25 Springer Series in Computational Mathematics

Galerkin Finite Element Methods for Parabolic Problems

Vidar Thomée

Second Edition

Springer

Springer Series in
Computational
Mathematics

25

Vidar Thomée

Galerkin Finite Element Methods for Parabolic Problems

Second Edition

Springer

Vidar Thomée

Department of Mathematics
Chalmers University of Technology
S-41296 Göteborg
Sweden
email: thomee@math.chalmers.se

Library of Congress Control Number: 2006925896

Mathematics Subject Classification (2000): 65M60, 65M12, 65M15

ISSN 0179-3632
ISBN-10 3-540-33121-2 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-33121-6 Springer Berlin Heidelberg New York
ISBN-10 3-540-63236-0 1st Edition Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media
springer.com

Printed in The Netherlands

Typesetting: by the author and techbooks using a Springer LaTeX macro package

Cover design: *design & production* GmbH, Heidelberg

Printed on acid-free paper SPIN: 11693536 46/techbooks 5 4 3 2 1 0

Preface

My purpose in this monograph is to present an essentially self-contained account of the mathematical theory of Galerkin finite element methods as applied to parabolic partial differential equations. The emphases and selection of topics reflects my own involvement in the field over the past 25 years, and my ambition has been to stress ideas and methods of analysis rather than to describe the most general and farreaching results possible. Since the formulation and analysis of Galerkin finite element methods for parabolic problems are generally based on ideas and results from the corresponding theory for stationary elliptic problems, such material is often included in the presentation.

The basis of this work is my earlier text entitled *Galerkin Finite Element Methods for Parabolic Problems,* Springer Lecture Notes in Mathematics, No. 1054, from 1984. This has been out of print for several years, and I have felt a need and been encouraged by colleagues and friends to publish an updated version. In doing so I have included most of the contents of the 14 chapters of the earlier work in an updated and revised form, and added four new chapters, on semigroup methods, on multistep schemes, on incomplete iterative solution of the linear algebraic systems at the time levels, and on semilinear equations. The old chapters on fully discrete methods have been reworked by first treating the time discretization of an abstract differential equation in a Hilbert space setting, and the chapter on the discontinuous Galerkin method has been completely rewritten.

The following is an outline of the contents of the book:

In the introductory Chapter 1 we begin with a review of standard material on the finite element method for Dirichlet's problem for Poisson's equation in a bounded domain, and consider then the simplest Galerkin finite element methods for the corresponding initial-boundary value problem for the linear heat equation. The discrete methods are based on associated weak, or variational, formulations of the problems and employ first piecewise linear and then more general approximating functions which vanish on the boundary of the domain. For these model problems we demonstrate the basic error estimates in energy and mean square norms, in the parabolic case first for the semidiscrete problem resulting from discretization in the spatial variables only, and then also for the most commonly used fully discrete schemes obtained by discretization in both space and time, such as the backward Euler and Crank-Nicolson methods.

In the following five chapters we study several extensions and generalizations of the results obtained in the introduction in the case of the spatially semidiscrete approximation, and show error estimates in a variety of norms. First, in Chapter 2, we formulate the semidiscrete problem in terms of a more general approximate solution operator for the elliptic problem in a manner which does not require the approximating functions to satisfy the homogeneous boundary conditions. As an example of such a method we discuss a method of Nitsche based on a nonstandard weak formulation. In Chapter 3 more precise results are shown in the case of the homogeneous heat equation. These results are expressed in terms of certain function spaces $\dot{H}^s(\Omega)$ which are characterized by both smoothness and boundary behavior of its elements, and which will be used repeatedly in the rest of the book. We also demonstrate that the smoothing property for positive time of the solution operator of the initial value problem has an analogue in the semidiscrete situation, and use this to show that the finite element solution converges to full order even when the initial data are nonsmooth. The results of Chapters 2 and 3 are extended to more general linear parabolic equations in Chapter 4. Chapter 5 is devoted to the derivation of stability and error bounds with respect to the maximum-norm for our plane model problem, and in Chapter 6 negative norm error estimates of higher order are derived, together with related results concerning superconvergence.

In the next six chapters we consider fully discrete methods obtained by discretization in time of the spatially semidiscrete problem. First, in Chapter 7, we study the homogeneous heat equation and give analogues of our previous results both for smooth and for nonsmooth data. The methods used for time discretization are of one-step type and rely on rational approximations of the exponential, allowing the standard Euler and Crank-Nicolson procedures as special cases. Our approach here is to first discretize a parabolic equation in an abstract Hilbert space framework with respect to time, and then to apply the results obtained to the spatially semidiscrete problem. The analysis uses eigenfunction expansions related to the elliptic operator occurring in the parabolic equation, which we assume positive definite. In Chapter 8 we generalize the above abstract considerations to a Banach space setting and allow a more general parabolic equation, which we now analyze using the Dunford-Taylor spectral representation. The time discretization is interpreted as a rational approximation of the semigroup generated by the elliptic operator, i.e., the solution operator of the initial-value problem for the homogeneous equation. Application to maximum-norm estimates is discussed. In Chapter 9 we study fully discrete one-step methods for the inhomogeneous heat equation in which the forcing term is evaluated at a fixed finite number of points per time stepping interval. In Chapter 10 we apply Galerkin's method also for the time discretization and seek discrete solutions as piecewise polynomials in the time variable which may be discontinuous at the now not necessarily equidistant nodes. In this *discontinuous Galerkin* procedure the forcing term enters in integrated form rather than at a finite number of points. In Chapter 11 we consider multistep backward difference methods. We first study such methods with constant time steps of order at most 6, and show stability as well as smooth and nonsmooth data error estimates, and then discuss the second order backward difference method with variable time steps. In Chapter 12 we study the incomplete iterative solution of the finite dimensional linear systems of algebraic equations which need to be solved at each level of the time stepping procedure, and exemplify by the use of a V-cycle multigrid algorithm.

The next two chapters are devoted to nonlinear problems. In Chapter 13 we discuss the application of the standard Galerkin method to a model nonlinear parabolic equation. We show error estimates for the spatially semidiscrete problem as well as the fully discrete backward Euler and Crank-Nicolson methods, using piecewise linear finite elements, and then pay special attention to the formulation and analysis of time stepping procedures based on these, which are linear in the unknown functions. In Chapter 14 we derive various results in the case of semilinear equations, in particular concerning the extension of the analysis for nonsmooth initial data from the case of linear homogenous equations.

In the last four chapters we consider various modifications of the standard Galerkin finite element method. In Chapter 15 we analyze the so called lumped mass method for which in certain cases a maximum-principle is valid. In Chapter 16 we discuss the H^1 and H^{-1} methods. In the first of these, the Galerkin method is based on a weak formulation with respect to an inner product in H^1 and for the second, the method uses trial and test functions from different finite dimensional spaces. In Chapter 17, the approximation scheme is based on a mixed formulation of the initial boundary value problem in which the solution and its gradient are sought independently in different spaces. In the final Chapter 18 we consider a singular problem obtained by introducing polar coordinates in a spherically symmetric problem in a ball in $\mathbf{R}^3$ and discuss Galerkin methods based on two different weak formulations defined by two different inner products.

References to the literature where the reader may find more complete treatments of the different topics, and some historical comments, are given at the end of each chapter.

A desirable mathematical background for reading the text includes standard basic partial differential equations and functional analysis, including Sobolev spaces; for the convenience of the reader we often give references to the literature concerning such matters.

The work presented, first in the Lecture Notes and now in this monograph, has grown from courses, lecture series, summer-schools, and written material that I have been involved in over a long period of time. I wish to thank my students and colleagues in these various contexts for the inspiration and support they have provided, and for the help they have given me as discussion partners and critics. As regards this new version of my work I particularly address my thanks to Georgios Akrivis, Stig Larsson, and Per-Gunnar Martinsson, who have read the manuscript in various degrees of detail and are responsible for many improvements. I also want to express my special gratitude to Yumi Karlsson who typed a first version of the text from the old lecture notes, and to Gunnar Ekolin who generously furnished me with expert help with the intricacies of TEX.

Göteborg *Vidar Thomée*
July 1997

Preface to the Second Edition

I am pleased to have been given the opportunity to prepare a second edition of this book. In doing so, I have kept most of the text essentially unchanged, but after correcting a number or typographical errors and other minor inadequacies, I have also taken advantage of this possibility to include some new material representing work that I have been involved in since the time when the original version appeared about eight years ago.

This concerns in particular progress in the application of semigroup theory to stability and error analysis. Using the theory of analytic semigroups it is convenient to reformulate the stability and smoothing properties as estimates for the resolvent of the associated elliptic operator and its discrete analogue. This is particularly useful in deriving maximum-norm estimates, and has led to improvements for both spatially semidiscrete and fully discrete problems. For this reason a somewhat expanded review of analytic semigroups is given in the present Chapter 6, on maximum-norm estimates for the semidiscrete problem, where now resolvent estimates for piecewise linear finite elements are discussed in some detail. These changes have affected the chapter on single step time stepping methods, expressed as rational approximation of semigroups, now placed as Chapter 9. The new emphasis has led to certain modifications and additions also in other chapters, particularly in Chapter 10 on multistep methods and Chapter 15 on the lumped mass method.

I have also added two chapter at the end of the book on other topics of recent interest to me. The first of these, Chapter 19, concern problems in which the spatial domain is polygonal, with particular attention given to noncovex such domains. rather than with smooth boundary, as in most of the rest of the book. In this case the corners generate singularites in the exact solution, and we study the effect of these on the convergence of the finite element solution.

The second new chapter, Chapter 20, considers an alternative to time stepping as a method for discretization in time, which is based on representing the solution as an integral involving the resolvent of the elliptic operator along a smooth curve extending into the right half of the complex plane, and then applying an accurate quadrature rule to this integral. This reduces the parabolic problem to a finite set of elliptic problems that may be solved in parallel. The method is then combined with finite element discretization in the spatial variable. When applicable, this method gives very accurate approximations of the exact solution in an efficient way.

I would like to take this opportunity to express my warm gratitude to Georgios Akrivis for his generous help and support. He has critically read through the new material and made many valuable suggestions.

Göteborg *Vidar Thomée*
March 2006

Table of Contents

1. The Standard Galerkin Method

In this introductory chapter we shall study the standard Galerkin finite element method for the approximate solution of the model initial-boundary value problem for the heat equation,

$$(1.1) \qquad \begin{aligned} u_t - \Delta u &= f \quad \text{in } \Omega, \qquad \text{for } t > 0,\\ u &= 0 \quad \text{on } \partial\Omega, \quad \text{for } t > 0, \quad \text{with } u(\cdot, 0) = v \quad \text{in } \Omega, \end{aligned}$$

where Ω is a domain in $\mathbb{R}^d$ with smooth boundary $\partial\Omega$, and where $u = u(x,t)$, u_t denotes $\partial u/\partial t$, and $\Delta = \sum_{j=1}^d \partial^2/\partial x_j^2$ the Laplacian.

