Numerical Methods for Elliptic and Parabolic Partial Differential Equations

Peter Knabner
Lutz Angermann

Springer

Texts in Applied Mathematics 44

Springer
New York
Berlin
Heidelberg
Hong Kong
London
Milan
Paris
Tokyo

This page intentionally left blank

Peter Knabner Lutz Angermann

Numerical Methods for Elliptic and Parabolic Partial Differential Equations

With 67 Figures

Springer

Peter Knabner
Institute for Applied Mathematics
University of Erlangen
Martensstrasse 3
D-91058 Erlangen
Germany
knabner@am.uni-erlangen.de

Lutz Angermann
Institute for Mathematics
University of Clausthal
Erzstrasse 1
D-38678 Clausthal-Zellerfeld
Germany
angermann@math.tu-clausthal.de

Series Editors

J.E. Marsden
Control and Dynamical Systems, 107–81
California Institute of Technology
Pasadena, CA 91125
USA
marsden@cds.caltech.edu

L. Sirovich
Division of Applied Mathematics
Brown University
Providence, RI 02912
USA
chico@camelot.mssm.edu

S.S. Antman
Department of Mathematics
and
Institute for Physical Science
and Technology
University of Maryland
College Park, MD 20742-4015
USA
ssa@math.umd.edu

Mathematics Subject Classification (2000): 65Nxx, 65Mxx, 65F10, 65H10

Library of Congress Cataloging-in-Publication Data
Knabner, Peter.
[Numerik partieller Differentialgleichungen. English]
Numerical methods for elliptic and parabolic partial differential equations /
Peter Knabner, Lutz Angermann.
p. cm. – (Texts in applied mathematics ; 44)
Include bibliographical references and index.
ISBN 0-387-95449-X (alk. paper)
1. Differential equations, Partial–Numerical solutions. I. Angermann, Lutz. II. Title.
III. Series.
QA377.K575 2003
515′.353–dc21 2002044522

ISBN 0-387-95449-X Printed on acid-free paper.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1 SPIN 10867187

Typesetting: Pages created by the authors in LaTeX 2e using Springer's svsing6.cls macro.

www.springer-ny.com

Springer-Verlag New York Berlin Heidelberg
A member of BertelsmannSpringer Science+Business Media GmbH

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
Providence, Rhode Island — L. Sirovich
College Park, Maryland — S.S. Antman

This page intentionally left blank

Preface to the English Edition

Shortly after the appearance of the German edition we were asked by Springer to create an English version of our book, and we gratefully accepted. We took this opportunity not only to correct some misprints and mistakes that have come to our knowledge[1] but also to extend the text at various places. This mainly concerns the role of the finite difference and the finite volume methods, which have gained more attention by a slight extension of Chapters 1 and 6 and by a considerable extension of Chapter 7. Time-dependent problems are now treated with all three approaches (finite differences, finite elements, and finite volumes), doing this in a uniform way as far as possible. This also made a reordering of Chapters 6–8 necessary. Also, the index has been enlarged. To improve the direct usability in courses, exercises now follow each section and should provide enough material for homework.

This new version of the book would not have come into existence without our already mentioned team of helpers, who also carried out first versions of translations of parts of the book. Beyond those already mentioned, the team was enforced by Cecilia David, Basca Jadamba, Dr. Serge Kräutle, Dr. Wilhelm Merz, and Peter Mirsch. Alexander Prechtel now took charge of the difficult modification process. Prof. Paul DuChateau suggested improvements. We want to extend our gratitude to all of them. Finally, we

[1] Users of the German edition may consult
http://www.math.tu-clausthal.de/~mala/publications/errata.pdf

thank senior editor Achi Dosanjh, from Springer-Verlag New York, Inc., for her constant encouragement.

Remarks for the Reader and the Use in Lectures

The size of the text corresponds roughly to four hours of lectures per week over two terms. If the course lasts only one term, then a selection is necessary, which should be orientated to the audience. We recommend the following "cuts":

Chapter 0 may be skipped if the partial differential equations treated therein are familiar. Section 0.5 should be consulted because of the notation collected there. The same is true for Chapter 1; possibly Section 1.4 may be integrated into Chapter 3 if one wants to deal with Section 3.9 or with Section 7.5.

Chapters 2 and 3 are the core of the book. The inductive presentation that we preferred for some theoretical aspects may be shortened for students of mathematics. To the lecturer's taste and depending on the knowledge of the audience in numerical mathematics Section 2.5 may be skipped. This might impede the treatment of the ILU preconditioning in Section 5.3. Observe that in Sections 2.1–2.3 the treatment of the model problem is merged with basic abstract statements. Skipping the treatment of the model problem, in turn, requires an integration of these statements into Chapter 3. In doing so Section 2.4 may be easily combined with Section 3.5. In Chapter 3 the theoretical kernel consists of Sections 3.1, 3.2.1, 3.3–3.4.

Chapter 4 presents an overview of its subject, not a detailed development, and is an extension of the classical subjects, as are Chapters 6 and 9 and the related parts of Chapter 7.

In the extensive Chapter 5 one might focus on special subjects or just consider Sections 5.2, 5.3 (and 5.4) in order to present at least one practically relevant and modern iterative method.

Section 8.1 and the first part of Section 8.2 contain basic knowledge of numerical mathematics and, depending on the audience, may be omitted.

The appendices are meant only for consultation and may complete the basic lectures, such as in analysis, linear algebra, and advanced mathematics for engineers.

Concerning related textbooks for supplementary use, to the best of our knowledge there is none covering approximately the same topics. Quite a few deal with finite element methods, and the closest one in spirit probably is [21], but also [6] or [7] have a certain overlap, and also offer additional material not covered here. From the books specialised in finite difference methods, we mention [32] as an example. The (node-oriented) finite volume method is popular in engineering, in particular in fluid dynamics, but to the best of our knowledge there is no presentation similar to ours in a

mathematical textbook. References to textbooks specialised in the topics of Chapters 4, 5 and 8 are given there.

Remarks on the Notation

Printing in *italics* emphasizes definitions of notation, even if this is not carried out as a numbered definition.

Vectors appear in different forms: Besides the "short" space vectors $x \in \mathbb{R}^d$ there are "long" representation vectors $\boldsymbol{u} \in \mathbb{R}^m$, which describe in general the degrees of freedom of a finite element (or volume) approximation or represent the values on grid points of a finite difference method. Here we choose **bold type**, also in order to have a distinctive feature from the generated functions, which frequently have the same notation, or from the grid functions.

Deviations can be found in Chapter 0, where vectorial quantities belonging to $\mathbb{R}^d$ are boldly typed, and in Chapters 5 and 8, where the unknowns of linear and nonlinear systems of equations, which are treated in a general manner there, are denoted by $x \in \mathbb{R}^m$.

Components of vectors will be designated by a subindex, creating a double index for indexed quantities. Sequences of vectors will be supplied with a superindex (in parentheses); only in an abstract setting do we use subindices.

Erlangen, Germany — Peter Knabner
Clausthal-Zellerfeld, Germany — Lutz Angermann
January 2002

This page intentionally left blank

Preface to the German Edition

This book resulted from lectures given at the University of Erlangen–Nuremberg and at the University of Magdeburg. On these occasions we often had to deal with the problem of a heterogeneous audience composed of students of mathematics and of different natural or engineering sciences. Thus the expectations of the students concerning the mathematical accuracy and the applicability of the results were widely spread. On the other hand, neither relevant models of partial differential equations nor some knowledge of the (modern) theory of partial differential equations could be assumed among the whole audience. Consequently, in order to overcome the given situation, we have chosen a selection of models and methods relevant for applications (which might be extended) and attempted to illuminate the whole spectrum, extending from the theory to the implementation, without assuming advanced mathematical background. Most of the theoretical obstacles, difficult for nonmathematicians, will be treated in an "inductive" manner. In general, we use an explanatory style without (hopefully) compromising the mathematical accuracy.

We hope to supply especially students of mathematics with the information necessary for the comprehension and implementation of finite element/finite volume methods. For students of the various natural or engineering sciences the text offers, beyond the possibly already existing knowledge concerning the application of the methods in special fields, an introduction into the mathematical foundations, which should facilitate the transformation of specific knowledge to other fields of applications.

We want to express our gratitude for the valuable help that we received during the writing of this book: Dr. Markus Bause, Sandro Bitterlich,

Dr. Christof Eck, Alexander Prechtel, Joachim Rang, and Dr. Eckhard Schneid did the proofreading and suggested important improvements. From the anonymous referees we received useful comments. Very special thanks go to Mrs. Magdalena Ihle and Dr. Gerhard Summ. Mrs. Ihle transposed the text quickly and precisely into TEX. Dr. Summ not only worked on the original script and on the TEX-form, he also organized the complex and distributed rewriting and extension procedure. The elimination of many inconsistencies is due to him. Additionally he influenced parts of Sections 3.4 and 3.8 by his outstanding diploma thesis. We also want to thank Dr. Chistoph Tapp for the preparation of the graphic of the title and for providing other graphics from his doctoral thesis [70].

Of course, hints concerning (typing) mistakes and general improvements are always welcome.

We thank Springer-Verlag for their constructive collaboration.

Last, but not least, we want to express our gratitude to our families for their understanding and forbearance, which were necessary for us especially during the last months of writing.

Erlangen, Germany — Peter Knabner
Magdeburg, Germany — Lutz Angermann
February 2000

Contents

Series Preface **v**

Preface to the English Edition **vii**

Preface to the German Edition **xi**

0 For Example: Modelling Processes in Porous Media with Differential Equations **1**
0.1 The Basic Partial Differential Equation Models 1
0.2 Reactions and Transport in Porous Media 5
0.3 Fluid Flow in Porous Media 7
0.4 Reactive Solute Transport in Porous Media 11
0.5 Boundary and Initial Value Problems 14

1 For the Beginning: The Finite Difference Method for the Poisson Equation **19**
1.1 The Dirichlet Problem for the Poisson Equation 19
1.2 The Finite Difference Method 21
1.3 Generalizations and Limitations of the Finite Difference Method 29
1.4 Maximum Principles and Stability 36

2 The Finite Element Method for the Poisson Equation **46**
2.1 Variational Formulation for the Model Problem 46

2.2 The Finite Element Method with Linear Elements 55
2.3 Stability and Convergence of the Finite Element Method 68
2.4 The Implementation of the Finite Element Method: Part 1 74
2.5 Solving Sparse Systems of Linear Equations by Direct Methods 82

3 The Finite Element Method for Linear Elliptic Boundary Value Problems of Second Order 92
3.1 Variational Equations and Sobolev Spaces 92
3.2 Elliptic Boundary Value Problems of Second Order . . . 100
3.3 Element Types and Affine Equivalent Triangulations 114
3.4 Convergence Rate Estimates 131
3.5 The Implementation of the Finite Element Method: Part 2 148
3.6 Convergence Rate Results in Case of Quadrature and Interpolation 155
3.7 The Condition Number of Finite Element Matrices . . . 163
3.8 General Domains and Isoparametric Elements 167
3.9 The Maximum Principle for Finite Element Methods . . 171

4 Grid Generation and A Posteriori Error Estimation 176
4.1 Grid Generation 176
4.2 A Posteriori Error Estimates and Grid Adaptation . . . 185

5 Iterative Methods for Systems of Linear Equations 198
5.1 Linear Stationary Iterative Methods 200
5.2 Gradient and Conjugate Gradient Methods 217
5.3 Preconditioned Conjugate Gradient Method 227
5.4 Krylov Subspace Methods for Nonsymmetric Systems of Equations 233
5.5 The Multigrid Method 238
5.6 Nested Iterations 251

6 The Finite Volume Method 255
6.1 The Basic Idea of the Finite Volume Method 256
6.2 The Finite Volume Method for Linear Elliptic Differential Equations of Second Order on Triangular Grids . . . 262

7 Discretization Methods for Parabolic Initial Boundary Value Problems 283
7.1 Problem Setting and Solution Concept 283
7.2 Semidiscretization by the Vertical Method of Lines . . . 293

7.3 Fully Discrete Schemes 311
7.4 Stability 315
7.5 The Maximum Principle for the One-Step-Theta Method 323
7.6 Order of Convergence Estimates 330

8 Iterative Methods for Nonlinear Equations 342
8.1 Fixed-Point Iterations 344
8.2 Newton's Method and Its Variants 348
8.3 Semilinear Boundary Value Problems for Elliptic and Parabolic Equations 360

9 Discretization Methods for Convection-Dominated Problems 368
9.1 Standard Methods and Convection-Dominated Problems 368
9.2 The Streamline-Diffusion Method 375
9.3 Finite Volume Methods 383
9.4 The Lagrange–Galerkin Method 387

A Appendices 390
A.1 Notation 390
A.2 Basic Concepts of Analysis 393
A.3 Basic Concepts of Linear Algebra 394
A.4 Some Definitions and Arguments of Linear Functional Analysis 399
A.5 Function Spaces 404

References: Textbooks and Monographs 409

References: Journal Papers 412

Index 415

This page intentionally left blank

0

For Example: Modelling Processes in Porous Media with Differential Equations

This chapter illustrates the scientific context in which differential equation models may occur, in general, and also in a specific example. Section 0.1 reviews the fundamental equations, for some of them discretization techniques will be developed and investigated in this book. In Sections 0.2 – 0.4 we focus on reaction and transport processes in porous media. These sections are independent of the remaining parts and may be skipped by the reader. Section 0.5, however, should be consulted because it fixes some notation to be used later on.

0.1 The Basic Partial Differential Equation Models

Partial differential equations are equations involving some partial derivatives of an unknown function u in several independent variables. Partial differential equations which arise from the modelling of spatial (and temporal) processes in nature or technology are of particular interest. Therefore, we assume that the variables of u are $x = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$ for $d \geq 1$, representing a spatial point, and possibly $t \in \mathbb{R}$, representing time. Thus the minimal set of variables is (x_1, x_2) or (x_1, t), otherwise we have ordinary differential equations. We will assume that $x \in \Omega$, where Ω is a bounded domain, e.g., a metal workpiece, or a groundwater aquifer, and $t \in (0, T]$ for some (time horizon) $T > 0$. Nevertheless also processes acting in the whole $\mathbb{R}^d \times \mathbb{R}$, or in unbounded subsets of it, are of interest. One may consult the Appendix for notations from analysis etc. used here. Often the function u

represents, or is related to, the volume density of an extensive quantity like mass, energy, or momentum, which is conserved. In their original form all quantities have dimensions that we denote in accordance with the International System of Units (SI) and write in square brackets [ ]. Let a be a symbol for the unit of the extensive quantity, then its volume density is assumed to have the form $S = S(u)$, i.e., the unit of $S(u)$ is $\mathsf{a/m}^3$. For example, for mass conservation $\mathsf{a} = \mathsf{kg}$, and $S(u)$ is a concentration. For describing the conservation we consider an arbitrary "not too bad" subset $\tilde{\Omega} \subset \Omega$, the *control volume*. The time variation of the total extensive quantity in $\tilde{\Omega}$ is then

$$\partial_t \int_{\tilde{\Omega}} S(u(x,t))dx \,. \tag{0.1}$$

If this function does not vanish, only two reasons are possible due to conservation:

— There is an internally distributed source density $Q = Q(x,t,u)$ $[\mathsf{a/m}^3\mathsf{/s}]$, being positive if $S(u)$ is produced, and negative if it is destroyed, i.e., one term to balance (0.1) is $\int_{\tilde{\Omega}} Q(x,t,u(x,t))dx$.

— There is a net flux of the extensive quantity over the boundary $\partial\tilde{\Omega}$ of $\tilde{\Omega}$. Let $\boldsymbol{J} = \boldsymbol{J}(x,t)$ $[\mathsf{a/m}^2\mathsf{/s}]$ denote the flux density, i.e., $\boldsymbol{J}_i$ is the amount, that passes a unit square perpendicular to the ith axis in one second in the direction of the ith axis (if positive), and in the opposite direction otherwise. Then another term to balance (0.1) is given by

$$-\int_{\partial\Omega} \boldsymbol{J}(x,t)\cdot\nu(x)d\sigma \,,$$

where ν denotes the outer unit normal on $\partial\Omega$. Summarizing the conservation reads

$$\partial_t \int_{\tilde{\Omega}} S(u(x,t))dx = -\int_{\partial\tilde{\Omega}} \boldsymbol{J}(x,t)\cdot\nu(x)d\sigma + \int_{\tilde{\Omega}} Q(x,t,u(x,t))dx \,. \tag{0.2}$$

The integral theorem of Gauss (see (2.3)) and an exchange of time derivative and integral leads to

$$\int_{\tilde{\Omega}} [\partial_t S(u(x,t)) + \nabla\cdot\boldsymbol{J}(x,t) - Q(x,t,u(x,t))]dx = 0\,,$$

and, as $\tilde{\Omega}$ is arbitrary, also to

$$\partial_t S(u(x,t)) + \nabla\cdot\boldsymbol{J}(x,t) = Q(x,t,u(x,t)) \text{ for } x\in\Omega,\ t\in(0,T] \,. \tag{0.3}$$

All manipulations here are formal assuming that the functions involved have the necessary properties. The partial differential equation (0.3) is the basic pointwise conservation equation, (0.2) its corresponding integral form. Equation (0.3) is one requirement for the two unknowns u and $\boldsymbol{J}$, thus it

has to be closed by a (phenomenological) *constitutive law*, postulating a relation between $\boldsymbol{J}$ and u.

Assume Ω is a container filled with a fluid in which a substance is dissolved. If u is the concentration of this substance, then $S(u) = u$ and $\mathsf{a} = \mathsf{kg}$. The description of $\boldsymbol{J}$ depends on the processes involved. If the fluid is at rest, then flux is only possible due to *molecular diffusion*, i.e., a flux from high to low concentrations due to random motion of the dissolved particles. Experimental evidence leads to

$$\boldsymbol{J}^{(1)} = -K\nabla u \tag{0.4}$$

with a parameter $K > 0$ $[\mathsf{m}^2/\mathsf{s}]$, the *molecular diffusivity*. Equation (0.4) is called *Fick's law*.

In other situations, like heat conduction in a solid, a similar model occurs. Here, u represents the temperature, and the underlying principle is energy conservation. The constitutive law is *Fourier's law*, which also has the form (0.4), but as K is a material parameter, it may vary with space or, for anisotropic materials, be a matrix instead of a scalar.

Thus we obtain the *diffusion equation*

$$\partial_t u - \nabla \cdot (\boldsymbol{K}\nabla u) = Q\,. \tag{0.5}$$

If K is scalar and constant — let $K = 1$ by scaling —, and $f := Q$ is independent of u, the equation simplifies further to

$$\partial_t u - \Delta u = f\,,$$

where $\Delta u := \nabla\cdot(\nabla u)$. We mentioned already that this equation also occurs in the modelling of heat conduction, therefore this equation or (0.5) is also called the *heat equation*.

If the fluid is in motion with a (given) velocity $\boldsymbol{c}$ then *(forced) convection* of the particles takes place, being described by

$$\boldsymbol{J}^{(2)} = u\boldsymbol{c}\,, \tag{0.6}$$

i.e., taking both processes into account, the model takes the form of the *convection-diffusion equation*

$$\partial_t u - \nabla \cdot (\boldsymbol{K}\nabla u - \boldsymbol{c}u) = Q\,. \tag{0.7}$$

The relative strength of the two processes is measured by the Péclet number (defined in Section 0.4). If convection is dominating one may ignore diffusion and only consider the *transport equation*

$$\partial_t u + \nabla \cdot (\boldsymbol{c}u) = Q\,. \tag{0.8}$$

The different nature of the two processes has to be reflected in the models, therefore, adapted discretization techniques will be necessary. In this book we will consider models like (0.7), usually with a significant contribution of diffusion, and the case of dominating convection is studied in Chapter 9. The pure convective case like (0.8) will not be treated.

In more general versions of (0.7) $\partial_t u$ is replaced by $\partial_t S(u)$, where S depends linearly or nonlinearly on u. In the case of heat conduction S is the internal energy density, which is related to the temperature u via the factors mass density and specific heat. For some materials the specific heat depends on the temperature, then S is a nonlinear function of u.

Further aspects come into play by the source term Q if it depends linearly or nonlinearly on u, in particular due to (chemical) reactions. Examples for these cases will be developed in the following sections. Since equation (0.3) and its examples describe conservation in general, it still has to be adapted to a concrete situation to ensure a unique solution u. This is done by the specification of an initial condition

$$S(u(x,0)) = S_0(x) \quad \text{for} \quad x \in \Omega\,,$$

and by boundary conditions. In the example of the water filled container no mass flux will occur across its walls, therefore, the following boundary condition

$$\boldsymbol{J} \cdot \nu(x,t) = 0 \quad \text{for} \quad x \in \partial\Omega,\ t \in (0,T) \tag{0.9}$$

is appropriate, which — depending on the definition of $\boldsymbol{J}$ — prescribes the normal derivative of u, or a linear combination of it and u. In Section 0.5 additional situations are depicted.

If a process is stationary, i.e. time-independent, then equation (0.3) reduces to

$$\nabla \cdot \boldsymbol{J}(x) = Q(x, u(x)) \quad \text{for} \quad x \in \Omega\,,$$

which in the case of diffusion and convection is specified to

$$-\nabla \cdot (\boldsymbol{K}\nabla u - \boldsymbol{c}u) = Q\,.$$

For constant K — let $K = 1$ by scaling —, $c = 0$, and $f := Q$, being independent of u, this equation reduces to

$$-\Delta u = f \quad \text{in} \quad \Omega\,,$$

the *Poisson equation*.

Instead of the boundary condition (0.9), one can prescribe the values of the function u at the boundary:

$$u(x) = g(x) \quad \text{for} \quad x \in \partial\Omega\,.$$

For models , where u is a concentration or temperature, the physical realisation of such a boundary condition may raise questions, but in mechanical models, where u is to interpreted as a displacement, such a boundary condition seems reasonable. The last boundary value problem will be the first model, whose discretization will be discussed in Chapters 1 and 2.

Finally it should be noted that it is advisable to non-dimensionalise the final model before numerical methods are applied. This means that both the independent variables x_i (and t), and the dependent one u, are replaced

by $x_i/x_{i,\mathrm{ref}}$, t/t_{ref}, and u/u_{ref}, where $x_{i,\mathrm{ref}}$, t_{ref}, and u_{ref} are fixed reference values of the same dimension as x_i, t, and u, respectively. These reference values are considered to be of typical size for the problems under investigation. This procedure has two advantages: On the one hand, the typical size is now 1, such that there is an absolute scale for (an error in) a quantity to be small or large. On the other hand, if the reference values are chosen appropriately a reduction in the number of equation parameters like $\boldsymbol{K}$ and $\boldsymbol{c}$ in (0.7) might be possible, having only fewer algebraic expressions of the original material parameters in the equation. This facilitates numerical parameter studies.

0.2 Reactions and Transport in Porous Media

A *porous medium* is a heterogeneous material consisting of a *solid matrix* and a *pore space* contained therein. We consider the pore space (of the porous medium) as connected; otherwise, the transport of *fluids* in the pore space would not be possible. Porous media occur in nature and manufactured materials. Soils and aquifers are examples in geosciences; porous catalysts, chromatographic columns, and ceramic foams play important roles in chemical engineering. Even the human skin can be considered a porous medium. In the following we focus on applications in the geosciences. Thus we use a terminology referring to the natural soil as a porous medium. On the *micro* or *pore scale* of a single grain or pore, i.e., in a range of µm to mm, the fluids constitute different phases in the thermodynamic sense. Thus we name this system in the case of k fluids including the solid matrix as $(k+1)$*-phase system* or we speak of k*-phase flow*.

We distinguish three classes of fluids with different affinities to the solid matrix. These are an aqueous phase, marked with the index "w" for water, a nonaqueous phase liquid (like oil or gasoline as natural resources or contaminants), marked with the index "o," and a gaseous phase, marked with the index "g" (e.g., soil air). Locally, at least one of these phases has always to be present; during a transient process phases can locally disappear or be generated. These fluid phases are in turn *mixtures* of several *components*. In applications of the earth sciences, for example, we do not deal with pure water but encounter different species in true or colloidal solution in the *solvent* water. The wide range of chemical components includes plant nutrients, mineral nutrients from salt domes, organic decomposition products, and various organic and inorganic chemicals. These substances are normally not inert, but are subject to reactions and transformation processes. Along with diffusion, *forced convection* induced by the motion of the fluid is the essential driving mechanism for the transport of solutes. But we also encounter *natural convection* by the coupling of the dynamics of the substance to the fluid flow. The description level at the microscale

that we have used so far is not suitable for processes at the laboratory or technical scale, which take place in ranges of `cm` to `m`, or even for processes in a catchment area with units of `km`. For those *macroscales* new models have to be developed, which emerge from averaging procedures of the models on the microscale. There may also exist principal differences among the various macroscales that let us expect different models, which arise from each other by *upscaling*. But this aspect will not be investigated here further. For the transition of micro to macro scales the engineering sciences provide the heuristic method of *volume averaging*, and mathematics the rigorous (but of only limited use) approach of *homogenization* (see [36] or [19]). None of the two possibilities can be depicted here completely. Where necessary we will refer to volume averaging for (heuristic) motivation.

Let $\Omega \subset \mathbb{R}^d$ be the domain of interest. All subsequent considerations are formal in the sense that the admissibility of the analytic manipulations is supposed. This can be achieved by the assumption of sufficient smoothness for the corresponding functions and domains.

Let $V \subset \Omega$ be an admissible *representative elementary volume* in the sense of volume averaging around a point $x \in \Omega$. Typically the shape and the size of a representative elementary volume are selected in such a manner that the averaged values of all geometric characteristics of the microstructure of the pore space are independent of the size of V but depend on the location of the point x. Then we obtain for a given variable ω_α in the phase α (after continuation of ω_α with 0 outside of α) the corresponding macroscopic quantities, assigned to the location x, as the *extrinsic phase average*

$$\langle \omega_\alpha \rangle := \frac{1}{|V|} \int_V \omega_\alpha$$

or as the *intrinsic phase average*

$$\langle \omega_\alpha \rangle^\alpha := \frac{1}{|V_\alpha|} \int_{V_\alpha} \omega_\alpha \, .$$

Here V_α denotes the subset of V corresponding to α. Let $t \in (0, T)$ be the time at which the process is observed. The notation $x \in \Omega$ means the vector in Cartesian coordinates, whose coordinates are referred to by x, y, and $z \in \mathbb{R}$. Despite this ambiguity the meaning can always be clearly derived from the context.

Let the index "s" (for solid) stand for the solid phase; then

$$\phi(x) := |V \setminus V_\mathrm{s}| \, / \, |V| > 0$$

denotes the *porosity*, and for every liquid phase α,

$$S_\alpha(x,t) := |V_\alpha| \, / \, |V \setminus V_\mathrm{s}| \geq 0$$

is the *saturation* of the phase α. Here we suppose that the solid phase is stable and immobile. Thus

$$\langle \omega_\alpha \rangle = \phi S_\alpha \langle \omega_\alpha \rangle^\alpha$$

for a fluid phase α and

$$\sum_{\alpha:\text{fluid}} S_\alpha = 1 \,. \tag{0.10}$$

So if the fluid phases are *immiscible* on the micro scale, they may be miscible on the macro scale, and the immiscibility on the macro scale is an additional assumption for the model.

As in other disciplines the differential equation models are derived here from conservation laws for the *extensive quantities* mass, impulse, and energy, supplemented by *constitutive relationships*, where we want to focus on the mass.

0.3 Fluid Flow in Porous Media

Consider a liquid phase α on the micro scale. In this chapter, for clarity, we write "short" vectors in $\mathbb{R}^d$ also in bold with the exception of the coordinate vector x. Let $\tilde{\varrho}_\alpha$ [$\mathtt{kg/m}^3$] be the (microscopic) *density*, $\tilde{\boldsymbol{q}}_\alpha := \left(\sum_\eta \tilde{\varrho}_\eta \tilde{\boldsymbol{v}}_\eta\right) / \tilde{\varrho}_\alpha$ [$\mathtt{m/s}$] the *mass average mixture velocity* based on the *particle velocity* $\tilde{\boldsymbol{v}}_\eta$ of a component η and its concentration in solution $\tilde{\varrho}_\eta$ [$\mathtt{kg/m}^3$]. The transport theorem of Reynolds (see, for example, [10]) leads to the mass conservation law

$$\partial_t \tilde{\varrho}_\alpha + \nabla \cdot (\tilde{\varrho}_\alpha \tilde{\boldsymbol{q}}_\alpha) = \tilde{f}_\alpha \tag{0.11}$$

with a distributed *mass source density* $\tilde{f}_\alpha$. By averaging we obtain from here the mass conservation law

$$\partial_t (\phi S_\alpha \varrho_\alpha) + \nabla \cdot (\varrho_\alpha \boldsymbol{q}_\alpha) = f_\alpha \tag{0.12}$$

with ϱ_α, the density of phase α, as the intrinsic phase average of $\tilde{\varrho}_\alpha$ and $\boldsymbol{q}_\alpha$, the *volumetric fluid velocity* or *Darcy velocity* of the phase α, as the extrinsic phase average of $\tilde{\boldsymbol{q}}_\alpha$. Correspondingly, f_α is an average mass source density.

Before we proceed in the general discussion, we want to consider some specific situations: The area between the groundwater table and the impermeable body of an *aquifer* is characterized by the fact that the whole pore space is occupied by a fluid phase, the soil water. The corresponding saturation thus equals 1 everywhere, and with omission of the index equation (0.12) takes the form

$$\partial_t (\phi \varrho) + \nabla \cdot (\varrho \boldsymbol{q}) = f \,. \tag{0.13}$$

If the density of water is assumed to be constant, due to neglecting the mass of solutes and compressibility of water, equation (0.13) simplifies further to the stationary equation

$$\nabla \cdot \boldsymbol{q} = f \,, \tag{0.14}$$

where f has been replaced by the volume source density f/ϱ, keeping the same notation. This equation will be completed by a relationship that can be interpreted as the macroscopic analogue of the conservation of momentum, but should be accounted here only as an experimentally derived constitutive relationship. This relationship is called *Darcy's law*, which reads as

$$\boldsymbol{q} = -\boldsymbol{K}\,(\nabla p + \varrho g \boldsymbol{e}_z) \tag{0.15}$$

and can be applied in the range of laminar flow. Here p [$\mathtt{N}/\mathtt{m}^2$] is the intrinsic average of the *water pressure*, g [$\mathtt{m}/\mathtt{s}^2$] the gravitational acceleration, $\boldsymbol{e}_z$ the unit vector in the z-direction oriented against the gravitation,

$$\boldsymbol{K} = \boldsymbol{k}/\mu \,, \tag{0.16}$$

a quantity, which is given by the *permeability* $\boldsymbol{k}$ determined by the solid phase, and the *viscosity* μ determined by the fluid phase. For an *anisotropic* solid, the matrix $\boldsymbol{k} = \boldsymbol{k}(x)$ is a symmetric positive definite matrix.

Inserting (0.15) in (0.14) and replacing $\boldsymbol{K}$ by $\boldsymbol{K}\varrho g$, known as *hydraulic conductivity* in the literature, and keeping the same notation gives the following linear equation for

$$h(x,t) := \frac{1}{\varrho g}\, p(x,t) + z \;,$$

the *piezometric head* h [$\mathtt{m}$]:

$$-\nabla \cdot (\boldsymbol{K} \nabla h) = f \;. \tag{0.17}$$

The resulting equation is stationary and linear. We call a differential equation model *stationary* if it depends only on the location x and not on the time t, and *instationary* otherwise. A differential equation and corresponding boundary conditions (cf. Section 0.5) are called *linear* if the sum or a scalar multiple of a solution again forms a solution for the sum, respectively the scalar multiple, of the sources.

If we deal with an *isotropic* solid matrix, we have $\boldsymbol{K} = K\boldsymbol{I}$ with the $d \times d$ unit matrix $\boldsymbol{I}$ and a scalar function K. Equation (0.17) in this case reads

$$-\nabla \cdot (K \nabla h) = f \;. \tag{0.18}$$

Finally if the solid matrix is homogeneous, i.e., K is constant, we get from division by K and maintaining the notation f the *Poisson equation*

$$-\Delta h = f \;, \tag{0.19}$$

which is termed the *Laplace equation* for $f = 0$. This model and its more general formulations occur in various contexts. If, contrary to the above assumption, the solid matrix is compressible under the pressure of the water, and if we suppose (0.13) to be valid, then we can establish a relationship

$$\phi = \phi(x,t) = \phi_0(x)\phi_f(p)$$

with $\phi_0(x) > 0$ and a monotone increasing ϕ_f such that with $S(p) := \phi_f'(p)$ we get the equation

$$\phi_0\, S(p)\, \partial_t p + \nabla \cdot \boldsymbol{q} = f$$

and the instationary equations corresponding to (0.17)–(0.19), respectively. For constant $S(p) > 0$ this yields the following linear equation:

$$\phi_0\, S\, \partial_t h - \nabla \cdot (\boldsymbol{K} \nabla h) = f\ , \tag{0.20}$$

which also represents a common model in many contexts and is known from corresponding fields of application as the *heat conduction equation*.

