(L^{-1}\|v\|^2_{L_2(\Omega)} + L\|\nabla v\|^2_{L_2(\Omega)}\Big)^{1/2}, \quad \forall v \in \mathcal{C}^1(\bar{\Omega}).$$

Hint: Apply (A.26) with $\hat{\Omega} = (0, 1) \times (0, 1)$ and use the scaling identities in Problem A.14.

Problem A.16. Let Ω be the unit square in $\mathbf{R}^2$. Prove the trace inequality in the form

$$\|v\|^2_{L_2(\Gamma)} \leq C\big(\|v\|^2_{L_2(\Omega)} + \|v\|_{L_2(\Omega)}\|\nabla v\|_{L_2(\Omega)}\big) \leq C\|v\|\,\|v\|_1.$$

Hint: Start from

$$v(0, y_2)^2 = v(y_1, y_2)^2 - \int_0^{y_1} \frac{\partial}{\partial x_1} v(s, y_2)^2 \,\mathrm{d}s.$$

Problem A.17. It is a fact from linear algebra that all norms on a finite-dimensional space V are equivalent. Illustrate this by proving the following norm equivalences in $V = \mathbf{R}^N$:

(A.35) $$\|v\|_{l_2} \leq \|v\|_{l_1} \leq \sqrt{N}\|v\|_{l_2},$$

(A.36) $$\|v\|_{l_\infty} \leq \|v\|_{l_2} \leq \sqrt{N}\|v\|_{l_\infty},$$

(A.37) $$\|v\|_{l_\infty} \leq \|v\|_{l_1} \leq N\|v\|_{l_\infty},$$

where

$$\|v\|_{l_p} = \Big(\sum_{j=1}^{N} |v_j|^p\Big)^{1/p} \quad \text{for } 1 \le p < \infty, \quad \|v\|_{l_\infty} = \max_{1\le j\le N} |v_j|.$$

Note that the equivalence constants tend to infinity as $N \to \infty$.

Problem A.18. Prove (A.33) and (A.34).

Problem A.19. Prove that the Fourier transform of $v(x) = \mathrm{e}^{-|x|^2}$ is $\hat{v}(\xi) = \pi^{d/2}\mathrm{e}^{-|\xi|^2/4}$.

Problem A.20. Assume that Sobolev's inequality in (A.27) has been proved for all $v \in \mathcal{C}^k(\bar{\Omega})$ with $k > d/2$. Prove Sobolev's imbedding $H^k(\Omega) \subset \mathcal{C}(\bar{\Omega})$. In other words, for each $v \in H^k(\Omega)$ show that there is $w \in \mathcal{C}(\bar{\Omega})$ such that $v = w$ almost everywhere, i.e., $\|v - w\|_{L_2} = 0$. Hint: $\mathcal{C}^k(\bar{\Omega})$ is dense in $H^k(\Omega)$ and $\mathcal{C}(\bar{\Omega})$ is a Banach space.

B Orientation on Numerical Linear Algebra

Both finite difference and finite element methods for elliptic problems lead to linear algebraic systems of equations of the form

$$AU = b, \tag{B.1}$$

where A is a nonsingular square matrix of order N. Also in time-stepping methods for evolution equations, problems of elliptic type need to be solved in the successive time steps. To solve such systems efficiently therefore becomes an important part of numerical analysis. When the dimension of the computational domain is at least 2 this is normally not possible by direct methods, and, except in special cases, one therefore turns to iterative methods. These take advantage of the fact that the matrices involved are sparse, i.e., most of their elements are zero, and have other special features. In this appendix we give a short overview, without proofs, of the most commonly used methods.