Before we start to discuss this problem we shall briefly review some basic relevant material about the finite element method for the corresponding stationary problem, the Dirichlet problem for Poisson's equation,

$$(1.2) \qquad -\Delta u = f \quad \text{in } \Omega, \quad \text{with } u = 0 \quad \text{on } \partial\Omega.$$

Using a variational formulation of this problem, we shall define an approximation of the solution u of (1.2) as a function u_h which belongs to a finite-dimensional linear space S_h of functions of x with certain properties. This function, in the simplest case a continuous, piecewise linear function on some partition of Ω, will be a solution of a finite system of linear algebraic equations. We show basic error estimates for this approximate solution in energy and least square norms.

We shall then turn to the parabolic problem (1.1) which we first write in a weak form. We then proceed to discretize this problem, first in the spatial variable x, which results in an approximate solution $u_h(\cdot, t)$ in the finite element space S_h, for $t \geq 0$, as a solution of an initial value problem for a finite-dimensional system of ordinary differential equations. We then define the fully discrete approximation by application of some finite difference time stepping method to this finite dimensional initial value problem. This yields an approximate solution $U = U_h$ of (1.1) which belongs to S_h at discrete time levels. Error estimates will be derived for both the spatially and fully discrete solutions.

For a general $\Omega \subset \mathbb{R}^d$ we denote below by $\|\cdot\|$ the norm in $L_2 = L_2(\Omega)$ and by $\|\cdot\|_r$ that in the Sobolev space $H^r = H^r(\Omega) = W_2^r(\Omega)$, so that for real-valued functions v,

$$\|v\| = \|v\|_{L_2} = \Big(\int_\Omega v^2 \, dx \Big)^{1/2},$$

and, for r a positive integer,

$$\|v\|_r = \|v\|_{H^r} = \Big(\sum_{|\alpha| \le r} \|D^\alpha v\|^2 \Big)^{1/2}, \tag{1.3}$$

where, with $\alpha = (\alpha_1, \dots, \alpha_d)$, $D^\alpha = (\partial/\partial x_1)^{\alpha_1} \cdots (\partial/\partial x_d)^{\alpha_d}$ denotes an arbitrary derivative with respect to x of order $|\alpha| = \sum_{j=1}^d \alpha_j$, so that the sum in (1.3) contains all such derivatives of order at most r. We recall that for functions in $H_0^1 = H_0^1(\Omega)$, i.e., the functions v with $\nabla v = \text{grad}\ v$ in L_2 and which vanish on $\partial\Omega$, $\|\nabla v\|$ and $\|v\|_1$ are equivalent norms (Friedrichs' lemma, see, e.g., [42] or [51]), and that

$$c\|v\|_1 \le \|\nabla v\| \le \|v\|_1, \quad \forall v \in H_0^1, \qquad \text{with } c > 0. \tag{1.4}$$

Throughout this book c and C will denote positive constants, not necessarily the same at different occurrences, which are independent of the parameters and functions involved.

We shall begin by recalling some basic facts concerning the Dirichlet problem (1.2). We first write this problem in a weak, or variational, form: We multiply the elliptic equation by a smooth function φ which vanishes on $\partial\Omega$ (it suffices to require $\varphi \in H_0^1$), integrate over Ω, and apply Green's formula on the left-hand side, to obtain

$$(\nabla u, \nabla \varphi) = (f, \varphi), \quad \forall \varphi \in H_0^1, \tag{1.5}$$

where we have used the L_2 inner products,

$$(v, w) = \int_\Omega vw \, dx, \quad (\nabla v, \nabla w) = \int_\Omega \sum_{j=1}^d \frac{\partial v}{\partial x_j} \frac{\partial w}{\partial x_j} \, dx. \tag{1.6}$$

A function $u \in H_0^1$ which satisfies (1.5) is called a variational solution of (1.2). It is an easy consequence of the Riesz representation theorem that a unique such solution exists if $f \in H^{-1}$, the dual space of H_0^1. In this case (f, φ) denotes the value of the functional f at φ. Further, since we have assumed the boundary $\partial\Omega$ to be smooth, the solution u is smoother by two derivatives in L_2 than the right hand side f, which may be expressed in the form of the *elliptic regularity* inequality

$$\|u\|_{m+2} \le C\|\Delta u\|_m = C\|f\|_m, \quad \text{for any } m \ge -1. \tag{1.7}$$

In particular, using also Sobolev's embeddning theorem, this shows that the solution u belongs to C^∞ if f belongs to C^∞. We refer to, e.g., Evans [96] for such material.

We remark for later reference that, for $m = -1, 0$, (1.7) holds also in the case of a convex polygonal domain Ω, but that this is not true for nonconvex polygonal domains.

As a preparation for the definition of the finite element solution of (1.2), we consider briefly the approximation of smooth functions in Ω which vanish on $\partial\Omega$. For concreteness, we shall exemplify by piecewise linear functions in a convex plane domain.

Let thus Ω be a convex domain in the plane with smooth boundary $\partial\Omega$, and let $\mathcal{T}_h$ denote a partition of Ω into disjoint triangles τ such that no vertex of any triangle lies on the interior of a side of another triangle and such that the union of the triangles determine a polygonal domain $\Omega_h \subset \Omega$ with boundary vertices on $\partial\Omega$.

Let h denote the maximal length of the sides of the triangulation $\mathcal{T}_h$. Thus h is a parameter which decreases as the triangulation is made finer. We shall assume that the angles of the triangulations are bounded below by a positive constant, independently of h, and sometimes also that the triangulations are *quasiuniform* in the sense that the triangles of $\mathcal{T}_h$ are of essentially the same size, which we express by demanding that the area of $\tau \in \mathcal{T}_h$ is bounded below by ch^2, with $c > 0$, independent of h.

Let now S_h denote the continuous functions on the closure $\bar{\Omega}$ of Ω which are linear in each triangle of $\mathcal{T}_h$ and which vanish outside Ω_h. Let $\{P_j\}_{j=1}^{N_h}$ be the interior vertices of $\mathcal{T}_h$. A function in S_h is then uniquely determined by its values at the points P_j and thus depends on N_h parameters. Let Φ_j be the *pyramid function* in S_h which takes the value 1 at P_j but vanishes at the other vertices. Then $\{\Phi_j\}_{j=1}^{N_h}$ forms a *basis* for S_h, and every χ in S_h admits a unique representation

$$\chi(x) = \sum_{j=1}^{N_h} \alpha_j \Phi_j(x), \quad \text{with } \alpha_j = \chi(P_j).$$

A given smooth function v on Ω which vanishes on $\partial\Omega$ may now be approximated by, for instance, its interpolant $I_h v$ in S_h, which we define as the function in S_h which agrees with v at the interior vertices of $\mathcal{T}_h$, i.e.,

$$I_h v(x) = \sum_{j=1}^{N_h} v(P_j)\Phi_j(x). \tag{1.8}$$

Using this notation in our plane domain Ω, the following error estimates for the interpolant defined in (1.8) are well known (see, e.g., [42] or [51]), namely, for $v \in H^2 \cap H_0^1$,

$$\|I_h v - v\| \le Ch^2\|v\|_2 \quad \text{and} \quad \|\nabla(I_h v - v)\| \le Ch\|v\|_2. \tag{1.9}$$

They may be derived by showing the corresponding estimate for each $\tau \in \mathcal{T}_h$ and then taking squares and adding. For an individual $\tau \in \mathcal{T}_h$ the proof is achieved by means of the Bramble-Hilbert lemma (cf. [42] or [51]), noting that $I_h v - v$ vanishes on τ for v linear.

We shall now return to the general case of a domain Ω in $\mathbb{R}^d$ and assume that we are given a family $\{S_h\}$ of finite-dimensional subspaces of H_0^1 such that, for some integer $r \geq 2$ and small h,

$$\inf_{\chi \in S_h} \{\|v - \chi\| + h\|\nabla(v - \chi)\|\} \leq Ch^s \|v\|_s, \quad \text{for } 1 \leq s \leq r, \tag{1.10}$$

when $v \in H^s \cap H_0^1$. The number r is referred to as the order of accuracy of the family $\{S_h\}$. The above example of piecewise linear functions in a plane domain corresponds to $d = r = 2$. In the case $r > 2, S_h$ often consists of piecewise polynomials of degree at most $r - 1$ on a triangulation $\mathcal{T}_h$ as above. For instance, $r = 4$ in the case of piecewise cubic polynomial subspaces. Also, in the general situation estimates such as (1.10) may often be obtained by exhibiting an *interpolation operator* $I_h : H^r \cap H_0^1 \to S_h$ such that

$$\|I_h v - v\| + h\|\nabla(I_h v - v)\| \leq Ch^s \|v\|_s, \quad \text{for } 1 \leq s \leq r. \tag{1.11}$$

When $\partial\Omega$ is curved and $r > 2$ there are difficulties in the construction and analysis of such operators near the boundary, but this may be accomplished, in principle, by mapping a curved triangle onto a straight-edged one (isoparametric elements). We shall not dwell on this here, but return in Chapter 2 to this problem.

We remark for later reference that if the family $\{S_h\}$ is based on a family of *quasiuniform* triangulations $\mathcal{T}_h$ and S_h consists of piecewise polynomials of degree at most $r - 1$, then one may show the *inverse inequality*

$$\|\nabla\chi\| \leq Ch^{-1}\|\chi\|, \quad \forall \chi \in S_h. \tag{1.12}$$

This follows by taking squares and adding from the corresponding inequality for each triangle $\tau \in \mathcal{T}_h$, which in turn is obtained by a transformation to a fixed reference triangle, and using the fact that all norms on a finite dimensional space are equivalent, see, e.g., [51].

The optimal orders to which functions and their gradients may be approximated under our assumption (1.10) are $O(h^r)$ and $O(h^{r-1})$, respectively, and we shall now construct approximations to these orders of the solution of the Dirichlet problem (1.2) by the finite element method. The approximate problem is then to find a function $u_h \in S_h$ such that, cf., (1.5),

$$(\nabla u_h, \nabla\chi) = (f, \chi), \quad \forall \chi \in S_h. \tag{1.13}$$

This way of defining an approximate solution in terms of the variational formulation of the problem is referred to as Galerkin's method, after the Russian applied mathematician Boris Grigorievich Galerkin (1871-1945).