We consider single phase flow further, but now we will consider gas as fluid phase. Because of the compressibility, the density is a function of the pressure, which is invertible due to its strict monotonicity to

$$p = P(\varrho)\ .$$

Together with (0.13) and (0.15) we get a nonlinear variant of the heat conduction equation in the unknown ϱ:

$$\partial_t(\phi\varrho) - \nabla \cdot \big(\boldsymbol{K}(\varrho \nabla P(\varrho) + \varrho^2 g \boldsymbol{e}_z)\big) = f\ , \tag{0.21}$$

which also contains derivatives of first order in space. If $P(\varrho) = \ln(\alpha\varrho)$ holds for a constant $\alpha > 0$, then $\varrho \nabla P(\varrho)$ simplifies to $\alpha \nabla \varrho$. Thus for horizontal flow we again encounter the heat conduction equation. For the relationship $P(\varrho) = \alpha\varrho$ suggested by the universal gas law, $\alpha\varrho\nabla\varrho = \frac{1}{2}\alpha\nabla\varrho^2$ remains nonlinear. The choice of the variable $u := \varrho^2$ would result in $u^{1/2}$ in the time derivative as the only nonlinearity. Thus in the formulation in ϱ the coefficient of $\nabla\varrho$ disappears in the divergence of $\varrho = 0$. Correspondingly, the coefficient $S(u) = \frac{1}{2}\phi u^{-1/2}$ of $\partial_t u$ in the formulation in u becomes unbounded for $u = 0$. In both versions the equations are *degenerate*, whose treatment is beyond the scope of this book. A variant of this equation has gained much attention as the *porous medium equation* (with convection) in the field of analysis (see, for example, [42]).

Returning to the general framework, the following generalization of Darcy's law can be justified experimentally for several liquid phases:

$$\boldsymbol{q}_\alpha = -\frac{k_{r\alpha}}{\mu_\alpha}\, \boldsymbol{k}\, (\nabla p_\alpha + \varrho_\alpha g \boldsymbol{e}_z)\ .$$

Here the *relative permeability* $k_{r\alpha}$ of the phase α depends upon the saturations of the present phases and takes values in $[0,1]$.

At the interface of two liquid phases α_1 and α_2 we observe a difference of the pressures, the so-called *capillary pressure*, that turns out experimentally to be a function of the saturations:

$$p_{c\alpha_1\alpha_2} := p_{\alpha_1} - p_{\alpha_2} = F_{\alpha_1\alpha_2}(S_w, S_o, S_g) \,. \tag{0.22}$$

A general model for multiphase flow, formulated for the moment in terms of the variables p_α, S_α, is thus given by the equations

$$\partial_t(\phi S_\alpha \varrho_\alpha) - \nabla \cdot (\varrho_\alpha \lambda_\alpha \boldsymbol{k}(\nabla p_\alpha + \varrho_\alpha g \boldsymbol{e}_z)) = f_\alpha \tag{0.23}$$

with the *mobilities* $\lambda_\alpha := k_{r\alpha}/\mu_\alpha$, and the equations (0.22) and (0.10), where one of the S_α's can be eliminated. For two liquid phases w and g, e.g., water and air, equations (0.22) and (0.10) for $\alpha = w, g$ read $p_c = p_g - p_w = F(S_w)$ and $S_g = 1 - S_w$. Apparently, this is a time-dependent, nonlinear model in the variables p_w, p_g, S_w, where one of the variables can be eliminated. Assuming constant densities ϱ_α, further formulations based on

$$\nabla \cdot \big(\boldsymbol{q}_w + \boldsymbol{q}_g\big) = f_w/\varrho_w + f_g/\varrho_g \tag{0.24}$$

can be given as consequences of (0.10). These equations consist of a stationary equation for a new quantity, the *global pressure*, based on (0.24), and a time-dependent equation for one of the saturations (see Exercise 0.2). In many situations it is justified to assume a gaseous phase with constant pressure in the whole domain and to scale this pressure to $p_g = 0$. Thus for $\psi := p_w = -p_c$ we have

$$\phi \partial_t S(\psi) - \nabla \cdot (\lambda(\psi) \boldsymbol{k}(\nabla \psi + \varrho g \boldsymbol{e}_z)) = f_w/\varrho_w \tag{0.25}$$

with constant pressure $\varrho := \varrho_w$, and $S(\psi) := F^{-1}(-\psi)$ as a strictly monotone increasing nonlinearity as well as λ.

With the convention to set the value of the air pressure to 0, the pressure in the aqueous phase is in the *unsaturated state*, where the gaseous phase is also present, and represented by negative values. The water pressure $\psi = 0$ marks the transition from the unsaturated to the *saturated* zone. Thus in the unsaturated zone, equation (0.25) represents a nonlinear variant of the heat conduction equation for $\psi < 0$, the *Richards equation*. As most functional relationships have the property $S'(0) = 0$, the equation degenerates in the absence of a gaseous phase, namely to a stationary equation in a way that is different from above.

Equation (0.25) with $S(\psi) := 1$ and $\lambda(\psi) := \lambda(0)$ can be continued in a consistent way with (0.14) and (0.15) also for $\psi \geq 0$, i.e., for the case of a sole aqueous phase. The resulting equation is also called Richards equation or a model of *saturated-unsaturated flow*.

0.4 Reactive Solute Transport in Porous Media

In this chapter we will discuss the transport of a single component in a liquid phase and some selected reactions. We will always refer to water as liquid phase explicitly. Although we treat *inhomogeneous reactions* in terms of surface reactions with the solid phase, we want to ignore exchange processes between the fluid phases. On the microscopic scale the mass conservation law for a single component η is, in the notation of (0.11) by omitting the phase index w,

$$\partial_t \tilde{\varrho}_\eta + \nabla \cdot (\tilde{\varrho}_\eta \tilde{\boldsymbol{q}}) + \nabla \cdot \boldsymbol{J}_\eta = \tilde{Q}_\eta \,,$$

where

$$\boldsymbol{J}_\eta := \tilde{\varrho}_\eta \, (\tilde{\boldsymbol{v}}_\eta - \tilde{\boldsymbol{q}}) \, [\mathtt{kg/m^2/s}] \tag{0.26}$$

represents the *diffusive mass flux* of the component η and $\tilde{Q}_\eta$ $[\mathtt{kg/m^3/s}]$ is its *volumetric production rate*. For a description of reactions via the *mass action law* it is appropriate to choose the mole as the unit of mass. The diffusive mass flux requires a phenomenological description. The assumption that solely binary molecular diffusion, described by *Fick's law*, acts between the component η and the solvent, means that

$$\boldsymbol{J}_\eta = -\tilde{\varrho} D_\eta \nabla \, (\tilde{\varrho}_\eta / \tilde{\varrho}) \tag{0.27}$$

with a *molecular diffusivity* $D_\eta > 0$ $[\mathtt{m^2/s}]$. The averaging procedure applied on (0.26), (0.27) leads to

$$\partial_t(\Theta c_\eta) + \nabla \cdot (\boldsymbol{q} c_\eta) + \nabla \cdot \boldsymbol{J}^{(1)} + \nabla \cdot \boldsymbol{J}^{(2)} = Q_\eta^{(1)} + Q_\eta^{(2)}$$

for the *solute concentration* of the component η, c_η $[\mathtt{kg/m^3}]$, as intrinsic phase average of $\tilde{\varrho}_\eta$. Here, we have $\boldsymbol{J}^{(1)}$ as the average of $\boldsymbol{J}_\eta$ and $\boldsymbol{J}^{(2)}$, the mass flux due to *mechanical dispersion*, a newly emerging term at the macroscopic scale. Analogously, $Q_\eta^{(1)}$ is the intrinsic phase average of $\tilde{Q}_\eta$, and $Q_\eta^{(2)}$ is a newly emerging term describing the exchange between the liquid and solid phases.

The *volumetric water content* is given by $\Theta := \phi S_\mathrm{w}$ with the water saturation S_w. Experimentally, the following phenomenological descriptions are suggested:

$$\boldsymbol{J}^{(1)} = -\Theta \tau D_\eta \nabla c_\eta$$

with a *tortuosity factor* $\tau \in (0, 1]$,

$$\boldsymbol{J}^{(2)} = -\Theta \boldsymbol{D}_\mathrm{mech} \nabla c_\eta \,, \tag{0.28}$$

and a symmetric positive definite *matrix of mechanical dispersion* $\boldsymbol{D}_\mathrm{mech}$, which depends on q/Θ. Consequently, the resulting differential equation reads

$$\partial_t(\Theta c_\eta) + \nabla \cdot (\boldsymbol{q} c_\eta - \Theta \boldsymbol{D} \nabla c_\eta) = Q_\eta \tag{0.29}$$

with $\boldsymbol{D} := \tau \boldsymbol{D}_\eta + \boldsymbol{D}_{\text{mech}}$, $Q_\eta := Q_\eta^{(1)} + Q_\eta^{(2)}$.

Because the mass flux consists of $\boldsymbol{q}c_\eta$, a part due to *forced convection*, and of $\boldsymbol{J}^{(1)} + \boldsymbol{J}^{(2)}$, a part that corresponds to a generalized Fick's law, an equation like (0.29) is called a *convection-diffusion equation*. Accordingly, for the part with first spatial derivatives like $\nabla \cdot (\boldsymbol{q}c_\eta)$ the term *convective part* is used, and for the part with second spatial derivatives like $-\nabla \cdot (\Theta \boldsymbol{D} \nabla c_\eta)$ the term *diffusive part* is used. If the first term determines the character of the solution, the equation is called *convection-dominated*. The occurrence of such a situation is measured by the quantity Pe, the *global Péclet number*, that has the form $\text{Pe} = \|\boldsymbol{q}\| L / \|\Theta \boldsymbol{D}\|$ [-]. Here L is a characteristic length of the domain Ω. The extreme case of purely convective transport results in a conservation equation of first order. Since the common models for the dispersion matrix lead to a bound for Pe, the reduction to the purely convective transport is not reasonable. However, we have to take convection-dominated problems into consideration.

Likewise, we speak of diffusive parts in (0.17) and (0.20) and of (nonlinear) diffusive and convective parts in (0.21) and (0.25). Also, the multiphase transport equation can be formulated as a nonlinear convection-diffusion equation by use of (0.24) (see Exercise 0.2), where convection often dominates. If the production rate Q_η is independent of c_η, equation (0.29) is linear.

In general, in case of a surface reaction of the component η, the kinetics of the reaction have to be described . If this component is not in competition with the other components, one speaks of *adsorption*. The kinetic equation thus takes the general form

$$\partial_t s_\eta(x,t) = k_\eta f_\eta(x, c_\eta(x,t), s_\eta(x,t)) \tag{0.30}$$

with a rate parameter k_η for the *sorbed concentration* s_η [kg/kg], which is given in reference to the mass of the solid matrix. Here, the components in sorbed form are considered spatially immobile. The conservation of the total mass of the component undergoing sorption gives

$$Q_\eta^{(2)} = -\varrho_b \partial_t s_\eta \tag{0.31}$$

with the *bulk density* $\varrho_b = \varrho_{\text{s}}(1-\phi)$, where ϱ_{s} denotes the density of the solid phase. With (0.30), (0.31) we have a system consisting of an instationary partial and an ordinary differential equation (with $x \in \Omega$ as parameter). A widespread model by *Langmuir* reads

$$f_\eta = k_a c_\eta (\overline{s}_\eta - s_\eta) - k_d s_\eta$$

with constants k_a, k_d that depend upon the temperature (among other factors), and a *saturation concentration* $\overline{s}_\eta$ (cf. for example [24]). If we assume $f_\eta = f_\eta(x, c_\eta)$ for simplicity, we get a scalar nonlinear equation in c_η,

$$\partial_t(\Theta c_\eta) + \nabla \cdot (\boldsymbol{q}c_\eta - \Theta \boldsymbol{D} \nabla c_\eta) + \varrho_b k_\eta f_\eta(\cdot, c_\eta) = Q_\eta^{(1)}\,, \tag{0.32}$$

and s_η is decoupled and extracted from (0.30). If the time scales of transport and reaction differ greatly, and the limit case $k_\eta \to \infty$ is reasonable, then (0.30) is replaced by

$$f_\eta(x, c_\eta(x,t), s_\eta(x,t)) = 0 \; .$$

If this equation is solvable for s_η, i.e.,

$$s_\eta(x,t) = \varphi_\eta(x, c_\eta(x,t)) \; ,$$

the following scalar equation for c_η with a nonlinearity in the time derivative emerges:

$$\partial_t(\Theta c_\eta + \varrho_b \varphi_\eta(\cdot, c_\eta)) + \nabla \cdot (\boldsymbol{q} c_\eta - \Theta \boldsymbol{D} \nabla c_\eta) = Q_\eta^{(1)} \; .$$

If the component η is in competition with other components in the surface reaction, as, e.g., in ion exchange, then f_η has to be replaced by a nonlinearity that depends on the concentrations of all involved components $c_1, \ldots, c_N$, $s_1, \ldots, s_N$. Thus we obtain a coupled system in these variables. Finally, if we encounter *homogeneous reactions* that take place solely in the fluid phase, an analogous statement is true for the source term $Q_\eta^{(1)}$.

Exercises

0.1 Give a geometric interpretation for the matrix condition of $\boldsymbol{k}$ in (0.16) and $\boldsymbol{D}_{\text{mech}}$ in (0.28).

0.2 Consider the two-phase flow (with constant ϱ_α, $\alpha \in \{\text{w}, \text{g}\}$)

$$\begin{aligned} \partial_t(\phi S_\alpha) + \nabla \cdot \boldsymbol{q}_\alpha &= f_\alpha \; , \\ \boldsymbol{q}_\alpha &= -\lambda_\alpha \boldsymbol{k} \left(\nabla p_\alpha + \varrho_\alpha g \boldsymbol{e}_z \right) \; , \\ S_\text{w} + S_\text{g} &= 1 \, , \\ p_\text{g} - p_\text{w} &= p_\text{c} \end{aligned}$$

with coefficient functions

$$p_\text{c} = p_\text{c}(S_\text{w}) \; , \quad \lambda_\alpha = \lambda_\alpha(S_\text{w}) \; , \quad \alpha \in \{\text{w}, \text{g}\}.$$

Starting from equation (0.23), perform a transformation to the new variables

$$\begin{aligned} \boldsymbol{q} &= \boldsymbol{q}_\text{w} + \boldsymbol{q}_\text{g} \; , && \text{“total flow,”} \\ p &= \frac{1}{2}(p_\text{w} + p_\text{g}) + \frac{1}{2} \int_{S_\text{c}}^{S} \frac{\lambda_\text{g} - \lambda_\text{w}}{\lambda_\text{g} + \lambda_\text{w}} \frac{dp_\text{c}}{d\xi} \, d\xi \; , && \text{“global pressure,”} \end{aligned}$$

and the water saturation S_w. Derive a representation of the phase flows in the new variables.

0.3 A frequently employed model for mechanical dispersion is

$$D_{\text{mech}} = \lambda_{\text{L}}|\boldsymbol{v}|_2 P_{\boldsymbol{v}} + \lambda_{\text{T}}|\boldsymbol{v}|_2(I - P_{\boldsymbol{v}})$$

with parameters $\lambda_{\text{L}} > \lambda_{\text{T}}$, where $\boldsymbol{v} = \boldsymbol{q}/\Theta$ and $P_{\boldsymbol{v}} = \boldsymbol{v}\boldsymbol{v}^T/|\boldsymbol{v}|_2^2$. Here λ_{L} and λ_{T} are the *longitudinal* and *transversal dispersion lengths*. Give a geometrical interpretation.

0.5 Boundary and Initial Value Problems

The differential equations that we derived in Sections 0.3 and 0.4 have the common form

$$\partial_t S(u) + \nabla \cdot (\boldsymbol{C}(u) - \boldsymbol{K}(\nabla u)) = Q(u) \tag{0.33}$$

with a *source term* S, a convective part $\boldsymbol{C}$, a diffusive part $\boldsymbol{K}$, i.e., a total flux $\boldsymbol{C} - \boldsymbol{K}$ and a source term Q, which depend linearly or nonlinearly on the unknown u. For simplification, we assume u to be a scalar. The nonlinearities $S, \boldsymbol{C}, \boldsymbol{K}$, and Q may also depend on x and t, which shall be suppressed in the notation in the following. Such an equation is said to be in *divergence form* or in *conservative form*; a more general formulation is obtained by differentiating $\nabla \cdot \boldsymbol{C}(u) = \frac{\partial}{\partial u}\boldsymbol{C}(u) \cdot \nabla u + (\nabla \cdot \boldsymbol{C})(u)$ or by introducing a generalized "source term" $Q = Q(u, \nabla u)$. Up to now we have considered differential equations pointwise in $x \in \Omega$ (and $t \in (0, T)$) under the assumption that all occurring functions are well-defined. Due to the applicability of the integral theorem of Gauss on $\tilde{\Omega} \subset \Omega$ (cf. (3.10)), the *integral form* of the conservation equation follows straightforwardly from the above:

$$\int_{\tilde{\Omega}} \partial_t S(u)\, dx + \int_{\partial\tilde{\Omega}} (\boldsymbol{C}(u) - \boldsymbol{K}(\nabla u)) \cdot \nu \, d\sigma = \int_{\tilde{\Omega}} Q(u, \nabla u)\, dx \tag{0.34}$$

with the outer unit normal ν (see Theorem 3.8) for a fixed time t or also in t integrated over $(0, T)$. Indeed, this equation (on the microscopic scale) is the primary description of the conservation of an extensive quantity: Changes in time through storage and sources in $\tilde{\Omega}$ are compensated by the normal flux over $\partial\tilde{\Omega}$. Moreover, for $\partial_t S$, $\nabla \cdot (\boldsymbol{C} - \boldsymbol{K})$, and Q continuous on the closure of $\tilde{\Omega}$, (0.33) follows from (0.34). If, on the other hand, F is a hyperplane in $\tilde{\Omega}$ where the material properties may rapidly change, the *jump condition*

$$[(\boldsymbol{C}(u) - \boldsymbol{K}(\nabla u)) \cdot \nu] = 0 \tag{0.35}$$

for a fixed unit normal ν on F follows from (0.34), where $[\,\cdot\,]$ denotes the difference of the one-sided limits (see Exercise 0.4).

Since the differential equation describes conservation only in general, it has to be supplemented by initial and boundary conditions in order to

specify a particular situation where a unique solution is expected. Boundary conditions are specifications on $\partial\Omega$, where ν denotes the outer unit normal

- of the normal component of the flux (inwards):

$$-(\boldsymbol{C}(u) - \boldsymbol{K}(\nabla u)) \cdot \nu = g_1 \quad \text{on } \Gamma_1 \tag{0.36}$$

(flux boundary condition),

- of a linear combination of the normal flux and the unknown itself:

$$-(\boldsymbol{C}(u) - \boldsymbol{K}(\nabla u)) \cdot \nu + \alpha u = g_2 \quad \text{on } \Gamma_2 \tag{0.37}$$

(mixed boundary condition),

- of the unknown itself:

$$u = g_3 \quad \text{on } \Gamma_3 \tag{0.38}$$

(Dirichlet boundary condition).

Here $\Gamma_1, \Gamma_2, \Gamma_3$ form a disjoint decomposition of $\partial\Omega$:

$$\partial\Omega = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3 \,, \tag{0.39}$$

where Γ_3 is supposed to be a closed subset of $\partial\Omega$. The *inhomogeneities* g_i and the factor α in general depend on $x \in \Omega$, and for nonstationary problems (where $S(u) \neq 0$ holds) on $t \in (0, T)$. The boundary conditions are linear if the g_i do not depend (nonlinearly) on u (see below). If the g_i are zero, we speak of *homogeneous*, otherwise of *inhomogeneous, boundary conditions*.

Thus the pointwise formulation of a nonstationary equation (where S does not vanish) requires the validity of the equation in the *space-time cylinder*

$$Q_T := \Omega \times (0, T)$$

and the boundary conditions on the *lateral surface* of the space-time cylinder

$$S_T := \partial\Omega \times (0, T) \,.$$

Different types of boundary conditions are possible with decompositions of the type (0.39). Additionally, an *initial condition* on the *bottom* of the space-time cylinder is necessary:

$$S(u(x, 0)) = S_0(x) \quad \text{for } x \in \Omega \,. \tag{0.40}$$

These are so-called *initial-boundary value problems*; for stationary problems we speak of *boundary value problems*. As shown in (0.34) and (0.35) flux boundary conditions have a natural relationship with the differential equation (0.33). For a linear diffusive part $\boldsymbol{K}(\nabla u) = \boldsymbol{K}\nabla u$ alternatively we may require

$$\partial_{\nu_K} u := \boldsymbol{K}\nabla u \cdot \nu = g_1 \quad \text{on } \Gamma_1 \,, \tag{0.41}$$

and an analogous mixed boundary condition. This boundary condition is the so-called *Neumann boundary condition*. Since $\boldsymbol{K}$ is symmetric, $\partial_{\nu_K} u = \nabla u \cdot \boldsymbol{K}\nu$ holds; i.e., $\partial_{\nu_K} u$ is the derivative in direction of the *conormal* $\boldsymbol{K}\nu$. For the special case $\boldsymbol{K} = \boldsymbol{I}$ the normal derivative is given.

In contrast to ordinary differential equations, there is hardly any general theory of partial differential equations. In fact, we have to distinguish different types of differential equations according to the various described physical phenomena. These determine, as discussed, different (initial-) boundary value specifications to render the problem *well-posed*. *Well-posedness* means that the problem possesses a unique solution (with certain properties yet to be defined) that depends continuously (in appropriate norms) on the data of the problem, in particular on the (initial and) boundary values. There exist also *ill-posed* boundary value problems for partial differential equations, which correspond to physical and technical applications. They require special techniques and shall not be treated here.

The classification into different types is simple if the problem is linear and the differential equation is of second order as in (0.33). By *order* we mean the highest order of the derivative with respect to the variables $(x_1, \ldots, x_d, t)$ that appears, where the time derivative is considered to be like a spatial derivative. Almost all differential equations treated in this book will be of second order, although important models in elasticity theory are of fourth order or certain transport phenomena are modelled by systems of first order.

The differential equation (0.33) is generally *nonlinear* due to the nonlinear relationships $S, \boldsymbol{C}, \boldsymbol{K}$, and Q. Such an equation is called *quasilinear* if all derivatives of the highest order are linear, i.e., we have

$$\boldsymbol{K}(\nabla u) = \boldsymbol{K}\nabla u \tag{0.42}$$

with a matrix $\boldsymbol{K}$, which may also depend (nonlinearly) on x, t, and u. Furthermore, (0.33) is called *semilinear* if nonlinearities are present only in u, but not in the derivatives, i.e., if in addition to (0.42) with $\boldsymbol{K}$ being independent of u, we have

$$S(u) = Su\,, \quad \boldsymbol{C}(u) = u\boldsymbol{c} \tag{0.43}$$

with scalar and vectorial functions S and $\boldsymbol{c}$, respectively, which may depend on x and t. Such variable factors standing before u or differential terms are called *coefficients* in general.

Finally, the differential equation is *linear* if we have, in addition to the above requirements,

$$Q(u) = -ru + f$$

with functions r and f of x and t.

In the case $f = 0$ the linear differential equation is termed *homogeneous*, otherwise *inhomogeneous*. A linear differential equation obeys the *superposition principle*: Suppose u_1 and u_2 are solutions of (0.33) with the

source terms f_1 and f_2 and otherwise identical coefficient functions. Then $u_1 + \gamma u_2$ is a solution of the same differential equation with the source term $f_1 + \gamma f_2$ for arbitrary $\gamma \in \mathbb{R}$. The same holds for linear boundary conditions. The term *solution* of an (initial-) boundary value problem is used here in a classical sense, yet to be specified, where all the quantities occurring should satisfy pointwise certain regularity conditions (see Definition 1.1 for the Poisson equation). However, for variational solutions (see Definition 2.2), which are appropriate in the framework of finite element methods, the above statements are also valid.

Linear differential equations of second order in two variables (x, y) (including possibly the time variable) can be classified in different *types* as follows:

To the homogeneous differential equation

$$\begin{aligned} Lu &= a(x,y)\frac{\partial^2}{\partial x^2}u + b(x,y)\frac{\partial^2}{\partial x \partial y}u + c(x,y)\frac{\partial^2}{\partial y^2}u \\ &\quad + d(x,y)\frac{\partial}{\partial x}u + e(x,y)\frac{\partial}{\partial y}u + f(x,y)u = 0 \end{aligned} \tag{0.44}$$

the following quadratic form is assigned:

$$(\xi, \eta) \mapsto a(x,y)\xi^2 + b(x,y)\xi\eta + c(x,y)\eta^2. \tag{0.45}$$

According to its eigenvalues, i.e., the eigenvalues of the matrix

$$\begin{pmatrix} a(x,y) & \frac{1}{2}b(x,y) \\ \frac{1}{2}b(x,y) & c(x,y) \end{pmatrix}, \tag{0.46}$$

we classify the types. In analogy with the classification of conic sections, which are described by (0.45) (for fixed (x, y)), the differential equation (0.44) is called *at the point* (x, y)

- *elliptic* if the eigenvalues of (0.46) are not 0 and have the same sign,
- *hyperbolic* if one eigenvalue is positive and the other is negative,
- *parabolic* if exactly one eigenvalue is equal to 0.

For the corresponding generalization of the terms for $d + 1$ variables and arbitrary order, the stationary boundary value problems we treat in this book will be elliptic, of second order, and — except in Chapter 8 — also linear; the nonstationary initial-boundary value problems will be parabolic.

Systems of hyperbolic differential equations of first order require particular approaches, which are beyond the scope of this book. Nevertheless, we dedicate Chapter 9 to convection-dominated problems, i.e., elliptic or parabolic problems close to the hyperbolic limit case.

The different discretization strategies are based on various formulations of the (initial-) boundary value problems: The *finite difference method*, which is presented in Section 1, and further outlined for nonstationary problems in Chapter 7, has the pointwise formulation of (0.33), (0.36)–(0.38)

(and (0.40)) as a starting point. The *finite element method*, , which lies in the focus of our book (Chapters 2, 3, and 7), is based on an integral formulation of (0.33) (which we still have to depict) that incorporates (0.36) and (0.37). The conditions (0.38) and (0.40) have to be enforced additionally. Finally, the *finite volume method* (Chapters 6 and 7) will be derived from the integral formulation (0.34), where also initial and boundary conditions come along as in the finite element approach.

Exercises

0.4 Derive (formally) (0.35) from (0.34).

0.5 Derive the orders of the given differential operators and differential equations, and decide in every case whether the operator is linear or nonlinear, and whether the linear equation is homogeneous or inhomogeneous:

(a) $Lu := u_{xx} + xu_y$,
(b) $Lu := u_x + uu_y$,
(c) $Lu := \sqrt{1+x^2}(\cos y)u_x + u_{yxy} - \left(\arctan \frac{x}{y}\right) u = \ln(x^2+y^2)$,
(d) $Lu := u_t + u_{xxxx} + \sqrt{1+u} = 0$,
(e) $u_{tt} - u_{xx} + x^2 = 0$.

0.6 (a) Determine the type of the given differential operator:
(i) $Lu := u_{xx} - u_{xy} + 2u_y + u_{yy} - 3u_{yx} + 4u$,
(ii) $Lu = 9u_{xx} + 6u_{xy} + u_{yy} + u_x$.

(b) Determine the parts of the plane where the differential operator $Lu := yu_{xx} - 2u_{xy} + xu_{yy}$ is elliptic, hyperbolic, or parabolic.

(c) (i) Determine the type of $Lu := 3u_y + u_{xy}$.
(ii) Compute the general solution of $Lu = 0$.

0.7 Consider the equation $Lu = f$ with a linear differential operator of second order, defined for functions in d variables ($d \in \mathbb{N}$) in $x \in \Omega \subset \mathbb{R}^d$. The transformation $\Phi : \Omega \to \Omega' \subset \mathbb{R}^d$ has a continuously differentiable, nonsingular Jacobi matrix $D\Phi := \frac{\partial \Phi}{\partial x}$.

Show that the partial differential equation does not change its type if it is written in the new coordinates $\xi = \Phi(x)$.

1

For the Beginning: The Finite Difference Method for the Poisson Equation

1.1 The Dirichlet Problem for the Poisson Equation

In this section we want to introduce the finite difference method using the Poisson equation on a rectangle as an example. By means of this example and generalizations of the problem, advantages and limitations of the approach will be elucidated. Also, in the following section the Poisson equation will be the main topic, but then on an arbitrary domain. For the spatial basic set of the differential equation $\Omega \subset \mathbb{R}^d$ we assume as minimal requirement that Ω is a domain, where a *domain* is a nonempty, open, and connected set. The boundary of this domain will be denoted by $\partial\Omega$, the closure $\Omega \cup \partial\Omega$ by $\overline{\Omega}$ (see Appendix A.2). The *Dirichlet problem for the Poisson equation* is then defined as follows: Given functions $g : \partial\Omega \to \mathbb{R}$ and $f : \Omega \to \mathbb{R}$, we are looking for a function $u : \overline{\Omega} \to \mathbb{R}$ such that

$$-\sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2} u = f \quad \text{in } \Omega\,, \tag{1.1}$$

$$u = g \quad \text{on } \partial\Omega\,. \tag{1.2}$$

This differential equation model has already appeared in (0.19) and (0.38) and beyond this application has an importance in a wide spectrum of disciplines. The unknown function u can be interpreted as an electromagnetic potential, a displacement of an elastic membrane, or a temperature. Similar to the multi-index notation to be introduced in (2.16) (but with

indices at the top) from now on for partial derivatives we use the following notation.

Notation: For $u : \Omega \subset \mathbb{R}^d \to \mathbb{R}$ we set

$$\begin{aligned} \partial_i u &:= \frac{\partial}{\partial x_i} u \qquad \text{for } i = 1, \ldots, d\,, \\ \partial_{ij} u &:= \frac{\partial^2}{\partial x_i \, \partial x_j} u \qquad \text{for } i, j = 1, \cdots, d\,, \\ \Delta u &:= (\partial_{11} + \ldots + \partial_{dd})\, u\,. \end{aligned}$$

The expression Δu is called the *Laplace operator*. By means of this, (1.1) can be written in abbreviated form as

$$-\Delta u = f \quad \text{in } \Omega\,. \tag{1.3}$$

We could also define the Laplace operator by

$$\Delta u = \nabla \cdot (\nabla u)\ ,$$

where $\nabla u = (\partial_1 u, \ldots, \partial_d u)^T$ denotes the *gradient* of a function u, and $\nabla \cdot v = \partial_1 v_1 + \cdots + \partial_d v_d$ the *divergence* of a vector field v. Therefore, an alternative notation exists, which will not be used in the following: $\Delta u = \nabla^2 u$. The incorporation of the minus sign in the left-hand side of (1.3), which looks strange at first glance, is related to the monotonicity and definiteness properties of $-\Delta$ (see Sections 1.4 and 2.1, respectively).

The notion of a solution for (1.1), (1.2) still has to specified more precisely. Considering the equations in a pointwise sense, which will be pursued in this chapter, the functions in (1.1), (1.2) have to exist, and the equations have to be satisfied pointwise. Since (1.1) is an equation on an open set Ω, there are no implications for the behaviour of u up to the boundary $\partial\Omega$. To have a real requirement due to the boundary condition, u has to be at least continuous up to the boundary, that is, on $\overline{\Omega}$. These requirements can be formulated in a compact way by means of corresponding function spaces. The function spaces are introduced more precisely in Appendix A.5. Some examples are

$$\begin{aligned} C(\Omega) &:= \left\{ u : \Omega \to \mathbb{R} \;\middle|\; u \text{ continuous in } \Omega \right\}\,, \\ C^1(\Omega) &:= \left\{ u : \Omega \to \mathbb{R} \;\middle|\; u \in C(\Omega)\,,\ \partial_i u \text{ exists in } \Omega\,, \right. \\ &\qquad\qquad \left. \partial_i u \in C(\Omega) \text{ for all } i = 1, \ldots, d \right\}\,. \end{aligned}$$

The spaces $C^k(\Omega)$ for $k \in \mathbb{N}$, $C(\overline{\Omega})$, and $C^k(\overline{\Omega})$, as well as $C(\partial\Omega)$, are defined analogously. In general, the requirements related to the (continuous) existence of derivatives are called, a little bit vaguely, *smoothness requirements*.

In the following, in view of the finite difference method, f and g will also be assumed continuous in Ω and $\partial\Omega$, respectively.

Definition 1.1 Assume $f \in C(\Omega)$ and $g \in C(\partial\Omega)$. A function u is called a (*classical*) *solution* of (1.1), (1.2) if $u \in C^2(\Omega) \cap C(\overline{\Omega})$, (1.1) holds for all $x \in \Omega$, and (1.2) holds for all $x \in \partial\Omega$.