B.1 Direct Methods

We consider first the case that the system (B.1) derives from the standard finite difference approximation (4.3) of the two-point boundary value problem (4.1). In this case A is a tridiagonal matrix, and it is easy to see that A may then be factored in $O(N)$ algebraic operations as $A = LR$, where L is bidiagonal and lower triangular and R is bidiagonal, upper triangular. The system may thus be written

$$LRU = b,$$

and one may now first solve $LG = b$ for $G = RU$ in $O(N)$ operations and then solve the latter equation for U, also in $O(N)$ operations. Altogether this is a direct method for (B.1), which requires $O(N)$ operations. Since the number of unknowns is N, this is the smallest possible order for any method.

Consider now an elliptic problem in a domain $\Omega \subset \mathbf{R}^d$ with $d \geq 2$. Using either finite differences or finite elements based on a quasi-uniform family of meshes, the dimension N of the corresponding finite dimensional problem is of order $O(h^{-d})$, where h is the mesh-size, and for $d \geq 2$ direct solution by Gauss elimination is normally not feasible as this method requires

$O(N^3) = O(h^{-3d})$ algebraic operations. Except in special cases one therefore turns to iterative methods.

One case when a direct method can be used, however, is provided by the model problem (4.11) with the five-point finite difference scheme on the unit square, which may be solved directly by using the discrete Fourier transform, defined by

$$\hat{b}_m = \sum_j b_j \mathrm{e}^{-2\pi \mathrm{i}\, m\cdot j\, h}, \quad m = (m_1, m_2),\ j = (j_1, j_2).$$

In fact, we then have $(-\Delta_h U)\hat{}_m = 2\pi^2 |m|^2 \hat{U}_m$, hence $\hat{U}_m = (2\pi^2|m|^2)^{-1}\hat{b}_m$, so that by the inverse discrete Fourier transform

$$U^j = \sum_m (2\pi^2|m|^2)^{-1} \hat{b}_m \mathrm{e}^{2\pi \mathrm{i}\, m\cdot j\, h}.$$

Using the Fast Fourier Transform (FFT) both $\hat{b}_m$ and U^j may be calculated in $O(N \log N)$ operations.

B.2 Iterative Methods. Relaxation, Overrelaxation, and Acceleration

As a basic iterative method for (B.1) we consider the Richardson method

$$U^{n+1} = U^n - \tau(AU^n - b) \quad \text{for } n \geq 0, \quad \text{with } U^0 \text{ given}, \tag{B.2}$$

where τ is a positive parameter. With U the exact solution of (B.1) we have

$$U^n - U = R(U^{n-1} - U) = \cdots = R^n(U^0 - U), \quad \text{where } R = I - \tau A,$$

and hence the rate of convergence of the method depends on $\|R^n\|$, where $\|M\| = \max_{\|x\|=1} \|Mx\|$ is the matrix norm subordinate to the Euclidean norm $\|\cdot\|$ in $\mathbf{R}^N$. When A is symmetric positive definite (SPD) and has eigenvalues $\{\lambda_j\}_{j=1}^N$, then, since $\{1 - \tau\lambda_j\}_{j=1}^N$ are the eigenvalues of R, we have

$$\|R^n\| = \rho^n, \quad \text{where } \rho = \rho(R) = \max_i |1 - \tau\lambda_i|,$$

and (B.2) converges if $\rho < 1$. The choice of τ which gives the smallest value of ρ is $\tau = 2/(\lambda_1 + \lambda_N)$, in which case $\rho = (\kappa - 1)/(\kappa + 1)$, where $\kappa = \kappa(A) = \lambda_N/\lambda_1$ is the condition number or A. We note, however, that this choice of τ requires knowledge of λ_1 and λ_N which is not normally available. In applications to second order elliptic problems one often has $\kappa = O(h^{-2})$ so that $\rho \leq 1 - ch^2$ with $c > 0$. Hence with the optimal choice of τ the number of iterations required to reduce the error to a small $\epsilon > 0$ is of order $O(h^{-2}|\log \epsilon|)$. Since each iteration uses $O(h^{-d})$ operations in the application

of $I - \tau A$, this shows that the total number of operations needed to reduce the error to a given tolerance is of order $O(h^{-d-2})$, which is smaller than for the direct solution by Gauss elimination when $d \geq 2$.