Note that, as a result of (1.5) and (1.13),

$$(\nabla(u_h - u), \nabla\chi) = 0, \quad \forall \chi \in S_h, \tag{1.14}$$

that is, the error in the discrete solution is orthogonal to S_h with respect to the Dirichlet inner product $(\nabla v, \nabla w)$.

In terms of a basis $\{\Phi_j\}_1^{N_h}$ for the finite element space S_h, our discrete problem may also be formulated: Find the coefficients α_j in $u_h(x) = \sum_{j=1}^{N_h} \alpha_j \Phi_j(x)$ such that

$$\sum_{j=1}^{N_h} \alpha_j (\nabla \Phi_j, \nabla \Phi_k) = (f, \Phi_k), \quad \text{for } k = 1, \dots, N_h.$$

In matrix notation this may be expressed as

$$\mathcal{A}\alpha = \widetilde{f},$$

where $\mathcal{A} = (a_{jk})$ is the *stiffness matrix* with elements $a_{jk} = (\nabla \Phi_j, \nabla \Phi_k)$, $\widetilde{f} = (f_k)$ the vector with entries $f_k = (f, \Phi_k)$, and α the vector of unknowns α_j. The dimensions of all of these arrays then equal N_h, the dimension of S_h (which equals the number of interior vertices in our plane example above). The stiffness matrix $\mathcal{A}$ is a Gram matrix and thus in particular positive definite and invertible so that our discrete problem has a unique solution. To see that $\mathcal{A} = (a_{jk})$ is positive definite, we note that

$$\sum_{j,k=1}^{d} a_{jk} \xi_j \xi_k = \|\nabla \Big(\sum_{j=1}^{d} \xi_j \Phi_j \Big)\|^2 \geq 0.$$

Here equality holds only if $\nabla(\sum_{j=1}^{d} \xi_j \Phi_j) \equiv 0$, so that $\sum_{j=1}^{d} \xi_j \Phi_j = 0$ by (1.4), and hence $\xi_j = 0, \ j = 1, \dots, N_h$.

When S_h consists of piecewise polynomial functions, the elements of the matrix $\mathcal{A}$ may be calculated exactly. However, unless f has a particularly simple form, the elements (f, Φ_j) of $\widetilde{f}$ have to be computed by some quadrature formula.

We shall prove the following estimate for the error between the solutions of the discrete and continuous problems. Note that these estimates are of optimal order as defined by our assumption (1.10). Here, as will always be the case in the sequel, the statements of the inequalities assume that u is sufficiently regular for the norms on the right to be finite.

We remark that since we require $\partial\Omega$ to be smooth, according to the elliptic regularity estimate (1.7), the solution of (1.2) can be guaranteed to have any degree of smoothness required by assuming the right hand side f to be sufficiently smooth. In particular, $u \in H^r \cap H_0^1$ if $f \in H^{r-2}$, and the solution u belongs to C^∞ if $\partial\Omega \in C^\infty$ and $f \in C^\infty$.

Theorem 1.1 *Assume that* (1.10) *holds, and let* u_h *and* u *be the solutions of* (1.13) *and* (1.2)*, respectively. Then, for* $1 \leq s \leq r$,

$$\|u_h - u\| \leq Ch^s \|u\|_s \quad \textit{and} \quad \|\nabla u_h - \nabla u\| \leq Ch^{s-1} \|u\|_s.$$

Proof. We start with the estimate for the error in the gradient. Since by (1.14) u_h is the orthogonal projection of u onto S_h with respect to the Dirichlet inner product, we have by (1.10)

$$\|\nabla(u_h - u)\| \le \inf_{\chi \in S_h} \|\nabla(\chi - u)\| \le Ch^{s-1}\|u\|_s. \tag{1.15}$$

For the error bound in L_2–norm we proceed by a duality argument. Let φ be arbitrary in L_2, take $\psi \in H^2 \cap H_0^1$ as the solution of

$$-\Delta\psi = \varphi \quad \text{in } \Omega, \quad \text{with } \psi = 0 \quad \text{on } \partial\Omega, \tag{1.16}$$

and recall the fact that by (1.7) the solution ψ of (1.16) is smoother by two derivatives in L_2 than the right hand side φ, so that

$$\|\psi\|_2 \le C\|\Delta\psi\| = C\|\varphi\|. \tag{1.17}$$

For any $\psi_h \in S_h$ we have

$$\begin{aligned}(u_h - u, \varphi) &= -(u_h - u, \Delta\psi) = (\nabla(u_h - u), \nabla\psi)\\ &= (\nabla(u_h - u), \nabla(\psi - \psi_h)) \le \|\nabla(u_h - u)\|\,\|\nabla(\psi - \psi_h)\|,\end{aligned} \tag{1.18}$$

and hence, using (1.15) together with (1.10) with $s = 2$ and (1.7) with $m = 0$,

$$(u_h - u, \varphi) \le Ch^{s-1}\|u\|_s\, h\|\psi\|_2 \le Ch^s\|u\|_s\|\varphi\|.$$

Choosing $\varphi = u_h - u$ completes the proof. □

After these preparations we now turn to the initial-boundary value problem (1.1) for the heat equation. As in the elliptic case we begin by writing the problem in weak form: We multiply the heat equation by a smooth function φ which vanishes on $\partial\Omega$ (or $\varphi \in H_0^1$), integrate over Ω, and apply Green's formula to the second term, to obtain, with (v, w) and $(\nabla v, \nabla w)$ as in (1.6),

$$(u_t, \varphi) + (\nabla u, \nabla\varphi) = (f, \varphi), \quad \forall \varphi \in H_0^1,\ t > 0. \tag{1.19}$$

We say that a function $u = u(x, t)$ is a weak solution of (1.1) on $[0, \bar{t}]$ if (1.19) holds with $u \in L_2(0, \bar{t}; H_0^1)$ and $u_t \in L_2(0, \bar{t}; H^{-1})$, and if $u(\cdot, 0) = v$. Again, since the boundary $\partial\Omega$ is smooth, such a solution is smooth provided the data are smooth functions, and in this case also satisfy certain compatibility conditions at $t = 0$. Similarly to (1.7) this may be expressed by a *parabolic regularity* estimate such as, cf. [96], with $u^{(j)} = (\partial/\partial t)^j u$ and $C = C_{m,\bar{t}}$,

$$\sum_{j=0}^{m+1} \int_0^{\bar{t}} \|u^{(j)}\|^2_{2(m-j)+2} dt \le C\Big(\|v\|^2_{2m+1} + \sum_{j=0}^{m} \int_0^{\bar{t}} \|f^{(j)}\|^2_{2(m-j)} dt\Big), \tag{1.20}$$

for $m \ge 0$. The compatibility conditions required express that the different conditions imposed in (1.1) at $\partial\Omega$ are consistent with each other. The first such condition, required for $m = 0$, is that since $u(t) = 0$ on $\partial\Omega$ for $t > 0$, then $u(0) = v$ also has to vanish on $\partial\Omega$. Next, for $m = 1$, since $u_t(t) = 0$ on $\partial\Omega$ for $t > 0$, smoothness requires that $u_t(0) = g := \Delta v + f(0) = 0$ on $\partial\Omega$, and similarly for $u^{(m)}(0)$ with $m \geq 2$. Again we refer to, e.g., Evans [96] for details.

As indicated above it is convenient to proceed in two steps with the derivation and analysis of the approximate solution of (1.1). In the first step we approximate $u(x,t)$ by means of a function $u_h(x,t)$ which, for each fixed t, belongs to a finite-dimensional linear space S_h of functions of x of the type considered above. This function will be a solution of an h-dependent finite system of ordinary differential equations in time and is referred to as a *spatially discrete*, or *semidiscrete*, solution. As in the elliptic case just considered, the spatially discrete problem is based on a weak formulation of (1.1). We then proceed to discretize this system in the time variable to obtain produce a *fully discrete* approximation of the solution of (1.1) by a *time stepping* method. In our basic schemes this discretization in time will be accomplished by a finite difference approximation of the time derivative.

We thus first pose the spatially semidiscrete problem, based on the weak formulation (1.19), to find $u_h(t) = u_h(\cdot,t)$, belonging to S_h for $t \geq 0$, such that

$$(1.21)\quad (u_{h,t},\chi) + (\nabla u_h, \nabla\chi) = (f,\chi), \quad \forall\, \chi \in\ S_h,\ t > 0, \quad \text{with } u_h(0) = v_h,$$

where v_h is some approximation of v in S_h.

In terms of the basis $\{\Phi_j\}_1^{N_h}$ for S_h, our semidiscrete problem may be stated: Find the coefficients $\alpha_j(t)$ in $u_h(x,t) = \sum_{j=1}^{N_h} \alpha_j(t)\Phi_j(x)$ such that

$$\sum_{j=1}^{N_h} \alpha_j'(t)(\Phi_j,\Phi_k) + \sum_{j=1}^{N_h} \alpha_j(t)(\nabla\Phi_j,\nabla\Phi_k) = (f,\Phi_k), \quad k = 1,\dots,N_h,$$

and, with γ_j the components of the given initial approximation v_h, $\alpha_j(0) = \gamma_j$ for $j = 1,\dots,N_h$. In matrix notation this may be expressed as

$$\mathcal{B}\alpha'(t) + \mathcal{A}\alpha(t) = \widetilde{f}(t), \quad \text{for } t > 0, \quad \text{with } \alpha(0) = \gamma,$$

where $\mathcal{B} = (b_{jk})$ is the *mass matrix* with elements $b_{jk} = (\Phi_j,\Phi_k)$, $\mathcal{A} = (a_{jk})$ the stiffness matrix with $a_{jk} = (\nabla\Phi_j,\nabla\Phi_k)$, $\widetilde{f} = (f_k)$ the vector with entries $f_k = (f,\Phi_k)$, $\alpha(t)$ the vector of unknowns $\alpha_j(t)$, and $\gamma = (\gamma_k)$. The dimension of all these items equals N_h, the dimension of S_h.

Since, like the stiffness matrix $\mathcal{A}$, the mass matrix $\mathcal{B}$ is a Gram matrix, and thus in particular positive definite and invertible, the above system of ordinary differential equations may be written

$$\alpha'(t) + \mathcal{B}^{-1}\mathcal{A}\alpha(t) = \mathcal{B}^{-1}\widetilde{f}(t), \quad \text{for } t > 0, \quad \text{with } \alpha(0) = \gamma,$$

and hence obviously has a unique solution for t positive.

Our first aim is to prove the following estimate in L_2 for the error between the solutions of the semidiscrete and continuous problems.