1.2 The Finite Difference Method

The finite difference method is based on the following approach: We are looking for an approximation to the solution of a boundary value problem at a finite number of points in $\overline{\Omega}$ (the *grid points*). For this reason we substitute the derivatives in (1.1) by difference quotients, which involve only function values at grid points in Ω and require (1.2) only at grid points. By this we obtain algebraic equations for the approximating values at grid points. In general, such a procedure is called the *discretization* of the boundary value problem. Since the boundary value problem is linear, the system of equations for the approximate values is also linear. In general, for other (differential equation) problems and other discretization approaches we also speak of the *discrete problem* as an *approximation* of the *continuous problem*. The aim of further investigations will be to estimate the resulting error and thus to judge the quality of the approximative solution.

Generation of Grid Points
In the following, for the beginning, we will restrict our attention to problems in two space dimensions ($d = 2$). For simplification we consider the case of a constant *step size* (or *mesh width*) $h > 0$ in both space directions. The quantity h here is the *discretization parameter*, which in particular determines the dimension of the discrete problem.

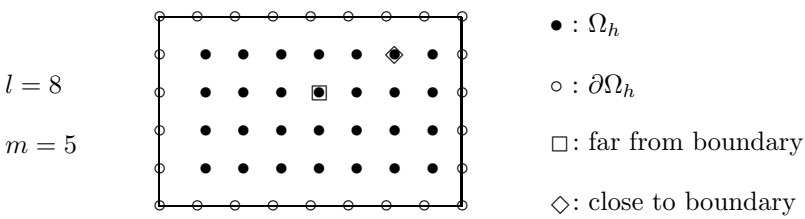


Figure 1.1. Grid points in a square domain.

For the time being, let Ω be a rectangle, which represents the simplest case for the finite difference method (see Figure 1.1). By translation of the coordinate system the situation can be reduced to $\Omega = (0, a) \times (0, b)$ with $a, b > 0$. We assume that the lengths a, b, and h are such that

$$a = lh, \quad b = mh \quad \text{for certain } l, m \in \mathbb{N}. \tag{1.4}$$

We define

$$\begin{aligned} \Omega_h &:= \left\{(ih, jh) \;\middle|\; i = 1, \ldots, l-1\,,\; j = 1, \ldots, m-1\right\} \\ &= \left\{(x, y) \in \Omega \;\middle|\; x = ih\,,\; y = jh \text{ with } i, j \in \mathbb{Z}\right\} \end{aligned} \tag{1.5}$$

as a set of *grid points in* Ω in which an approximation of the differential equation has to be satisfied. In the same way,

$$\begin{aligned} \partial\Omega_h &:= \left\{(ih, jh) \;\middle|\; i \in \{0, l\}\,,\; j \in \{0, \ldots, m\} \text{ or } i \in \{0, \ldots, l\}\,,\; j \in \{0, m\}\right\} \\ &= \left\{(x, y) \in \partial\Omega \;\middle|\; x = ih\,,\; y = jh \text{ with } i, j \in \mathbb{Z}\right\} \end{aligned}$$

defines the *grid points on* $\partial\Omega$ in which an approximation of the boundary condition has to be satisfied. The union of grid points will be denoted by

$$\overline{\Omega}_h := \Omega_h \cup \partial\Omega_h.$$

Setup of the System of Equations

Lemma 1.2 *Let* $\Omega := (x-h, x+h)$ *for* $x \in \mathbb{R}$, $h > 0$. *Then there exists a quantity* R, *depending on* u *and* h, *the absolute value of which can be bounded independently of* h *and such that*

(1) for $u \in C^2(\overline{\Omega})$:

$$u'(x) = \frac{u(x+h) - u(x)}{h} + hR \quad \textit{and} \quad |R| \le \frac{1}{2}\|u''\|_\infty\,,$$

(2) for $u \in C^2(\overline{\Omega})$:

$$u'(x) = \frac{u(x) - u(x-h)}{h} + hR \quad \textit{and} \quad |R| \le \frac{1}{2}\|u''\|_\infty\,,$$

(3) for $u \in C^3(\overline{\Omega})$:

$$u'(x) = \frac{u(x+h) - u(x-h)}{2h} + h^2R \quad \textit{and} \quad |R| \le \frac{1}{6}\|u'''\|_\infty\,,$$

(4) for $u \in C^4(\overline{\Omega})$:

$$u''(x) = \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} + h^2R \textit{ and } |R| \le \frac{1}{12}\|u^{(4)}\|_\infty\,.$$

Here the maximum norm $\|\cdot\|_\infty$ *(see Appendix A.5) has to be taken over the interval of the involved points* $(x, x+h)$, $(x-h, x)$, *or* $(x-h, x+h)$.

Proof: The proof follows immediately by Taylor expansion. As an example we consider statement 3: From

$$u(x \pm h) = u(x) \pm hu'(x) + \frac{h^2}{2}u''(x) \pm \frac{h^3}{6}u'''(x \pm \xi_\pm) \text{ for certain } \xi_\pm \in (0, h)$$

the assertion follows by linear combination. □

Notation: The quotient in statement 1 is called the *forward difference quotient*, and it is denoted by $\partial^+ u(x)$. The quotient in statement 2 is called the *backward difference quotient* ($\partial^- u(x)$), and the one in statement 3 the *symmetric difference quotient* ($\partial^0 u(x)$). The quotient appearing in statement 4 can be written as $\partial^- \partial^+ u(x)$ by means of the above notation.

In order to use statement 4 in every space direction for the approximation of $\partial_{11} u$ and $\partial_{22} u$ in a grid point (ih, jh), in addition to the conditions of Definition 1.1, the further smoothness properties $\partial^{(3,0)} u, \partial^{(4,0)} u \in C(\overline{\Omega})$ and analogously for the second coordinate are necessary. Here we use, e.g., the notation $\partial^{(3,0)} u := \partial^3 u / \partial x_1^3$ (see (2.16)).

Using these approximations for the boundary value problem (1.1), (1.2), at each grid point $(ih, jh) \in \Omega_h$ we get

$$\begin{aligned} -\Big(& \frac{u\left((i+1)h, jh\right) - 2u(ih, jh) + u\left((i-1)h, jh\right)}{h^2} \\ & + \frac{u\left(ih, (j+1)h\right) - 2u(ih, jh) + u\left(ih, (j-1)h\right)}{h^2} \Big) = \\ & = \quad f(ih, jh) + R(ih, jh) h^2 . \end{aligned} \tag{1.6}$$

Here R is as described in statement 4 of Lemma 1.2, a function depending on the solution u and on the step size h, but the absolute value of which can be bounded independently of h. In cases where we have less smoothness of the solution u, we can nevertheless formulate the approximation (1.6) for $-\Delta u$, but the size of the error in the equation is unclear at the moment.

For the grid points $(ih, jh) \in \partial\Omega_h$ no approximation of the boundary condition is necessary:

$$u(ih, jh) = g(ih, jh) \; .$$

If we neglect the term Rh^2 in (1.6), we get a system of linear equations for the approximating values u_{ij} for $u(x, y)$ at points $(x, y) = (ih, jh) \in \overline{\Omega}_h$. They have the form

$$\frac{1}{h^2}\big(- u_{i,j-1} - u_{i-1,j} + 4u_{ij} - u_{i+1,j} - u_{i,j+1}\big) = f_{ij} \tag{1.7}$$

$$\text{for } i = 1, \ldots, l-1 \; , \; j = 1, \ldots, m-1 \; ,$$

$$u_{ij} = g_{ij} \qquad \text{if } i \in \{0, l\}, \, j = 0, \ldots, m \text{ or } j \in \{0, m\}, \, i = 0, \ldots, l \; . \tag{1.8}$$

Here we used the abbreviations

$$f_{ij} := f(ih, jh), \quad g_{ij} := g(ih, jh) \; . \tag{1.9}$$

Therefore, for each unknown grid value u_{ij} we get an equation. The grid points (ih, jh) and the approximating values u_{ij} located at these have a natural two-dimensional indexing.

In equation (1.7) for a grid point (i, j) only the *neighbours* at the four cardinal points of the compass appear, as it is displayed in Figure 1.2. This

interconnection is also called the *five-point stencil* of the difference method and the method the *five-point stencil discretization*.

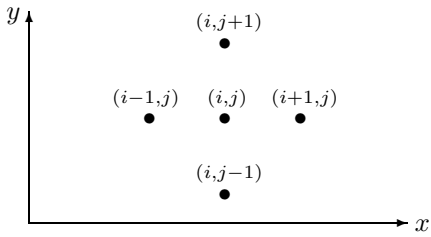


Figure 1.2. Five-point stencil.

At the interior grid points $(x, y) = (ih, jh) \in \Omega_h$, two cases can be distinguished:

(1) (i, j) has a position such that its all neighbouring grid points lie in Ω_h (*far from the boundary*).

(2) (i, j) has a position such that at least one neighbouring grid point (r, s) lies on $\partial\Omega_h$ (*close to the boundary*). Then in equation (1.7) the value u_{rs} is known due to (1.8) ($u_{rs} = g_{rs}$), and (1.7) can be modified in the following way:
Remove the values u_{rs} with $(rh, sh) \in \partial\Omega_h$ in the equations for (i, j) close to the boundary and add the value g_{rs}/h^2 to the right-hand side of (1.7). The set of equations that arises by this elimination of boundary unknowns by means of Dirichlet boundary conditions we call $(1.7)^*$; it is equivalent to (1.7), (1.8).

Instead of considering the values u_{ij}, $i = 1, \ldots, l-1$, $j = 1, \ldots, m-1$, one also speaks of the *grid function* $u_h : \Omega_h \to \mathbb{R}$, where $u_h(ih, jh) = u_{ij}$ for $i = 1, \ldots, l-1$, $j = 1, \ldots, m-1$. Grid functions on $\partial\Omega_h$ or on $\overline{\Omega}_h$ are defined analogously. Thus we can formulate the finite difference method in the following way: Find a grid function u_h on $\overline{\Omega}_h$ such that equations (1.7), (1.8) hold, or, equivalently find a grid function u_h on Ω_h such that equations $(1.7)^*$ hold.

Structure of the System of Equations
After choosing an ordering of the u_{ij} for $i = 0, \ldots, l$, $j = 0, \ldots, m$, the system of equations $(1.7)^*$ takes the following form:

$$A_h \boldsymbol{u}_h = \boldsymbol{q}_h \tag{1.10}$$

with $A_h \in \mathbb{R}^{M_1, M_1}$ and $\boldsymbol{u}_h, \boldsymbol{q}_h \in \mathbb{R}^{M_1}$, where $M_1 = (l-1)(m-1)$.

This means that nearly identical notations for the grid function and its representing vector are chosen for a fixed numbering of the grid points. The only difference is that the representing vector is printed in bold. The ordering of the grid points may be arbitrary, with the restriction that the

points in Ω_h are enumerated by the first M_1 indices, and the points in $\partial\Omega_h$ are labelled with the subsequent $M_2 = 2(l+m)$ indices. The structure of A_h is not influenced by this restriction.

Because of the described elimination process, the right-hand side $\boldsymbol{q}_h$ has the following form:

$$\boldsymbol{q}_h = -\hat{A}_h \boldsymbol{g} + \boldsymbol{f} \ , \tag{1.11}$$

where $\boldsymbol{g} \in \mathbb{R}^{M_2}$ and $\boldsymbol{f} \in \mathbb{R}^{M_1}$ are the vectors representing the grid functions

$$f_h : \Omega_h \to \mathbb{R} \quad \text{and} \quad g_h : \partial\Omega_h \to \mathbb{R}$$

according to the chosen numbering with the values defined in (1.9). The matrix $\hat{A}_h \in \mathbb{R}^{M_1, M_2}$ has the following form:

$$(\hat{A}_h)_{ij} = \begin{cases} -\dfrac{1}{h^2} & \text{if the node } i \text{ is close to the boundary} \\ & \text{and } j \text{ is a neighbour in the five-point stencil,} \\ 0 & \text{otherwise .} \end{cases} \tag{1.12}$$

For any ordering, only the diagonal element and at most four further entries in a row of A_h, defined by (1.7), are different from 0; that is, the matrix is *sparse* in a strict sense, as is assumed in Chapter 5.

An obvious ordering is the *rowwise* numbering of Ω_h according to the following scheme:

$$\begin{array}{ccccc} \substack{(h,b-h)\\(l-1)(m-2)+1} & \substack{(2h,b-h)\\(l-1)(m-2)+2} & \cdots & \cdots & \substack{(a-h,b-h)\\(l-1)(m-1)} \\ \substack{(h,b-2h)\\(l-1)(m-3)+1} & \substack{(2h,b-2h)\\(l-1)(m-3)+2} & \cdots & \cdots & \substack{(a-h,b-2h)\\(l-1)(m-2)} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \substack{(h,2h)\\l} & \substack{(2h,2h)\\l+1} & \cdots & \cdots & \substack{(a-h,2h)\\2l-2} \\ \substack{(h,h)\\1} & \substack{(2h,h)\\2} & \cdots & \cdots & \substack{(a-h,h)\\l-1} \end{array} \ . \tag{1.13}$$

Another name of the above scheme is *lexicographic* ordering. (However, this name is better suited to the *columnwise* numbering.)

In this case the matrix A_h has the following form of an $(m-1)\times(m-1)$ block tridiagonal matrix:

$$A_h = h^{-2} \begin{pmatrix} T & -I & & & & \\ -I & T & -I & & \mathbf{0} & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & \mathbf{0} & & -I & T & -I \\ & & & & -I & T \end{pmatrix} \tag{1.14}$$

with the unit matrix $I \in \mathbb{R}^{l-1,l-1}$ and

$$T = \begin{pmatrix} 4 & -1 & & & & \\ -1 & 4 & -1 & & 0 & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & -1 & 4 & -1 \\ & & & & -1 & 4 \end{pmatrix} \in \mathbb{R}^{l-1,l-1} .$$

We return to the consideration of an arbitrary numbering. In the following we collect several properties of the matrix $A_h \in \mathbb{R}^{M_1,M_1}$ and the extended matrix

$$\tilde{A}_h := \left(A_h \mid \hat{A}_h \right) \in \mathbb{R}^{M_1,M} ,$$

where $M := M_1 + M_2$. The matrix $\tilde{A}_h$ takes into account all the grid points in $\overline{\Omega}_h$. It has no relevance with the resolution of (1.10), but with the stability of the discretization, which will be investigated in Section 1.4.

- $(A_h)_{rr} > 0$ for all $r = 1, \ldots, M_1$,
- $(\tilde{A}_h)_{rs} \le 0$ for all $r = 1, \ldots, M_1,\ s = 1, \ldots, M$ such that $r \neq s$,
- $$\sum_{s=1}^{M_1} (A_h)_{rs} \begin{cases} \ge 0 & \text{for all } r = 1, \ldots, M_1, \\ > 0 & \text{if } r \text{ belongs to a grid point close to the boundary,} \end{cases} \qquad (1.15)$$
- $\sum_{s=1}^{M} (\tilde{A}_h)_{rs} = 0$ for all $r = 1, \ldots, M_1$,
- A_h is irreducible,
- A_h is regular.

Therefore, the matrix A_h is weakly row diagonally dominant (see Appendix A.3 for definitions from linear algebra). The irreducibility follows from the fact that two arbitrary grid points may be connected by a path consisting of corresponding neighbours in the five-point stencil. The regularity follows from the irreducible diagonal dominance. From this we can conclude that (1.10) can be solved by Gaussian elimination without pivot search. In particular, if the matrix has a band structure, this will be preserved. This fact will be explained in more detail in Section 2.5.

The matrix A_h has the following further properties:

- A_h is symmetric,
- A_h is positive definite.

It is sufficient to verify these properties for a fixed ordering, for example the rowwise one, since by a change of the ordering matrix, A_h is transformed to PA_hP^T with some regular matrix P, by which neither symmetry nor

positive definiteness is destroyed. Nevertheless, the second assertion is not obvious. One way to verify it is to compute eigenvalues and eigenvectors explicitly, but we refer to Chapter 2, where the assertion follows naturally from Lemma 2.13 and (2.36). The eigenvalues and eigenvectors are specified in (5.24) for the special case $l = m = n$ and also in (7.60). Therefore, (1.10) can be resolved by Cholesky's method, taking into account the bandedness.

Quality of the Approximation by the Finite Difference Method
We now address the following question: To what accuracy does the grid function u_h corresponding to the solution $\boldsymbol{u}_h$ of (1.10) approximate the solution u of (1.1), (1.2)?

To this end we consider the grid function $U : \Omega_h \to \mathbb{R}$, which is defined by

$$U(ih, jh) := u(ih, jh). \tag{1.16}$$

To measure the size of $U - u_h$, we need a norm (see Appendix A.4 and also A.5 for the subsequently used definitions). Examples are the *maximum norm*

$$\|u_h - U\|_\infty := \max_{\substack{i=1,\dots,l-1 \\ j=1,\dots,m-1}} |(u_h - U)(ih, jh)| \tag{1.17}$$

and the *discrete L^2-norm*

$$\|u_h - U\|_{0,h} := h \left(\sum_{i=1}^{l-1} \sum_{j=1}^{m-1} ((u_h - U)(ih, jh))^2 \right)^{1/2} . \tag{1.18}$$

Both norms can be conceived as the application of the continuous norms $\|\cdot\|_\infty$ of the function space $L^\infty(\Omega)$ or $\|\cdot\|_0$ of the function space $L^2(\Omega)$ to piecewise constant prolongations of the grid functions (with a special treatment of the area close to the boundary). Obviously, we have

$$\|v_h\|_{0,h} \le \sqrt{ab}\, \|v_h\|_\infty$$

for a grid function v_h, but the reverse estimate does not hold uniformly in h, so that $\|\cdot\|_\infty$ is a stronger norm. In general, we are looking for a norm $\|\cdot\|_h$ in the space of grid functions in which the method *converges* in the sense

$$\|u_h - U\|_h \to 0 \quad \text{for } h \to 0$$

or even has an *order of convergence* $p > 0$, by which we mean the existence of a constant $C > 0$ independent of h such that

$$\|u_h - U\|_h \le C\, h^p .$$

Due to the construction of the method, for a solution $u \in C^4(\overline{\Omega})$ we have

$$A_h \boldsymbol{U} = \boldsymbol{q}_h + h^2 \boldsymbol{R} ,$$

where $\boldsymbol{U}$ and $\boldsymbol{R} \in \mathbb{R}^{M_1}$ are the representations of the grid functions U and R according to (1.6) in the selected ordering. Therefore, we have:

$$A_h(\boldsymbol{u}_h - \boldsymbol{U}) = -h^2 \boldsymbol{R}$$

and thus

$$|A_h(\boldsymbol{u}_h - \boldsymbol{U})|_\infty = h^2 |\boldsymbol{R}|_\infty = Ch^2$$

with a constant $C(= |\boldsymbol{R}|_\infty) > 0$ independent of h.

From Lemma 1.2, 4. we conclude that

$$C = \frac{1}{12} \left(\|\partial^{(4,0)} u\|_\infty + \|\partial^{(0,4)} u\|_\infty \right) .$$

That is, for a solution $u \in C^4(\overline{\Omega})$ the method is *consistent* with the boundary value problem with an *order of consistency* 2. More generally, the notion takes the following form:

Definition 1.3 Let (1.10) be the system of equations that corresponds to a (finite difference) approximation on the grid points Ω_h with a discretization parameter h. Let $\boldsymbol{U}$ be the representation of the grid function that corresponds to the solution u of the boundary value problem according to (1.16). Furthermore, let $\|\cdot\|_h$ be a norm in the space of grid functions on Ω_h, and let $|\cdot|_h$ be the corresponding vector norm in the space $\mathbb{R}^{M_{1h}}$, where M_{1h} is the number of grid points in Ω_h. The approximation is called *consistent* with respect to $\|\cdot\|_h$ if

$$|A_h \boldsymbol{U} - \boldsymbol{q}_h|_h \to 0 \quad \text{for} \quad h \to 0 .$$

The approximation has the *order of consistency* $p > 0$ if

$$|A_h \boldsymbol{U} - \boldsymbol{q}_h|_h \leq Ch^p$$

with a constant $C > 0$ independent of h.

Thus the *consistency* or *truncation error* $A_h \boldsymbol{U} - \boldsymbol{q}_h$ measures the quality of how the exact solution satisfies the approximating equations. As we have seen, in general it can be determined easily by Taylor expansion, but at the expense of unnaturally high smoothness assumptions. But one has to be careful in expecting the error $|\boldsymbol{u}_h - \boldsymbol{U}|_h$ to behave like the consistency error. We have

$$\left|\boldsymbol{u}_h - \boldsymbol{U}\right|_h = \left|A_h^{-1} A_h(\boldsymbol{u}_h - \boldsymbol{U})\right|_h \leq \left\|A_h^{-1}\right\|_h \left|A_h(\boldsymbol{u}_h - \boldsymbol{U})\right|_h , \qquad (1.19)$$

where the matrix norm $\|\cdot\|_h$ has to be chosen to be compatible with the vector norm $|\cdot|_h$. The error behaves like the consistency error asymptotically in h if $\left\|A_h^{-1}\right\|_h$ can be bounded independently of h; that is if the method is *stable* in the following sense:

Definition 1.4 In the situation of Definition 1.3, the approximation is called *stable* with respect to $\|\cdot\|_h$ if there exists a constant $C > 0$

independent of h such that

$$\left\|A_h^{-1}\right\|_h \leq C \, .$$

From the above definition we can obviously conclude, with (1.19), the following result:

Theorem 1.5 *A consistent and stable method is convergent, and the order of convergence is at least equal to the order of consistency.*

Therefore, specifically for the five-point stencil discretization of (1.1), (1.2) on a rectangle, stability with respect to $\|\cdot\|_\infty$ is desirable. In fact, it follows from the structure of A_h: Namely, we have

$$\left\|A_h^{-1}\right\|_\infty \leq \frac{1}{16}(a^2+b^2) \, . \tag{1.20}$$

This follows from more general considerations in Section 1.4 (Theorem 1.14). Putting the results together we have the following theorem:

Theorem 1.6 *Let the solution u of* (1.1), (1.2) *on a rectangle Ω be in $C^4(\overline{\Omega})$. Then the five-point stencil discretization has an order of convergence* 2 *with respect to $\|\cdot\|_\infty$, more precisely,*

$$|\boldsymbol{u}_h - \boldsymbol{U}|_\infty \leq \frac{1}{192}(a^2+b^2)\left(\|\partial^{(4,0)}u\|_\infty + \|\partial^{(0,4)}u\|_\infty\right) h^2 \, .$$

Exercises

1.1 Complete the proof of Lemma 1.2 and also investigate the error of the respective difference quotients, assuming only $u \in C^2[x-h, x+h]$.

1.2 Generalize the discussion concerning the five-point stencil discretization (including the order of convergence) of (1.1), (1.2) on a rectangle for $h_1 > 0$ in the x_1 direction and $h_2 > 0$ in the x_2 direction.

1.3 Show that an irreducible weakly row diagonally dominant matrix cannot have vanishing diagonal elements.

1.3 Generalizations and Limitations of the Finite Difference Method

We continue to consider the boundary value problem (1.1), (1.2) on a rectangle Ω. The five-point stencil discretization developed may be interpreted as a mapping $-\Delta_h$ from functions on $\overline{\Omega}_h$ into grid functions on Ω_h, which

is defined by

$$-\Delta_h v_h(x_1, x_2) := \sum_{i,j=-1}^{1} c_{ij} v_h(x_1 + ih, x_2 + jh) \,, \tag{1.21}$$

where $c_{0,0} = 4/h^2$, $c_{0,1} = c_{1,0} = c_{0,-1} = c_{-1,0} = -1/h^2$, and $c_{ij} = 0$ for all other (i, j). For the description of such a difference stencil as defined in (1.21) the points of the compass (in two space dimensions) may also be involved. In the five-point stencil only the main points of the compass appear.

The question of whether the *weights* c_{ij} can be chosen differently such that we gain an approximation of $-\Delta u$ with higher order in h has to be answered negatively (see Exercise 1.7). In this respect the five-point stencil is optimal. This does not exclude that other difference stencils with more entries, but of the same order of convergence, might be worthwhile to consider. An example, which will be derived in Exercise 3.11 by means of the finite element method, has the following form:

$$c_{0,0} = \frac{8}{3h^2} \,, \quad c_{ij} = -\frac{1}{3h^2} \quad \text{for all other } i, j \in \{-1, 0, 1\} \,. \tag{1.22}$$

This nine-point stencil can be interpreted as a linear combination of the five-point stencil and a five-point stencil for a coordinate system rotated by $\frac{\pi}{4}$ (with step size $\sqrt{2}\,h$), using the weights $\frac{1}{3}$ and $\frac{2}{3}$ in this linear combination. Using a general nine-point stencil a method with order of consistency greater than 2 can be constructed only if the right-hand side f at the point (x_1, x_2) is approximated not by the evaluation $f(x_1, x_2)$, but by applying a more general stencil. The *mehrstellen method* ("Mehrstellenverfahren") defined by Collatz is such an example (see, for example, [15, p. 66]).

Methods of higher order can be achieved by larger stencils, meaning that the summation indices in (1.21) have to be replaced by k and $-k$, respectively, for $k \in \mathbb{N}$. But already for $k = 2$ such difference stencils cannot be used for grid points close to the boundary, so that there one has to return to approximations of lower order.

If we consider the five-point stencil to be a suitable discretization for the Poisson equation, the high smoothness assumption for the solution in Theorem 1.6 should be noted. This requirement cannot be ignored, since in general it does not hold true. On the one hand, for a smoothly bounded domain (see Appendix A.5 for a definition of a domain with C^l-boundary) the smoothness of the solution is determined only by the smoothness of the data f and g (see for example [13, Theorem 6.19]), but on the other hand, corners in the domain reduce this smoothness the more, the more reentrant the corners are. Let us consider the following examples:

For the boundary value problem (1.1), (1.2) on a rectangle $(0, a) \times (0, b)$ we choose $f = 1$ and $g = 0$; this means arbitrarily smooth functions. Nevertheless, for the solution u, the statement $u \in C^2(\overline{\Omega})$ cannot hold, because otherwise, $-\Delta u(0, 0) = 1$ would be true, but on the other hand,

we have $\partial_{1,1}u(x,0) = 0$ because of the boundary condition and hence also $\partial_{1,1}u(0,0) = 0$ and $\partial_{2,2}u(0,y) = 0$ analogously. Therefore, $\partial_{2,2}u(0,0) = 0$. Consequently, $-\Delta u(0,0) = 0$, which contradicts the assumption above. Therefore, Theorem 1.6 is not applicable here.

In the second example we consider the domain with reentrant corner (see Figure 1.3)

$$\Omega = \left\{(x,y) \in \mathbb{R}^2 \;\middle|\; x^2 + y^2 < 1\,,\ x < 0 \text{ or } y > 0\right\}\,.$$

In general, if we identify $\mathbb{R}^2$ and $\mathbb{C}$, this means $(x,y) \in \mathbb{R}^2$ and $z = x + iy \in \mathbb{C}$, we have that if $w : \mathbb{C} \to \mathbb{C}$ is analytic (holomorphic), then both the real and the imaginary parts $\Re w, \Im w : \mathbb{C} \to \mathbb{R}$ are *harmonic*, which means that they solve $-\Delta u = 0$.

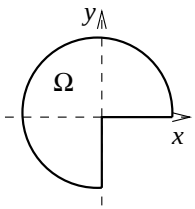


Figure 1.3. Domain Ω with reentrant corner.

We choose $w(z) := z^{2/3}$. Then the function $u(x,y) := \Im\left((x+iy)^{2/3}\right)$ solves the equation

$$-\Delta u = 0 \quad \text{in} \quad \Omega\,.$$

In polar coordinates, $x = r\cos\varphi$, $y = r\sin\varphi$, the function u takes the form

$$u(x,y) = \Im\left(\left(re^{i\varphi}\right)^{2/3}\right) = r^{2/3}\sin\left(\frac{2}{3}\varphi\right)\,.$$

Therefore, u satisfies the boundary conditions

$$\begin{aligned} u\left(e^{i\varphi}\right) &= \sin\left(\frac{2}{3}\varphi\right) \quad \text{for } 0 \le \varphi \le \frac{3\pi}{2}\,, \\ u(x,y) &= 0 \quad \text{otherwise on } \partial\Omega\,. \end{aligned} \tag{1.23}$$

But note that $w'(z) = \frac{2}{3}z^{-1/3}$ is unbounded for $z \to 0$, so that $\partial_1 u, \partial_2 u$ are unbounded for $(x,y) \to 0$. Therefore, in this case we do not even have $u \in C^1(\overline{\Omega})$.

The examples do not show that the five-point stencil discretization is not suitable for the boundary value problems considered, but they show the necessity of a theory of convergence, which requires only as much smoothness as was to be expected.

In the following we discuss some generalizations of the boundary value problems considered so far.

General Domains Ω

We continue to consider (1.1), (1.2) but on a general domain in $\mathbb{R}^2$, for which the parts of the boundary are not necessarily aligned to the coordinate axes. Therefore we can keep the second equation in (1.5) as the definition of Ω_h, but have to redefine the set of boundary grid points $\partial\Omega_h$.

For example, if for some point $(x, y) \in \Omega_h$ we have

$$(x - h, y) \notin \Omega\,,$$

then there exists a number $s \in (0, 1]$ such that

$$(x - \vartheta h, y) \in \Omega \quad \text{for all } \vartheta \in [0, s) \quad \text{and} \quad (x - sh, y) \notin \Omega\,.$$

Then $(x - sh, y) \in \partial\Omega$, and therefore we define

$$(x - sh, y) \in \partial\Omega_h\,.$$

The other main points of the compass are treated analogously. In this way the grid spacing in the vicinity of the boundary becomes variable; in particular, it can be smaller than h.

For the quality of the approximation we have the following result:

Lemma 1.7 *Let $\Omega = (x - h_1, x + h_2)$ for $x \in \mathbb{R}$, $h_1, h_2 > 0$.*

(1) Then for $u \in C^3(\overline{\Omega})$,

$$\begin{aligned} u''(x) &= \frac{2}{h_1 + h_2}\left(\frac{u(x + h_2) - u(x)}{h_2} - \frac{u(x) - u(x - h_1)}{h_1}\right) \\ &\quad + \max\{h_1, h_2\}\, R\,, \end{aligned}$$

where R is bounded independently of h_1, h_2.

(2) There are no $\alpha, \beta, \gamma \in \mathbb{R}$ such that

$$u''(x) = \alpha\, u(x - h_1) + \beta\, u(x) + \gamma\, u(x + h_2) + R_1 h_1^2 + R_2 h_2^2$$

for all polynomials u of degree 3 if $h_1 \neq h_2$.

Proof: Exercises 1.4 and 1.5. □

This leads to a discretization that is difficult to set up and for which the order of consistency and order of convergence are not easily determined.

Other Boundary Conditions

We want to consider the following example. Let $\partial\Omega = \Gamma_1 \cup \Gamma_3$ be divided into two disjoint subsets. We are looking for a function u such that

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega\,, \\ \partial_\nu u := \nabla u \cdot \nu &= g \quad \text{on } \Gamma_1\,, \\ u &= 0 \quad \text{on } \Gamma_3\,, \end{aligned} \tag{1.24}$$

where $\nu : \partial\Omega \to \mathbb{R}^d$ is the outer unit normal, and thus $\partial_\nu u$ is the normal derivative of u.

For a part of the boundary oriented in a coordinate direction, $\partial_\nu u$ is just a positive or negative partial derivative. But if only grid points in $\overline{\Omega}_h$ are to be used, only $\pm\partial^+ u$ and $\pm\partial^- u$ respectively (in the coordinates orthogonal to the direction of the boundary) are available directly from the above approximations with a corresponding reduction of the order of consistency. For a boundary point without these restrictions the question of how to approximate $\partial_\nu u$ appropriately is open.

As an example we consider (1.24) for a rectangle $\Omega = (0,a)\times(0,b)$, where

$$\Gamma_1 := \{(a,y) \mid y \in (0,b)\}\,, \quad \Gamma_3 := \Gamma \setminus \Gamma_1\,. \tag{1.25}$$

At the boundary grid points (a, jh), $j = 1,\ldots,m-1$, $\partial_2 u = \nabla u \cdot \nu$ is prescribed, which can be approximated directly only by $\partial^- u$. Due to Lemma 1.2, 2 this leads to a reduction in the consistency order (see Exercise 1.8). The resulting system of equations may include the Neumann boundary grid points in the set of unknowns, for which an equation with the entries $1/h$ in the diagonal and $-1/h$ in an off-diagonal corresponding to the eastern neighbour $(a-h, jh)$ has to be added. Alternatively, those boundary points can be eliminated, leading for the eastern neighbour to a modified difference stencil (multiplied by h^2)

$$\begin{array}{ccc} & -1 & \\ -1 & 3 & \\ & -1 & \end{array} \tag{1.26}$$

for the right-hand side $h^2 f(a-h, jh) + hg(a, jh)$. In both cases the matrix properties of the system of equations as collected in (1.15) still hold, with the exception of $\sum_{s=1}^{M_1} (A_h)_{rs} = 0$, both for the Neumann boundary points and their neighbours, if no Dirichlet boundary point is involved in their stencil. Thus the term "close to the boundary" has to be interpreted as "close to the Dirichlet boundary."