The early more refined methods were designed for finite difference methods of positive type for second order elliptic equations, particularly for the five-point scheme (4.12). The corresponding matrix may then be written $A = D - E - F$, where D is diagonal and E and F are (elementwise) nonnegative and strictly lower and upper triangular. Examples of more efficient methods are then the Jacobi and Gauss-Seidel methods which are defined by

$$U^{n+1} = U^n - B(AU^n - b) = RU^n + Bb, \quad \text{with } R = I - BA, \tag{B.3}$$

in which $B = B_{\mathrm{J}} = D^{-1}$ or $B = B_{\mathrm{GS}} = (D - E)^{-1}$, so that $R = R_{\mathrm{J}} = D^{-1}(E + F)$ and $R = R_{\mathrm{GS}} = (D - E)^{-1}F$, respectively. In the application to the model problem (4.9) in the unit square, using the five-point operator, the equations may be normalized so that $D = 4I$ and the application of R_{J} then simply means that the new value in the iteration step at any interior mesh-point x_j is obtained by taking the average of the old values at the four neighboring points $x_{j\pm e_l}$. The Gauss-Seidel method also takes averages, but with the mesh-points taken in a given order, and successively uses the values already obtained in forming the averages. The methods are therefore also referred to as the methods of simultaneous and successive displacements, respectively. For the model problem one may easily determine the eigenvalues and eigenvectors of A and show that with $h = 1/M$ one has $\rho(R_{\mathrm{J}}) = \cos(\pi h) = 1 - \frac{1}{2}\pi^2 h^2 + O(h^4)$ and $\rho(R_{\mathrm{GS}}) = \rho(R_{\mathrm{J}})^2 = 1 - \pi^2 h^2 + O(h^4)$, so that the number of iterations needed to reduce the error to ϵ is of the orders $2h^{-2}\pi^2|\log\epsilon|$ and $h^{-2}\pi^2|\log\epsilon|$, respectively. The Gauss-Seidel method thus requires about half as many iterations as the Jacobi method.

Forming the averages in the Jacobi and Gauss-Seidel methods may be thought as relaxation. It turns out that one may obtain better results than those described above by overrelaxation, i.e., by choosing

$$B_\omega = (D - \omega E)^{-1} \quad \text{and} \quad R_\omega = (D - \omega E)^{-1}((1 - \omega)E + F), \quad \text{with } \omega > 1.$$

It may be shown that for the model problem the optimal choice of the parameter is

$$\omega_{\mathrm{opt}} = 2/(1 + \sqrt{1 - \rho^2}), \quad \text{where } \rho = \rho(B_{\mathrm{J}}) = \cos(\pi h),$$

i.e., $\omega_{\mathrm{opt}} = 2/(1 + \sin(\pi h)) = 2 - 2\pi h + O(h^2)$, and that correspondingly

$$\rho(R_{\omega_{\mathrm{opt}}}) = \omega_{\mathrm{opt}} - 1 = 1 - 2\pi h + O(h^2).$$

The number of iterations required is thus then of order $O(h^{-1})$, which is significantly smaller than for the above methods. This is the method of successive overrelaxation (SOR).