Theorem 1.2 *Let* u_h *and* u *be the solutions of* (1.21) *and* (1.1), *and assume* $v = 0$ *on* $\partial\Omega$. *Then*

$$\|u_h(t) - u(t)\| \le \|v_h - v\| + Ch^r\big(\|v\|_r + \int_0^t \|u_t\|_r \, ds\big), \quad \textit{for } t \ge 0.$$

Here as earlier we require that the solution of the continuous problem has the regularity implicitly assumed by the presence of the norms on the right. Note also that if (1.11) holds and $v_h = I_h v$, then the first term on the right is dominated by the second. This also holds if $v_h = P_h v$, where P_h denotes the orthogonal projection of v onto S_h with respect to the inner product in L_2, since this choice is the best approximation of v in S_h with respect to the L_2 norm, and thus at least as good as $I_h v$.

Another such optimal order choice for v_h is the so-called *elliptic* or *Ritz projection* R_h onto S_h which we define as the orthogonal projection with respect to the inner product $(\nabla v, \nabla w)$, so that

$$(1.22) \qquad (\nabla R_h v, \nabla \chi) = (\nabla v, \nabla \chi), \quad \forall\, \chi \in S_h, \quad \text{for } v \in H_0^1.$$

In view of (1.14), this definition may be expressed by saying that $R_h v$ is the finite element approximation of the solution of the corresponding elliptic problem with exact solution v. A pervading strategy throughout the error analysis in the rest of this book is to write the error in the parabolic problem as a sum of two terms,

$$(1.23) \quad u_h(t) - u(t) = \theta(t) + \rho(t), \quad \text{where } \theta = u_h - R_h u, \quad \rho = R_h u - u,$$

which are then bounded separately. The second term, $\rho(t)$, is thus the error in an elliptic problem and may be handled as such, whereas the first term $\theta(t)$ will be the main object of the analysis.

It follows at once from setting $\chi = R_h v$ in (1.22) that the Ritz projection is stable in H_0^1, or

$$(1.24) \qquad \|\nabla R_h v\| \le \|\nabla v\|, \quad \forall\, v \in H_0^1.$$

As an immediate consequence of Theorem 1.1 we have the following error estimate for $R_h v$.

Lemma 1.1 *Assume that* (1.10) *holds. Then, with* R_h *defined by* (1.22) *we have*

$$\|R_h v - v\| + h\|\nabla(R_h v - v)\| \le Ch^s\|v\|_s, \quad \textit{for } v \in H^s \cap H_0^1, \ 1 \le s \le r.$$

Proof of Theorem 1.2. We write the error according to (1.23) and obtain easily by Lemma 1.1 and obvious estimates

$$\|\rho(t)\| \le Ch^r\|u(t)\|_r \le Ch^r\big(\|v\|_r + \int_0^t \|u_t\|_r\, ds\big). \tag{1.25}$$

In order to bound θ, we note that by our definitions

$$\begin{aligned}(\theta_t,\chi) &+ (\nabla\theta,\nabla\chi)\\ &= (u_{h,t},\chi) + (\nabla u_h,\nabla\chi) - (R_h u_t,\chi) - (\nabla R_h u,\nabla\chi)\\ &= (f,\chi) - (R_h u_t,\chi) - (\nabla u,\nabla\chi) = (u_t - R_h u_t,\chi),\end{aligned} \tag{1.26}$$

or

$$(\theta_t,\chi) + (\nabla\theta,\nabla\chi) = -(\rho_t,\chi), \quad \forall\chi\in S_h,\ t>0, \tag{1.27}$$

where we have used the easily established fact that the operator R_h commutes with time differentiation. Since θ belongs to S_h, we may choose $\chi=\theta$ in (1.27) and conclude

$$(\theta_t,\theta) + \|\nabla\theta\|^2 = -(\rho_t,\theta), \quad \text{for } t>0. \tag{1.28}$$

Here the second is nonnegative, and we obtain thus

$$\tfrac12\frac{d}{dt}\|\theta\|^2 = \|\theta\|\,\frac{d}{dt}\|\theta\| \le \|\rho_t\|\,\|\theta\|,$$

and hence, after cancellation of one factor $\|\theta\|$ (the case that $\|\theta(t)\|=0$ for some t may easily be handled), and integration,

$$\|\theta(t)\| \le \|\theta(0)\| + \int_0^t \|\rho_t\|\, ds. \tag{1.29}$$

Here, using Lemma 1.1, we find

$$\|\theta(0)\| = \|v_h - R_h v\| \le \|v_h - v\| + \|R_h v - v\| \le \|v_h - v\| + Ch^r\|v\|_r,$$

and since

$$\|\rho_t\| = \|R_h u_t - u_t\| \le Ch^r\|u_t\|_r,$$

the desired bound for $\|\theta(t)\|$ now follows. □

In the above proof we have made use in (1.28) of the fact that $\|\nabla\theta\|^2$ is nonnegative, and simply discarded this term. By using it in a somewhat more careful way, one may demonstrate that the effect of the initial data upon the error tends to zero exponentially as t tends to ∞. In fact, with λ_1 the smallest eigenvalue of $-\Delta$, with Dirichlet boundary data, we have

$$\|\nabla v\|^2 \ge \lambda_1\|v\|^2, \quad \forall v\in H_0^1, \tag{1.30}$$

and hence (1.28) yields

$$\tfrac{1}{2}\frac{d}{dt}\|\theta\|^2 + \lambda_1\|\theta\|^2 \le \|\rho_t\|\|\theta\|.$$

It follows as above that

$$\frac{d}{dt}\|\theta\| + \lambda_1\|\theta\| \le \|\rho_t\|,$$

and hence

$$\begin{aligned}\|\theta(t)\| &\le e^{-\lambda_1 t}\|\theta(0)\| + \int_0^t e^{-\lambda_1(t-s)}\|\rho_t(s)\|\,ds \\ &\le e^{-\lambda_1 t}\|v_h - v\| + Ch^r\Big(e^{-\lambda_1 t}\|v\|_r + \int_0^t e^{-\lambda_1(t-s)}\|u_t(s)\|_r\,ds\Big).\end{aligned} \tag{1.31}$$

Using the first part of (1.25) we conclude that with v_h appropriately chosen

$$\|u_h(t) - u(t)\| \le Ch^r\Big(e^{-\lambda_1 t}\|v\|_r + \|u(t)\|_r + \int_0^t e^{-\lambda_1(t-s)}\|u_t(s)\|_r\,ds\Big).$$

We shall not pursue the error analysis for large t below.

We shall now briefly look at another way of expressing the argument in the proof of Theorem 1.2, which consists in working with the equation for θ in operator form. We first recall that by Duhamel's principle, the solution of (1.1) may be written

$$u(t) = E(t)v + \int_0^t E(t-s)f(s)\,ds. \tag{1.32}$$

Here $E(t)$ is the solution operator of the homogeneous equation, the case $f \equiv 0$ of (1.1), i.e., the operator which takes the initial values $u(0) = v$ into the solution $u(t)$ at time t. This operator may also be thought of as the *semigroup* $e^{\Delta t}$ on L_2 generated by the Laplacian, considered as defined in $\mathcal{D}(\Delta) = H^2 \cap H_0^1$. We now introduce a *discrete Laplacian* $\Delta_h : S_h \to S_h$ by

$$(\Delta_h\psi, \chi) = -(\nabla\psi, \nabla\chi), \quad \forall \psi, \chi \in S_h; \tag{1.33}$$

this analogue of Green's formula clearly defines $\Delta_h\psi = \sum_{j=1}^{N_h} d_j\Phi_j$ by

$$\sum_{j=1}^{N_h} d_j(\Phi_j, \Phi_k) = -(\nabla\psi, \nabla\Phi_k), \quad \text{for } k = 1, \dots, N_h,$$

since the matrix of this system is the positive definite mass matrix encountered above. The operator $-\Delta_h$ is easily seen to be selfadjoint and positive definite in S_h with respect to $(\cdot,\cdot)$. Note that Δ_h is related to our other operators by

$$\Delta_h R_h = P_h \Delta. \tag{1.34}$$

For, by our definitions,

$$(\Delta_h R_h v, \chi) = -(\nabla R_h v, \nabla\chi) = -(\nabla v, \nabla\chi) = (\Delta v, \chi) = (P_h \Delta v, \chi), \ \forall \chi \in S_h.$$

With this notation the semidiscrete equation takes the form

$$(u_{h,t}, \chi) - (\Delta_h u_h, \chi) = (P_h f, \chi), \quad \forall \chi \in S_h, \quad t > 0,$$

and thus, since the factors on the left are all in S_h, (1.21) may be written as

$$u_{h,t} - \Delta_h u_h = P_h f, \quad \text{for } t > 0, \quad \text{with } u_h(0) = v_h. \tag{1.35}$$

Using (1.34) we hence obtain for θ

$$\begin{aligned} \theta_t - \Delta_h \theta &= (u_{h,t} - \Delta_h u_h) - (R_h u_t - \Delta_h R_h u) \\ &= P_h f + (P_h - R_h) u_t - P_h(u_t - \Delta u) = P_h (I - R_h) u_t = -P_h \rho_t, \end{aligned}$$

or

$$\theta_t - \Delta_h \theta = -P_h \rho_t, \quad \text{for } t > 0, \quad \text{with } \theta(0) = v_h - R_h v. \tag{1.36}$$

We now denote by $E_h(t)$ the discrete analogue of the operator $E(t)$ introduced above, the solution operator of the homogeneous semidiscrete problem (1.35). The analogue of (1.32), together with (1.36), then shows

$$\theta(t) = E_h(t)\theta(0) - \int_0^t E_h(t-s) P_h \rho_t(s)\, ds. \tag{1.37}$$

We now note that $E_h(t)$ is stable in L_2, or, more precisely, as in the proof of (1.31),

$$\|E_h(t) v_h\| \le e^{-\lambda_1 t} \|v_h\| \le \|v_h\|, \quad \text{for } v_h \in S_h, \ t \ge 0. \tag{1.38}$$

Since obviously P_h has unit norm in L_2, (1.37) implies (1.29), from which Theorem 1.2 follows as above. The desired estimate for θ is thus a consequence of the stability estimate for $E_h(t)$ combined with the error estimate for the elliptic problem applied to $\rho_t = (R_h - I)u_t$.

In a similar way we may prove the following estimate for the error in the gradient.