If one wants to take advantage of the symmetric difference quotient $\partial^0 u$, then "artificial" values at new external grid points $(a+h, jh)$ appear.

To keep the balance of unknowns and equations, it can be assumed that the differential equation also holds at (a, jh), and thus it is discretized with the five-point stencil there. If one attributes the discrete boundary condition to the external grid point, then again the properties (1.15) hold with the abovementioned interpretation. Alternatively, the external grid points can be eliminated, leading to a modified difference stencil (multiplied by h^2) at (a, jh):

$$\begin{array}{ccc} & -1 & \\ -2 & 4 & \\ & -1 & \end{array} \tag{1.27}$$

for the right-hand side $h^2 f(a, jh) + 2hg(a, jh)$, with the same interpretation of properties (1.15).

More General Differential Equations

As an example we consider the differential equation

$$-\nabla \cdot (k\,\nabla u) = f \quad \text{on } \Omega \tag{1.28}$$

with a continuous coefficient function $k : \Omega \to \mathbb{R}$, which is bounded from below by a positive constant on Ω. This equation states the conservation of an extensive quantity u whose flux is $-k\nabla u$ (see Section 0.5). This should be respected by the discretization, and therefore the form of (1.28) obtained by working out the derivatives is not recommended as a basis for the discretization. The differential expression in (1.28) can be discretized by a successive application of central difference quotients, but then again the order of consistency has to be investigated.

In addition, one has to take into account the fact that the smoothness of u depends on the smoothness of k. If processes in heterogeneous materials have to be described, then k is often discontinuous. In the simplest example k is assumed to take two different values: Let $\Omega = \Omega_1 \cup \Omega_2$ and

$$k|_{\Omega_1} = k_1 > 0\,, \quad k|_{\Omega_2} = k_2 > 0$$

with constants $k_1 \neq k_2$.

As worked out in Section 0.5, on the interior boundary $S := \overline{\Omega}_1 \cap \overline{\Omega}_2$ a *transmission condition* has to be imposed:

- u is continuous,
- $(k\nabla u) \cdot \nu$ is continuous, where ν is the outer normal on $\partial\Omega_1$, for example.

This leads to the following conditions on u_i, being the restrictions of u on $\overline{\Omega}_i$ for $i = 1, 2$:

$$\begin{aligned} -k_1 \Delta u_1 &= f \quad \text{in } \Omega_1\,, && (1.29)\\ -k_2 \Delta u_2 &= f \quad \text{in } \Omega_2\,, \\ u_1 &= u_2 \quad \text{on } S\,, && (1.30)\\ k_1 \partial_\nu u_1 &= k_2 \partial_\nu u_2 \quad \text{on } S\,. \end{aligned}$$

In this case the question of an appropriate discretization is also open.

Summarizing, we have the following catalogue of requirements: We are looking for a notion of solution for (general) boundary value problems with nonsmooth coefficients and right-hand sides such that, for example, the transmission condition is fulfilled automatically.

We are looking for a discretization on general domains such that, for example, the (order of) convergence can also be assured for less smooth solutions and also Neumann boundary conditions as in (1.24) can be treated easily.

The finite element method in the subsequent chapters will fulfil these requirements to a large extent.

Exercises

1.4 Prove Lemma 1.7, 1.

1.5 Under the assumption that $u : \Omega \subset \mathbb{R} \to \mathbb{R}$ is a sufficiently smooth function, determine in the ansatz

$$\alpha u(x - h_1) + \beta u(x) + \gamma u(x + h_2) \,, \quad h_1, h_2 > 0 \,,$$

the coefficients $\alpha = \alpha(h_1, h_2)$, $\beta = \beta(h_1, h_2)$, $\gamma = \gamma(h_1, h_2)$, such that

(a) for $x \in \Omega$, $u'(x)$ will be approximated with the order as high as possible,

(b) for $x \in \Omega$, $u''(x)$ will be approximated with the order as high as possible,

and in particular, prove 1.7, 2.

Hint: Determine the coefficients such that the formula is exact for polynomials with the degree as high as possible.

1.6 Let $\Omega \subset \mathbb{R}^2$ be a bounded domain. For a sufficiently smooth function $u : \Omega \to \mathbb{R}$ determine the difference formula with an order as high as possible to approximate $\partial_{11} u(x_1, x_2)$, using the 9 values $u(x_1 + \gamma_1 h, x_2 + \gamma_2 h)$, where $\gamma_1, \gamma_2 \in \{-1, 0, 1\}$.

1.7 Let $\Omega \subset \mathbb{R}^2$ be a bounded domain. Show that in (1.21) there exists no choice of c_{ij} such that for an arbitrary smooth function $u : \Omega \to \mathbb{R}$,

$$|\Delta u(x) - \Delta_h u(x)| \leq C h^3$$

is valid with a constant C independent of h.

1.8 For the example (1.24), (1.25), investigate the order of consistency both for the discretization (1.26) and (1.27) in the maximum norm. Are there improvements possible considering the discrete L^2-norm? (See (1.18).)

1.9 Consider example (1.24) with

$$\Gamma_1 := \{(a, y) \mid y \in (0, b)\} \cup \{(x, b) \mid x \in (0, a]\},$$
$$\Gamma_3 := \Gamma \setminus \Gamma_1,$$

and discuss the applicability of the one-sided and the symmetric difference quotients for the approximation of the Neumann boundary condition, in particular with respect to properties (1.15). In which way does the boundary condition at (a, b), where no unique normal exists, have to be interpreted?

1.10 Generalize the discussion concerning the five-point stencil discretization (including the order of convergence) to the boundary value problem

$$\begin{aligned} -\Delta u + ru &= f \quad \text{in } \Omega, \\ u &= g \quad \text{on } \partial\Omega, \end{aligned}$$

for $r > 0$ and $\Omega := (0,a) \times (0,b)$. To approximate the reactive term ru, the following schemes in the notation of (1.21) are to be used:

(a) $c_{0,0} = 1, \quad c_{ij} = 0$ otherwise,

(b) $c_{0,0} > 0, \quad c_{0,1}, c_{1,0}, c_{0,-1}, c_{-1,0} \geq 0, \quad c_{ij} = 0$ otherwise, and $\sum_{i,j=-1}^{1} c_{ij} = 1$.

1.4 Maximum Principles and Stability

In this section the proof of the stability estimate (1.20), which is still missing, will be given. For this reason we develop a more general framework, in which we will then also discuss the finite element method (see Section 3.9) and the time-dependent problems (see Section 7.5). The boundary value problem (1.1), (1.2) satisfies a (*weak*) *maximum principle* in the following sense: If f is continuous and $f(x) \leq 0$ for all $x \in \Omega$ (for short $f \leq 0$), then

$$\max_{x \in \overline{\Omega}} u(x) \leq \max_{x \in \partial\Omega} u(x) \, .$$

This *maximum principle* is also *strong* in the following sense: The maximum of u on $\overline{\Omega}$ can be attained in Ω only if u is constant (see, for example, [13], also for the following assertions). By exchanging u, f, g by $-u, -f, -g$, respectively, we see that there is an analogous (*strong*) *minimum principle*. The same holds for more general linear differential equations as in (1.28), which may also contain convective parts (this means first-order derivatives). But if the equation contains a reactive part (this means without derivatives), as in the example

$$-\Delta u + ru = f \quad \text{in } \Omega$$

with a continuous function $r : \Omega \to \mathbb{R}$ such that $r(x) \geq 0$ for $x \in \Omega$, there is a weak maximum principle only in the following form: If $f \leq 0$, then

$$\max_{x \in \overline{\Omega}} u(x) \leq \max \left\{ \max_{x \in \partial\Omega} u(x), 0 \right\} .$$

The weak maximum principle directly implies assertions about the dependence of the solution u of the boundary value problem on the data f and g; this means *stability properties*. One can also follow this method in investigating the discretization. For the basic example we have

Theorem 1.8 *Let u_h be a grid function on $\overline{\Omega}_h$ defined by* (1.7), (1.8) *and suppose $f_{ij} \le 0$ for all $i = 1, \ldots, l-1$, $j = 1, \ldots, m-1$. Then if u_h attains its maximum on $\Omega_h \cup \partial\Omega_h^*$ at a point $(i_0 h, j_0 h) \in \Omega_h$, then the following holds:*

$$u_h \text{ is constant on } \Omega_h \cup \partial\Omega_h^* \,.$$

Here

$$\partial\Omega_h^* := \partial\Omega_h \setminus \{(0,0), (a,0), (0,b), (a,b)\} \,.$$

In particular, we have

$$\max_{(x,y)\in\Omega_h} u_h(x,y) \le \max_{(x,y)\in\partial\Omega_h^*} u_h(x,y) \,.$$

Proof: Let $\bar{u} := u_h(i_0 h, j_0 h)$. Then because of (1.7) and $f_{ij} \le 0$ we have

$$4\bar{u} \le \sum_{(k,l)\in N_{(i_0,j_0)}} u_h(kh, lh) \le 4\bar{u} \,,$$

since in particular $u_h(kh, lh) \le \bar{u}$ for $(k,l) \in N_{(i_0,j_0)}$. Here we used the notation

$$N_{(i_0,j_0)} = \{((i_0-1), j_0), ((i_0+1), j_0), (i_0, (j_0+1)), (i_0, (j_0-1))\}$$

for the set of indices of neighbours of $(i_0 h, j_0 h)$ in the five-point stencil. From these inequalities we conclude that

$$u_h(kh, lh) = \bar{u} \quad \text{for} \quad (k,l) \in N_{(i_0,j_0)} \,.$$

If we apply this argument to the neighbours in $\overline{\Omega}_h$ of the grid points (kh, lh) for $(k,l) \in N_{(i_0,j_0)}$ and then continue in the same way to the sets of neighbours in $\overline{\Omega}_h$ arising in every such step, then finally, for each grid point $(ih, jh) \in \Omega_h \cup \partial\Omega_h^*$ the claimed identity $u_h(ih, jh) = \bar{u}$ is achieved. □

The exceptional set of vertices $\partial\Omega_h \setminus \partial\Omega_h^*$ does not participate in any difference stencil, so that the values there are of no relevance for u_h.

We want to generalize this result and therefore consider a system of equations as in (1.10), (1.11):

$$A_h \boldsymbol{u}_h = \boldsymbol{q}_h = -\hat{A}_h \hat{\boldsymbol{u}}_h + \boldsymbol{f} \,, \tag{1.31}$$

where $A_h \in \mathbb{R}^{M_1, M_1}$ as in (1.10), $\hat{A}_h \in \mathbb{R}^{M_1, M_2}$ as in (1.11), $\boldsymbol{u}_h, \boldsymbol{f} \in \mathbb{R}^{M_1}$, and $\hat{\boldsymbol{u}}_h \in \mathbb{R}^{M_2}$. This may be interpreted as the discretization of a boundary value problem obtained by the finite difference method or any other approach and without restrictions on the dimensionality of the domain. At least on one part of the boundary Dirichlet boundary conditions are required. Then the entries of the vector $\boldsymbol{u}_h$ can be interpreted as the unknown values at the grid points in $\Omega_h \cup \partial\Omega_h^{(1)}$, where $\partial\Omega_h^{(1)}$ correspond to a part of $\partial\Omega$ (with flux or mixed boundary condition). Analogously, the vector $\hat{\boldsymbol{u}}_h$

(indexed from $M_1 + 1$ to $M_1 + M_2$) corresponds to the values fixed by the Dirichlet boundary conditions on $\partial\Omega_h^{(2)}$. Again let $M = M_1 + M_2$ and

$$\tilde{A}_h := \left(A_h \mid \hat{A}_h \right) \in \mathbb{R}^{M_1, M} .$$

This means in particular that the dimensions M_1 and M_2 are not fixed, but are in general unbounded for $h \to 0$.

Oriented on (1.15) we require the following general assumptions for the rest of the section:

$$\begin{aligned}
&(1) && (A_h)_{rr} > 0 \quad \text{for all } r = 1, \ldots, M_1 , \\
&(2) && (A_h)_{rs} \le 0 \quad \text{for all } r, s = 1, \ldots, M_1 \text{ such that } r \ne s , \\
&(3)\ (\text{i}) && \sum_{s=1}^{M_1} (A_h)_{rs} \ge 0 \quad \text{for all } r = 1, \ldots, M_1 , \\
&\quad (\text{ii}) && \text{for at least one index the strict inequality holds} , \\
&(4) && A_h \text{ is irreducible} , \\
&(5) && (\hat{A}_h)_{rs} \le 0 \quad \text{for all } r = 1, \ldots, M_1 , \ s = M_1 + 1, \ldots, M , \\
&(6) && \sum_{s=1}^{M} (\tilde{A}_h)_{rs} \ge 0 \quad \text{for all } r = 1, \ldots, M_1 , \\
&(7) && \text{for every } s = M_1 + 1, \ldots, M \text{ there exists } r \in \{1, \ldots, M_1\} \\
& && \text{such that } (\hat{A}_h)_{rs} \ne 0 .
\end{aligned} \tag{1.32}$$

Generalizing the notation above for $r \in \{1, \ldots, M_1\}$, the indices $s \in \{1, \ldots, M\} \setminus \{r\}$ are called *neighbours*, for which $(\tilde{A}_h)_{rs} \ne 0$, and they are assembled to form the set N_r. Therefore, the irreducibility of A_h means that arbitrary $r, s \in \{1, \ldots, M_1\}$ can be connected by neighbourhood relationships.

The condition (7) is not a restriction: It only avoids the inclusion of known values $(\hat{\boldsymbol{u}}_h)_s$ that do not influence the solution of (1.31) at all. For the five-point stencil on the rectangle, these are the values at the corner points. Because of the condition (7), every index $r \in \{M_1 + 1, \ldots, M\}$ is connected to every index $s \in \{1, \ldots, M_1\}$ by means of neighbourhood relationships.

The conditions (2) and (3) imply the weak diagonal dominance of A_h. Note that the conditions are formulated redundantly: The condition (3) also follows from (5) through (7).

To simplify the notation we will use the following conventions, where $\boldsymbol{u}$, $\boldsymbol{v}$ and A, B are vectors and matrices, respectively, of suitable dimensions:

$$\begin{array}{lcll}
\boldsymbol{u} \ge \boldsymbol{0} & \text{if and only if} & (\boldsymbol{u})_i \ge 0 & \text{for all indices } i , \\
\boldsymbol{u} \ge \boldsymbol{v} & \text{if and only if} & \boldsymbol{u} - \boldsymbol{v} \ge 0 , & \\
A \ge 0 & \text{if and only if} & (A)_{ij} \ge 0 & \text{for all indices } (i, j) , \\
A \ge B & \text{if and only if} & A - B \ge 0 . &
\end{array} \tag{1.33}$$

Theorem 1.9 *We consider* (1.31) *under the assumptions* (1.32). *Furthermore, let* $\boldsymbol{f} \leq \mathbf{0}$. *Then a* strong maximum principle *holds: If the components of* $\tilde{\boldsymbol{u}}_h = \binom{\boldsymbol{u}_h}{\hat{\boldsymbol{u}}_h}$ *attain a nonnegative maximum for some index* $r \in \{1, \ldots, M_1\}$, *then all the components are equal. In particular, a* weak maximum principle *is fulfilled:*

$$\max_{r\in\{1,\ldots,M\}} (\tilde{\boldsymbol{u}}_h)_r \leq \max\left\{0, \max_{r\in\{M_1+1,\ldots,M\}} (\hat{\boldsymbol{u}}_h)_r\right\} . \tag{1.34}$$

Proof: Let $\bar{u} = \max_{s\in\{1,\ldots,M\}}(\tilde{\boldsymbol{u}}_h)_s$, and $\bar{u} = (\boldsymbol{u}_h)_r$ where $r \in \{1, \ldots, M_1\}$. Because of (1.32) (2), (5), (6) the rth row of (1.31) implies

$$\begin{aligned} (A_h)_{rr}\bar{u} &\leq -\sum_{s\in N_r} \left(\tilde{A}_h\right)_{rs} (\tilde{\boldsymbol{u}}_h)_s = \sum_{s\in N_r} \left|\left(\tilde{A}_h\right)_{rs}\right| (\tilde{\boldsymbol{u}}_h)_s \\ &\leq \sum_{s\in N_r} \left|\left(\tilde{A}_h\right)_{rs}\right| \bar{u} \leq (A_h)_{rr}\bar{u} , \end{aligned} \tag{1.35}$$

where the assumption $\bar{u} \geq 0$ is used in the last estimate. Therefore, everywhere equality has to hold. Since the second inequality is valid also for every single term and $(\tilde{A}_h)_{rs} \neq 0$ by the definition of N_r, we finally conclude that

$$(\tilde{\boldsymbol{u}}_h)_s = \bar{u} \quad \text{for all } s \in N_r .$$

This allows us to apply this argument to all $s \in N_r \cap \{1, \ldots, M_1\}$, then to the corresponding sets of neighbours, and so on, until the assertion is proven. □

The requirement of irreducibility can be weakened if instead of (1.32) (6) we have

$$(6)^* \quad \sum_{s=1}^{M} \left(\tilde{A}_h\right)_{rs} = 0 \quad \text{for all } r = 1, \ldots, M_1 .$$

Then condition (4) can be replaced by the requirement

$(4)^*$ For every $r_1 \in \{1, \ldots, M_1\}$ such that

$$\sum_{s=1}^{M_1} (A_h)_{r_1 s} = 0 \tag{1.36}$$

there are indices $r_2, \ldots, r_{l+1}$ such that

$$(A_h)_{r_i r_{i+1}} \neq 0 \quad \text{for } i = 1, \ldots, l$$

and

$$\sum_{s=1}^{M_1} (A_h)_{r_{l+1} s} > 0 . \tag{1.37}$$

These modified conditions without (7) will be denoted by $(1.32)^*$.

Motivated by the example above we call a point $r \in \{1, \ldots, M_1\}$ *far from the boundary* if (1.36) holds, and *close to the boundary* if (1.37) holds, and the points $r \in \{M_1 + 1, \ldots, M\}$ are called *boundary points*.

Theorem 1.10 *We consider* (1.31) *under the assumption* $(1.32)^*$. *If* $\boldsymbol{f} \leq \mathbf{0}$, *then*

$$\max_{r \in \{1,\ldots,M\}} (\tilde{\boldsymbol{u}}_h)_r \leq \max_{r \in \{M_1+1,\ldots,M\}} (\hat{\boldsymbol{u}}_h)_r \,. \tag{1.38}$$

Proof: We use the same notation and the same arguments as in the proof of Theorem 1.9. In (1.35) in the last estimate equality holds, so that no sign conditions for $\bar{u}$ are necessary. Because of $(4)^*$ the maximum will also be attained at a point close to the boundary and therefore also at its neighbours. Because of $(6)^*$ a boundary point also belongs to these neighbours, which proves the assertion. □

From the maximum principles we immediately conclude a *comparison principle*:

Lemma 1.11 *We assume* (1.32) *or* $(1.32)^*$. *Let* $\boldsymbol{u}_{h1}, \boldsymbol{u}_{h2} \in \mathbb{R}^{M_1}$ *be solutions of*

$$A_h \boldsymbol{u}_{hi} = -\hat{A}_h \hat{\boldsymbol{u}}_{hi} + \boldsymbol{f}_i \quad \text{for } i = 1, 2$$

for given $\boldsymbol{f}_1, \boldsymbol{f}_2 \in \mathbb{R}^{M_1}$, $\hat{\boldsymbol{u}}_{h1}, \hat{\boldsymbol{u}}_{h2} \in \mathbb{R}^{M_2}$, *which satisfy* $\boldsymbol{f}_1 \leq \boldsymbol{f}_2$, $\hat{\boldsymbol{u}}_{h1} \leq \hat{\boldsymbol{u}}_{h2}$. *Then*

$$\boldsymbol{u}_{h1} \leq \boldsymbol{u}_{h2} \,.$$

Proof: From $A_h(\boldsymbol{u}_{h1} - \boldsymbol{u}_{h2}) = -\hat{A}_h(\hat{\boldsymbol{u}}_{h1} - \hat{\boldsymbol{u}}_{h2}) + \boldsymbol{f}_1 - \boldsymbol{f}_2$ we can conclude with Theorem 1.9 or 1.10 that

$$\max_{r \in \{1,\ldots,M_1\}} (\boldsymbol{u}_{h1} - \boldsymbol{u}_{h2})_r \leq 0 \,.$$

□

This implies in particular the uniqueness of a solution of (1.31) for arbitrary $\hat{\boldsymbol{u}}_h$ and $\boldsymbol{f}$ and also the regularity of A_h.

In the following we denote by $\mathbf{0}$ and 0 the zero vector and the zero matrix, respectively, where all components are equal to 0. An immediate consequence of Lemma 1.11 is the following

Theorem 1.12 *Let* $A_h \in \mathbb{R}^{M_1, M_1}$ *be a matrix with the properties* (1.32) (1)–(3) (i), $(4)^*$, *and* $\boldsymbol{u}_h \in \mathbb{R}^{M_1}$. *Then*

$$A_h \boldsymbol{u}_h \geq \mathbf{0} \quad \textit{implies} \quad \boldsymbol{u}_h \geq \mathbf{0} \,. \tag{1.39}$$

Proof: To be able to apply Lemma 1.11, one has to construct a matrix $\hat{A}_h \in \mathbb{R}^{M_1, M_2}$ such that $(1.32)^*$ holds. Obviously, this is possible. Then one

can choose

$$\begin{array}{lll} \boldsymbol{u}_{h2} := \boldsymbol{u}_h \,, & \boldsymbol{f}_2 := A_h \boldsymbol{u}_{h2} \,, & \hat{\boldsymbol{u}}_{h2} := \boldsymbol{0} \,, \\ \boldsymbol{u}_{h1} := \boldsymbol{0} \,, & \boldsymbol{f}_1 := \boldsymbol{0} \,, & \hat{\boldsymbol{u}}_{h1} := \boldsymbol{0} \end{array}$$

to conclude the assertion. Because of $\hat{\boldsymbol{u}}_{hi} := \boldsymbol{0}$ for $i = 1, 2$ the specific definition of $\hat{A}_h$ plays no role. □

A matrix with the property (1.39) is called *inverse monotone*. An equivalent requirement is

$$\boldsymbol{v}_h \geq \boldsymbol{0} \quad \Rightarrow \quad A_h^{-1} \boldsymbol{v}_h \geq \boldsymbol{0} \,,$$

and therefore by choosing the unit vectors as $\boldsymbol{v}_h$,

$$A_h^{-1} \geq 0 \,.$$

Inverse monotone matrices that also satisfy (1.32) (1), (2) are called *M-matrices*.

Finally, we can weaken the assumptions for the validity of the comparison principle.

Corollary 1.13 *Suppose that* $A_h \in \mathbb{R}^{M_1, M_1}$ *is inverse monotone and* (1.32) (5) *holds. Let* $\boldsymbol{u}_{h1}, \boldsymbol{u}_{h2} \in \mathbb{R}^{M_1}$ *be solutions of*

$$A_h \boldsymbol{u}_{hi} = -\hat{A}_h \hat{\boldsymbol{u}}_{hi} + \boldsymbol{f}_i \quad \text{for } i = 1, 2$$

for given $\boldsymbol{f}_1, \boldsymbol{f}_2 \in \mathbb{R}^{M_1}$, $\hat{\boldsymbol{u}}_{h1}, \hat{\boldsymbol{u}}_{h2} \in \mathbb{R}^{M_2}$ *that satisfy* $\boldsymbol{f}_1 \leq \boldsymbol{f}_2$, $\hat{\boldsymbol{u}}_{h1} \leq \hat{\boldsymbol{u}}_{h2}$. *Then*

$$\boldsymbol{u}_{h1} \leq \boldsymbol{u}_{h2} \,.$$

Proof: Multiplying the equation

$$A_h(\boldsymbol{u}_{h1} - \boldsymbol{u}_{h2}) = -\hat{A}_h(\hat{\boldsymbol{u}}_{h1} - \hat{\boldsymbol{u}}_{h2}) + \boldsymbol{f}_1 - \boldsymbol{f}_2$$

from the left by the matrix A_h^{-1}, we get

$$\boldsymbol{u}_{h1} - \boldsymbol{u}_{h2} = -\underbrace{A_h^{-1}}_{\geq 0} \underbrace{\hat{A}_h}_{\leq 0} \underbrace{(\hat{\boldsymbol{u}}_{h1} - \hat{\boldsymbol{u}}_{h2})}_{\leq 0} + \underbrace{A_h^{-1}}_{\geq 0} \underbrace{(\boldsymbol{f}_1 - \boldsymbol{f}_2)}_{\leq 0} \leq 0 \,.$$

□

The importance of Corollary 1.13 lies in the fact that there exist discretization methods, for which the matrix $\tilde{A}_h$ does not satisfy, e.g., condition (1.32) (6), or (6)* but $A_h^{-1} \geq 0$. A typical example of such a method is the finite volume method described in Chapter 6.

In the following we denote by $\mathbf{1}$ a vector (of suitable dimension) whose components are *all* equal to 1.

Theorem 1.14 *We assume* (1.32) (1)–(3), (4)*, (5). *Furthermore, let* $\boldsymbol{w}_h^{(1)}, \boldsymbol{w}_h^{(2)} \in \mathbb{R}^{M_1}$ *be given such that*

$$A_h \boldsymbol{w}_h^{(1)} \geq \mathbf{1} \,, \quad A_h \boldsymbol{w}_h^{(2)} \geq -\hat{A}_h \mathbf{1} \,. \tag{1.40}$$

Then a solution of $A_h \boldsymbol{u}_h = -\hat{A}_h \hat{\boldsymbol{u}}_h + \boldsymbol{f}$ *satisfies*

(1) $-\big(|\boldsymbol{f}|_\infty \boldsymbol{w}_h^{(1)} + |\hat{\boldsymbol{u}}_h|_\infty \boldsymbol{w}_h^{(2)}\big) \le \boldsymbol{u}_h \le |\boldsymbol{f}|_\infty \boldsymbol{w}_h^{(1)} + |\hat{\boldsymbol{u}}_h|_\infty \boldsymbol{w}_h^{(2)}$,

(2) $|\boldsymbol{u}_h|_\infty \le \big|\boldsymbol{w}_h^{(1)}\big|_\infty |\boldsymbol{f}|_\infty + \big|\boldsymbol{w}_h^{(2)}\big|_\infty |\hat{\boldsymbol{u}}_h|_\infty$.

Under the assumptions (1.32) (1)–(3), (4)*, and* (1.40) *the matrix norm* $\|\cdot\|_\infty$ *induced by* $|\cdot|_\infty$ *satisfies*

$$\big\|A_h^{-1}\big\|_\infty \le \big|\boldsymbol{w}_h^{(1)}\big|_\infty .$$

Proof: Since $-|\boldsymbol{f}|_\infty \mathbf{1} \le \boldsymbol{f} \le |\boldsymbol{f}|_\infty \mathbf{1}$ and the analogous statement for $\hat{\boldsymbol{u}}_h$ is valid, the vector $\boldsymbol{v}_h := |\boldsymbol{f}|_\infty \boldsymbol{w}_h^{(1)} + |\hat{\boldsymbol{u}}_h|_\infty \boldsymbol{w}_h^{(2)} - \boldsymbol{u}_h$ satisfies

$$A_h \boldsymbol{v}_h \ge |\boldsymbol{f}|_\infty \mathbf{1} - \boldsymbol{f} - \hat{A}_h \left(|\hat{\boldsymbol{u}}_h|_\infty \mathbf{1} - \hat{\boldsymbol{u}}_h\right) \ge \mathbf{0} ,$$

where we have also used $-\hat{A}_h \ge 0$ in the last estimate. Therefore, the right inequality of (1) implies from Theorem 1.12 that the left inequality can be proven analogously. The further assertions follow immediately from (1). □

Because of the inverse monotonicity and from (1.32) (5) the vectors postulated in Theorem 1.14 have to satisfy $\boldsymbol{w}_h^{(i)} \ge \mathbf{0}$ necessarily for $i = 1, 2$. Thus stability with respect to $\|\cdot\|_\infty$ of the method defined by (1.31) assuming (1.32) (1)–(3), (4)* is guaranteed if a vector $\mathbf{0} \le \boldsymbol{w}_h \in \mathbb{R}^{M_1}$ and a constant $C > 0$ independent of h can be found such that

$$A_h \boldsymbol{w}_h \ge \mathbf{1} \quad \text{and} \quad |\boldsymbol{w}_h|_\infty \le C . \tag{1.41}$$

Finally, this will be proven for the five-point stencil discretization (1.1), (1.2) on the rectangle $\Omega = (0, a) \times (0, b)$ for $C = \frac{1}{16}(a^2 + b^2)$.

For this reason we define polynomials of second degree w_1, w_2 by

$$w_1(x) := \frac{1}{4}\, x(a - x) \quad \text{and} \quad w_2(y) := \frac{1}{4}\, y(b - y) . \tag{1.42}$$

It is clear that $w_1(x) \ge 0$ for all $x \in [0, a]$ and $w_2(y) \ge 0$ for all $y \in [0, b]$. Furthermore, we have $w_1(0) = 0 = w_1(a)$ and $w_2(0) = 0 = w_2(b)$, and

$$w_1''(x) = -\frac{1}{2} \quad \text{and} \quad w_2''(y) = -\frac{1}{2} .$$

Therefore w_1 and w_2 are strictly concave and attain their maximum in $\frac{a}{2}$ and $\frac{b}{2}$, respectively. Thus the function $w(x, y) := w_1(x) + w_2(x)$ satisfies

$$\begin{aligned} -\Delta w &= 1 \quad \text{in } \Omega , \\ w &\ge 0 \quad \text{on } \partial\Omega . \end{aligned} \tag{1.43}$$

Now let $\boldsymbol{w}_h \in \mathbb{R}^{M_1}$ be, for a fixed ordering, the representation of the grid function w_h defined by

$$(w_h)(ih, jh) := w(ih, jh) \quad \text{for } i = 1, \ldots, l - 1 ,\ j = 1, \ldots, m - 1 .$$

Analogously, let $\hat{\boldsymbol{w}}_h \in \mathbb{R}^{M_2}$ be the representation of the function $\hat{w}_h$ defined on $\partial\Omega_h^*$. As can be seen from the error representation in Lemma 1.2, statement 4, the difference quotient $\partial^-\partial^+ u(x)$ is exact for polynomials of second degree. Therefore, we conclude from (1.43) that

$$A_h \boldsymbol{w}_h = -\hat{A}_h \hat{\boldsymbol{w}}_h + \mathbf{1} \geq \mathbf{1}\,,$$

which finally implies

$$|\boldsymbol{w}_h|_\infty = \|w_h\|_\infty \leq \|w\|_\infty = w_1\Big(\frac{a}{2}\Big) + w_2\Big(\frac{b}{2}\Big) = \frac{1}{16}(a^2+b^2)\,.$$

This example motivates the following general procedure to construct $\boldsymbol{w}_h \in \mathbb{R}^{M_1}$ and a constant C such that (1.41) is fulfilled.

Assume that the boundary value problem under consideration reads in an abstract form

$$\begin{array}{lcll} (Lu)(x) & = & f(x) & \text{for} \quad x \in \Omega\,, \\ (Ru)(x) & = & g(x) & \text{for} \quad x \in \partial\Omega\,. \end{array} \tag{1.44}$$

Similar to (1.43) we can consider — in case of existence — a solution w of (1.44) for some f, g, such that $f(x) \geq 1$ for all $x \in \Omega$, $g(x) \geq 0$ for all $x \in \Omega$. If w is bounded on Ω, then

$$(\boldsymbol{w}_h)_i := w(x_i), \quad i = 1, \ldots, M_1,$$

for the (non-Dirichlet) grid points x_i, is a candidate for $\boldsymbol{w}_h$. Obviously,

$$|\boldsymbol{w}_h|_\infty \leq \|w\|_\infty\,.$$

Correspondingly, we set

$$(\hat{\boldsymbol{w}}_h)_i = w(x_i) \geq 0\,, \quad i = M_1+1, \ldots, M_2\,,$$

for the Dirichlet-boundary grid points.