We consider again an iterative method of the form (B.3) with $\rho(R) < 1$. For the purpose of accelerating the convergence we now introduce the new sequence $V^n = \sum_{j=0}^n \beta_{nj} U^j$, where the β_{nj} are real numbers. Setting $p_n(\lambda) = \sum_{j=0}^n \beta_{nj}\lambda^j$, and assuming $p_n(1) = \sum_{j=0}^n \beta_{nj} = 1$ for $n \geq 0$, we obtain easily $V^n - U = p_n(R)(U^0 - U)$, where U is the solution of (B.1). For V^n to converge fast to U one therefore wants to choose the β_{nj} in such a way that the spectral radius $\rho(p_n(R))$ becomes small with n. By the Cayley-Hamilton theorem for matrices one has $p_N(R) = 0$, if p_N is the characteristic polynomial of R, and hence $V^N = U$, but this requires a prohibitively large number of iterations. For $n < N$ we have by the spectral mapping theorem that $\rho(p_n(R)) = \max_i |p_n(\mu_i)|$, where $\{\mu_i\}_{i=1}^N$ are the eigenvalues of R. In particular, if R is symmetric and $\rho = \rho(R)$, so that $|\mu_i| \leq \rho$ for all i, then one may show that, taking the maximum instead over $[-\rho, \rho] \supset \sigma(R)$, the optimal polynomial is $p_n(\lambda) = T_n(\lambda/\rho)/T_n(1/\rho)$, where T_n is the nth Chebyshev polynomial, and the corresponding value of $\rho(p_n(R))$ is bounded by

$$\begin{aligned} T_n(1/\rho)^{-1} &= 2\Big\{ \Big(\frac{1+\sqrt{1-\rho^2}}{\rho}\Big)^n + \Big(\frac{1+\sqrt{1-\rho^2}}{\rho}\Big)^{-n} \Big\}^{-1} \\ &\leq 2\Big(\frac{\rho}{1+\sqrt{1-\rho^2}}\Big)^n. \end{aligned}$$

For the model problem using the Gauss-Seidel basic iteration we have as above $\rho = 1 - \pi^2 h^2 + O(h^4)$ and we find that the average error reduction factor per iteration step in our present method is bounded by $1 - \sqrt{2}\pi h + O(h^2)$, which is of the same order of magnitude as for SOR.

B.3 Alternating Direction Methods

We now describe the Peaceman-Rachford alternating direction implicit iterative method for the model problem (4.9) on the unit square, using the five-point discrete elliptic equation (4.11) with $h = 1/M$. In this case we may write $A = H + V$, where H and V correspond to the horizontal and vertical difference operators $-h^2\partial_1\bar{\partial}_1$ and $-h^2\partial_2\bar{\partial}_2$. Note that H and V are positive definite and commute. Introducing an acceleration parameter τ and an intermediate value $U^{n+1/2}$, we may consider the scheme defining U^{n+1} from U^n by

$$\begin{aligned} (\tau I + H)U^{n+1/2} &= (\tau I - V)U^n + b, \\ (\tau I + V)U^{n+1} &= (\tau I - H)U^{n+1/2} + b, \end{aligned} \tag{B.4}$$

or after elimination, with G_τ appropriate and using that H and V commute,

$$U^{n+1} = R_\tau U^n + G_\tau, \quad \text{where } R_\tau = (\tau I - H)(\tau I + H)^{-1}(\tau I - V)(\tau I + V)^{-1}.$$

Note that the equations in (B.4) have tridiagonal matrices and may be solved in $O(N)$ operations, as we have indicated earlier. The error satisfies $U^n - U = R_\tau^n(U^0 - U)$, and with μ_i the (common) eigenvalues of H and V, we have $\|R_\tau\| \le \max_i |(\tau-\mu_i)/(\tau+\mu_i)|^2 < 1$, where it is easy to see that the maximum occurs for $i = 1$ or M. With $\mu_1 = 4\sin^2(\frac{1}{2}\pi h)$, $\mu_M = 4\cos^2(\frac{1}{2}\pi h)$ the optimal τ is $\tau_{\text{opt}} = (\mu_1\mu_M)^{1/2}$ with the maximum for $i = 1$, so that, with $\kappa = \kappa(H) = \kappa(V) = \mu_M/\mu_1$,

$$\|R_{\tau_{\text{opt}}}\| \le \Big(\frac{(\mu_1\mu_M)^{1/2} - \mu_1}{(\mu_1\mu_M)^{1/2} + \mu_1}\Big)^{1/2} = \frac{\kappa^{1/2} - 1}{\kappa^{1/2} + 1} = 1 - \pi h + O(h^2).$$

This again shows the same order of convergence as for SOR.