Theorem 1.3 *Under the assumptions of Theorem 1.2 we have*

$$\begin{aligned} \|\nabla u_h(t) - \nabla u(t)\| &\le C\|\nabla v_h - \nabla v\| \\ &\quad + Ch^{r-1}\Big(\|v\|_r + \|u(t)\|_r + \big(\int_0^t \|u_t\|_{r-1}^2\, ds\big)^{1/2}\Big), \quad \textit{for } t \ge 0. \end{aligned}$$

Proof. As before we write the error in the form (1.23). Here, by Lemma 1.1,

$$\|\nabla\rho(t)\| = \|\nabla(R_h u(t) - u(t))\| \le Ch^{r-1}\|u(t)\|_r.$$

In order to estimate $\nabla\theta$, we use again (1.27), now with $\chi = \theta_t$. We obtain

$$\|\theta_t\|^2 + \tfrac{1}{2}\frac{d}{dt}\|\nabla\theta\|^2 = -(\rho_t, \theta_t) \le \tfrac{1}{2}\|\rho_t\|^2 + \tfrac{1}{2}\|\theta_t\|^2,$$

so that $(d/dt)\|\nabla\theta\|^2 \le \|\rho_t\|^2$ or

$$\begin{aligned}\|\nabla\theta(t)\|^2 &\le \|\nabla\theta(0)\|^2 + \int_0^t \|\rho_t\|^2\,ds \\ &\le \big(\|\nabla(v_h - v)\| + \|\nabla(R_h v - v)\|\big)^2 + \int_0^t \|\rho_t\|^2\,ds.\end{aligned} \tag{1.39}$$

Hence, in view of Lemma 1.1,

$$\|\nabla\theta(t)\|^2 \le 2\|\nabla(v_h - v)\|^2 + Ch^{2r-2}\Big(\|v\|_r^2 + \int_0^t \|u_t\|_{r-1}^2\,ds\Big), \tag{1.40}$$

which completes the proof. □

Note that if $v_h = I_h v$ or $v_h = R_h v$, then, by Lemma 1.1 or (1.11), respectively, the first term on the right hand side in Theorem 1.3 is again bounded by the second.

In the case of a quasiuniform family of triangulations $\mathcal{T}_h$ of a plane domain, or, more generally, when the inverse estimate (1.12) holds, an estimate for the error in the gradient may also be obtained directly from the result of Theorem 1.2. In fact, we obtain then, for χ arbitrary in S_h,

$$\begin{aligned}\|\nabla u_h(t) - \nabla u(t)\| &\le \|\nabla(u_h(t) - \chi)\| + \|\nabla\chi - \nabla u(t)\| \\ &\le Ch^{-1}\|u_h(t) - \chi\| + \|\nabla\chi - \nabla u(t)\| \\ &\le Ch^{-1}\|u_h(t) - u(t)\| + Ch^{-1}(\|\chi - u(t)\| + h\|\nabla\chi - \nabla u(t)\|).\end{aligned} \tag{1.41}$$

Here, by our approximation assumption (1.10), we have, with suitable $\chi \in S_h$,

$$\|\chi - u(t)\| + h\|\nabla\chi - \nabla u(t)\| \le Ch^r\|u(t)\|_r,$$

and hence, bounding the first term on the right in (1.41) by Theorem 1.2, for the appropriate choice of χ,

$$\|\nabla u_h(t) - \nabla u(t)\| \le Ch^{r-1}\Big(\|v\|_r + \int_0^t \|u_t(s)\|_r\,ds\Big).$$

We make the following observation concerning the gradient of the term $\theta = u_h - R_h u$ in (1.23): Assume that we have chosen $v_h = R_h v$ so that $\theta(0) = 0$. Then, in addition to (1.40), we have from (1.39)

$$\|\nabla\theta(t)\| \le C\Big(\int_0^t \|\rho_t\|^2\,ds\Big)^{1/2} \le Ch^r\Big(\int_0^t \|u_t\|_r^2\,ds\Big)^{1/2}. \tag{1.42}$$

Hence $\nabla\theta(t)$ is of order $O(h^r)$, whereas the gradient of the total error can only be $O(h^{r-1})$. 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* .

Because the formulation of Galerkin's method is posed in terms of L_2 inner products, the most natural error estimates are expressed in L_2-based norms. Error analyses in other norms have also been pursued in the literature, and for later reference we quote the following *maximum-norm error estimate*, for piecewise linear approximating functions in a plane domain Ω, see, e.g., [42]. Here we write $L_\infty = L_\infty(\Omega)$ and $W_\infty^r = W_\infty^r(\Omega)$, with

$$\|v\|_{L_\infty} = \sup_{x\in\Omega} |u(x)|, \quad \|v\|_{W_\infty^r} = \max_{|\alpha|\le r} \|D^\alpha v\|_{L_\infty}.$$

We note first that the error in the interpolant introduced above is second order also in maximum-norm, so that (cf. (1.9))

$$\|I_h v - v\|_{L_\infty} \le Ch^2\|v\|_{W_\infty^2}, \quad \text{for } v \in W_\infty^2 \cap H_0^1. \tag{1.43}$$

The error estimate for the elliptic finite element problem is then the following.

Theorem 1.4 *Let $\Omega \subset \mathbb{R}^2$ and assume that S_h consists of piecewise linear finite element functions, and that the family $\mathcal{T}_h$ is quasiuniform. Let u_h and u be the solutions of* (1.13) *and* (1.2)*, respectively. Then*

$$\|u_h - u\|_{L_\infty} \le Ch^2\ell_h\|u\|_{W_\infty^2}, \quad \textit{where } \ell_h = \max(1, \log(1/h)) \tag{1.44}$$

We note that, in view of (1.43), this error estimate is nonoptimal, but it has been shown, see Haverkamp [116], that the logarithmic factor in (1.44) cannot be removed. Note that although ℓ_h is unbounded for small h, it is of moderate size for realistic values of h.

Recall the definition (1.22) of the Ritz projection $R_h : H_0^1 \to S_h$, and its stability in H_0^1. When the family of triangulations is quasiuniform, this projection is known to have the almost maximum-norm stability property

$$\|R_h v\|_{L_\infty} \le C\ell_h\|v\|_{L_\infty}. \tag{1.45}$$

The proof of this is relatively difficult, and will not be included here. We remark that in contrast to (1.24) and (1.45), R_h is not bounded in L_2. The error bound of Theorem 1.4 is now an easy consequence of this stability result and the interpolation error estimate of (1.43), since

$$\|R_h v - v\|_{L_\infty} \le \|R_h(v - I_h v)\|_{L_\infty} + \|I_h v - v\|_{L_\infty} \le Ch^2\ell_h\|v\|_{W_\infty^2}.$$

As a simple example of an application of the superconvergent order estimate (1.42), we shall indicate briefly how it may be used to show an essentially optimal order error bound for the parabolic problem in the maximum-norm. Consider thus the concrete situation described in the beginning of this chapter with Ω a plane smooth convex domain and S_h consisting of piecewise linear functions ($d = r = 2$) on quasiuniform triangulations of Ω. Then, by Theorem 1.4,

$$\|\rho(t)\|_{L_\infty} = \|R_h u(t) - u(t)\|_{L_\infty} \le Ch^2 \ell_h \|u(t)\|_{W^2_\infty}. \tag{1.46}$$

In two dimensions, Sobolev's inequality almost bounds the maximum-norm by the norm in H^1, and it may be shown (cf. Lemma 6.4 below) that for functions in the subspace S_h,

$$\|\chi\|_{L_\infty} \le C\ell_h^{1/2} \|\nabla\chi\|, \quad \forall \chi \in S_h.$$

Applied to θ this shows, by (1.42) (with $r = 2$), that

$$\|\theta(t)\|_{L_\infty} \le Ch^2 \ell_h^{1/2} \Big(\int_0^t \|u_t\|_2^2 \, ds \Big)^{1/2},$$

and we may thus conclude for the error in the parabolic problem that

$$\|u_h(t) - u(t)\|_{L_\infty} \le \|\rho(t)\|_{L_\infty} + \|\theta(t)\|_{L_\infty} \le C(t,u) h^2 \ell_h.$$

We now turn our attention to some simple schemes for discretization also with respect to the time variable. We introduce a time step k and the time levels $t = t_n = nk$, where n is a nonnegative integer, and denote by $U^n = U_h^n \in S_h$ the approximation of $u(t_n)$ to be determined.

We begin by the *backward Euler Galerkin method*, which is defined by replacing the time derivative in (1.21) by a backward difference quotient, or, if $\bar\partial U^n = (U^n - U^{n-1})/k$,

$$(\bar\partial U^n, \chi) + (\nabla U^n, \nabla\chi) = (f(t_n), \chi), \quad \forall \chi \in S_h, \ n \ge 1, \quad U^0 = v_h. \tag{1.47}$$

For U^{n-1} given this defines U^n implicitly from the equation

$$(U^n, \chi) + k(\nabla U^n, \nabla\chi) = (U^{n-1} + kf(t_n), \chi), \quad \forall \chi \in S_h,$$

which is the finite element formulation of an elliptic equation of the form $(I - k\Delta)u = g$. With matrix notation as in the semidiscrete situation, this may be written

$$(\mathcal{B} + k\mathcal{A})\alpha^n = \mathcal{B}\alpha^{n-1} + k\widetilde{f}(t_n),$$

where $\mathcal{B} + k\mathcal{A}$ is positive definite and hence, in particular, invertible.

We shall prove the following error estimate.

Theorem 1.5 *With U^n and u the solutions of* (1.47) *and* (1.1)*, respectively, we have, if $\|v_h - v\| \le Ch^r\|v\|_r$ and $v = 0$ on $\partial\Omega$,*

$$\|U^n - u(t_n)\| \le Ch^r\big(\|v\|_r + \int_0^{t_n} \|u_t\|_r\,ds\big) + k\int_0^{t_n} \|u_{tt}\|\,ds, \quad \textit{for } n \ge 0.$$

Proof. In analogy with (1.23) we write

$$U^n - u(t_n) = (U^n - R_h u(t_n)) + (R_h u(t_n) - u(t_n)) = \theta^n + \rho^n,$$

and here $\rho^n = \rho(t_n)$ is bounded as claimed in (1.25). This time, a calculation corresponding to (1.26) yields

$$(1.48) \qquad (\bar\partial\theta^n, \chi) + (\nabla\theta^n, \nabla\chi) = -(\omega^n, \chi), \quad \forall\chi \in S_h,\ n \ge 1,$$

where

$$\omega^n = R_h\bar\partial u(t_n) - u_t(t_n) = (R_h - I)\bar\partial u(t_n) + (\bar\partial u(t_n) - u_t(t_n)) = \omega_1^n + \omega_2^n.$$