The exact fulfillment of the discrete equations by $\boldsymbol{w}_h$ cannot be expected anymore, but in case of consistency the residual can be made arbitrarily small for small h. This leads to

Theorem 1.15 *Assume that a solution $w \in C(\overline{\Omega})$ of (1.44) exists for data $f \geq 1$ and $g \geq 0$. If the discretization of the form (1.31) is consistent with (1.44) (for these data), and there exists $\tilde{H} > 0$ so that for some $\tilde{\alpha} > 0$:*

$$-\hat{A}_h \hat{\boldsymbol{w}}_h + \boldsymbol{f} \geq \tilde{\alpha}\mathbf{1} \quad \textit{for} \quad h \leq \tilde{H}\,, \tag{1.45}$$

then for every $0 < \alpha < \tilde{\alpha}$ there exists $H > 0$, so that

$$A_h \boldsymbol{w}_h \geq \alpha \mathbf{1} \quad \textit{for} \quad h \leq H\,.$$

Proof: Set

$$\boldsymbol{\tau}_h := A_h \boldsymbol{w}_h + \hat{A}_h \hat{\boldsymbol{w}}_h - \boldsymbol{f}$$

for the consistency error, then

$$|\boldsymbol{\tau}_h|_\infty \to 0 \quad \text{for} \quad h \to 0\,.$$

Thus

$$\begin{aligned} A_h \boldsymbol{w}_h &= \boldsymbol{\tau}_h - \hat{A}_h \hat{\boldsymbol{w}}_h + \boldsymbol{f} \\ &\geq -|\boldsymbol{\tau}_h|_\infty \mathbf{1} + \tilde{\alpha}\mathbf{1} \quad \text{for} \quad h \leq \tilde{H} \\ &\geq \alpha \mathbf{1} \quad \text{for} \quad h \leq H \end{aligned}$$

and some appropriate $H > 0$. □

Thus a proper choice in (1.41) is

$$\frac{1}{\alpha}\boldsymbol{w}_h \quad \text{and} \quad C := \frac{1}{\alpha}\|w\|_\infty\,. \tag{1.46}$$

The condition (1.45) is not critical: In case of Dirichlet boundary conditions and (1.32) (5) (for corresponding rows i of $\hat{A}_h$) then, due to $(\boldsymbol{f})_i \geq 1$, we can even choose $\tilde{\alpha} = 1$. The discussion of Neumann boundary conditions following (1.24) shows that the same can be expected.

Theorem 1.15 shows that for a discretization with an inverse monotone system matrix consistency already implies stability.

To conclude this section let us discuss the various ingredients of (1.32) or (1.32)* that are sufficient for a range of properties from the inverse monotonicity up to a strong maximum principle: For the five-point stencil on a rectangle all the properties are valid for Dirichlet boundary conditions. If partly Neumann boundary conditions appear, the situation is the same, but now *close* and *far* from the boundary refers to its Dirichlet part. In the interpretation of the implications one has to take into account that the heterogeneities of the Neumann boundary condition are now part of the right-hand side $\boldsymbol{f}$, as seen, e.g., in (1.26). If mixed boundary conditions are applied, as

$$\partial_\nu u + \alpha u = g \quad \text{on} \quad \Gamma_2 \tag{1.47}$$

for some $\Gamma_2 \subset \Gamma$ and $\alpha = \alpha(x) > 0$, then the situation is the same again if αu is approximated just by evaluation, at the cost that (4)* no longer holds. The situation is similar if reaction terms appear in the differential equation (see Exercise 1.10).

Exercises

1.11 Give an example of a matrix $\hat{A}_h \in \mathbb{R}^{M_1, M_2}$ that can be used in the proof of Theorem 1.12.

1.12 Show that the transposition of an M-matrix is again an M-matrix.

1.13 In the assumptions of Theorem 1.9 substitute (1.32) (4) by (4)* and amend (6) to

$(6)^{\#}$ Condition (1.32) (6) is valid and
$\sum_{s=1}^{M_1} (A_h)_{rs} > 0 \Rightarrow$ there exists $s \in \{M_1, \ldots, M\}$ such that $(\hat{A}_h)_{rs} < 0$.

Under these conditions prove a weak maximum principle as in Theorem 1.9.

1.14 Assuming the existence of $\boldsymbol{w}_h \in \mathbb{R}^{M_1}$ such that $A_h \boldsymbol{w}_h \geq \mathbf{1}$ and $|\boldsymbol{w}_h|_\infty \leq C$ for some constant C independent of h, show directly (without Theorem 1.14) a refined order of convergence estimate on the basis of an order of consistency estimate in which also the shape of $\boldsymbol{w}_h$ appears.

2

The Finite Element Method for the Poisson Equation

The finite element method, frequently abbreviated by FEM, was developed in the fifties in the aircraft industry, after the concept had been independently outlined by mathematicians at an earlier time. Even today the notions used reflect that one origin of the development lies structural mechanics. Shortly after this beginning, the finite element method was applied to problems of heat conduction and fluid mechanics, which form the application background of this book.

An intensive mathematical analysis and further development was started in the later sixties. The basics of this mathematical description and analysis are to be developed in this and the following chapter. The homogeneous Dirichlet boundary value problem for the Poisson equation forms the paradigm of this chapter, but more generally valid considerations will be emphasized. In this way the abstract foundation for the treatment of more general problems in Chapter 3 is provided. In spite of the importance of the finite element method for structural mechanics, the treatment of the linear elasticity equations will be omitted. But we note that only a small expense is necessary for the application of the considerations to these equations. We refer to [11], where this is realized with a very similar notation.

2.1 Variational Formulation for the Model Problem

We will develop a new solution concept for the boundary value problem (1.1), (1.2) as a theoretical foundation for the finite element method. For

such a solution, the validity of the differential equation (1.1) is no longer required pointwise but in the sense of some integral average with "arbitrary" weighting functions φ. In the same way, the boundary condition (1.2) will be weakened by the renunciation of its pointwise validity.

For the present, we want to confine the considerations to the case of homogeneous boundary conditions (i.e., $g \equiv 0$), and so we consider the following homogeneous Dirichlet problem for the Poisson equation: Given a function $f : \Omega \to \mathbb{R}$, find a function $u : \overline{\Omega} \to \mathbb{R}$ such that

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega\,, && (2.1)\\ u &= 0 \quad \text{on } \partial\Omega\,. && (2.2) \end{aligned}$$

In the following let Ω be a domain such that the integral theorem of Gauss is valid, i.e. for any vector field $\boldsymbol{q} : \Omega \to \mathbb{R}^d$ with components in $C(\overline{\Omega}) \cap C^1(\Omega)$ it holds

$$\int_\Omega \nabla \cdot \boldsymbol{q}(x)\, dx = \int_{\partial\Omega} \nu(x) \cdot \boldsymbol{q}(x)\, d\sigma\,. \tag{2.3}$$

Let the function $u : \overline{\Omega} \to \mathbb{R}$ be a classical solution of (2.1), (2.2) in the sense of Definition 1.1, which additionally satisfies $u \in C^1(\overline{\Omega})$ to facilitate the reasoning. Next we consider arbitrary $v \in C_0^\infty(\Omega)$ as so-called *test functions*. The smoothness of these functions allows all operations of differentiation, and furthermore, all derivatives of a function $v \in C_0^\infty(\Omega)$ vanish on the boundary $\partial\Omega$. We multiply equation (2.1) by v, integrate the result over Ω, and obtain

$$\begin{aligned} \langle f, v\rangle_0 &= \int_\Omega f(x)v(x)\, dx \;=\; -\int_\Omega \nabla \cdot (\nabla u)(x)\, v(x)\, dx \\ &= \int_\Omega \nabla u(x) \cdot \nabla v(x)\, dx - \int_{\partial\Omega} \nabla u(x) \cdot \nu(x)\, v(x)\, d\sigma \qquad (2.4)\\ &= \int_\Omega \nabla u(x) \cdot \nabla v(x)\, dx\,. \end{aligned}$$

The equality sign at the beginning of the second line of (2.4) is obtained by integration by parts using the integral theorem of Gauss with $\boldsymbol{q} = v\nabla u$. The boundary integral vanishes because $v = 0$ holds on $\partial\Omega$.

If we define, for $u \in C^1(\overline{\Omega})$, $v \in C_0^\infty(\Omega)$, a real-valued mapping a by

$$a(u, v) := \int_\Omega \nabla u(x) \cdot \nabla v(x)\, dx\,,$$

then the classical solution of the boundary value problem satisfies the identity

$$a(u, v) = \langle f, v\rangle_0 \quad \text{for all } v \in C_0^\infty(\Omega)\,. \tag{2.5}$$

The mapping a defines a scalar product on $C_0^\infty(\Omega)$ that induces the norm

$$\|u\|_a := \sqrt{a(u,u)} = \left\{ \int_\Omega |\nabla u|^2 \, dx \right\}^{1/2} \tag{2.6}$$

(see Appendix A.4 for these notions). Most of the properties of a scalar product are obvious. Only the definiteness (A4.7) requires further considerations. Namely, we have to show that

$$a(u,u) = \int_\Omega (\nabla u \cdot \nabla u)(x) \, dx = 0 \iff u \equiv 0 \,.$$

To prove this assertion, first we show that $a(u,u) = 0$ implies $\nabla u(x) = 0$ for all $x \in \Omega$. To do this, we suppose that there exists some point $\bar{x} \in \Omega$ such that $\nabla u(\bar{x}) \neq 0$. Then $(\nabla u \cdot \nabla u)(\bar{x}) = |\nabla u|^2(\bar{x}) > 0$. Because of the continuity of ∇u, a small neighbourhood G of $\bar{x}$ exists with a positive measure $|G|$ and $|\nabla u|(x) \geq \alpha > 0$ for all $x \in G$. Since $|\nabla u|^2(x) \geq 0$ for all $x \in \Omega$, it follows that

$$\int_\Omega |\nabla u|^2 (x) \, dx \geq \alpha^2 \, |G| > 0 \,,$$

which is in contradiction to $a(u,u) = 0$. Consequently, $\nabla u(x) = 0$ holds for all $x \in \Omega$; i.e., u is constant in Ω. Since $u(x) = 0$ for all $x \in \partial\Omega$, the assertion follows.

Unfortunately, the space $C_0^\infty(\Omega)$ is too small to play the part of the basic space because the solution u does not belong to $C_0^\infty(\Omega)$ in general. The identity (2.4) is to be satisfied for a larger class of functions, which include, as an example for v, the solution u and the finite element approximation to u to be defined later.

For the present we define as the basic space V,

$$V := \big\{ u : \Omega \to \mathbb{R} \;\big|\; u \in C(\bar{\Omega}) \,,\ \partial_i u \text{ exists and is piecewise continuous for all } i = 1, \ldots, d,\ u = 0 \text{ on } \partial\Omega \big\}. \tag{2.7}$$

To say that $\partial_i u$ is *piecewise continuous* means that the domain Ω can be decomposed as follows:

$$\bar{\Omega} = \bigcup_j \bar{\Omega}_j \,,$$

with a finite number of open sets Ω_j, with $\Omega_j \cap \Omega_k = \emptyset$ for $j \neq k$, and $\partial_i u$ is continuous on Ω_j and it can continuously be extended on $\bar{\Omega}_j$.

Then the following properties hold:

- a is a scalar product also on V,
- $C_0^\infty(\Omega) \subset V$,
- $C_0^\infty(\Omega)$ is *dense* in V with respect to $\|\cdot\|_a$; i.e., for any $u \in V$ a sequence $(u_n)_{n\in\mathbb{N}}$ in $C_0^\infty(\Omega)$ exists such that $\|u_n - u\|_a \to 0$ for $n \to \infty$, (2.8)

- $C_0^\infty(\Omega)$ is dense in V with respect to $\|\cdot\|_0$. (2.9)

The first and second statements are obvious. The two others require a certain technical effort. A more general statement will be formulated in Theorem 3.7.

With that, we obtain from (2.5) the following result:

Lemma 2.1 *Let u be a classical solution of* (2.1), (2.2) *and let $u \in C^1(\bar{\Omega})$. Then*

$$a(u,v) = \langle f, v\rangle_0 \quad \textit{for all } v \in V\,. \tag{2.10}$$

Equation (2.10) is also called a variational equation.

Proof: Let $v \in V$. Then $v_n \in C_0^\infty(\Omega)$ exist with $v_n \to v$ with respect to $\|\cdot\|_0$ and also to $\|\cdot\|_a$. Therefore, it follows from the continuity of the bilinear form with respect to $\|\cdot\|_a$ (see (A4.22)) and the continuity of the functional defined by the right-hand side $v \mapsto \langle f, v\rangle_0$ with respect to $\|\cdot\|_0$ (because of the Cauchy–Schwarz inequality in $L^2(\Omega)$) that

$$\langle f, v_n\rangle_0 \to \langle f, v\rangle_0 \qquad \text{and} \qquad a(u, v_n) \to a(u, v) \quad \text{for } n \to \infty\,.$$

Since $a(u, v_n) = \langle f, v_n\rangle_0$, we get $a(u, v) = \langle f, v\rangle_0$. □

The space V in the identity (2.10) can be further enlarged as long as (2.8) and (2.9) will remain valid. This fact will be used later to give a correct definition.

Definition 2.2 A function $u \in V$ is called a *weak* (or *variational*) *solution* of (2.1), (2.2) if the following variational equation holds:

$$a(u,v) = \langle f, v\rangle_0 \quad \text{for all } v \in V\,.$$

If u models e.g. the displacement of a membrane, this relation is called the *principle of virtual work*.

Lemma 2.1 guarantees that a classical solution u is a weak solution.
The weak formulation has the following properties:

- It requires less smoothness: $\partial_i u$ has to be only piecewise continuous.
- The validity of the boundary condition is guaranteed by the definition of the function space V.

We now show that the variational equation (2.10) has exactly the same solution(s) as a minimization problem:

Lemma 2.3 *The variational equation* (2.10) *has the same solutions $u \in V$ as the minimization problem*

$$F(v) \to \min \quad \textit{for all } v \in V\,, \tag{2.11}$$

where

$$F(v) := \frac{1}{2}a(v,v) - \langle f, v\rangle_0 \quad \left(= \frac{1}{2}\|v\|_a^2 - \langle f, v\rangle_0\right).$$

Proof: (2.10) ⇒ (2.11):
Let u be a solution of (2.10) and let $v \in V$ be chosen arbitrarily. We define $w := v - u \in V$ (because V is a vector space), i.e., $v = u + w$. Then, using the bilinearity and symmetry, we have

$$\begin{aligned} F(v) &= \frac{1}{2}a(u+w,u+w) - \langle f, u+w\rangle_0 \\ &= \frac{1}{2}a(u,u) + a(u,w) + \frac{1}{2}a(w,w) - \langle f,u\rangle_0 - \langle f,w\rangle_0 \quad (2.12) \\ &= F(u) + \frac{1}{2}a(w,w) \geq F(u)\,, \end{aligned}$$

where the last inequality follows from the positivity of a; i.e., (2.11) holds.

(2.10) ⇐ (2.11):
Let u be a solution of (2.11) and let $v \in V$, $\varepsilon \in \mathbb{R}$ be chosen arbitrarily. We define $g(\varepsilon) := F(u + \varepsilon v)$ for $\varepsilon \in \mathbb{R}$. Then

$$g(\varepsilon) = F(u+\varepsilon v) \geq F(u) = g(0) \quad \text{for all } \varepsilon \in \mathbb{R}\,,$$

because $u + \varepsilon v \in V$; i.e., g has a global minimum at $\varepsilon = 0$.

It follows analogously to (2.12):

$$g(\varepsilon) = \frac{1}{2}a(u,u) - \langle f,u\rangle_0 + \varepsilon\left(a(u,v) - \langle f,v\rangle_0\right) + \frac{\varepsilon^2}{2}a(v,v)\,.$$

Hence the function g is a quadratic polynomial in ε, and in particular, $g \in C^1(\mathbb{R})$ is valid. Therefore we obtain the necessary condition

$$0 = g'(\varepsilon) = a(u,v) - \langle f,v\rangle_0$$

for the existence of a minimum at $\varepsilon = 0$. Thus u solves (2.10), because $v \in V$ has been chosen arbitrarily. □

For applications e.g. in structural mechanics as above, the minimization problem is called the *principle of minimal potential energy.*

Remark 2.4 Lemma 2.3 holds for general vector spaces V if a is a symmetric, positive bilinear form and the right-hand side $\langle f, v\rangle_0$ is replaced by $b(v)$, where $b : V \to \mathbb{R}$ is a linear mapping, a *linear functional.* Then the variational equation reads as

$$\text{find } u \in V \quad \text{with} \quad a(u,v) = b(v) \quad \text{for all } v \in V\,, \tag{2.13}$$

and the minimization problem as

$$\text{find } u \in V \quad \text{with} \quad F(u) = \min\left\{F(v) \;\middle|\; v \in V\right\}, \tag{2.14}$$

where $F(v) := \frac{1}{2}a(v,v) - b(v)$.

Lemma 2.5 *The weak solution according to* (2.10) (*or* (2.11)) *is unique.*

Proof: Let u_1, u_2 be two weak solutions, i.e.,

$$\begin{aligned} a(u_1, v) &= \langle f, v\rangle_0 \,, \\ a(u_2, v) &= \langle f, v\rangle_0 \,, \end{aligned} \qquad \text{for all } v \in V\,.$$

By subtraction, it follows that

$$a(u_1 - u_2, v) = 0 \quad \text{for all } v \in V.$$

Choosing $v = u_1 - u_2$ implies $a(u_1 - u_2, u_1 - u_2) = 0$ and consequently $u_1 = u_2$, because a is definite. □

Remark 2.6 Lemma 2.5 is generally valid if a is a definite bilinear form and b is a linear form.

So far, we have defined two different norms on V: $\|\cdot\|_a$ and $\|\cdot\|_0$. The difference between these norms is essential because they are not equivalent on the vector space V defined by (2.7), and consequently, they generate different convergence concepts, as will be shown by the following example:

Example 2.7 Let $\Omega = (0,1)$, i.e.

$$a(u,v) := \int_0^1 u'v'\,dx\,,$$

and let $v_n : \Omega \to \mathbb{R}$ for $n \geq 2$ be defined by (cf. Figure 2.1)

$$v_n(x) = \begin{cases} nx\,, & \text{for} \quad 0 \leq x \leq \frac{1}{n}\,, \\ 1\,, & \text{for} \quad \frac{1}{n} \leq x \leq 1 - \frac{1}{n}\,, \\ n - nx\,, & \text{for} \quad 1 - \frac{1}{n} \leq x \leq 1\,. \end{cases}$$

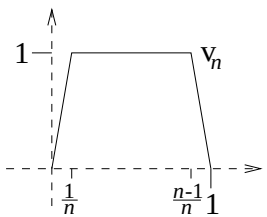


Figure 2.1. The function v_n.

Then

$$\|v_n\|_0 \quad \leq \quad \left\{\int_0^1 1\,dx\right\}^{1/2} = 1\,,$$

$$\|v_n\|_a \;=\; \left\{ \int_0^{\frac{1}{n}} n^2\,dx + \int_{1-\frac{1}{n}}^{1} n^2\,dx \right\}^{1/2} = \sqrt{2n} \to \infty \text{ for } n \to \infty\,.$$

Therefore, there exists no constant $C > 0$ such that $\|v\|_a \le C\|v\|_0$ for all $v \in V$.

However, as we will show in Theorem 2.18, there exists a constant $C > 0$ such that the estimate

$$\|v\|_0 \le C\|v\|_a \quad \text{for all } v \in V$$

holds; i.e., $\|\cdot\|_a$ is the stronger norm.

It is possible to enlarge the basic space V without violating the previous statements. The enlargement is also necessary because, for instance, the proof of the existence of a solution of the variational equation (2.13) or the minimization problem (2.14) requires in general the completeness of V. However, the actual definition of V does not imply the completeness, as the following example shows:

Example 2.8 Let $\Omega = (0,1)$ again and therefore

$$a(u,v) := \int_0^1 u'v'\,dx\,.$$

For $u(x) := x^\alpha(1-x)^\alpha$ with $\alpha \in \left(\frac{1}{2}, 1\right)$ we consider the sequence of functions

$$u_n(x) := \begin{cases} u(x) & \text{for} \quad x \in \left[\frac{1}{n}, 1-\frac{1}{n}\right]\,, \\ n\,u(\frac{1}{n})\,x & \text{for} \quad x \in \left[0, \frac{1}{n}\right]\,, \\ n\,u(1-\frac{1}{n})\,(1-x) & \text{for} \quad x \in \left[1-\frac{1}{n}, 1\right]\,. \end{cases}$$

Then

$$\begin{aligned} \|u_n - u_m\|_a &\to 0 \qquad \text{for } n, m \to \infty\,, \\ \|u_n - u\|_a &\to 0 \qquad \text{for } n \to \infty\,, \end{aligned}$$

but $u \notin V$, where V is defined analogously to (2.7) with $d = 1$.

In Section 3.1 we will see that a vector space $\tilde{V}$ normed with $\|\cdot\|_a$ exists such that $u \in \tilde{V}$ and $V \subset \tilde{V}$. Therefore, V is not complete with respect to $\|\cdot\|_a$; otherwise, $u \in V$ must be valid. In fact, there exists a (unique) completion of V with respect to $\|\cdot\|_a$ (see Appendix A.4, especially (A4.26)), but we have to describe the new "functions" added by this process. Besides, integration by parts must be valid such that a classical solution continues to be also a weak solution (compare with Lemma 2.1). Therefore, the following idea is unsuitable.

Attempt of a correct definition of V:
Let V be the set of all u with the property that $\partial_i u$ exists for all $x \in \Omega$ without any requirements on $\partial_i u$ in the sense of a function.

For instance, there exists *Cantor's function* with the following properties: $f : [0,1] \to \mathbb{R}$, $f \in C([0,1])$, $f \neq 0$, f is not constant, $f'(x)$ exists with $f'(x) = 0$ for all $x \in [0,1]$.

Here the fundamental theorem of calculus, $f(x) = \int_0^x f'(s)\,ds + f(0)$, and thus the principle of integration by parts, are no longer valid.

Consequently, additional conditions for $\partial_i u$ are necessary.

To prepare an adequate definition of the space V, we extend the definition of derivatives by means of their action on averaging procedures. In order to do this, we introduce the *multi-index* notation.

A vector $\alpha = (\alpha_1, \ldots, \alpha_d)$ of nonnegative integers $\alpha_i \in \{0, 1, 2, \ldots\}$ is called a *multi-index*. The number $|\alpha| := \sum_{i=1}^d \alpha_i$ denotes the *order* (or *length*) of α.
For $x \in \mathbb{R}^d$ let

$$x^\alpha := x_1^{\alpha_1} \cdots x_d^{\alpha_d}\,. \tag{2.15}$$

A shorthand notation for the differential operations can be adopted by this: For an appropriately differentiable function u let

$$\partial^\alpha u := \partial_1^{\alpha_1} \cdots \partial_d^{\alpha_d} u\,. \tag{2.16}$$

We can obtain this definition from (2.15) by replacing x by the symbolic vector

$$\nabla := (\partial_1, \ldots, \partial_d)^T$$

of the first partial derivatives.

For example, if $d = 2$ and $\alpha = (1, 2)$, then $|\alpha| = 3$ and

$$\partial^\alpha u = \partial_1 \partial_2^2 u = \frac{\partial^3 u}{\partial x_1 \partial x_2^2}\,.$$

Now let α be a multi-index of length k and let $u \in C^k(\Omega)$. We then obtain for arbitrary test functions $\varphi \in C_0^\infty(\Omega)$ by integration by parts

$$\int_\Omega \partial^\alpha u\, \varphi\, dx = (-1)^k \int_\Omega u\, \partial^\alpha \varphi\, dx\,.$$

The boundary integrals vanish because $\partial^\beta \varphi = 0$ on $\partial\Omega$ for all multi-indices β.

Therefore, we make the following definition:

Definition 2.9 $v \in L^2(\Omega)$ is called the *weak* (or *generalized*) derivative $\partial^\alpha u$ of $u \in L^2(\Omega)$ for the multi-index α if for all $\varphi \in C_0^\infty(\Omega)$,

$$\int_\Omega v\, \varphi\, dx = (-1)^{|\alpha|} \int_\Omega u\, \partial^\alpha \varphi\, dx\,.$$

The weak derivative is well-defined because it is unique: Let $v_1, v_2 \in L^2(\Omega)$ be two weak derivatives of u. It follows that

$$\int_\Omega (v_1 - v_2)\,\varphi\,dx = 0 \quad \text{for all } \varphi \in C_0^\infty(\Omega)\,.$$

Since $C_0^\infty(\Omega)$ is dense in $L^2(\Omega)$, we can furthermore conclude that

$$\int_\Omega (v_1 - v_2)\,\varphi\,dx = 0 \quad \text{for all } \varphi \in L^2(\Omega)\,.$$

If we now choose specifically $\varphi = v_1 - v_2$, we obtain

$$\|v_1 - v_2\|_0^2 = \int_\Omega (v_1 - v_2)\,(v_1 - v_2)\;dx = 0\,,$$

and $v_1 = v_2$ (a.e.) follows immediately. In particular, $u \in C^k(\bar\Omega)$ has weak derivatives $\partial^\alpha u$ for α with $|\alpha| \le k$, and the weak derivatives are identical to the classical (pointwise) derivatives.

Also the differential operators of vector calculus can be given a weak definition analogous to Definition 2.9. For example, for a vector field $\boldsymbol{q}$ with components in $L^2(\Omega)$, $v \in L^2(\Omega)$ is the *weak divergence* $v = \nabla \cdot \boldsymbol{q}$ if for all $\varphi \in C_0^\infty(\Omega)$

$$\int_\Omega v\varphi\,dx = -\int_\Omega \boldsymbol{q} \cdot \nabla\varphi\,dx\,.$$

The **correct choice of the space** V is the space $H_0^1(\Omega)$, which will be defined below. First we define

$$\begin{aligned} H^1(\Omega) \;:=\; \big\{ u : \Omega \to \mathbb{R} \;\big|\; & u \in L^2(\Omega)\,,\ u \text{ has weak derivatives} \\ & \partial_i u \in L^2(\Omega) \text{ for all } i = 1, \dots, d \big\}\,. \end{aligned} \tag{2.17}$$

A scalar product on $H^1(\Omega)$ is defined by

$$\langle u, v \rangle_1 := \int_\Omega u(x)v(x)\,dx + \int_\Omega \nabla u(x) \cdot \nabla v(x)\,dx \tag{2.18}$$

with the norm

$$\|u\|_1 := \sqrt{\langle u, u \rangle_1} = \left\{ \int_\Omega |u(x)|^2\,dx + \int_\Omega |\nabla u(x)|^2\,dx \right\}^{1/2} \tag{2.19}$$

induced by this scalar product.

The above "temporary" definition (2.7) of V takes care of the boundary condition $u = 0$ on $\partial\Omega$ by conditions for the functions. I.e. we want to choose the basic space V analogously as:

$$H_0^1(\Omega) := \big\{ u \in H^1(\Omega) \;\big|\; u = 0 \quad \text{on } \partial\Omega \big\}\,. \tag{2.20}$$

Here $H^1(\Omega)$ and $H_0^1(\Omega)$ are special cases of so-called *Sobolev spaces*.

For $\Omega \subset \mathbb{R}^d$, $d \ge 2$, $H^1(\Omega)$ may contain unbounded functions. In particular, we have to examine carefully the meaning of $u|_{\partial\Omega}$ ($\partial\Omega$ has the

d-dimensional measure 0) and, in particular, $u = 0$ on $\partial\Omega$. This will be described in Section 3.1.

Exercises

2.1

(a) Consider the interval $(-1, 1)$; prove that the function $u(x) = |x|$ has the generalized derivative $u'(x) = \text{sign}(x)$.

(b) Does $\text{sign}(x)$ have a generalized derivative?

2.2 Let $\overline{\Omega} = \bigcup_{l=1}^{N} \overline{\Omega}_l$, $N \in \mathbb{N}$, where the bounded subdomains $\Omega_l \subset \mathbb{R}^2$ are pairwise disjoint and possess piecewise smooth boundaries. Show that a function $u \in C(\overline{\Omega})$ with $u|_{\Omega_l} \in C^1(\overline{\Omega}_l)$, $1 \le l \le N$, has a weak derivative $\partial_i u \in L^2(\Omega)$, $i = 1, 2$, that coincides in $\bigcup_{l=1}^{N} \Omega_l$ with the classical one.

2.3 Let V be the set of functions that are continuous and piecewise continuously differentiable on $[0, 1]$ and that satisfy the additional conditions $u(0) = u(1) = 0$. Show that there exist infinitely many elements in V that minimize the functional

$$F(u) := \int_0^1 \left\{1 - [u'(x)]^2\right\}^2 dx.$$

2.2 The Finite Element Method with Linear Elements

The weak formulation of the boundary value problem (2.1), (2.2) leads to particular cases of the following general, here equivalent, problems:

Let V be a vector space, let $a : V \times V \to \mathbb{R}$ be a bilinear form, and let $b : V \to \mathbb{R}$ be a linear form.

Variational equation:

$$\text{Find } u \in V \quad \text{with} \quad a(u, v) = b(v) \quad \text{for all } v \in V\,. \tag{2.21}$$

Minimization problem:

$$\begin{aligned} &\text{Find } u \in V \quad \text{with} \quad F(u) = \min\left\{F(v) \;\middle|\; v \in V\right\}\,, \\ &\text{where} \quad F(v) = \frac{1}{2}a(v, v) - b(v)\,. \end{aligned} \tag{2.22}$$

The *discretization approach* consists in the following procedure: Replace V by a finite-dimensional subspace V_h; i.e., solve instead of (2.21) the finite-dimensional variational equation,

$$\text{find } u_h \in V_h \quad \text{with} \quad a(u_h, v) = b(v) \quad \text{for all } v \in V_h\,. \tag{2.23}$$

This approach is called the *Galerkin method*. Or solve instead of (2.22) the finite-dimensional minimization problem,

$$\text{find } u_h \in V_h \quad \text{with} \quad F(u_h) = \min \left\{ F(v) \;\middle|\; v \in V_h \right\} . \tag{2.24}$$

This approach is called the *Ritz method*.

It is clear from Lemma 2.3 and Remark 2.4 that the Galerkin method and the Ritz method are equivalent for a positive and symmetric bilinear form. The finite-dimensional subspace V_h is called an *ansatz space*.

The finite element method can be interpreted as a Galerkin method (and in our example as a Ritz method, too) for an ansatz space with special properties. In the following, these properties will be extracted by means of the simplest example.

Let V be defined by (2.7) or let $V = H_0^1(\Omega)$.
The weak formulation of the boundary value problem (2.1), (2.2) corresponds to the choice

$$a(u,v) := \int_\Omega \nabla u \cdot \nabla v \, dx \,, \quad b(v) := \int_\Omega f v \, dx \,.$$

Let $\Omega \subset \mathbb{R}^2$ be a domain with a polygonal boundary; i.e., the boundary Γ of Ω consists of a finite number of straight-line segments as shown in Figure 2.2.

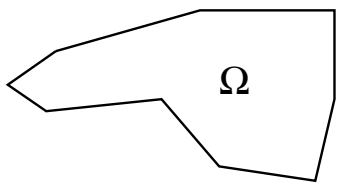


Figure 2.2. Domain with a polygonal boundary.

Let $\mathcal{T}_h$ be a partition of Ω into closed triangles K (i.e., including the boundary ∂K) with the following properties:

(1) $\overline{\Omega} = \cup_{K \in \mathcal{T}_h} K$;

(2) For $K, K' \in \mathcal{T}_h$, $K \neq K'$,

$$\text{int}\,(K) \cap \text{int}\,(K') = \emptyset \,, \tag{2.25}$$

where $\text{int}\,(K)$ denotes the open triangle (without the boundary ∂K).

(3) If $K \neq K'$ but $K \cap K' \neq \emptyset$, then $K \cap K'$ is either a point or a common edge of K and K' (cf. Figure 2.3).

A partition of Ω with the properties (1), (2) is called a *triangulation* of Ω. If, in addition, a partition of Ω satisfies property (3), it is called a *conforming triangulation* (cf. Figure 2.4).

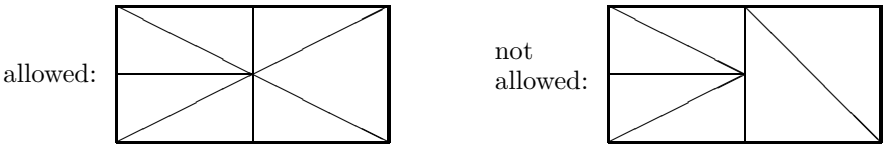


Figure 2.3. Triangulations.

The triangles of a triangulation will be numbered $K_1, \ldots, K_N$. The subscript h indicates the fineness of the triangulation, e.g.,

$$h := \max \left\{ \operatorname{diam}(K) \;\middle|\; K \in \mathcal{T}_h \right\} ,$$

where $\operatorname{diam}(K) := \sup \left\{ |x - y| \;\middle|\; x, y \in K \right\}$ denotes the diameter of K. Thus here h is the maximum length of the edges of all the triangles. Sometimes, $K \in \mathcal{T}_h$ is also called a (geometric) *element* of the partition.

The vertices of the triangles are called the *nodes*, and they will be numbered

$$a_1, a_2, \ldots, a_M ,$$

i.e., $a_i = (x_i, y_i)$, $i = 1, \ldots, M$, where $M = M_1 + M_2$ and

$$\begin{aligned} a_1, \ldots, a_{M_1} &\in \Omega , \\ a_{M_1+1}, \ldots, a_M &\in \partial\Omega . \end{aligned} \tag{2.26}$$

This kind of arrangement of the nodes is chosen only for the sake of simplicity of the notation and is not essential for the following considerations.