A more efficient procedure is obtained by using varying acceleration parameters τ_j, $j = 1, 2, \ldots$, corresponding to the n step error reduction matrix $\tilde{R}_n = \prod_{j=1}^n R_{\tau_j}$. It can be shown that the τ_j can be chosen cyclically with period m in such a way that $m \approx c\log\kappa \approx c\log(1/h)$, so that the average error reduction rate is

$$\|\tilde{R}_m\|^{1/m} = \max_{1\le i\le M}\Big(\prod_{j=0}^{m-1}\Big|\frac{\tau_j - \mu_i}{\tau_j + \mu_i}\Big|\Big)^{2/m} \le 1 - c(\log(1/h))^{-1}, \quad c > 0.$$

The analysis indicated depends strongly on the fact that H and V commute, which only happens for rectangles and constant coefficients, but the method may be defined and shown convergent in more general cases.

B.4 Preconditioned Conjugate Gradient Methods

We now turn to some iterative methods for systems mainly associated with the emergence of the finite element method. We begin by describing the conjugate gradient method, and assume that A is SPD. Considering the iterative method for (B.1) defined by

$$U^{n+1} = (I - \tau_n A)U^n + \tau_n b \quad \text{for } n \ge 0, \quad \text{with } U^0 = 0,$$

we find at once that, for any choice of the parameters τ_j, U^n belongs to the so-called Krylov space $K_n(A; b) = \text{span}\{b, Ab, \ldots, A^{n-1}b\}$, i.e., consisting of linear combinations of the $A^i b$, $i = 0, \ldots, n-1$. The conjugate gradient method defines these parameters so that U^n is the best approximation of the exact solution U of (B.1) in $K_n(A; b)$ with respect to the norm defined by $|U| = (AU, U)^{1/2}$, i.e., U^n is the orthogonal projection of U onto $K_n(A; b)$ with respect to the inner product (AV, W). By our above discussion it follows that, with $\kappa = \kappa(A)$ the condition number of A,

$$|U^n - U| \le (T_n(1/\rho))^{-1}|U| \le 2\Big(\frac{\kappa^{1/2} - 1}{\kappa^{1/2} + 1}\Big)^n |U|. \tag{B.5}$$

The computation of U^n can be carried out by a two term recurrence relation, for instance, in the following form using the residuals $r^n = b - AU^n$ and the auxiliary vectors $q^n \in K_{n+1}(A; b)$, orthogonal to $K_n(A; b)$,

$$U^{n+1} = U^n + \frac{(r^n, q^n)}{(Aq^n, q^n)} q^n, \quad q^{n+1} = r^{n+1} - \frac{(Ar^{n+1}, q^n)}{(Aq^n, q^n)} q^n, \quad U^0 = 0,\ q^0 = b.$$

In the preconditioned conjugate gradient (PCG) method the conjugate gradient method is applied to equation (B.1) after multiplication by some SPD approximation B of A^{-1}, which is easier to determine than A^{-1}, so that the equation (B.1) may be written $BAU = Bb$. We note that BA is SPD with respect to the inner product $(B^{-1}V, W)$. The error estimate (B.5) is now valid in the corresponding norm with $\kappa = \kappa(BA)$; B would be chosen so that this condition number is smaller than $\kappa(A)$. For the recursion formulas the only difference is that now $r^n = B(b - AU^n)$ and $q^0 = Bb$.

B.5 Multigrid and Domain Decomposition Methods

In the case that the system (B.1) comes from a standard finite element problem, one way of defining a preconditioner as an approximate inverse of A is by means of the multigrid method. This method is based on the observation that large components of the errors are associated with low frequencies in a spectral representation. The basic idea is then to work in a systematic way with a sequence of triangulations and to reduce the low frequency errors on coarse triangulations, which corresponds to small size problems, and to reduce the higher frequency, or oscillatory, residual errors on finer triangulations by a smoothing operator, such as a step of the Jacobi method, which is relatively inexpensive.