Choosing $\chi = \theta^n$ in (1.48), we have $(\bar\partial\theta^n, \theta^n) \le \|\omega^n\|\,\|\theta^n\|$, or

$$\|\theta^n\|^2 - (\theta^{n-1}, \theta^n) \le k\|\omega^n\|\,\|\theta^n\|,$$

so that

$$(1.49) \qquad \|\theta^n\| \le \|\theta^{n-1}\| + k\|\omega^n\|,$$

and, by repeated application,

$$(1.50) \qquad \|\theta^n\| \le \|\theta^0\| + k\sum_{j=1}^n \|\omega^j\| \le \|\theta^0\| + k\sum_{j=1}^n \|\omega_1^j\| + k\sum_{j=1}^n \|\omega_2^j\|.$$

Here, as before, $\theta^0 = \theta(0)$ is bounded as desired. We write

$$(1.51) \qquad \omega_1^j = (R_h - I)k^{-1}\int_{t_{j-1}}^{t_j} u_t\,ds = k^{-1}\int_{t_{j-1}}^{t_j} (R_h - I)u_t\,ds,$$

and obtain

$$k\sum_{j=1}^n \|\omega_1^j\| \le \sum_{j=1}^n \int_{t_{j-1}}^{t_j} Ch^r\|u_t\|_r\,ds = Ch^r\int_0^{t_n} \|u_t\|_r\,ds.$$

Further,

$$(1.52) \qquad k\,\omega_2^j = u(t_j) - u(t_{j-1}) - ku_t(t_j) = -\int_{t_{j-1}}^{t_j} (s - t_{j-1})u_{tt}(s)\,ds,$$

so that

$$k\sum_{j=1}^n \|\omega_2^j\| \le \sum_{j=1}^n \|\int_{t_{j-1}}^{t_j} (s - t_{j-1})u_{tt}(s)\,ds\| \le k\int_0^{t_n} \|u_{tt}\|\,ds.$$

Together our estimates complete the proof of the theorem. □

In order to show an estimate for $\nabla\theta^n$ we may choose instead $\chi = \bar\partial\theta^n$ in (1.48) to obtain $\bar\partial\|\nabla\theta^n\|^2 \le \|\omega^n\|^2$, or, if $\nabla\theta^0 = 0$,

$$\|\nabla\theta^n\|^2 \le k\sum_{j=1}^{n}\|\omega^j\|^2 \le Ch^{2s}\int_0^{t_n}\|u_t\|_s^2\,dt + Ck^2\int_0^{t_n}\|u_{tt}\|^2\,dt, \tag{1.53}$$

for $1 \le s \le r$. Together with the standard estimate for $\nabla\rho$ this shows, with $s = r-1$ in (1.53),

$$\|\nabla(U^n - u(t_n))\| \le C(u)(h^{r-1} + k).$$

If one uses Theorem 1.5 together with the inverse inequality (1.12) one now obtains the weaker estimate $\|\nabla(U^n - u(t_n))\| \le C(u)(h^{r-1} + kh^{-1})$. We also note that with $s = r$ in (1.53) one may conclude the maximum-norm estimate

$$\|U^n - u(t_n)\|_{L_\infty} \le C(u)\ell_h(h^r + k).$$

Note that because of the nonsymmetric choice of the discretization in time, the backward Euler Galerkin method is only first order in k. We therefore now turn to the *Crank-Nicolson Galerkin method.* Here the semidiscrete equation is discretized in a symmetric fashion around the point $t_{n-\frac12} = (n-\frac12)k$, which will produce a second order accurate method in time. More precisely, we set $\widehat{U}^n = \frac12(U^n + U^{n-1})$ and define $U^n \in S_h$ by

$$(\bar\partial U^n, \chi) + (\nabla\widehat{U}^n, \nabla\chi) = (f(t_{n-\frac12}), \chi), \quad \forall\chi \in S_h, \quad \text{for } n \ge 1, \tag{1.54}$$

with $U^0 = v_h$. Here the equation for U^n may be written in matrix form as

$$(\mathcal{B} + \tfrac12 k\mathcal{A})\alpha^n = (\mathcal{B} - \tfrac12 k\mathcal{A})\alpha^{n-1} + k\widetilde{f}(t_{n-\frac12}),$$

with a positive definite matrix $\mathcal{B} + \frac12 k\mathcal{A}$. Now the error estimate reads as follows.

Theorem 1.6 *Let U^n and u be the solutions of* (1.54) *and* (1.1)*, respectively, and let $\|v_h - v\| \le Ch^r\|v\|_r$ and $v = 0$ on $\partial\Omega$. Then we have, for $n \ge 0$,*

$$\|U^n - u(t_n)\| \le Ch^r\Big(\|v\|_r + \int_0^{t_n}\|u_t\|_r\,ds\Big) + Ck^2\int_0^{t_n}(\|u_{ttt}\| + \|\Delta u_{tt}\|)\,ds.$$

Proof. With ρ^n bounded as above, we only need to consider θ^n. We have

$$(\bar\partial\theta^n, \chi) + (\nabla\widehat{\theta}^n, \nabla\chi) = -(\omega^n, \chi), \quad \text{for } \chi \in S_h,\ n \ge 1, \tag{1.55}$$

where now

$$(1.56)\qquad \begin{aligned}\omega^n =&(R_h - I)\bar\partial u(t_n) + (\bar\partial u(t_n) - u_t(t_{n-\frac12}))\\ &+ \Delta\big(u(t_{n-\frac12}) - \tfrac12(u(t_n)+u(t_{n-1}))\big) = \omega_1^n + \omega_2^n + \omega_3^n.\end{aligned}$$

Choosing this time $\chi = \widehat{\theta}^n$ in (1.55), we find

$$(\bar\partial\theta^n, \widehat{\theta}^n) \le \tfrac12\|\omega^n\|(\|\theta^n\| + \|\theta^{n-1}\|),$$

or

$$\|\theta^n\|^2 - \|\theta^{n-1}\|^2 \le k\|\omega^n\|(\|\theta^n\| + \|\theta^{n-1}\|),$$

so that, after cancellation of a common factor,

$$\|\theta^n\| \le \|\theta^{n-1}\| + k\|\omega^n\|, \quad \text{for } n \ge 1.$$

After repeated application this yields

$$\|\theta^n\| \le \|\theta^0\| + k\sum_{j=1}^n \big(\|\omega_1^j\| + \|\omega_2^j\| + \|\omega_3^j\|\big).$$

With θ^0 and ω_1^j estimated as before, it remains to bound the terms in ω_2^j and ω_3^j. We have

$$\begin{aligned}k\|\omega_2^j\| &= \|u(t_j) - u(t_{j-1}) - ku_t(t_{j-\frac12})\|\\ &= \tfrac12\Big\|\int_{t_{j-1}}^{t_{j-\frac12}} (s-t_{j-1})^2 u_{ttt}(s)\,ds + \int_{t_{j-\frac12}}^{t_j} (s-t_j)^2 u_{ttt}(s)\,ds\Big\|\\ &\le Ck^2\int_{t_{j-1}}^{t_j} \|u_{ttt}\|\,ds,\end{aligned}$$

and similarly,

$$k\|\omega_3^j\| = k\|\Delta\big(u(t_{j-\frac12}) - \tfrac12(u(t_j)+u(t_{j-1}))\big)\| \le Ck^2\int_{t_{j-1}}^{t_j} \|\Delta u_{tt}\|\,ds.$$

Altogether,

$$(1.57)\qquad k\sum_{j=1}^n(\|\omega_2^j\| + \|\omega_3^j\|) \le Ck^2\int_0^{t_n}(\|u_{ttt}\| + \|\Delta u_{tt}\|)\,ds,$$

which completes the proof. □

Another way to attain second order accuracy in the discretization in time is to approximate the time derivative in the differential equation by a second order backward difference quotient. Setting

$$\bar{D}U^n = \bar{\partial}U^n + \tfrac{1}{2}k\bar{\partial}^2U^n = (\tfrac{3}{2}U^n - 2U^{n-1} + \tfrac{1}{2}U^{n-2})/k,$$

we have at once by Taylor expansion, for a smooth function u,

$$\bar{D}u(t_n) = u_t(t_n) + O(k^2), \quad \text{as } k \to 0.$$

We therefore pose the discrete problem

$$(\bar{D}U^n, \chi) + (\nabla U^n, \nabla\chi) = (f(t_n), \chi), \quad \forall \chi \in S_h,\ n \geq 2. \tag{1.58}$$

Note that for n fixed this equation employs three time levels rather than the two of our previous methods. We therefore have to restrict its use to $n \geq 2$, because we do not want to use U^n with n negative. With $U^0 = v_h$ given, we then also need to define U^1 in some way, and we choose to do so by employing one step of the backward Euler method, i.e., we set

$$(\bar{\partial}U^1, \chi) + (\nabla U^1, \nabla\chi) = (f(t_1), \chi), \quad \forall \chi \in S_h. \tag{1.59}$$

We note that in our earlier matrix notation, (1.58) may be written as

$$(\tfrac{3}{2}\mathcal{B} + k\mathcal{A})\alpha^n = 2\mathcal{B}\alpha^{n-1} - \tfrac{1}{2}\mathcal{B}\alpha^{n-2} + k\widetilde{f}(t_n), \quad \text{for } n \geq 2,$$

with the matrix coefficient of α^n again positive definite.

We have this time the following $O(h^r + k^2)$ error estimate.

Theorem 1.7 *Let U^n and u be the solutions of* (1.58) *and* (1.1), *with $U^0 = v_h$ and U^1 defined by* (1.59). *Then, if $\|v_h - v\| \leq Ch^r\|v\|_r$ and $v = 0$ on $\partial\Omega$, we have*

$$\begin{aligned}\|U^n - u(t_n)\| \leq Ch^r\Big(\|v\|_r + \int_0^{t_n} \|u_t\|_r\, ds\Big)& \\ + Ck\int_0^k \|u_{tt}\|\, ds + Ck^2\int_0^{t_n}\|u_{ttt}\|\, ds,& \quad \textit{for } n \geq 0.\end{aligned}$$

Proof. Writing again $U^n - u(t_n) = \theta^n + \rho^n$ we only need to bound θ^n, which now satisfies

$$\begin{aligned}(\bar{D}\theta^n, \chi) + (\nabla\theta^n, \nabla\chi) &= -(\omega^n, \chi), \quad \text{for } n \geq 2,\\ (\bar{\partial}\theta^1, \chi) + (\nabla\theta^1, \nabla\chi) &= -(\omega^1, \chi),\end{aligned} \tag{1.60}$$

where

$$\omega^n = \bar{D}R_hu^n - u_t^n = (R_h - I)\bar{D}u^n + (\bar{D}u^n - u_t^n) = \omega_1^n + \omega_2^n, \quad n \geq 2,$$

$$\omega^1 = (R_h - I)\bar{\partial}u^1 + (\bar{\partial}u^1 - u_t^1) = \omega_1^1 + \omega_2^1.$$

We shall show the inequality

$$\|\theta^n\| \le \|\theta^0\| + 2k\sum_{j=2}^{n}\|\omega^j\| + \tfrac{5}{2}k\|\omega^1\|, \quad \text{for } n \ge 1. \tag{1.61}$$

Assuming this for a moment, we need to bound the errors ω_1^j and ω_2^j. Using Taylor expansions with the appropriate remainder terms in integral form we find easily, for $j \ge 2$,

$$k\|\omega_1^j\| \le Ch^r k\|\bar{D}u^j\|_r \le Ch^r \int_{t_{j-2}}^{t_j} \|u_t\|_r\,ds, \quad k\|\omega_2^j\| \le Ck^2 \int_{t_{j-2}}^{t_j} \|u_{ttt}\|\,ds.$$

As for the backward Euler method we have

$$k\|\omega_1^1\| + k\|\omega_2^1\| \le Ch^r \int_0^k \|u_t\|_r\,ds + k\int_0^k \|u_{tt}\|\,ds,$$

and we hence conclude

$$k\sum_{j=1}^{n}\|\omega^j\| \le Ch^r \int_0^{t_n} \|u_t\|_r\,ds + k\int_0^k \|u_{tt}\|\,ds + Ck^2\int_0^{t_n}\|u_{ttt}\|\,ds.$$

Together with our earlier estimate for θ^0, this completes the proof of the estimate for θ^n and thus of the theorem.