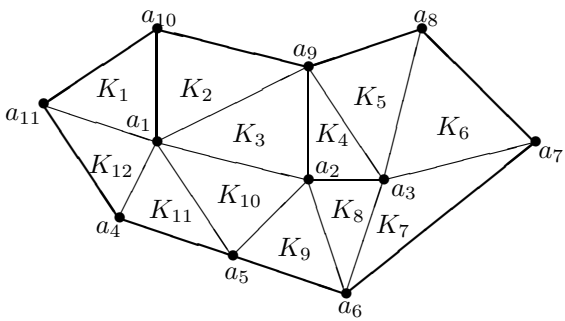


Figure 2.4. A conforming triangulation with $N = 12$, $M = 11$, $M_1 = 3$, $M_2 = 8$.

An approximation of the boundary value problem (2.1), (2.2) with *linear finite elements* on a given triangulation $\mathcal{T}_h$ of Ω is obtained if the ansatz space V_h is defined as follows:

$$V_h := \left\{ u \in C(\bar{\Omega}) \;\middle|\; u|_K \in \mathcal{P}_1(K) \text{ for all } K \in \mathcal{T}_h,\ u = 0 \text{ on } \partial\Omega \right\} . \tag{2.27}$$

Here $\mathcal{P}_1(K)$ denotes the set of polynomials of first degree (in 2 variables) on K; i.e., $p \in \mathcal{P}_1(K) \Leftrightarrow p(x,y) = \alpha + \beta x + \gamma y$ for all $(x,y) \in K$ and for fixed $\alpha, \beta, \gamma \in \mathbb{R}$.

Since $p \in \mathcal{P}_1(K)$ is also defined on the space $\mathbb{R} \times \mathbb{R}$, we use the short but inaccurate notation $\mathcal{P}_1 = \mathcal{P}_1(K)$; according to the context, the domain of definition will be given as $\mathbb{R} \times \mathbb{R}$ or as a subset of it.

We have

$$V_h \subset V\,.$$

This is clear for the case of definition of V by (2.7) because $\partial_x u|_K = \text{const}$, $\partial_y u|_K = \text{const}$ for $K \in \mathcal{T}_h$ for all $u \in V_h$. If $V = H_0^1(\Omega)$, then this inclusion is not so obvious. A proof will be given in Theorem 3.20 below.

An element $u \in V_h$ is determined uniquely by the values $u(a_i)$, $i = 1, \ldots, M_1$ (the *nodal values*).

In particular, the given nodal values already enforce the continuity of the piecewise linear composed functions. Correspondingly, the homogeneous Dirichlet boundary condition is satisfied if the nodal values at the boundary nodes are set to zero.

In the following, we will demonstrate these properties by an unnecessarily involved proof. The reason is that this proof will introduce all of the considerations that will lead to analogous statements for the more general problems of Section 3.4.

Let X_h be the larger ansatz space consisting of continuous, piecewise linear functions but regardless of any boundary conditions, i.e.,

$$X_h := \left\{ u \in C(\bar{\Omega}) \;\middle|\; u|_K \in \mathcal{P}_1(K) \quad \text{for all } K \in \mathcal{T}_h \right\}\,.$$

Lemma 2.10 *For given values at the nodes $a_1, \ldots, a_M$, the interpolation problem in X_h is uniquely solvable. That is, if the values $u_1, \ldots, u_M$ are given, then there exists a uniquely determined element*

$$u \in X_h \quad \textit{such that} \quad u(a_i) = u_i\,, \quad i = 1, \ldots, M\,.$$

If $u_j = 0$ for $j = M_1 + 1, \ldots, M$, then it is even true that

$$u \in V_h\,.$$

Proof: **(1)** For any arbitrary $K \in \mathcal{T}_h$ we consider the *local interpolation problem*:

$$\text{Find } p = p_K \in \mathcal{P}_1 \quad \text{such that} \quad p(a_i) = u_i\,,\ i = 1,2,3\,, \tag{2.28}$$

where a_i, $i = 1,2,3$, denote the vertices of K, and the values u_i, $i = 1,2,3$, are given. First we show that problem (2.28) is uniquely solvable for a particular triangle.

A solution of (2.28) for the so-called *reference element* $\hat{K}$ (cf. Figure 2.5) with the vertices $\hat{a}_1 = (0,0)$, $\hat{a}_2 = (1,0)$, $\hat{a}_3 = (0,1)$ is given by

$$p(x,y) = u_1 N_1(x,y) + u_2 N_2(x,y) + u_3 N_3(x,y)$$

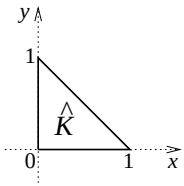


Figure 2.5. Reference element $\hat{K}$.

with the *shape functions*

$$\begin{aligned} N_1(x,y) &= 1-x-y\,, \\ N_2(x,y) &= x\,, \\ N_3(x,y) &= y\,. \end{aligned} \tag{2.29}$$

Evidently, $N_i \in \mathcal{P}_1$, and furthermore,

$$N_i\left(\hat{a}_j\right) = \delta_{ij} = \begin{cases} 1 & \text{for } i=j\,, \\ 0 & \text{for } i\neq j\,, \end{cases} \qquad \text{for} \quad i,j=1,2,3\,,$$

and thus

$$p\left(\hat{a}_j\right) = \sum_{i=1}^{3} u_i N_i\left(\hat{a}_j\right) = u_j \quad \text{for all} \quad j=1,2,3\,.$$

The uniqueness of the solution can be seen in the following way: If p_1, p_2 satisfy the interpolation problem (2.28) for the reference element, then for $p := p_1 - p_2 \in \mathcal{P}_1$ we have

$$p\left(\hat{a}_i\right) = 0\,, \quad i=1,2,3\,.$$

Here p is given in the form $p(x,y) = \alpha + \beta x + \gamma y$. If we fix the second variable $y=0$, we obtain a polynomial function of one variable

$$p(x,0) = \alpha + \beta x =: q(x) \in \mathcal{P}_1(\mathbb{R})\,.$$

The polynomial q satisfies $q(0) = 0 = q(1)$, and $q \equiv 0$ follows by the uniqueness of the polynomial interpolation in one variable; i.e., $\alpha = \beta = 0$. Analogously, we consider

$$q(y) := p(0,y) = \alpha + \gamma y = \gamma y\,,$$

and we obtain from $q(1) = 0$ that $\gamma = 0$ and consequently $p \equiv 0$.

In fact, this additional proof of uniqueness is not necessary, because the uniqueness already follows from the solvability of the interpolation problem because of $\dim \mathcal{P}_1 = 3$ (compare with Section 3.3).

Now we turn to the case of a general triangle K. A general triangle K is mapped onto $\hat{K}$ by an affine transformation (cf. Figure 2.6)

$$F : \hat{K} \to K\,, \quad F(\hat{x}) = B\hat{x} + d\,, \tag{2.30}$$

where $B \in \mathbb{R}^{2,2}$, $d \in \mathbb{R}^2$ are such that $F\left(\hat{a}_i\right) = a_i$.

$B = (b_1, b_2)$ and d are determined by the vertices a_i of K as follows:

$$\begin{aligned} a_1 &= F(\hat{a}_1) = F(0) = d\,, \\ a_2 &= F(\hat{a}_2) = b_1 + d = b_1 + a_1\,, \\ a_3 &= F(\hat{a}_3) = b_2 + d = b_2 + a_1\,; \end{aligned}$$

i.e., $b_1 = a_2 - a_1$ and $b_2 = a_3 - a_1$. The matrix B is regular because $a_2 - a_1$ and $a_3 - a_1$ are linearly independent, ensuring $F(\hat{a}_i) = a_i$.

Since

$$K = \text{conv}\,\{a_1, a_2, a_3\} := \left\{ \sum_{i=1}^{3} \lambda_i a_i \;\Big|\; 0 \le \lambda_i \le 1\,,\; \sum_{i=1}^{3} \lambda_i = 1 \right\}$$

and especially $\hat{K} = \text{conv}\,\{\hat{a}_1, \hat{a}_2, \hat{a}_3\}$, $F[\hat{K}] = K$ follows from the fact that the affine-linear mapping F satisfies

$$F\left(\sum_{i=1}^{3} \lambda_i \hat{a}_i \right) = \sum_{i=1}^{3} \lambda_i F(\hat{a}_i) = \sum_{i=1}^{3} \lambda_i a_i$$

for $0 \le \lambda_i \le 1$, $\sum_{i=1}^{3} \lambda_i = 1$.

In particular, the edges (where one λ_i is equal to 0) of $\hat{K}$ are mapped onto the edges of K.

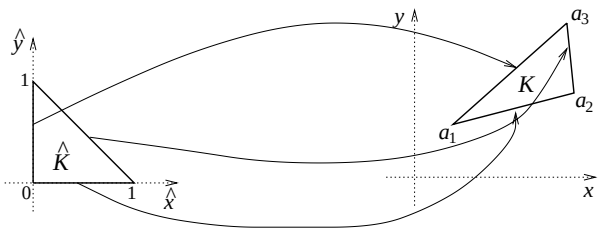


Figure 2.6. Affine-linear transformation.

Analogously, the considerations can be applied to the space $\mathbb{R}^d$ word for word by replacing the set of indices $\{1, 2, 3\}$ by $\{1, \ldots, d+1\}$. This will be done in Section 3.3.

The polynomial space $\mathcal{P}_1$ does not change under the affine transformation F.

(2) We now prove that the local functions $u|_K$ can be composed continuously:

For every $K \in \mathcal{T}_h$, let $p_K \in \mathcal{P}_1$ be the unique solution of (2.28), where the values u_1, u_2, u_3 are the values $u_{i_1}, u_{i_2}, u_{i_3} (i_1, i_2, i_3 \in \{1, \ldots, M\})$ that have to be interpolated at these nodes.

Let $K, K' \in \mathcal{T}_h$ be two different elements that have a common edge E. Then $p_K = p_{K'}$ on E is to be shown. This is valid because E can be mapped onto $[0, 1] \times \{0\}$ by an affine transformation (cf. Figure 2.7). Then

$q_1(x) = p_K(x, 0)$ and $q_2(x) := p_{K'}(x, 0)$ are elements of $\mathcal{P}_1(\mathbb{R})$, and they solve the same interpolation problem at the points $x = 0$ and $x = 1$; thus $q_1 \equiv q_2$.

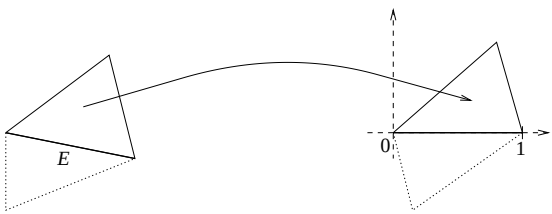


Figure 2.7. Affine-linear transformation of E on the reference element $[0, 1]$.

Therefore, the definition of u by means of

$$u(x) = p_K(x) \quad \text{for } x \in K \in \mathcal{T}_h \tag{2.31}$$

is unique, and this function satisfies $u \in C(\bar{\Omega})$ and $u \in X_h$.

(3) Finally, we will show that $u = 0$ on $\partial\Omega$ for u defined by (2.31) if $u_i = 0\,(i = M_1 + 1, \ldots, M)$ for the boundary nodes.
The boundary $\partial\Omega$ consists of edges of elements $K \in \mathcal{T}_h$. Let E be such an edge; i.e., E has the vertices a_{i_1}, a_{i_2} with $i_j \in \{M_1 + 1, \ldots, M\}$. The given boundary values yield $u(a_{i_j}) = 0$ for $j = 1, 2$. By means of an affine transformation analogously to the above one we obtain that $u|_E$ is a polynomial of first degree in one variable and that $u|_E$ vanishes at two points. So $u|_E = 0$, and the assertion follows. □

The following statement is an important consequence of the unique solvability of the interpolation problem in X_h irrespective of its particular definition: The interpolation conditions

$$\varphi_i(a_j) = \delta_{ij}\,, \quad j = 1, \ldots, M\,, \tag{2.32}$$

uniquely determine functions $\varphi_i \in X_h$ for $i = 1, \ldots, M$. For any $u \in X_h$, we have

$$u(x) = \sum_{i=1}^{M} u(a_i)\varphi_i(x) \quad \text{for} \quad x \in \Omega\,, \tag{2.33}$$

because both the left-hand side and the right-hand side functions belong to X_h and are equal to $u(a_i)$ at $x = a_i$.

The representation $u = \sum_{i=1}^{M} \alpha_i\varphi_i$ is unique, too, for otherwise, a function $w \in X_h$, $w \neq 0$, such that $w(a_i) = 0$ for all $i = 1, \ldots, M$ would exist. Thus $\{\varphi_1, \ldots, \varphi_M\}$ is a basis of X_h, especially $\dim X_h = M$. This basis is called a *nodal basis* because of (2.33). For the particular case of a piecewise linear ansatz space on triangles, the basis functions are called

pyramidal functions because of their shape. If the set of indices is restricted to $\{1, \ldots, M_1\}$; i.e., we omit the basis functions corresponding to the boundary nodes, then a basis of V_h will be obtained and $\dim V_h = M_1$.

Summary: The function values $u(a_i)$ at the nodes $a_1, \ldots, a_M$ are the *degrees of freedom* of $u \in X_h$, and the values at the interior points $a_1, \ldots, a_{M_1}$ are the *degrees of freedom* of $u \in V_h$.

The following consideration is valid for an arbitrary ansatz space V_h with a basis $\{\varphi_1, \ldots, \varphi_M\}$. The Galerkin method (2.23) reads as follows: Find $u_h = \sum_{i=1}^{M} \xi_i \varphi_i \in V_h$ such that $a(u_h, v) = b(v)$ for all $v \in V_h$. Since $v = \sum_{i=1}^{M} \eta_i \varphi_i$ for $\eta_i \in \mathbb{R}$, this is equivalent to

$$\begin{aligned} a(u_h, \varphi_i) &= b(\varphi_i) \quad \text{for all } i = 1, \ldots, M \iff \\ a\left(\sum_{j=1}^{M} \xi_j \varphi_j, \varphi_i\right) &= b(\varphi_i) \quad \text{for all } i = 1, \ldots, M \iff \\ \sum_{j=1}^{M} a(\varphi_j, \varphi_i)\, \xi_j &= b(\varphi_i) \quad \text{for all } i = 1, \ldots, M \iff \\ A_h \boldsymbol{\xi} &= \boldsymbol{q}_h \end{aligned} \tag{2.34}$$

with $A_h = (a(\varphi_j, \varphi_i))_{ij} \in \mathbb{R}^{M,M}$, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_M)^T$ and $\boldsymbol{q}_h = (b(\varphi_i))_i$. Therefore, the Galerkin method is equivalent to the system of equations (2.34).

The considerations for deriving (2.34) show that, in the case of equivalence of the Galerkin method with the Ritz method, the system of equations (2.34) is equivalent to the minimization problem

$$F_h(\boldsymbol{\xi}) = \min \left\{ F_h(\boldsymbol{\eta}) \;\middle|\; \boldsymbol{\eta} \in \mathbb{R}^M \right\}, \tag{2.35}$$

where

$$F_h(\boldsymbol{\eta}) = \frac{1}{2} \boldsymbol{\eta}^T A_h \boldsymbol{\eta} - \boldsymbol{q}_h^T \boldsymbol{\eta}\,.$$

Because of the symmetry and positive definiteness, the equivalence of (2.34) and (2.35) can be easily proven, and it forms the basis for the CG methods that will be discussed in Section 5.2.

Usually, A_h is called *stiffness matrix*, and $\boldsymbol{q}_h$ is called the *load vector*. These names originated from mechanics. For our model problem, we have

$$\begin{aligned} (A_h)_{ij} &= a(\varphi_j, \varphi_i) = \int_\Omega \nabla \varphi_j \cdot \nabla \varphi_i \, dx\,, \\ (\mathbf{q}_h)_i &= b(\varphi_i) = \int_\Omega f \varphi_i \, dx\,. \end{aligned}$$

By applying the finite element method, we thus have to perform the following steps:

(1) Determination of $A_h, \boldsymbol{q}_h$. This step is called *assembling*.

(2) Solution of $A_h \boldsymbol{\xi} = \boldsymbol{q}_h$.

If the basis functions φ_i have the property $\varphi_i(a_j) = \delta_{ij}$, then the solution of system (2.34) satisfies the relation $\xi_i = u_h(a_i)$, i.e., we obtain the vector of the nodal values of the finite element approximation.

Using only the properties of the bilinear form a, we obtain the following properties of A_h:

- A_h is symmetric for an arbitrary basis $\{\varphi_i\}$ because a is symmetric.
- A_h is positive definite for an arbitrary basis $\{\varphi_i\}$ because for $u = \sum_{i=1}^M \xi_i \varphi_i$,

$$\begin{aligned} \boldsymbol{\xi}^T A_h \boldsymbol{\xi} &= \sum_{i,j=1}^M \xi_j a(\varphi_j, \varphi_i)\xi_i = \sum_{j=1}^M \xi_j a\left(\varphi_j, \sum_{i=1}^M \xi_i \varphi_i\right) \\ &= a\left(\sum_{j=1}^M \xi_j \varphi_j, \sum_{i=1}^M \xi_i \varphi_i\right) = a(u,u) > 0 \end{aligned} \tag{2.36}$$

for $\boldsymbol{\xi} \neq 0$ and therefore $u \not\equiv 0$.
Here we have used only the positive definiteness of a.

Thus we have proven the following lemma.

Lemma 2.11 *The Galerkin method* (2.23) *has a unique solution if a is a symmetric, positive definite bilinear form and if b is a linear form.*

In fact, as we will see in Theorem 3.1, the symmetry of a is not necessary.

- For a special basis (i.e., for a specific finite element method), A_h is a sparse matrix, i.e., only a few entries $(A_h)_{ij}$ do not vanish. Evidently,

$$(A_h)_{ij} \neq 0 \quad \Leftrightarrow \quad \int_\Omega \nabla\varphi_j \cdot \nabla\varphi_i \, dx \neq 0 \,.$$

This can happen only if $\operatorname{supp}\varphi_i \cap \operatorname{supp}\varphi_j \neq \emptyset$, as this property is again necessary for $\operatorname{supp}\nabla\varphi_i \cap \operatorname{supp}\nabla\varphi_j \neq \emptyset$ because of

$$(\operatorname{supp}\nabla\varphi_i \cap \operatorname{supp}\nabla\varphi_j) \subset (\operatorname{supp}\varphi_i \cap \operatorname{supp}\varphi_j) \,.$$

The basis function φ_i vanishes on an element that does not contain the node a_i because of the uniqueness of the solution of the local interpolation problem. Therefore,

$$\operatorname{supp}\varphi_i = \bigcup_{\substack{K \in \mathcal{T}_h \\ a_i \in K}} K \,,$$

cf. Figure (2.8), and thus

$$(A_h)_{ij} \neq 0 \quad \Rightarrow \quad a_i, a_j \in K \text{ for some } K \in \mathcal{T}_h \,; \tag{2.37}$$

i.e., a_i, a_j are *neighbouring* nodes.

If we use the piecewise linear ansatz space on triangles and if a_i is an interior node in which L elements meet, then there exist at most L nondiagonal entries in the ith row of A_h. This number is determined only by the type of the triangulation, and it is independent of the fineness h, i.e., of the number of unknowns of the system of equations.

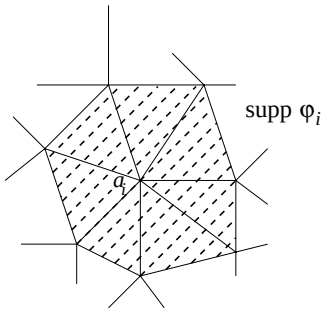


Figure 2.8. Support of the nodal basis function.

Example 2.12 We consider again the boundary value problem (2.1), (2.2) on $\Omega = (0, a) \times (0, b)$ again, i.e.

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega\,, \\ u &= 0 \quad \text{on } \partial\Omega\,, \end{aligned}$$

under the condition (1.4). The triangulation on which the method is based is created by a partition of Ω into squares with edges of length h and by a subsequent uniform division of each square into two triangles according to a fixed rule (*Friedrichs–Keller triangulation*). In order to do this, two possibilities (a) and (b) (see Figures 2.9 and 2.10) exist.

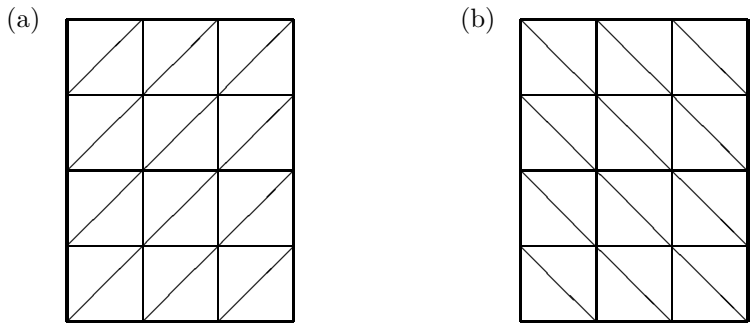


Figure 2.9. Possibilities of Friedrichs–Keller triangulation.

In both cases, a node a_Z belongs to six elements, and consequently, it has at most six neighbours:

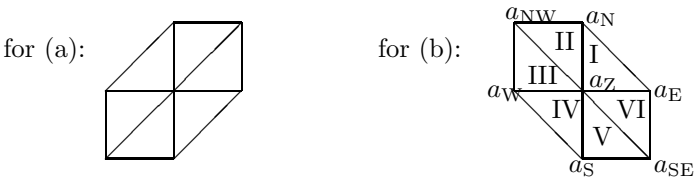


Figure 2.10. Support of the basis function.

Case (a) becomes case (b) by the transformation $x \mapsto a - x, y \mapsto y$. This transformation leaves the differential equation or the weak formulation, respectively, unchanged. Thus the Galerkin method with the ansatz space V_h according to (2.27) does not change, because $\mathcal{P}_1$ is invariant with respect to the above transformation. Therefore, the discretization matrices A_h according to (2.34) are seen to be identical by taking into account the renumbering of the nodes by the transformation.

Thus it is sufficient to consider only one case, say (b). A node which is far away from the boundary has 6 neighbouring nodes in $\{a_1, \ldots, a_{M_1}\}$, a node close to the boundary has less. The entries of the matrix in the row corresponding to a_{Z} depend on the derivatives of the basis function φ_{Z} as well as on the derivatives of the basis functions corresponding to the neighbouring nodes. The values of the partial derivatives of φ_{Z} in elements having the common vertex a_{Z} are listed in Table 2.1, where these elements are numbered according to Figure 2.10.

	I	II	III	IV	V	VI
$\partial_1\varphi_{\mathrm{Z}}$	$-\frac{1}{h}$	0	$\frac{1}{h}$	$\frac{1}{h}$	0	$-\frac{1}{h}$
$\partial_2\varphi_{\mathrm{Z}}$	$-\frac{1}{h}$	$-\frac{1}{h}$	0	$\frac{1}{h}$	$\frac{1}{h}$	0

Table 2.1. Derivatives of the basis functions.

Thus for the entries of the matrix in the row corresponding to a_{Z} we have

$$(A_h)_{\mathrm{Z,Z}} = a(\varphi_{\mathrm{Z}}, \varphi_{\mathrm{Z}}) = \int_{\mathrm{I}\cup\ldots\cup\mathrm{VI}} |\nabla\varphi_{\mathrm{Z}}|^2 \, dx = 2\int_{\mathrm{I}\cup\mathrm{II}\cup\mathrm{III}} \left[(\partial_1\varphi_{\mathrm{Z}})^2 + (\partial_2\varphi_{\mathrm{Z}})^2 \right] dx,$$

because the integrands are equal on I and IV, on II and V, and on III and VI. Therefore

$$(A_h)_{\mathrm{Z,Z}} = 2\int_{\mathrm{I}\cup\mathrm{III}} (\partial_1\varphi_{\mathrm{Z}})^2 \, dx + 2\int_{\mathrm{I}\cup\mathrm{II}} (\partial_2\varphi_{\mathrm{Z}})^2 \, dx = 2h^{-2}h^2 + 2h^{-2}h^2 = 4\,,$$

$$\begin{aligned}(A_h)_{\mathrm{Z,N}} = a(\varphi_{\mathrm{N}}, \varphi_{\mathrm{Z}}) &= \int_{\mathrm{I}\cup\mathrm{II}} \nabla\varphi_{\mathrm{N}} \cdot \nabla\varphi_{\mathrm{Z}} \, dx \\ &= \int_{\mathrm{I}\cup\mathrm{II}} \partial_2\varphi_{\mathrm{N}}\partial_2\varphi_{\mathrm{Z}} \, dx = \int_{\mathrm{I}\cup\mathrm{II}} \left(-h^{-1}\right) h^{-1} dx = -1\,,\end{aligned}$$

because $\partial_1 \varphi_{\mathrm{Z}} = 0$ on II and $\partial_1 \varphi_{\mathrm{N}} = 0$ on I. The element I for φ_{N} corresponds to the element V for φ_{Z}; i.e., $\partial_1 \varphi_{\mathrm{N}} = 0$ on I, analogously, it follows that $\partial_2 \varphi_{\mathrm{N}} = h^{-1}$ on I $\cup$ II. In the same way we get

$$(A_h)_{\mathrm{Z,E}} = (A_h)_{\mathrm{Z,W}} = (A_h)_{\mathrm{Z,S}} = -1$$

as well as

$$(A_h)_{\mathrm{Z,NW}} = a\,(\varphi_{\mathrm{NW}}, \varphi_{\mathrm{Z}}) = \int_{\mathrm{II}\cup\mathrm{III}} \partial_1 \varphi_{\mathrm{NW}}\, \partial_1 \varphi_{\mathrm{Z}} + \partial_2 \varphi_{\mathrm{NW}}\, \partial_2 \varphi_{\mathrm{Z}}\, dx = 0\,.$$

The last identity is due to $\partial_1 \varphi_{\mathrm{NW}} = 0$ on III and $\partial_2 \varphi_{\mathrm{NW}} = 0$ on III, because the elements V and VI for φ_{Z} agree with the elements III and II for φ_{NW}, respectively.

Analogously, we obtain for the remaining value

$$(A_h)_{\mathrm{Z,SE}} = 0\,,$$

such that only 5 (instead of the maximum 7) nonzero entries per row exist.

The way of assembling the stiffness matrix described above is called *node-based* assembling. However, most of the computer programs implementing the finite element method use an *element-based* assembling, which will be considered in Section 2.4.

If the nodes are numbered rowwise analogously to (1.13) and if the equations are divided by h^2, then $h^{-2} A_h$ coincides with the discretization matrix (1.14), which is known from the finite difference method. But here the right-hand side is given by

$$h^{-2}\,(\mathbf{q}_h)_i = h^{-2} \int_\Omega f \varphi_i\, dx = h^{-2} \int_{\mathrm{I}\cup\ldots\cup\mathrm{VI}} f \varphi_i\, dx$$

for $a_{\mathrm{Z}} = a_i$ and thus it is not identical to $f(a_i)$, the right-hand side of the finite difference method.

However, if the *trapezoidal rule*, which is exact for $g \in \mathcal{P}_1$, is applied to approximate the right-hand side according to

$$\int_K g(x)\, dx \approx \frac{1}{3} \mathrm{vol}\,(K) \sum_{i=1}^{3} g(a_i) \tag{2.38}$$

for a triangle K with the vertices a_i, $i = 1, 2, 3$ and with the area $\mathrm{vol}\,(K)$, then

$$\int_{\mathrm{I}} f \varphi_i\, dx \approx \frac{1}{3}\frac{1}{2} h^2\,(f(a_{\mathrm{Z}}) \cdot 1 + f(a_{\mathrm{O}}) \cdot 0 + f(a_{\mathrm{N}}) \cdot 0) = \frac{1}{6} h^2 f(a_{\mathrm{Z}}).$$

Analogous results are obtained for the other triangles, and thus

$$h^{-2} \int_{\mathrm{I}\cup\ldots\cup\mathrm{VI}} f \varphi_i\, dx \approx f(a_{\mathrm{Z}})\,.$$

In summary, we have the following result.

Lemma 2.13 *The finite element method with linear finite elements on a triangulation according to Figure* 2.9 *and with the trapezoidal rule to approximate the right-hand side yields the same discretization as the finite difference method from* (1.7), (1.8).

We now return to the general formulation (2.21)–(2.24). The approach of the Ritz method (2.24), instead of the Galerkin method (2.23), yields an identical approximation because of the following result.

Lemma 2.14 *If a is a symmetric and positive bilinear form and b is a linear form, then the Galerkin method* (2.23) *and the Ritz method* (2.24) *have identical solutions.*

Proof: Apply Lemma 2.3 with V_h instead of V. □

Hence the finite element method is the Galerkin method (and in our problem the Ritz method, too) for an *ansatz space* V_h with the following properties:

- The coefficients have a local interpretation (here as nodal values).

The basis functions have a small support such that:

- the discretization matrix is sparse,
- the entries of the matrix can be assembled locally.

Finally, for the boundary value problem (2.1), (2.2) with the corresponding weak formulation, we consider other ansatz spaces, which to some extent do not have these properties:

(1) In Section 3.2.1, (3.28), we will show that mixed boundary conditions need not be included in the ansatz space. Then we can choose the finite dimensional polynomial space $V_h = \operatorname{span}\left\{1, x, y, xy, x^2, y^2, \ldots\right\}$ for it. But in this case, A_h is a dense matrix and ill-conditioned. Such ansatz spaces yield the *classical* Ritz–Galerkin methods.

(2) Let $V_h = \operatorname{span}\{\varphi_1, \ldots, \varphi_N\}$ and let $\varphi_i \not\equiv 0$ satisfy, for some λ_i,

$$a(\varphi_i, v) = \lambda_i \langle \varphi_i, v \rangle_0 \quad \text{for all } v \in V\,,$$

i.e., the weak formulation of the eigenvalue problem

$$\begin{aligned} -\Delta u &= \lambda u && \text{in } \Omega\,, \\ u &= 0 && \text{on } \partial\Omega\,, \end{aligned}$$

for which eigenvalues $0 < \lambda_1 \leq \lambda_2 \leq \ldots$ and corresponding eigenfunctions φ_i exist such that $\langle \varphi_i, \varphi_j \rangle_0 = \delta_{ij}$ (e.g., see [12, p. 335]). For special domains Ω, (λ_i, φ_i) can be determined explicitly, and

$$(A_h)_{ij} = a(\varphi_j, \varphi_i) = \lambda_j \langle \varphi_j, \varphi_i \rangle_0 = \lambda_j \delta_{ij}$$

is obtained. Thus A_h is a diagonal matrix, and the system of equations $A_h \boldsymbol{\xi} = \boldsymbol{q}_h$ can be solved without too great expense. But this kind of assembling is possible with acceptable costs for special cases only.

(3) The (spectral) *collocation method* consists in the requirement that the equations (2.1), (2.2) be satisfied only at certain distinct points $x_i \in \overline{\Omega}$, called *collocation points,* for a special polynomial space V_h.

The above examples describe Galerkin methods without having the typical properties of a finite element method.

2.3 Stability and Convergence of the Finite Element Method

We consider the general case of a variational equation of the form (2.21) and the Galerkin method (2.23). Here let a be a bilinear form, which is not necessarily symmetric, and let b be a linear form.

Then, if

$$e := u - u_h \, (\in V)$$

denotes the error, the important *error equation*

$$a(e, v) = 0 \quad \text{for all } v \in V_h \tag{2.39}$$

is satisfied. To obtain this equation, it is sufficient to consider equation (2.21) only for $v \in V_h \subset V$ and then to subtract from the result the Galerkin equation (2.23).

If, in addition, a is symmetric and positive definite, i.e.,

$$a(u, v) = a(v, u)\,, \quad a(u, u) \geq 0\,, \quad a(u, u) = 0 \Leftrightarrow u = 0$$

(i.e., a is a scalar product), then the error is orthogonal to the space V_h with respect to the scalar product a.

Therefore, the relation (2.39) is often called the *orthogonality of the error (to the ansatz space).* In general, the element $u_h \in V_h$ with minimal distance to $u \in V$ with respect to the induced norm $\|\cdot\|_a$ is characterized by (2.39):

Lemma 2.15 *Let $V_h \subset V$ be a subspace, let a be a scalar product on V, and let $\|u\|_a := a(u,u)^{1/2}$ be the norm induced by a. Then for $u_h \in V_h$, it follows that*

$$a(u - u_h, v) = 0 \quad \textit{for all } v \in V_h \quad \Leftrightarrow \tag{2.40}$$

$$\|u - u_h\|_a = \min \left\{ \|u - v\|_a \;\middle|\; v \in V_h \right\} . \tag{2.41}$$

Proof: For arbitrary but fixed $u \in V$, let $b(v) := a(u, v)$ for $v \in V_h$. Then b is a linear form on V_h, so (2.40) is a variational formulation on V_h.