Assuming that Ω is a plane polygonal domain we may, for instance, proceed as follows. We first perform a coarse triangulation of Ω. Each of the triangles is then divided into four similar triangles, and this process is repeated, which after a finite number M of steps leads to a fine triangulation with each of the original triangles divided into 4^M small triangles. It is on this fine triangulation which we want to use the finite element method, and thus to define an iterative method. To find the next iterate U^{n+1} from U^n we start at the finest triangulation and go recursively from one level of fineness to the previous in three steps:

1. A preliminary smoothing on the finer of the present triangulations.
2. Correction on the coarser triangulation by solving a residual equation.
3. A postsmoothing on the finer triangulation.

The execution of step 2 is thus itself carried out in three steps, starting with a smoothing on the present level and going to step 2 on the next coarser level, until one arrives at the original coarse triangulation, where the corresponding

residual equation is solved exactly. Postsmoothing on successive finer levels then completes the algorithm for computing the next iterate U^{n+1}. This particular procedure is referred to as the V-cycle algorithm. It turns out that, under the appropriate assumptions, the error reduction matrix R satisfies $\|R\| \leq \rho < 1$, with ρ independent of M, i.e., of h, and that the number of operations is of order $O(N)$, where $N = O(h^{-2})$ is the dimension of the matrix associated with the finest triangulation.

A class of iterative methods that have attracted a lot of attention recently is the so called domain decomposition methods. These assume that the domain Ω in which we want to solve our elliptic problem may be decomposed into subdomains $\Omega_j,\ j = 1, \ldots, M$, which could overlap. The idea is to reduce the boundary value problem on Ω into problems on each of the Ω_j, which are then coupled by their values on the intersections. The problems on the Ω_j could be solved independently on parallel processors. This is particularly efficient when the individual problems may be solved very fast, e.g., by fast transform methods.

The domain decomposition methods go back to the Schwarz alternating procedure, in which $\Omega = \Omega_1 \cup \Omega_2$ for two overlapping domains Ω_1 and Ω_2. Considering the Dirichlet problem (1.1) and (1.2) on Ω (with $g = 0$ on Γ) one defines a sequence $\{u^k\}_{k=0}^{\infty}$ starting with a given u^0 vanishing on $\partial\Omega$, by

$$
\begin{aligned}
-\Delta u^{2k+1} &= f && \text{in } \Omega_1,\\
u^{2k+1} &= \begin{cases} u^{2k} & \text{on } \partial\Omega_1 \cap \Omega_2,\\ 0 & \text{on } \partial\Omega_1 \cap \partial\Omega, \end{cases}\\
-\Delta u^{2k+2} &= f && \text{in } \Omega_2,\\
u^{2k+2} &= \begin{cases} u^{2k+1} & \text{on } \partial\Omega_2 \cap \Omega_1,\\ 0 & \text{on } \partial\Omega_2 \cap \partial\Omega, \end{cases}
\end{aligned}
$$

and this procedure can be combined with numerical solution by, e.g., finite elements.

The following alternative approach may be pursued when Ω_1 and Ω_2 are disjoint but with a common interface $\partial\Omega_1 \cap \partial\Omega_2$: If u_j denotes the solution in $\Omega_j,\ j = 1, 2$, then the transmission conditions $u_1 = u_2,\ \partial u_1/\partial n = \partial u_2/\partial n$ have to be satisfied on the interface. One method is then to reduce the problem to an integral type equation on the interface and use this as a basis of an iterative method.

Bibliography

Partial Differential Equations

R. Dautray and 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 and 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 and R. C. Rogers, *An Introduction to Partial Differential Equations*, Springer-Verlag, New York, 1993.

Functional Analysis

L. Debnath and 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.

Finite Element Methods

D. Braess, *Finite Elements*, second ed., Cambridge University Press, Cambridge, 2001.

S. C. Brenner and 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, and C. Johnson, *Introduction to adaptive methods for differential equations*, Acta Numerica, 1995, Cambridge Univ. Press, Cambridge, 1995, pp. 105–158.