It remains to show (1.61). Introducing the difference operators $\delta_l\theta^n = \theta^n - \theta^{n-l}$ for $l = 1, 2$, we may write $k\bar{D}\theta^n = 2\delta_1\theta^n - \frac{1}{2}\delta_2\theta^n$. Since $2(\delta_l\theta^n, \theta^n) = \delta_l\|\theta^n\|^2 + \|\delta_l\theta^n\|^2$, we therefore have

$$k(\bar{D}\theta^n, \theta^n) = \delta_1\|\theta^n\|^2 - \tfrac{1}{4}\delta_2\|\theta^n\|^2 + \|\delta_1\theta^n\|^2 - \tfrac{1}{4}\|\delta_2\theta^n\|^2, \text{ for } n \ge 2.$$

Replacing n by j and then summing from 2 to n, we have

$$\sum_{j=2}^{n}(\delta_1\|\theta^j\|^2 - \tfrac{1}{4}\delta_2\|\theta^j\|^2) = \tfrac{3}{4}\|\theta^n\|^2 - \tfrac{1}{4}\|\theta^{n-1}\|^2 - \tfrac{3}{4}\|\theta^1\|^2 + \tfrac{1}{4}\|\theta^0\|^2,$$

and further, since $\delta_2\theta^n = \delta_1\theta^n + \delta_1\theta^{n-1}$, we obtain

$$\begin{aligned}\sum_{j=2}^{n}(\|\delta_1\theta^j\|^2 - \tfrac{1}{4}\|\delta_2\theta^j\|^2) &\ge \sum_{j=2}^{n}\big(\|\delta_1\theta^j\|^2 - \tfrac{1}{4}(\|\delta_1\theta^j\| + \|\delta_1\theta^{j-1}\|)^2\big)\\ &\ge \tfrac{1}{2}\sum_{j=2}^{n}(\|\delta_1\theta^j\|^2 - \|\delta_1\theta^{j-1}\|^2) = \tfrac{1}{2}(\|\delta_1\theta^n\|^2 - \|\delta_1\theta^1\|^2).\end{aligned}$$

Hence,

$$
\begin{aligned}
&k(\bar\partial\theta^1,\theta^1)+k\sum_{j=2}^{n}(\bar D\theta^j,\theta^j)\\
&\geq \tfrac12\big(\|\theta^1\|^2-\|\theta^0\|^2+\|\delta_1\theta^1\|^2\big)+\tfrac12\big(\|\delta_1\theta^n\|^2-\|\delta_1\theta^1\|^2\big)\\
&\quad+\big(\tfrac34\|\theta^n\|^2-\tfrac14\|\theta^{n-1}\|^2-\tfrac34\|\theta^1\|^2+\tfrac14\|\theta^0\|^2\big)\\
&\geq \tfrac34\|\theta^n\|^2-\tfrac14\|\theta^{n-1}\|^2-\tfrac14\|\theta^1\|^2-\tfrac14\|\theta^0\|^2.
\end{aligned}
\tag{1.62}
$$

But by (1.60) with $\chi=\theta^n$ we have

$$k(\bar\partial\theta^1,\theta^1)+k\sum_{j=2}^{n}(\bar D\theta^j,\theta^j)+k\sum_{j=1}^{n}(\nabla\theta^j,\nabla\theta^j)=-k\sum_{j=1}^{n}(\omega^j,\theta^j),$$

and by (1.62) this yields

$$\|\theta^n\|^2\leq \tfrac13\big(\|\theta^{n-1}\|^2+\|\theta^1\|^2+\|\theta^0\|^2\big)+\tfrac43 k\sum_{j=1}^{n}\|\omega^j\|\,\|\theta^j\|.$$

Suppose m is chosen so that $\|\theta^m\|=\max_{0\leq j\leq n}\|\theta^j\|$. Then

$$\|\theta^m\|^2\leq \tfrac13\Big(\|\theta^m\|+\|\theta^1\|+\|\theta^0\|+4k\sum_{j=1}^{n}\|\omega^j\|\Big)\|\theta^m\|,$$

whence

$$\|\theta^n\|\leq\|\theta^m\|\leq \tfrac12(\|\theta^1\|+\|\theta^0\|)+2k\sum_{j=1}^{n}\|\omega^j\|.$$

Since, as follows from above, $\|\theta^1\|\leq\|\theta^0\|+k\|\omega^1\|$, this completes the proof of (1.61) and thus of the theorem. □

In the above time discretization schemes we have used a constant time step k. We shall close this introductory discussion of fully discrete methods with an example of a variable time step version of the backward Euler method.

Let thus $0=t_0<t_1<\cdots<t_n<\cdots$ be a partition of the positive time axis and set $k_n=t_n-t_{n-1}$. We may then consider the approximation U^n of $u(t_n)$ defined by

$$(\bar\partial_n U^n,\chi)+(\nabla U^n,\nabla\chi)=(f(t_n),\chi),\quad \forall\chi\in S_h,\ n\geq1,\tag{1.63}$$

with $U^0=v_h$, where $\bar\partial_n U^n=(U^n-U^{n-1})/k_n$. We have the following error estimate which reduces to that of Theorem 1.5 for constant time steps.

Theorem 1.8 *Let U^n and u be the solutions of* (1.63) *and* (1.1)*, with $U^0=v_h$ such that $\|v_h-v\|\leq Ch^r\|v\|_r$ and $v=0$ on $\partial\Omega$. Then we have for $n\geq0$*

$$\|U^n-u(t_n)\|\leq Ch^r\Big(\|v\|_r+\int_0^{t_n}\|u_t\|_r\,ds\Big)+\sum_{j=1}^{n}k_j\int_{t_{j-1}}^{t_j}\|u_{tt}\|\,ds.$$

Proof. This time we have for θ^n,

$$(\bar{\partial}_n \theta^n, \chi) + (\nabla\theta^n, \nabla\chi) = -(\omega^n, \chi), \quad \forall \chi \in S_h, \ n \geq 1,$$

where now

$$\omega^n = (R_h - I)\bar{\partial}_n u^n + (\bar{\partial}_n u^n - u_t^n) = \omega_1^n + \omega_2^n.$$

Referring to the proof of Theorem 1.5, (1.49) will be replaced by $\|\theta^n\| \leq \|\theta^{n-1}\| + k_n\|\omega^n\|$, and hence (1.50) by

$$\|\theta^n\| \leq \|\theta^0\| + \sum_{j=1}^{n} k_j(\|\omega_1^j\| + \|\omega_2^j\|).$$

Now

$$\sum_{j=1}^{n} k_j\|\omega_1^j\| \leq \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} Ch^r \|u_t\|_r \, ds = Ch^r \int_0^{t_n} \|u_t\|_r \, ds,$$

and, since (1.52) still holds, with k replaced by k_j,

$$\sum_{j=1}^{n} k_j\|\omega_2^j\| \leq \sum_{j=1}^{n} \Big\| \int_{t_{j-1}}^{t_j} (s - t_{j-1}) u_{tt}(s) \, ds \Big\| \leq \sum_{j=1}^{n} k_j \int_{t_{j-1}}^{t_j} \|u_{tt}\| \, ds.$$

Together with the standard estimates for ρ^n and θ^0, this completes the proof of the theorem. □

We note that the form of the error bound in Theorem 1.8 suggests using shorter time steps when $\|u_{tt}\|$ is larger. We shall return to such considerations in later chapters.

We complete this introductory chapter with some short remarks about other initial boundary value problems for the heat equation than (1.1), and consider first a simple situation with Neumann rather than Dirichlet boundary conditions. Consider thus instead of (1.1) the initial boundary value problem

$$u_t - \Delta u + u = f \quad \text{in } \Omega, \qquad \text{for } t > 0,$$
$$\frac{\partial u}{\partial n} = 0 \quad \text{on } \partial\Omega, \quad \text{for } t > 0, \quad u(\cdot, 0) = v \quad \text{in } \Omega, \tag{1.64}$$

where $\partial u/\partial n$ denotes the derivative in the direction of the exterior normal to $\partial\Omega$. The corresponding stationary problem is then

$$-\Delta u + u = f \quad \text{in } \Omega, \quad \text{with } \frac{\partial u}{\partial n} = 0 \quad \text{on } \partial\Omega. \tag{1.65}$$

In order to formulate this in variational form, we now multiply by $\varphi \in H^1$, thus without requiring $\varphi = 0$ on $\partial\Omega$, integrate over Ω, and use Green's formula to obtain

$$(\nabla u, \nabla \varphi) + (u, \varphi) = (f, \varphi), \quad \forall \varphi \in H^1.$$

We note that if u is smooth, this in turn shows

$$(-\Delta u + u, \varphi) + \int_{\partial \Omega} \frac{\partial u}{\partial n} \varphi \, ds = (f, \varphi), \quad \forall \varphi \in H^1,$$

from which (1.65) follows since φ is arbitrary. In particular, the boundary condition is now a consequence of the variational formulation, in contrast to our earlier discussion when the boundary condition was enforced by looking for a solution in H_0^1. We therefore say that $\partial u / \partial n = 0$ is a *natural boundary condition*, whereas the Dirichlet boundary condition is referred to as an *essential boundary condition*. The lower order term in the differential equation was included to make (1.65) uniquely solvable; note that $\lambda = 0$ is an eigenvalue of $-\Delta$ under Neumann boundary conditions since $\Delta 1 \equiv 0$, whereas $-\Delta + I$ is positive definite.