According to Lemma 2.14 or Lemma 2.3, this variational formulation has the same solutions as

$$\begin{aligned} F(u_h) &= \min\left\{F(v) \;\middle|\; v \in V_h\right\} \\ \text{with} \quad F(v) &:= \frac{1}{2}a(v,v) - b(v) = \frac{1}{2}a(v,v) - a(u,v)\,. \end{aligned}$$

Furthermore, F has the same minima as the functional

$$\begin{aligned} \Big(2F(v) + a(u,u)\Big)^{1/2} &= \Big(a(v,v) - 2a(u,v) + a(u,u)\Big)^{1/2} \\ &= \Big(a(u-v,u-v)\Big)^{1/2} = \|u-v\|_a\,, \end{aligned}$$

because the additional term $a(u,u)$ is a constant. Therefore, F has the same minima as (2.41). □

If an approximation u_h of u is to be sought exclusively in V_h, then the element u_h, determined by the Galerkin method, is the optimal choice with respect to $\|\cdot\|_a$.

A general, not necessarily symmetric, bilinear form a is assumed to satisfy the following conditions, where $\|\cdot\|$ denotes a norm on V:

- a is *continuous* with respect to $\|\cdot\|$; i.e., there exists $M > 0$ such that

$$|a(u,v)| \le M\|u\|\|v\| \quad \text{for all } u, v \in V\,; \tag{2.42}$$

- a is *V-elliptic*; i.e., there exists $\alpha > 0$ such that

$$a(u,u) \ge \alpha\|u\|^2 \quad \text{for } u \in V\,. \tag{2.43}$$

If a is a scalar product, then (2.42) with $M = 1$ and (2.43) (as equality) with $\alpha = 1$ are valid for the induced norm $\|\cdot\| := \|\cdot\|_a$ due to the Cauchy–Schwarz inequality.

The V-ellipticity is an essential condition for the unique existence of a solution of the variational equation (2.21) and of the boundary value problem described by it, which will be presented in more detail in Sections 3.1 and 3.2. It also implies — without further conditions — the stability of the Galerkin approximation.

Lemma 2.16 *The Galerkin solution u_h according to* (2.23) *is* stable *in the following sense:*

$$\|u_h\| \le \frac{1}{\alpha}\|b\| \quad \textit{independently of } h\,, \tag{2.44}$$

where

$$\|b\| := \sup\left\{\frac{|b(v)|}{\|v\|} \;\middle|\; v \in V\,,\ v \neq 0\right\}.$$

Proof: In the case $u_h = 0$, there is nothing to prove. Otherwise, from $a(u_h, v) = b(v)$ for all $v \in V_h$, it follows that

$$\alpha \|u_h\|^2 \leq a(u_h, u_h) = b(u_h) \leq \frac{|b(u_h)|}{\|u_h\|} \|u_h\| \leq \|b\| \, \|u_h\|.$$

Dividing this relation by $\alpha \|u_h\|$, we get the assertion. □

Moreover, the approximation property (2.41) holds up to a constant:

Theorem 2.17 (Céa's lemma)
Assume (2.42), (2.43). *Then the following error estimate for the Galerkin solution holds:*

$$\|u - u_h\| \leq \frac{M}{\alpha} \min \left\{ \|u - v\| \;\middle|\; v \in V_h \right\}. \tag{2.45}$$

Proof: If $\|u - u_h\| = 0$, then there is nothing to prove. Otherwise, let $v \in V_h$ be arbitrary. Because of the error equation (2.39) and $u_h - v \in V_h$,

$$a(u - u_h, u_h - v) = 0\,.$$

Therefore, using (2.43) we have

$$\begin{aligned} \alpha \|u - u_h\|^2 &\leq a(u - u_h, u - u_h) = a(u - u_h, u - u_h) + a(u - u_h, u_h - v) \\ &= a(u - u_h, u - v)\,. \end{aligned}$$

Furthermore, by means of (2.42) we obtain

$$\alpha \|u - u_h\|^2 \leq a(u - u_h, u - v) \leq M \|u - u_h\| \, \|u - v\| \text{ for arbitrary } v \in V_h\,.$$

Thus the assertion follows by division by $\alpha \|u - u_h\|$. □

Therefore also in general, in order to get an asymptotic error estimate in h, it is sufficient to estimate the *best approximation error* of V_h, i.e.,

$$\min \left\{ \|u - v\| \;\middle|\; v \in V_h \right\}.$$

However, this consideration is meaningful only in those cases where M/α is not too large. Section 3.2 shows that this condition is no longer satisfied for convection-dominated problems. Therefore, the Galerkin approach has to be modified, which will be described in Chapter 9.

We want to apply the theory developed up to now to the weak formulation of the boundary value problem (2.1), (2.2) with V according to (2.7) or (2.20) and V_h according to (2.27). According to (2.4) the bilinear form a and the linear form b read as

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, dx\,, \quad b(v) = \int_\Omega f v \, dx\,.$$

In order to guarantee that the linear form b is well-defined on V, it is sufficient to assume that the right-hand side f of the boundary value problem belongs to $L^2(\Omega)$.

Since a is a scalar product on V,

$$\|u\| = \|u\|_a = \left(\int_\Omega |\nabla u|^2 \, dx \right)^{1/2}$$

is an appropriate norm. Alternatively, the norm introduced in (2.19) for $V = H_0^1(\Omega)$ can be taken as

$$\|u\|_1 = \left(\int_\Omega |u(x)|^2 \, dx + \int_\Omega |\nabla u(x)|^2 \, dx \right)^{1/2} .$$

In the latter case, the question arises whether the conditions (2.42) and (2.43) are still satisfied. Indeed,

$$|a(u,v)| \le \|u\|_a \|v\|_a \le \|u\|_1 \|v\|_1 \quad \text{for all } u, v \in V .$$

The first inequality follows from the Cauchy–Schwarz inequality for the scalar product a, and the second inequality follows from the trivial estimate

$$\|u\|_a = \left(\int_\Omega |\nabla u(x)|^2 \, dx \right)^{1/2} \le \|u\|_1 \quad \text{for all } u \in V .$$

Thus a is continuous with respect to $\| \cdot \|_1$ with $M = 1$.

The V-ellipticity of a, i.e., the property

$$a(u,u) = \|u\|_a^2 \ge \alpha \|u\|_1^2 \quad \text{for some } \alpha > 0 \text{ and all } u \in V ,$$

is not valid in general for $V = H^1(\Omega)$. However, in the present situation of $V = H_0^1(\Omega)$ it is valid because of the incorporation of the boundary condition into the definition of V :

Theorem 2.18 (Poincaré) *Let $\Omega \subset \mathbb{R}^n$ be open and bounded. Then a constant $C > 0$ exists (depending on Ω) such that*

$$\|u\|_0 \le C \left(\int_\Omega |\nabla u(x)|^2 \, dx \right)^{1/2} \quad \textit{for all } u \in H_0^1(\Omega) .$$

Proof: Cf. [13]. For a special case, see Exercise 2.5. □

Thus (2.43) is satisfied, for instance with

$$\alpha = \frac{1}{1 + C^2} ,$$

(see also (3.26) below) and thus in particular

$$\alpha \|u\|_1^2 \le a(u,u) = \|u\|_a^2 \le \|u\|_1^2 \quad \text{for all } u \in V , \tag{2.46}$$

i.e., the norms $\|\cdot\|_1$ and $\|\cdot\|_a$ are equivalent on $V = H_0^1(\Omega)$ and therefore they generate the same convergence concept:

$$u_h \to u \text{ with respect to } \|\cdot\|_1 \Leftrightarrow \|u_h - u\|_1 \to 0$$

$$\Leftrightarrow \|u_h - u\|_a \to 0 \Leftrightarrow u_h \to u \text{ with respect to } \|\cdot\|_a .$$

In summary the estimate (2.45) holds for $\|\cdot\| = \|\cdot\|_1$ with the constant $1/\alpha$.

Because of the Cauchy–Schwarz inequality for the scalar product on $L^2(\Omega)$ and

$$b(v) = \int_\Omega f(x)v(x)\,dx ,$$

i.e., $|b(v)| \le \|f\|_0 \|v\|_0 \le \|f\|_0 \|v\|_1$, and thus $\|b\| \le \|f\|_0$, the stability estimate (2.44) for a right-hand side $f \in L^2(\Omega)$ takes the particular form

$$\|u_h\|_1 \le \frac{1}{\alpha}\|f\|_0 .$$

Up to now, our considerations have been independent of the special form of V_h. Now we make use of the choice of V_h according to (2.27). In order to obtain an estimate of the approximation error of V_h, it is sufficient to estimate the term $\|u - \bar{v}\|$ for some special element $\bar{v} \in V_h$. For this element $\bar{v} \in V_h$, we choose the interpolant $I_h(u)$, where

$$\begin{aligned} I_h : \left\{ u \in C(\bar{\Omega}) \;\middle|\; u = 0 \text{ on } \partial\Omega \right\} &\to V_h , \\ u &\mapsto I_h(u) \text{ with } I_h(u)(a_i) = u(a_i) . \end{aligned} \tag{2.47}$$

This interpolant exists and is unique (Lemma 2.10). Obviously,

$$\min\left\{ \|u - v\|_1 \;\middle|\; v \in V_h \right\} \le \|u - I_h(u)\|_1 \quad \text{for } u \in C(\bar{\Omega}) \text{ and } u = 0 \text{ on } \partial\Omega .$$

If the weak solution u possesses weak derivatives of second order, then for certain sufficiently fine triangulations $\mathcal{T}_h$, i.e., $0 < h \le \bar{h}$ for some $\bar{h} > 0$, an estimate of the type

$$\|u - I_h(u)\|_1 \le Ch \tag{2.48}$$

holds, where C depends on u but is independent of h (cf. (3.88)). The proof of this estimate will be explained in Section 3.4, where also sufficient conditions on the family of triangulations $(\mathcal{T}_h)_h$ will be specified.

Exercises

2.4 Let $a(u,v) := \int_0^1 x^2 u'v'dx$ for arbitrary $u, v \in H_0^1(0,1)$.

(a) Show that there is no constant $C_1 > 0$ such that the inequality

$$a(u,u) \ge C_1 \int_0^1 (u')^2\,dx \quad \text{for all } u \in H_0^1(0,1)$$

is valid.

(b) Now let $\mathcal{T}_h := \{(x_{i-1}, x_i)\}_{i=1}^N$, $N \in \mathbb{N}$, be an equidistant partition of $(0,1)$ with the parameter $h = 1/N$ and $V_h := \operatorname{span}\{\varphi_i\}_{i=1}^{N-1}$, where

$$\varphi_i(x) := \begin{cases} (x - x_{i-1})/h & \text{in } (x_{i-1}, x_i)\,, \\ (x_{i+1} - x)/h & \text{in } (x_i, x_{i+1})\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

Does there exist a constant $C_2 > 0$ with

$$a(u_h, u_h) \geq C_2 \int_0^1 (u_h')^2 \, dx \quad \text{for all } u_h \in V_h\,?$$

2.5

(a) For $\Omega := (\alpha, \beta) \times (\gamma, \delta)$ and V according to (2.7), prove the *inequality of Poincaré:* There exists a positive constant C with

$$\|u\|_0 \leq C \|u\|_a \quad \text{for all } u \in V\,.$$

Hint: Start with the relation $u(x,y) = \int_\alpha^x \partial_x u(s,y)\, ds$.

(b) For $\Omega := (\alpha, \beta)$ and $v \in C([\alpha, \beta])$ with a piecewise continuous derivative v' and $v(\gamma) = 0$ for some $\gamma \in [\alpha, \beta]$, show that

$$\|v\|_0 \leq (\beta - \alpha) \|v'\|_0\,.$$

2.6 Let $\Omega := (0,1) \times (0,1)$. Given $f \in C(\overline{\Omega})$, discretize the boundary value problem $-\Delta u = f$ in Ω, $u = 0$ on $\partial\Omega$, by means of the usual five-point difference stencil as well as by means of the finite element method with linear elements. A quadratic grid as well as the corresponding Friedrichs–Keller triangulation will be used.

Prove the following stability estimates for the matrix of the linear system of equations:

$$\text{(a) } \|A_h^{-1}\|_\infty \leq \frac{1}{8}\,, \quad \text{(b) } \|A_h^{-1}\|_2 \leq \frac{1}{16}\,, \quad \text{(c) } \|A_h^{-1}\|_0 \leq 1\,,$$

where $\|\cdot\|_\infty, \|\cdot\|_2$ denote the maximum row sum norm and the spectral norm of a matrix, respectively, and $\|A_h^{-1}\|_0 := \sup_{v_h \in V_h} \|v_h\|_0^2 / \|v_h\|_a^2$ with $\|v_h\|_a^2 := \int_\Omega |\nabla v_h|^2\, dx$.

Comment: The constant in (c) is not optimal.

2.7 Let Ω be a domain with polygonal boundary and let $\mathcal{T}_h$ be a conforming triangulation of Ω. The nodes a_i of the triangulation are enumerated from 1 to M.

Let the triangulation satisfy the following assumption: There exist constants $C_1, C_2 > 0$ such that for all triangles $K \in \mathcal{T}_h$ the relation

$$C_1 h^2 \leq \operatorname{vol}(K) \leq C_2 h^2$$

is satisfied. h denotes the maximum of the diameters of all elements of $\mathcal{T}_h$.

(a) Show the equivalence of the following norms for $u_h \in V_h$ in the space V_h of continuous, piecewise linear functions over Ω :

$$\|u_h\|_0 := \left\{ \int_\Omega |u_h|^2 \, dx \right\}^{1/2}, \quad \|u_h\|_{0,h} := h \left\{ \sum_{i=1}^{M} u_h^2(a_i) \right\}^{1/2}.$$

(b) Consider the special case $\Omega := (0,1)\times(0,1)$ with the Friedrichs–Keller triangulation as well as the subspace $V_h \cap H_0^1(\Omega)$ and find "as good as possible" constants in the corresponding equivalence estimate.

2.4 The Implementation of the Finite Element Method: Part 1

In this section we will consider some aspects of the implementation of the finite element method using linear ansatz functions on triangles for the model boundary value problem (1.1), (1.2) on a polygonally bounded domain $\Omega \subset \mathbb{R}^2$. The case of inhomogeneous Dirichlet boundary conditions will be treated also to a certain extent as far as it is possible up to now.

2.4.1 Preprocessor

The main task of the preprocessor is to determine the triangulation.

An input file might have the following format:

Let the number of variables (including also the boundary nodes for Dirichlet boundary conditions) be M. We generate the following list:

x-coordinate of node 1	y-coordinate of node 1
...	...
x-coordinate of node M	y-coordinate of node M

Let the number of (triangular) elements be N. These elements will be listed in the *element-node table*. Here, every element is characterized by the indices of the nodes corresponding to this element in a well-defined order (e.g., counterclockwise); cf. Figure 2.11.

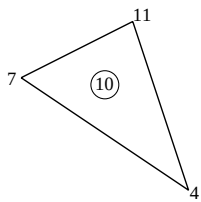


Figure 2.11. Element no. 10 with nodes nos. 4, 11, 7.

For example, the 10th row of the element-node table contains the entry

4 11 7

Usually, a triangulation is generated by a triangulation algorithm. A short overview on methods for the grid generation will be given in Section 4.1. One of the simplest versions of a grid generation algorithm has the following structure (cf. Figure 2.12):

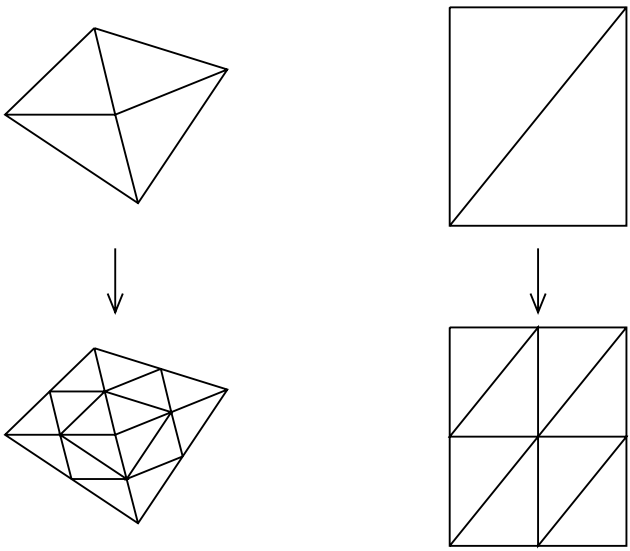


Figure 2.12. Refinement by quartering.

Prescribe a coarse triangulation (according to the above format) and refine this triangulation (repeatedly) by subdividing a triangle into 4 congruent triangles by connecting the midpoints of the edges with straight lines.

If this uniform refinement is done globally, i.e., for all triangles of the coarse grid, then triangles are created that have the same interior angles as the elements of the coarse triangulation. Thus the quality of the triangulation, indicated, for example, by the ratios of the diameters of an element and of its inscribed circle (see Definition 3.28), does not change. However, if the subdivision is performed only locally, the resulting triangulation is no longer admissible, in general. Such an inadmissible triangulation can be corrected by bisection of the corresponding neighbouring (unrefined) triangles. But this implies that some of the interior angles are bisected and consequently, the quality of the triangulation becomes poorer if the bisection step is performed too frequently. The following algorithm circumvents the depicted problem. It is due to R. Bank and is implemented, for example, in the PLTMG code (see [4]).

A Possible Refinement Algorithm
Let a (uniform) triangulation $\mathcal{T}$ be given (e.g., by repeated uniform refinement of a coarse triangulation). The edges of this triangulation are called *red edges*.

(1) Subdivide the edges according to a certain local refinement criterion (introduction of new nodes) by successive bisection (cf. Figure 2.13).

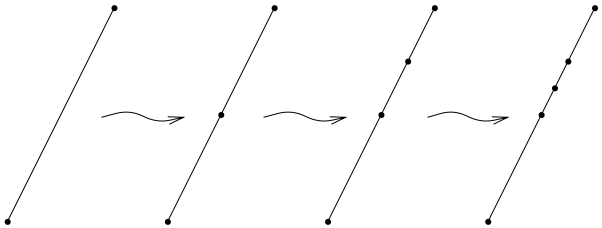


Figure 2.13. New nodes on edges.

(2) If a triangle $K \in \mathcal{T}$ has on its edges in addition to the vertices two or more nodes, then subdivide K into four congruent triangles. Iterate over step 2 (cf. Figure 2.14).

(3) Subdivide the triangles with nodes at the midpoints of the edges into 2 triangles by bisection. This step introduces the so-called *green edges*.

(4) If the refinement is to be continued, first remove the green edges.

2.4.2 Assembling

Denote by $\varphi_1, \ldots, \varphi_M$ the global basis functions. Then the stiffness matrix A_h has the following entries:

$$(A_h)_{ij} = \int_\Omega \nabla\varphi_j \cdot \nabla\varphi_i \, dx = \sum_{m=1}^{N} A_{ij}^{(m)}$$

with

$$A_{ij}^{(m)} = \int_{K_m} \nabla\varphi_j \cdot \nabla\varphi_i \, dx \,.$$

Let $a_1, \ldots, a_M$ denote the nodes of the triangulation. Because of the implication

$$A_{ij}^{(m)} \neq 0 \;\Rightarrow\; a_i, a_j \in K_m$$

(cf. (2.37)), the element K_m yields nonzero contributions for $A_{ij}^{(m)}$ only if $a_i, a_j \in K_m$ at best. Such nonzero contributions are called *element entries* of A_h. They add up to the *entries* of A_h.

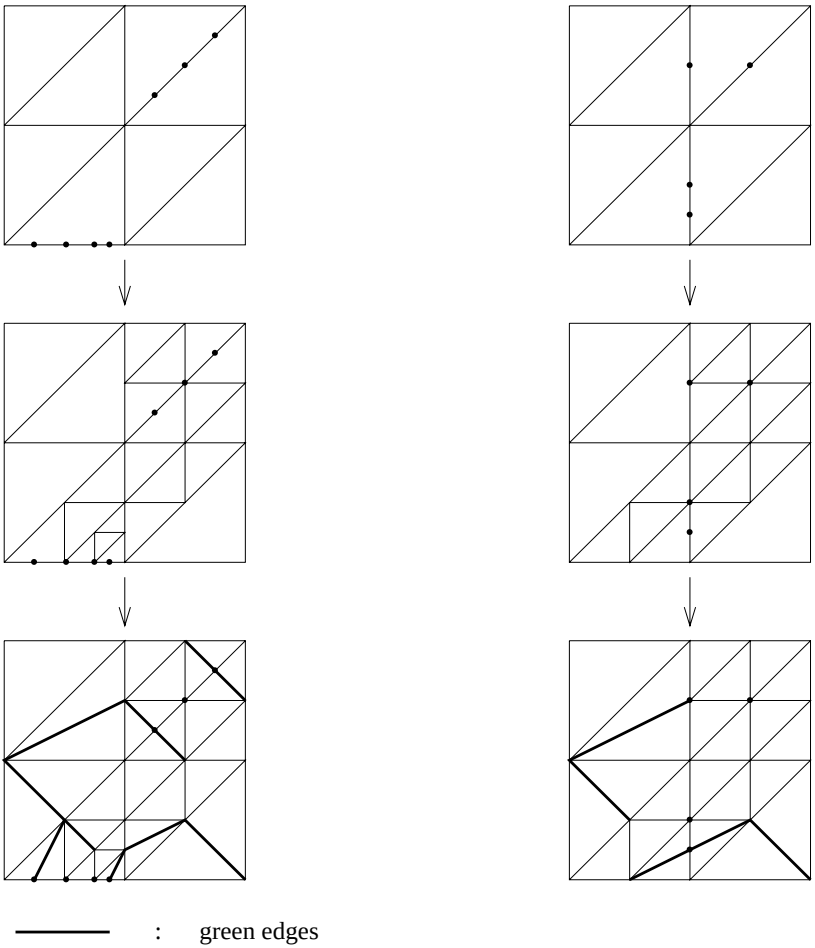


Figure 2.14. Two refinement sequences.

In Example 2.12 we explained a node-based assembling of the stiffness matrix. In contrast to this and on the basis of the above observations, in the following we will perform an *element-based assembling* of the stiffness matrix.

To assemble the entries of $A^{(m)}$, we will start from a local numbering (cf. Figure 2.15) of the nodes by assigning the local numbers 1, 2, 3 to the global node numbers r_1, r_2, r_3 (numbered counterclockwise). In contrast to the usual notation adopted in this book, here indices of vectors according to the local numbering are included in parentheses and written as superscripts.

Figure 2.15. Global (left) and local numbering.

Thus in fact, we generate

$$\left(A^{(m)}_{r_i r_j}\right)_{i,j=1,2,3} \quad \text{as} \quad \left(\tilde{A}^{(m)}_{ij}\right)_{i,j=1,2,3}.$$

To do this, we first perform a transformation of K_m onto some reference element and then we evaluate the integral on this element exactly.

Hence the entry of the *element stiffness matrix* reads as

$$\tilde{A}^{(m)}_{ij} = \int_{K_m} \nabla\varphi_{r_j} \cdot \nabla\varphi_{r_i}\, dx\,.$$

The reference element $\hat{K}$ is transformed onto the global element K_m by means of the relation $F(\hat{x}) = B\hat{x} + d$, therefore

$$D_{\hat{x}} u(F(\hat{x})) = D_x u(F(\hat{x}))\, D_{\hat{x}} F(\hat{x}) = D_x u(F(\hat{x}))\, B\,,$$

where $D_x u$ denotes the row vector $(\partial_1 u, \partial_2 u)$, i.e., the corresponding differential operator. Using the more standard notation in terms of gradients and taking into consideration the relation $B^{-T} := (B^{-1})^T$, we obtain

$$\nabla_x u\,(F(\hat{x})) = B^{-T} \nabla_{\hat{x}}\,(u\,(F(\hat{x}))) \tag{2.49}$$

and thus

$$\begin{aligned}
\tilde{A}^{(m)}_{ij} &= \int_{\hat{K}} \nabla_x \varphi_{r_j}\,(F(\hat{x})) \cdot \nabla_x \varphi_{r_i}\,(F(\hat{x}))\, |\det(DF(\hat{x}))|\; d\hat{x} \\
&= \int_{\hat{K}} B^{-T} \nabla_{\hat{x}} \left(\varphi_{r_j}\,(F(\hat{x}))\right) \cdot B^{-T} \nabla_{\hat{x}}\,(\varphi_{r_i}\,(F(\hat{x})))\, |\det(B)|\; d\hat{x} \\
&= \int_{\hat{K}} B^{-T} \nabla_{\hat{x}} \hat{\varphi}_{r_j}(\hat{x}) \cdot B^{-T} \nabla_{\hat{x}} \hat{\varphi}_{r_i}(\hat{x})\, |\det(B)|\; d\hat{x} \qquad (2.50) \\
&= \int_{\hat{K}} B^{-T} \nabla_{\hat{x}} N_j(\hat{x}) \cdot B^{-T} \nabla_{\hat{x}} N_i(\hat{x})\, |\det(B)|\; d\hat{x}\,,
\end{aligned}$$

where the transformed basis functions $\hat{\varphi}_{r_i}$, $\hat{\varphi}(\hat{x}) := \varphi(F(\hat{x}))$ coincide with the local basis functions on $\hat{K}$, i.e., with the shape functions N_i:

$$\hat{\varphi}_{r_i}(\hat{x}) = N_i(\hat{x}) \quad \text{for } \hat{x} \in \hat{K}\,.$$

The shape functions N_i have been defined in (2.29) (where (x, y) there must be replaced by $(\hat{x}_1, \hat{x}_2)$ here) for the standard reference element defined there.

Introducing the matrix $C := \left(B^{-1}\right)\left(B^{-1}\right)^T = \left(B^T B\right)^{-1}$, we can write

$$\tilde{A}_{ij}^{(m)} = \int_{\hat{K}} C\,\nabla_{\hat{x}} N_j(\hat{x}) \cdot \nabla_{\hat{x}} N_i(\hat{x})\ |\det(B)|\ d\hat{x}\,. \tag{2.51}$$

Denoting the matrix B by $B = \left(b^{(1)}, b^{(2)}\right)$, then it follows that

$$C = \begin{pmatrix} b^{(1)}\cdot b^{(1)} & b^{(1)}\cdot b^{(2)} \\ b^{(1)}\cdot b^{(2)} & b^{(2)}\cdot b^{(2)} \end{pmatrix}^{-1} = \frac{1}{\det(B)^2}\begin{pmatrix} b^{(2)}\cdot b^{(2)} & -b^{(1)}\cdot b^{(2)} \\ -b^{(1)}\cdot b^{(2)} & b^{(1)}\cdot b^{(1)} \end{pmatrix}$$

because $\det(B^T B) = \det(B)^2$. The previous considerations can be easily extended to the computation of the stiffness matrices of more general differential operators like

$$\int_\Omega K(x)\nabla\varphi_j(x)\cdot\nabla\varphi_i(x)\,dx$$

(cf. Section 3.5). For the standard reference element, which we use from now on, we have $b^{(1)} = a^{(2)} - a^{(1)}$, $b^{(2)} = a^{(3)} - a^{(1)}$. Here $a^{(i)}$, $i = 1, 2, 3$, are the locally numbered nodes of K interpreted as vectors of $\mathbb{R}^2$.

From now on we make also use of the special form of the stiffness matrix and obtain

$$\begin{aligned} \tilde{A}_{ij}^{(m)} &= \gamma_1 \int_{\hat{K}} \partial_{\hat{x}_1} N_j\, \partial_{\hat{x}_1} N_i\, d\hat{x} \\ &\quad + \gamma_2 \int_{\hat{K}} \partial_{\hat{x}_1} N_j\, \partial_{\hat{x}_2} N_i + \partial_{\hat{x}_2} N_j\, \partial_{\hat{x}_1} N_i\, d\hat{x} \\ &\quad + \gamma_3 \int_{\hat{K}} \partial_{\hat{x}_2} N_j\, \partial_{\hat{x}_2} N_i\, d\hat{x} \end{aligned} \tag{2.52}$$

with

$$\begin{aligned} \gamma_1 &:= c_{11}|\det(B)| = \frac{1}{|\det(B)|}\left(a^{(3)} - a^{(1)}\right)\cdot\left(a^{(3)} - a^{(1)}\right), \\ \gamma_2 &:= c_{12}|\det(B)| = -\frac{1}{|\det(B)|}\left(a^{(2)} - a^{(1)}\right)\cdot\left(a^{(3)} - a^{(1)}\right), \\ \gamma_3 &:= c_{22}|\det(B)| = \frac{1}{|\det(B)|}\left(a^{(2)} - a^{(1)}\right)\cdot\left(a^{(2)} - a^{(1)}\right). \end{aligned}$$

In the implementation it is advisable to compute the values γ_i just once from the local geometrical information given in the form of the vertices $a^{(i)} = a_{r_i}$ and to store them permanently.

Thus we obtain for the local stiffness matrix

$$\tilde{A}^{(m)} = \gamma_1 S_1 + \gamma_2 S_2 + \gamma_3 S_3 \tag{2.53}$$

with

$$S_1 := \left(\int_{\hat{K}} \partial_{\hat{x}_1} N_j \partial_{\hat{x}_1} N_i\, d\hat{x}\right)_{ij},$$

$$S_2 := \left(\int_{\hat{K}} \partial_{\hat{x}_1} N_j \partial_{\hat{x}_2} N_i + \partial_{\hat{x}_2} N_j \partial_{\hat{x}_1} N_i \, d\hat{x} \right)_{ij},$$
$$S_3 := \left(\int_{\hat{K}} \partial_{\hat{x}_2} N_j \partial_{\hat{x}_2} N_i \, d\hat{x} \right)_{ij}.$$

An explicit computation of the matrices S_i is possible because the integrands are constant, and also these matrices can be stored permanently:

$$S_1 = \frac{1}{2}\begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \; S_2 = \frac{1}{2}\begin{pmatrix} 2 & -1 & -1 \\ -1 & 0 & 1 \\ -1 & 1 & 0 \end{pmatrix}, \; S_3 = \frac{1}{2}\begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix}.$$

The right-hand side $(\boldsymbol{q}_h)_i = \int_\Omega f(x)\varphi_i(x)\,dx$ can be treated in a similar manner:

$$(\boldsymbol{q}_h)_i = \sum_{m=1}^{N} \left(\boldsymbol{q}^{(m)}\right)_i$$

with

$$\left(\boldsymbol{q}^{(m)}\right)_i = \int_{K_m} f(x)\varphi_i(x)\,dx \quad (\neq 0 \;\Rightarrow\; a_i \in K_m).$$

Again, we transform the global numbering $\left(q_{r_i}^{(m)}\right)_{i=1,2,3}$ for the triangle $K_m = \operatorname{conv}\{a_{r_1}, a_{r_2}, a_{r_3}\}$ into the local numbering $\left(\tilde{q}_i^{(m)}\right)_{i=1,2,3}$. Analogously to the determination of the entries of the stiffness matrix, we have

$$\begin{aligned} \tilde{q}_i^{(m)} &= \int_{\hat{K}} f\left(F(\hat{x})\right) \varphi_{r_i}\left(F(\hat{x})\right) |\det(B)|\, d\hat{x} \\ &= \int_{\hat{K}} \hat{f}(\hat{x})\, N_i(\hat{x})\, |\det(B)|\, d\hat{x}, \end{aligned}$$

where $\hat{f}(\hat{x}) := f(F(\hat{x}))$ for $\hat{x} \in \hat{K}$.

In general, this integral cannot be evaluated exactly. Therefore, it has to be approximated by a quadrature rule.

A *quadrature rule* for $\int_{\hat{K}} g(\hat{x})\,d\hat{x}$ is of the type

$$\sum_{k=1}^{R} \omega_k \, g\left(\hat{b}^{(k)}\right)$$

with certain *weights* ω_k and *quadrature points* $\hat{b}^{(k)}$. As an example, we take the *trapezoidal rule* (cf. (2.38)), where

$$\hat{b}^{(1)} = \hat{a}_1 = (0,0), \quad \hat{b}^{(2)} = \hat{a}_2 = (1,0), \quad \hat{b}^{(3)} = \hat{a}_3 = (0,1),$$
$$\omega_k = \tfrac{1}{6}, \quad k = 1,2,3.$$

Thus for arbitrary but fixed quadrature rules, we have

$$\tilde{q}_i^{(m)} \approx \sum_{k=1}^{R} \omega_k \, \hat{f}\big(\hat{b}^{(k)}\big) \, N_i\big(\hat{b}^{(k)}\big) \, |\det(B)| \,. \tag{2.54}$$

Of course, the application of different quadrature rules on different elements is possible, too. The values $N_i\big(\hat{b}^{(k)}\big)$, $i = 1, 2, 3$, $k = 1, \ldots, R$, should be evaluated just once and should be stored. The discussion on the use of quadrature rules will be continued in Sections 3.5.2 and 3.6.