G. Strang and 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 and R. L. Taylor, *The Finite Element Method. Vol. 1–3*, Fifth edition, Butterworth-Heinemann, Oxford, 2000.

Finite Difference Methods

G. E. Forsythe and W. R. Wasow, *Finite-Difference Methods for Partial Differential Equations*, John Wiley & Sons Inc., New York, 1960.

B. Gustafsson, H.-O. Kreiss, and J. Oliger, *Time Dependent Problems and Difference Methods*, John Wiley & Sons Inc., New York, 1995.

R. D. Richtmyer and 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.

Other Classes of Numerical Methods

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, and T. A. Zang, *Spectral Methods in Fluid Dynamics*, Springer-Verlag, New York, 1988.

G. Chen and J. Zhou, *Boundary Element Methods*, Computational Mathematics and Applications, Academic Press, London, 1992.

J. Douglas, Jr. and T. Dupont, *Collocation Methods for Parabolic Equations in a Single Space Variable*, Springer-Verlag, Berlin, 1974, Lecture Notes in Mathematics, Vol. 385.

D. Gottlieb and S. A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*, Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1977.

R. Li, Z. Chen, and 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 and 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.

Numerical Linear 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 and A. Hohmann, *Numerical Analysis in Modern Scientific Computing*, second ed., Springer, New York, 2003.

G. H. Golub and C. F. Van Loan, *Matrix Computations*, third ed., Johns Hopkins University Press, Baltimore, MD, 1996.

A. Quarteroni and A. Valli, *Domain Decomposition Methods for Partial Differential Equations*, Oxford University Press, New York, 1999.

B. F. Smith, P. E. Bjørstad, and W. D. Gropp, *Domain Decomposition*, Cambridge University Press, Cambridge, 1996.

L. N. Trefethen and 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

Texts in Applied Mathematics

(continued from page ii)

34. *Chicone:* Ordinary Differential Equations with Applications, Second Edition.
35. *Kevorkian:* Partial Differential Equations: Analytical Solution Techniques, Second Edition.
36. *Dullerud/Paganini:* A Course in Robust Control Theory: A Convex Approach.
37. *Quarteroni/Sacco/Saleri:* Numerical Mathematics.
38. *Gallier:* Geometric Methods and Applications: For Computer Science and Engineering.
39. *Atkinson/Han:* Theoretical Numerical Analysis: A Functional Analysis Framework, Second Edition.
40. *Brauer/Castillo-Chávez:* Mathematical Models in Population Biology and Epidemiology.
41. *Davies:* Integral Transforms and Their Applications, Third Edition.
42. *Deuflhard/Bornemann:* Scientific Computing with Ordinary Differential Equations.
43. *Deuflhard/Hohmann:* Numerical Analysis in Modern Scintific Computing: An Introduction, Second Edition.
44. *Knabner/Angermann:* Numerical Methods for Elliptic and Parabolic Partial Differential Equations.
45. *Larsson/Thomée:* Partial Differential Equations with Numerical Methods.
46. *Pedregal:* Introduction to Optimization.
47. *Ockendon/Ockendon:* Waves and Compressible Flow.
48. *Hinrichsen/Pritchard:* Mathematical Systems Theory I.
49. *Bullo/Lewis:* Geometric Control of Mechanical Systems: Modeling, Analysis, and Design for Simple Mechanical Control Systems.
50. *Verhulst:* Methods and Applications of Singular Perturbations: Boundary Layers and Multiple Timescale Dynamics.
51. *Bondeson/Rylander/Ingelström:* Computational Electromagnetics.
52. *Holmes:* Introduction to Numerical Methods in Differential Equations.
53. *Pavliotis/Stuart:* Multiscale Methods: Averaging and Homogenization.
54. *Hesthaven/Warburton:* Nodal Discontinuous Galerkin Methods.
55. *Allaire/Kaber:* Numerical Linear Algebra.