From the above variational formulation it is natural to assume now that the approximating space S_h is a subspace of H^1, without requiring its elements to vanish on $\partial \Omega$, and safisfies (1.10) when $v \in H^s$. The discrete stationary problem is then

$$(\nabla u_h, \nabla \chi) + (u_h, \chi) = (f, \chi), \quad \forall \chi \in S_h,$$

and this may be analyzed as in Theorem 1.1. The corresponding spatially discrete version of (1.64) is

$$(u_{h,t}, \chi) + (\nabla u_h, \nabla \chi) + (u_h, \chi) = (f, \chi), \ \forall \chi \in \ S_h, \ t > 0, \quad u_h(0) = v_h,$$

and the analysis of this method, and also of corresponding fully discrete ones, follow the same lines as in the case of Dirichlet boundary conditions.

We also mention the time periodic boundary value problem

$$\begin{aligned} &u_t - \Delta u = f \quad \text{in } \Omega, \qquad \text{for } 0 < t < \omega, \\ &\qquad\quad u = 0 \quad \text{on } \partial\Omega, \quad \text{for } 0 < t < \omega, \quad u(\cdot, 0) = u(\cdot, \omega) \quad \text{in } \Omega, \end{aligned} \tag{1.66}$$

where $\omega > 0$ is the period. Setting $u(0) = v$ we have by Duhamel's principle for a possible solution

$$v = u(\omega) = E(\omega) v + \int_0^\omega E(\omega - s) f(s) \, ds,$$

and since $\|E(\omega)\| < 1$ by (1.38), this equation has a unique solution v. Once v is known, (1.66) may be solved as an initial value problem. Spatially semidiscrete and fully discrete versions of the problem may be formulated in obvious ways and analyzed by the techniques developed here.

The *finite element method* originated in the engineering literature in the 1950s, when structural engineers combined the well established framework analysis with *variational methods* in continuum mechanics into a discretization method in which a structure is thought of as consisting of *elements* with locally defined strains or stresses; a standard reference from the engineering literature is Zienkiewicz [249]. In the mid 1960s, a number of papers appeared independently in the numerical analysis literature which were concerned with the construction and analysis of *finite difference* schemes for elliptic problems by *variational principles*, e.g., Céa [45], Demjanovič [68], Feng [98], Friedrichs and Keller [101], and Oganesjan and Ruchovets [187]. By considering approximating functions as defined at all points rather than at meshpoints, the mathematical theory of finite elements then became established through contributions such as Birkhoff, Schultz and Varga [27], where the theory of splines was brought to bear on the development, and Zlámal [250], with the first stringent error analysis of more complicated elements. The duality argument for the L_2 error estimate quoted in Theorem 1.1 was developed independently by Aubin [7], Nitsche [179] and Oganesjan and Ruchovets [188], and later maximum-norm error estimates such as (1.44) were shown by Scott [214], Natterer [175], and Nitsche [182], see Schatz and Wahlbin [208]. The sharpness of this estimate, with the logarithmic factor, was shown in Haverkamp [116].

General treatments of the mathematics of the finite element method for *elliptic* problems can be found in textbooks such as, e.g., Babuška and Aziz [11], Strang and Fix [221], Ciarlet [51] and Brenner and Scott [42], and we shall sometimes quote these for background material.

The development of the theory of finite elements for *parabolic* problems started around 1970. At this time finite difference analysis for such problems had reached a high level of refinement after the fundamental 1928 paper by Courant, Friedrichs and Lewy [52], and became the background and starting point for the finite element analysis of such problems. Names of particular distinction in the development of finite differences in the 50s and 60s are, e.g., F. John, D. G. Aronson, H. O. Kreiss, O. B. Widlund, J. Douglas, Jr., and collaborators, Russian researchers such as Samarskii, etc. (cf. the survey paper Thomée [230]).

The material presented in this introductary chapter is standard; some early references are Douglas and Dupont [74], Price and Varga [196] and Fix and Nassif [99]. An important step in the development was the introduction and exploitation by Wheeler [246] of the Ritz projection, which made it possible to improve earlier suboptimal L_2-norm error estimates to optimal order ones. The nucleus of the present survey is Thomée [229]. Several of the topics that have been touched upon only lightly in this chapter will be developed in more detail in the rest of the book where we will consider both more general equations and wider classes of discretization methods, as well as more detailed investigations of the dependence of the error bounds on the regularity of the exact solutions of our problems. Concerning the discretization of the time-periodic problem mentioned at the end, see Carasso [44], Bernardi [26], and Hansbo [114].

For standard material concerning the mathematical treatment of elliptic and parabolic differential equations we refer to Evans [96], cf. also Lions and Magenes [156] and, for parabolic equations, Friedman [100].

2. Methods Based on More General Approximations of the Elliptic Problem

In our above discussion of finite element approximation of the parabolic problem, the discretization in space was based on using a family of finite-dimensional spaces $S_h \subset H_0^1 = H_0^1(\Omega)$, such that, for some $r \geq 2$, the approximation property (1.10) holds. The most natural example of such a family in a plane domain Ω is to take for S_h the continuous functions which reduce to polynomials of degree at most $r-1$ on the triangles τ of a triangulation $\mathcal{T}_h$ of Ω of the type described in the beginning of Chapter 1, and which vanish on $\partial\Omega$. However, for $r > 2$ and in the case of a domain with smooth boundary, it is not possible, in general, to satisfy the homogeneous boundary conditions exactly for this choice. This difficulty occurs, of course, already for the elliptic problem, and several methods have been suggested to deal with it. In this chapter we shall consider, as a typical example, a method which was proposed by Nitsche for this purpose. This will serve as background for our subsequent discussion of the discretization of the parabolic problem. Another example, a so called mixed method, will be considered in Chapter 17 below.

Consider thus, with Ω a plane domain with smooth boundary, the Dirichlet problem

$$-\Delta u = f \quad \text{in } \Omega, \quad \text{with } u = 0 \quad \text{on } \partial\Omega. \tag{2.1}$$

Let now the $\mathcal{T}_h = \{\tau_j\}_{j=1}^{M_h}$ belong to a family of quasiuniform triangulations of Ω, with $\max_j \operatorname{diam}(\tau_j) \leq h$, where the boundary triangles are allowed to have one curved edge along $\partial\Omega$, and let S_h denote the finite-dimensional linear space of continuous functions on $\bar{\Omega}$ which reduce to polynomials of degree $\leq r-1$ on each triangle τ_j, without any boundary conditions imposed at $\partial\Omega$, i.e.,

$$S_h = \{\chi \in \mathcal{C}(\bar{\Omega});\ \chi\big|_{\tau_j} \in \Pi_{r-1}\}, \tag{2.2}$$

where Π_s denotes the set of polynomials of degree at most s.

In addition to the inner product in $L_2 = L_2(\Omega)$ we set

$$\langle \varphi, \psi \rangle = \int_{\partial\Omega} \varphi\psi \, ds, \quad \text{and} \quad |\varphi| = \langle \varphi, \varphi \rangle^{1/2} = \|\varphi\|_{L_2(\partial\Omega)},$$

and introduce the bilinear form

$$N_\gamma(\varphi,\psi) = (\nabla\varphi,\nabla\psi) - \langle \frac{\partial\varphi}{\partial n},\psi\rangle - \langle\varphi,\frac{\partial\psi}{\partial n}\rangle + \gamma h^{-1}\langle\varphi,\psi\rangle, \tag{2.3}$$

where γ is a positive constant to be fixed later and $\partial/\partial n$ denotes differentiation in the direction of the exterior normal to $\partial\Omega$.

Now let u be a solution of our Dirichlet problem (2.1). Then, using Green's formula, we have, since u vanishes on $\partial\Omega$,

$$\begin{aligned} N_\gamma(u,\chi) &= (\nabla u,\nabla\chi) - \langle \frac{\partial u}{\partial n},\chi\rangle - \langle u,\frac{\partial\chi}{\partial n}\rangle + \gamma h^{-1}\langle u,\chi\rangle \\ &= -(\Delta u,\chi) = (f,\chi), \quad \text{for } \chi\in S_h. \end{aligned} \tag{2.4}$$

With this in mind we define Nitsche's method for (2.1) to find $u_h \in S_h$ satisfying the variational equation

$$N_\gamma(u_h,\chi) = (f,\chi), \quad \forall\chi\in S_h. \tag{2.5}$$

We shall demonstrate below that if γ is appropriately chosen, then this problem admits a unique solution for which optimal order error estimates hold.

For our analysis we introduce, for φ appropriately smooth, the norm

$$|||\varphi||| = \big(\|\nabla\varphi\|^2 + h\Big|\frac{\partial\varphi}{\partial n}\Big|^2 + h^{-1}|\varphi|^2\big)^{1/2}.$$

We first note the following inverse property.

Lemma 2.1 *There is a constant C independent of h such that*

$$|||\chi||| \le Ch^{-1}\|\chi\|, \quad \forall\chi\in S_h.$$

Proof. Because of the quasiuniformity of the family of triangulations $\mathcal{T}_h$, $\nabla\chi$ is estimated by (1.12). Further,

$$\Big|\frac{\partial\chi}{\partial n}\Big|^2 \le C_0h^{-1}\|\nabla\chi\|^2, \quad \forall\chi\in S_h. \tag{2.6}$$

This follows easily by using for each boundary triangle τ_j a linear transformation to map it onto a unit size reference triangle $\widetilde{\tau}_j$ with vertices $(0,0),(1,0)$, and $(0,1)$, say, with the curved edge between $(0,1)$ and $(1,0)$, and noting that here $\|\eta\|_{L_2(\partial\widetilde{\tau}_j)} \le C\|\eta\|_{L_2(\widetilde{\tau}_j)}$ for $\eta = \partial\chi/\partial x_i$, since the right hand side is a norm on Π_{r-2}. Using the inverse of the linear transformation to map $\widetilde{\tau}_j$ back to τ_j, we obtain $\|\partial\chi/\partial x_i\|^2_{L_2(\partial\tau_j)} \le Ch^{-1}\|\partial\chi/\partial x_i\|^2_{L_2(\tau_j)}$, and (2.6) follows by summation over the boundary triangles. Using also (1.12) this bounds $\partial\chi/\partial n$ in the desired way. Finally, in the same way, $|\chi|^2 \le C_0h^{-1}\|\chi\|^2$ for $\chi\in S_h$. Together these estimates show the lemma. □

We now show that the bilinear form $N_\gamma(\cdot,\cdot)$ defined in (2.3) is continuous in terms of $|||\cdot|||$ and positive definite when restricted to S_h.