In summary, the following algorithm provides the assembling of the stiffness matrix and the right-hand side:

Loop over all elements $m = 1, \ldots, N$:

- Allocating a local numbering to the nodes based on the element-node table: $1 \mapsto r_1,\ 2 \mapsto r_2,\ 3 \mapsto r_3$.
- Assembling of the element stiffness matrix $\tilde{A}^{(m)}$ according to (2.51) or (2.53).
  Assembling of the right-hand side according to (2.54).
- Loop over $i, j = 1, 2, 3$:

$$\begin{aligned} (A_h)_{r_i r_j} &:= (A_h)_{r_i r_j} + \tilde{A}_{ij}^{(m)} \,, \\ (\boldsymbol{q}_h)_{r_i} &:= (\boldsymbol{q}_h)_{r_i} + \tilde{q}_i^{(m)} \,. \end{aligned}$$

For the sake of efficiency of this algorithm, it is necessary to adjust the memory structure to the particular situation; we will see how this can be done in Section 2.5.

2.4.3 Realization of Dirichlet Boundary Conditions: Part 1

Nodes where a Dirichlet boundary condition is prescribed must be labeled specially, here, for instance, by the convention $M = M_1 + M_2$, where the nodes numbered from $M_1 + 1$ to M correspond to the Dirichlet boundary nodes. In more general cases, other realizations are to be preferred.

In the first step of assembling of stiffness matrix and the load vector, the Dirichlet nodes are treated like all the other ones. After this, the Dirichlet nodes are considered separately. If such a node has the number j, the boundary condition is included by the following procedure:

Replace the jth row and the jth column (for conservation of the symmetry) of A_h by the jth unit vector and $(\boldsymbol{q}_h)_j$ by $g(a_j)$, if $u(x) = g(x)$ is prescribed for $x \in \partial\Omega$. If the jth column is replaced by the unit vector, the right-hand side $(\boldsymbol{q}_h)_i$ for $i \neq j$ must be modified to $(\boldsymbol{q}_h)_i - (A_h)_{ij} g(a_j)$. In other words, the contributions caused by the Dirichlet boundary condition are included into the right-hand side. This is exactly the elimination that led to the form (1.10), (1.11) in Chapter 1.

2.5 Solving Sparse Systems of Linear Equations by Direct Methods

Let A be an $M \times M$ matrix. Given a vector $\mathbf{q} \in \mathbb{R}^M$, we consider the system of linear equations

$$A\boldsymbol{\xi} = \mathbf{q}\,.$$

The matrices arising from the finite element discretization are *sparse*; i.e., they have a bounded number of nonzero entries per row independent of the dimension of the system of equations. For the simple example of Section 2.2, this bound is determined by the number of neighbouring nodes (see (2.37)). Methods for solving systems of equations should take advantage of the sparse structure. For iterative methods, which will be examined in Chapter 5, this is easier to reach than for direct methods. Therefore, the importance of direct methods has decreased. Nevertheless, in adapted form and for small or medium size problems, they are still the method of choice.

Elimination without Pivoting using Band Structure
In the general case, where the matrix A is assumed only to be nonsingular, there exist $M \times M$ matrices P, L, U such that

$$PA = LU\,.$$

Here P is a permutation matrix, L is a scaled lower triangular matrix, and U is an upper triangular matrix; i.e., they have the form

$$L = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ l_{ij} & & 1 \end{pmatrix}, \qquad U = \begin{pmatrix} u_{11} & & u_{ij} \\ & \ddots & \\ 0 & & u_{MM} \end{pmatrix}.$$

This decomposition corresponds to the Gaussian elimination method with pivoting. The method is very easy and has favourable properties with respect to the sparse structure, if pivoting is not necessary (i.e., $P = I$, $A = LU$). Then the matrix A is called *LU factorizable*.

Denote by A_k the leading principal submatrix of A of dimension $k \times k$, i.e.,

$$A_k := \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix},$$

and suppose that it already has been factorized as $A_k = L_k U_k$. This is obviously possible for $k = 1$: $A_1 = (a_{11}) = (1)(a_{11})$. The matrix A_{k+1} can be represented in the form of a block matrix

$$A_{k+1} = \left(\begin{array}{c|c} A_k & b \\ \hline c^T & d \end{array}\right)$$

with $b, c \in \mathbb{R}^k$, $d \in \mathbb{R}$.

Using the ansatz

$$L_{k+1} = \left(\begin{array}{c|c} L_k & 0 \\ \hline l^T & 1 \end{array}\right), \qquad U_{k+1} = \left(\begin{array}{c|c} U_k & u \\ \hline 0 & s \end{array}\right)$$

with unknown vectors $u, l \in \mathbb{R}^k$ and $s \in \mathbb{R}$, it follows that

$$A_{k+1} = L_{k+1}U_{k+1} \iff L_k u = b\,,\ U_k^T l = c\,,\ l^T u + s = d\,. \tag{2.55}$$

From this, we have the following result:

Let A be nonsingular. Then lower and upper triangular matrices L, U exist with $A = LU$ if and only if A_k is nonsingular for all $1 \le k \le M$. For this case, L and U are determined uniquely. (2.56)

Furthermore, from (2.55) we have the following important consequences: If the first l components of the vector b are equal to zero, then this is valid for the vector u, too:

$$\text{If } b = \begin{pmatrix} 0 \\ \beta \end{pmatrix}, \text{ then } u \text{ also has the structure } u = \begin{pmatrix} 0 \\ \varrho \end{pmatrix}.$$

Similarly,

$$c = \begin{pmatrix} 0 \\ \gamma \end{pmatrix} \text{ implies the structure } l = \begin{pmatrix} 0 \\ \lambda \end{pmatrix}.$$

For example, if the matrix A has a structure as shown in Figure 2.16, then the zeros outside of the surrounded entries are preserved after the LU factorization. Before we introduce appropriate definitions to generalize these results, we want to consider the special case of symmetric matrices.

$$A = \begin{pmatrix} * & 0 & * & 0 & 0 \\ 0 & * & * & 0 & * \\ * & * & * & * & * \\ 0 & 0 & * & * & 0 \\ 0 & * & * & 0 & * \end{pmatrix}$$

Figure 2.16. Profile of a matrix.

If A is as before nonsingular and LU factorizable, then $U = DL^T$ with a diagonal matrix $D = \operatorname{diag}(d_i)$, and therefore

$$A = LDL^T\,.$$

This is true because A has the form $A = LD\tilde{U}$, where the upper triangular matrix $\tilde{U}$ satisfies the scaling condition $\tilde{u}_{ii} = 1$ for all $i = 1, \ldots, M$. Such a factorization is unique, and thus

$$A = A^T \text{ implies } L^T = \tilde{U}\,, \text{ therefore } A = LDL^T\,.$$

If in particular A is symmetric and positive definite, then also $d_i > 0$ is valid. Thus exactly one matrix $\tilde{L}$ of the form

$$\tilde{L} = \begin{pmatrix} l_{11} & & 0 \\ & \ddots & \\ l_{ij} & & l_{MM} \end{pmatrix} \quad \text{with } l_{ii} > 0 \quad \text{for all } i$$

exists such that

$$A = \tilde{L}\tilde{L}^T, \quad \text{the so-called } \textit{Cholesky decomposition.}$$

We have

$$\tilde{L}_{\text{Chol}} = L_{\text{Gauss}}\sqrt{D}, \quad \text{where} \quad \sqrt{D} := \operatorname{diag}\left(\sqrt{d_i}\right).$$

This shows that the Cholesky method for the determination of the Cholesky factor $\tilde{L}$ also preserves certain zeros of A in the same way as the Gaussian elimination without pivoting.

In what follows, we want to specify the set of zeros that is preserved by Gaussian elimination without pivoting. We will not consider a symmetric matrix; but for the sake of simplicity we will consider a matrix with a symmetric distribution of its entries.

Definition 2.19 Let $A \in \mathbb{R}^{M\times M}$ be a matrix such that $a_{ii} \neq 0$ for $i = 1, \ldots, M$ and

$$a_{ij} \neq 0 \quad \text{if and only if} \quad a_{ji} \neq 0 \quad \text{for all } i, j = 1, \ldots, M\,. \tag{2.57}$$

We define, for $i = 1, \ldots, M$,

$$f_i(A) := \min\left\{j \;\middle|\; a_{ij} \neq 0\,,\ 1 \leq j \leq i\right\}.$$

Then

$$m_i(A) := i - f_i(A)$$

is called the *ith (left-hand side) row bandwidth* of A.

The *bandwidth* of a matrix A that satisfies (2.57) is the number

$$m(A) := \max_{1\leq i\leq M} m_i(A) = \max\left\{i - j \;\middle|\; a_{ij} \neq 0\,,\ 1 \leq j \leq i \leq M\right\}.$$

The *band* of the matrix A is

$$B(A) := \left\{(i,j),(j,i) \;\middle|\; i - m(A) \leq j \leq i\,,\ 1 \leq i \leq M\right\}.$$

The set

$$\operatorname{Env}(A) := \left\{(i,j),(j,i) \;\middle|\; f_i(A) \leq j \leq i\,,\ 1 \leq i \leq M\right\}$$

is called the *hull* or *envelope* of A. The number

$$p(A) := M + 2\sum_{i=1}^{M} m_i(A)$$

is called the *profile* of A.

The profile is the number of elements of $\mathrm{Env}(A)$.

For the matrix A in Figure 2.16 we have $(m_1(A), \ldots, m_5(A)) = (0, 0, 2, 1, 3)$, $m(A) = 3$, and $p(A) = 17$.

Summarizing the above considerations, we have proved the following theorem:

Theorem 2.20 *Let A be a matrix with the symmetric structure* (2.57). *Then the Cholesky method or the Gaussian elimination without pivoting preserves the hull and in particular the bandwidth.*

The hull may contain zeros that will be replaced by (nonzero) entries during the decomposition process. Therefore, in order to keep this *fill-in* small, the profile should be as small as possible.

Furthermore, in order to exploit the matrix structure for an efficient assembling and storage, this structure (or some estimate of it) should be known in advance, before the computation of the matrix entries is started.

For example, if A is a stiffness matrix with the entries

$$a_{ij} = a(\varphi_j, \varphi_i) = \int_\Omega \nabla\varphi_j \cdot \nabla\varphi_i \, dx \,,$$

then the property

$$a_{ij} \neq 0 \quad \Rightarrow \quad a_i, a_j \text{ are neighbouring nodes}$$

can be used for the definition of an (eventually too large) symmetric matrix structure. This is also valid for the case of a nonsymmetric bilinear form and thus a nonsymmetric stiffness matrix. Also in this case, the definition of $f_i(A)$ can be replaced by

$$f_i(A) := \min\left\{ j \;\middle|\; 1 \leq j \leq i \,,\ j \text{ is a neighbouring node of } i \right\}.$$

Since the characterization (2.56) of the possibility of the Gaussian elimination without pivoting cannot be checked directly, we have to specify sufficient conditions. Examples for such conditions are the following (see [34]):

- A is symmetric and positive definite,
- A is an M-matrix.

  Sufficient conditions for this property were given in (1.32) and $(1.32)^*$. In Section 3.9, geometrical conditions for the family of triangulations $(\mathcal{T}_h)_h$ will be derived that guarantee that the finite element discretization considered here creates an M-matrix.

Data Structures

For sparse matrices, it is appropriate to store only the components within the band or the hull. A symmetric matrix $A \in \mathbb{R}^{M \times M}$ with bandwidth m can be stored in $M(m+1)$ memory positions. By means of the index

conversion $a_{ik} \rightsquigarrow b_{i,k-i+m+1}$ for $k \le i$, the matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1,m+1} & & & \\ a_{21} & a_{22} & \cdots & \vdots & \ddots & 0 & \\ \vdots & \vdots & \ddots & \vdots & \ddots & \ddots & \\ a_{m+1,1} & a_{m+1,2} & \cdots & a_{m+1,m+1} & \ddots & \ddots & \ddots \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ & 0 & \ddots & \ddots & \ddots & \ddots & \ddots \\ & & & a_{M,M-m} & \cdots & a_{M,M-1} & a_{M,M} \end{pmatrix} \in \mathbb{R}^{M\times M}$$

is mapped to the matrix

$$B = \begin{pmatrix} 0 & \cdots & \cdots & 0 & a_{11} \\ 0 & \cdots & 0 & a_{21} & a_{22} \\ \vdots & & & \vdots & \vdots \\ 0 & a_{m,1} & \cdots & \cdots & a_{m,m} \\ a_{m+1,1} & \cdots & \cdots & a_{m+1,m} & a_{m+1,m+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{M,M-m} & \cdots & \cdots & a_{M,M-1} & a_{M,M} \end{pmatrix} \in \mathbb{R}^{M\times(m+1)}.$$

The unused elements of B, i.e., $(B)_{ij}$ for $i = 1,\dots,m$, $j = 1,\dots,m+1-i$, are here filled with zeros.

For a general band matrix, the matrix $B \in \mathbb{R}^{M\times(2m+1)}$ obtained by the above conversion has the following form:

$$B = \begin{pmatrix} 0 & \cdots & 0 & a_{11} & a_{12} & \cdots & a_{1,m+1} \\ 0 & \cdots & a_{21} & a_{22} & \cdots & \cdots & a_{2,m+2} \\ \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & a_{m,1} & \cdots & \cdots & \cdots & \cdots & a_{m,2m} \\ a_{m+1,1} & \cdots & \cdots & \cdots & \cdots & \cdots & a_{m+1,2m+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{M-m,M-2m} & \cdots & \cdots & \cdots & \cdots & \cdots & a_{M-m,M} \\ a_{M-m+1,M-2m+1} & \cdots & \cdots & \cdots & \cdots & a_{M-m+1,M} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ a_{M,M-m} & \cdots & \cdots & a_{M,M} & 0 & \cdots & 0 \end{pmatrix}.$$

Here, in the right lower part of the matrix, a further sector of unused elements arose, which is also filled with zeros.

If the storage is based on the hull, additionally a pointer field is needed, which points to the diagonal elements, for example. If the matrix is sym-

metric, again the storage of the lower triangular matrix is sufficient. For the matrix A from Figure 2.16 under the assumption that A is symmetric, the pointer field could act as shown in Figure 2.17.

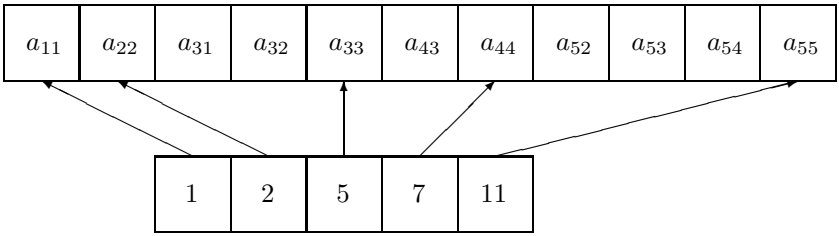


Figure 2.17. Linear storage of the hull.

Coupled Assembling and Decomposition

A formerly popular method, the so-called *frontal method*, performs simultaneously assembling and the Cholesky factorization.

We consider this method for the example of the stiffness matrix $A_h = (a_{ij}) \in \mathbb{R}^{M\times M}$ with bandwidth m (with the original numbering).

The method is based on the kth step of the Gaussian or Cholesky method (cf. Figure 2.18).

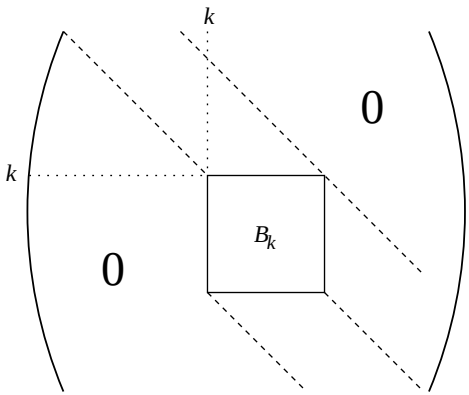


Figure 2.18. kth step of the Cholesky method.

Only the entries of B_k are to be changed, i.e., only those elements a_{ij} with $k \le i, j \le k + m$. The corresponding formula is

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)}\,, \quad i, j = k+1, \ldots, k+m\,. \tag{2.58}$$

Here, the upper indices indicate the steps of the elimination method, which we store in a_{ij}. The entries a_{ij} are generated by summation of entries of

the element stiffness matrix of those elements K that contain nodes with the indices i, j.

Furthermore, to perform the elimination step (2.58), only $a_{ik}^{(k)}, a_{kj}^{(k)}$ for $i, j = k, \ldots, k+m$ must be completely assembled; $a_{ij}^{(k)}, i, j = k+1, \ldots, k+m$, can be replaced by $\tilde{a}_{ij}^{(k)}$ if $a_{ij}^{(k+1)}$ is later defined by $a_{ij}^{(k+1)} := \tilde{a}_{ij}^{(k+1)} + a_{ij}^{(k)} - \tilde{a}_{ij}^{(k)}$. That is, for the present, a_{ij} needs to consist of only a few contributions of elements K with nodes i, j in K.

From these observations, the following algorithm is obtained. The kth step for $k = 1, \ldots, M$ reads as follows:

- Assemble all of the missing contributions of elements K that contain the node with index k.
- Compute $A^{(k+1)}$ by modification of the entries of B_k according to (2.58).
- Store the kth row of $A^{(k+1)}$, also out of the main memory.
- Define B_{k+1} (by a south-east shift).

Here the assembling is node-based and not element-based.

The advantage of this method is that A_h need not be completely assembled and stored in the main memory, but only a matrix $B_k \in \mathbb{R}^{(m+1)\times(m+1)}$. Of course, if M is not too large, there may be no advantage.

Bandwidth Reduction

The *complexity*, i.e., the number of operations, is crucial for the application of a particular method:

The Cholesky method, applied to a symmetric matrix $A \in \mathbb{R}^{M\times M}$ with bandwidth m, requires $O(m^2 M)$ operations in order to compute L.

However, the bandwidth m of the stiffness matrix depends on the numbering of the nodes. Therefore, a numbering is to be found where the number m is as small as possible.

We want to consider this again for the example of the Poisson equation on the rectangle with the discretization according to Figure 2.9. Let the interior nodes have the coordinates (ih, jh) with $i = 1, \ldots k-1$, $j = 1, \ldots, l-1$. The discretization corresponds to the finite difference method introduced beginning with (1.10); i.e., the bandwidth is equal to $k-1$ for a rowwise numbering or $l-1$ for a columnwise numbering.

For $k \ll l$ or $k \gg l$, this fact results in a large difference of the bandwidth m or of the profile (of the left triangle), which is of size $(k-1)(l-1)(m+1)$ except for a term of m^2. Therefore, the columnwise numbering is preferred for $k \gg l$; the rowwise numbering is preferred for $k \ll l$.

For a general domain Ω, a numbering algorithm based on a given triangulation $\mathcal{T}_h$ and on a basis $\{\varphi_i\}$ of V_h is necessary with the following properties:

The structure of A resulting from the numbering must be such that the band or the profile of A is as small as possible. Furthermore, the numbering algorithm should yield the numbers $m(A)$ or $f_i(A)$, $m_i(A)$ such that the matrix A can also be assembled using the element matrices $A^{(k)}$.

Given a triangulation $\mathcal{T}_h$ and a corresponding basis $\{\varphi_i \mid 1 \le i \le M\}$ of V_h, we start with the assignment of some graph G to this triangulation as follows:

The nodes of G coincide with the nodes $\{a_1, \ldots, a_M\}$ of the triangulation. The definition of its edges is:

$$(a_i, a_j) \text{ is an edge of } G \iff$$
$$\text{there exists a } K \in \mathcal{T}_h \text{ such that } \varphi_i|_K \not\equiv 0\,,\ \varphi_j|_K \not\equiv 0\,.$$

In Figure 2.19 some examples are given, where the example (2) will be introduced in Section 3.3.

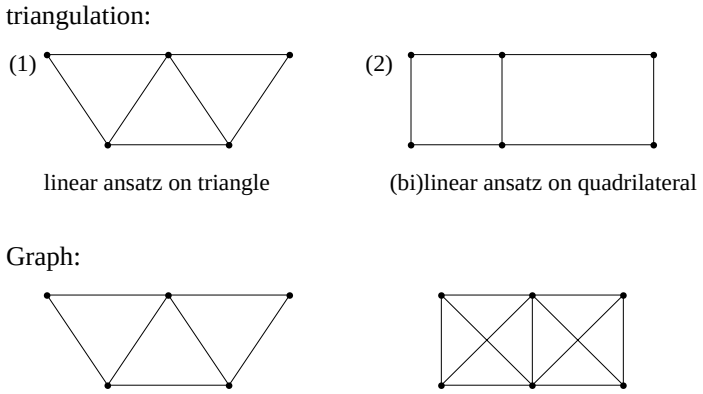


Figure 2.19. Triangulation and assigned graph.

If several degrees of freedom are assigned to some node of the triangulation $\mathcal{T}_h$, then also in G several nodes are assigned to it. This is the case, for example, if so-called Hermite elements are considered, which will be introduced in Section 3.3. The costs of administration are small if the same number of degrees of freedom is assigned to all nodes of the triangulation.

An often-used numbering algorithm is the *Cuthill–McKee method*. This algorithm operates on the graph G just defined. Two nodes a_i, a_j of G are called *neighboured* if (a_i, a_j) is an edge of G. The *degree* of a node a_i of G is defined as the number of neighbours of a_i.

The kth step of the algorithm for $k = 1, \ldots, M$ has the following form:

$k = 1$: Choose a starting node, which gets the number 1. This starting node forms the level 1.

$k > 1$: If all nodes are already numbered, the algorithm is terminated. Otherwise, the level k is formed by taking all the nodes that are not num-

bered yet and that are neighbours of a node of level $k-1$. The nodes of level k will be consecutively numbered.

Within a level, we can sort, for example, by the degree, where the node with the smallest degree is numbered first.

The *reverse Cuthill–McKee method* consists of the above method and the inversion of the numbering at the end; i.e.,

$$\text{new node number } = M + 1 - \text{old node number}\,.$$

This corresponds to a reflection of the matrix at the counterdiagonal. The bandwidth does not change by the inversion, but the profile may diminish drastically for many examples (cf. Figure 2.20).

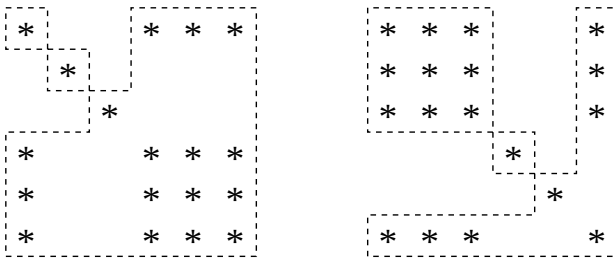


Figure 2.20. Change of the hull by reflection at the counterdiagonal.

The following estimate holds for the bandwidth m of the numbering created by the Cuthill–McKee algorithm:

$$\frac{D+i}{2} \le m \le \max_{2\le k\le \nu} (N_{k-1} + N_k - 1)\ .$$

Here D is the maximum degree of a node of G, ν is the number of levels, and N_k is the number of nodes of level k. The number i is equal to 0 if D is even, and i is equal to 1 if D is odd. The left-hand side of the above inequality is easy to understand by means of the following argument: To reach a minimal bandwidth, all nodes that are neighbours of a_i in the graph G should also be neighbours of a_i in the numbering. Then the best situation is given if the neighboured nodes would appear uniformly immediately before and after a_i. If D is odd, then one side has one node more than the other.

To verify the right-hand side, consider a node a_i that belongs to level $k-1$ as well as a node a_j that is a neighbour of a_i in the graph G and that is not yet numbered in level $k-1$. Therefore, a_j will get a number in the kth step. The largest bandwidth is obtained if a_i is the first node of the numbering of level $k-1$ and if a_j is the last node of level k. Hence exactly $(N_{k-1}-1)+(N_k-1)$ nodes lie between both of these; i.e., their distance in the numbering is $N_{k-1}+N_k-1$.

It is favourable if the number ν of levels is as large as possible and if all the numbers N_k are of the same size, if possible. Therefore, the starting node should be chosen "at one end" of the graph G if possible; if all the

starting nodes are to be checked, the expense will be $O(M\tilde{M})$, where $\tilde{M}$ is the number of edges of G. One possibility consists in choosing a node with minimum degree for the starting node. Another possibility is to let the algorithm run once and then to choose the last-numbered node as the starting node.

If a numbering is created by the (reverse) Cuthill–McKee algorithm, we can try to improve it "locally", i.e., by exchanging particular nodes.

Exercise

2.8 Show that the number of arithmetic operations for the Cholesky method for an $M \times M$ matrix with bandwidth m has order $Mm^2/2$; additionally, M square roots have to be calculated.

3

The Finite Element Method for Linear Elliptic Boundary Value Problems of Second Order

3.1 Variational Equations and Sobolev Spaces

We now continue the definition and analysis of the "correct" function spaces that we began in (2.17)–(2.20). An essential assumption ensuring the existence of a solution of the variational equation (2.13) is the completeness of the basic space $(V, \|\cdot\|)$. In the concrete case of the Poisson equation the "preliminary" function space V according to (2.7) can be equipped with the norm $\|\cdot\|_1$, defined in (2.19), which has been shown to be equivalent to the norm $\|\cdot\|_a$, given in (2.6) (see (2.46)). If we consider the minimization problem (2.14), which is equivalent to the variational equation, the functional F is bounded from below such that the infimum assumes a finite value and there exists a *minimal sequence* $(v_n)_n$ in V, that is, a sequence with the property

$$\lim_{n\to\infty} F(v_n) = \inf\left\{F(v) \;\middle|\; v \in V\right\} .$$

The form of F also implies that $(v_n)_n$ is a Cauchy sequence. If this sequence converges to an element $v \in V$, then, due to the continuity of F with respect to $\|\cdot\|$, it follows that v is a solution of the minimization problem. This completeness of V with respect to $\|\cdot\|_a$, and hence with respect to $\|\cdot\|_1$, is not satisfied in the definition (2.7), as Example 2.8 has shown. Therefore, an extension of the basic space V, as formulated in (2.20), is necessary. This space will turn out to be "correct," since it is complete with respect to $\|\cdot\|_1$.

In what follows we use the following **general assumption**:

$$\begin{aligned} &V \text{ is a vector space with scalar product } \langle\cdot,\cdot\rangle \text{ and the norm } \|\cdot\| \\ &\text{induced by } \langle\cdot,\cdot\rangle \text{ (for this, } \|v\| := \langle v,v\rangle^{1/2} \text{ for } v \in V \mathit{is satisfied})\,; \\ &V \text{ is complete with respect to } \|\cdot\|, \text{ i.e. a Hilbert space}\,; \\ &a : V \times V \to \mathbb{R} \text{ is a (not necessarily symmetric) bilinear form}\,; \\ &b : V \to \mathbb{R} \text{ is a linear form}\,. \end{aligned} \tag{3.1}$$

The following theorem generalizes the above consideration to nonsymmetric bilinear forms:

Theorem 3.1 (Lax–Milgram) *Suppose the following conditions are satisfied:*

- *a is continuous (cf. (2.42)); that is, there exists some constant $M > 0$ such that*

$$|a(u,v)| \;\leq\; M\|u\|\,\|v\| \quad \textit{for all } u, v \in V\,; \tag{3.2}$$

- *a is V-elliptic (cf. (2.43)); that is, there exists some constant $\alpha > 0$ such that*

$$a(u,u) \;\geq\; \alpha\|u\|^2 \quad \textit{for all } u \in V\,; \tag{3.3}$$

- *b is continuous; that is, there exists some constant $C > 0$ such that*

$$|b(u)| \;\leq\; C\|u\| \quad \textit{for all } u \in V\,. \tag{3.4}$$

Then the variational equation (2.21), namely,

$$\textit{find } \bar{u} \in V \textit{ such that} \quad a(\bar{u}, v) = b(v) \quad \textit{for all } v \in V, \tag{3.5}$$

has one and only one solution.

Here, one cannot avoid the assumptions (3.1) and (3.2)–(3.4) in general.

Proof: See, for example, [26]; for an alternative proof see Exercise 3.1. □

Now returning to the example above, the assumptions (3.2) and (3.3) are obviously satisfied for $\|\cdot\| = \|\cdot\|_a$. However, the "preliminary" definition of the function space V of (2.7) with norm $\|\cdot\|_a$ defined in (2.19) is insufficient, since $(V, \|\cdot\|_a)$ is not complete. Therefore, the space V must be extended. Indeed, it is not the norm on V that has been chosen incorrectly, since V is also not complete with respect to another norm $\|\cdot\|$ that satisfies (3.2) and (3.3). In this case the norms $\|\cdot\|$ and $\|\cdot\|_a$ would be equivalent (cf. (2.46)), and consequently,

$$(V, \|\cdot\|_a) \text{ complete} \quad \Longleftrightarrow \quad (V, \|\cdot\|) \text{ complete}\,.$$

Now we extend the space V and thereby generalize definition (2.17).

Definition 3.2 Suppose $\Omega \subset \mathbb{R}^d$ is a (bounded) domain.
The *Sobolev space* $H^k(\Omega)$ is defined by

$$H^k(\Omega) := \left\{ v : \Omega \to \mathbb{R} \;\middle|\; v \in L^2(\Omega)\,,\ \text{the weak derivatives } \partial^\alpha v \text{ exist in } L^2(\Omega) \text{ and for all multi-indices } \alpha \text{ with } |\alpha| \le k \right\}.$$

A scalar product $\langle \cdot, \cdot \rangle_k$ and the resulting norm $\| \cdot \|_k$ in $H^k(\Omega)$ are defined as follows:

$$\langle v, w \rangle_k := \int_\Omega \sum_{\substack{\alpha \text{ multi-index} \\ |\alpha| \le k}} \partial^\alpha v \, \partial^\alpha w \, dx\,, \tag{3.6}$$

$$\begin{aligned} \|v\|_k := \langle v, v \rangle_k^{1/2} &= \left(\int_\Omega \sum_{\substack{\alpha \text{ multi-index} \\ |\alpha| \le k}} \left| \partial^\alpha v \right|^2 dx \right)^{1/2} \\ &= \left(\sum_{\substack{\alpha \text{ multi-index} \\ |\alpha| \le k}} \int_\Omega \left| \partial^\alpha v \right|^2 dx \right)^{1/2} = \left(\sum_{\substack{\alpha \text{ multi-index} \\ |\alpha| \le k}} \left\| \partial^\alpha v \right\|_0^2 \right)^{1/2}. \end{aligned} \tag{3.7}$$

Greater flexibility with respect to the smoothness properties of the functions that are contained in the definition is obtained by requiring that v and its weak derivatives should belong not to $L^2(\Omega)$ but to $L^p(\Omega)$. In the norm denoted by $\| \cdot \|_{k,p}$ the $L^2(\Omega)$ and ℓ_2 norms (for the vector of the derivative norms) have to be replaced by the $L^p(\Omega)$ and ℓ_p norms, respectively (see Appendices A.3 and A.5). However, the resulting space, denoted by $W_p^k(\Omega)$, can no longer be equipped with a scalar product for $p \neq 2$. Although these spaces offer greater flexibility, we will not use them except in Sections 3.6, 6.2, and 9.3.

Besides the norms $\| \cdot \|_k$, there are seminorms $| \cdot |_l$ for $0 \le l \le k$ in $H^k(\Omega)$, defined by

$$|v|_l = \left(\sum_{\substack{\alpha \text{ multi-index} \\ |\alpha| = l}} \left\| \partial^\alpha v \right\|_0^2 \right)^{1/2},$$

such that

$$\|v\|_k = \left(\sum_{l=0}^{k} |v|_l^2 \right)^{1/2},$$

In particular, these definitions are compatible with those in (2.18),

$$\langle v, w \rangle_1 := \int_\Omega vw + \nabla v \cdot \nabla w \, dx\,,$$

and with the notation $\| \cdot \|_0$ for the $L^2(\Omega)$ norm, giving a meaning to this one.

The above definition contains some assertions that are formulated in the following theorem:

Theorem 3.3 *The bilinear form $\langle \cdot, \cdot \rangle_k$ is a scalar product on $H^k(\Omega)$; that is, $\| \cdot \|_k$ is a norm on $H^k(\Omega)$.*

$H^k(\Omega)$ is complete with respect to $\| \cdot \|_k$, and is thus a Hilbert space.

Proof: See, for example, [37]. □

Obviously,

$$H^k(\Omega) \subset H^l(\Omega) \quad \text{for } k \geq l\,,$$

and the embedding is continuous, since

$$\|v\|_l \leq \|v\|_k \quad \text{for all } v \in H^k(\Omega)\,. \tag{3.8}$$

In the one-dimensional case ($d = 1$) $v \in H^1(\Omega)$ is necessarily continuous:

Lemma 3.4

$$H^1(a,b) \subset C[a,b]\,,$$

and the embedding is continuous, where $C[a,b]$ is equipped with the norm $\| \cdot \|_\infty$; that is, there exists some constant $C > 0$ such that

$$\|v\|_\infty \leq C\|v\|_1 \quad \textit{for all } v \in H^1(a,b)\,. \tag{3.9}$$

Proof: See Exercise 3.2. □

Since the elements of $H^k(\Omega)$ are first of all only square integrable functions, they are determined only up to points of a set of (d-dimensional) measure zero. Therefore, a result as in Lemma 3.4 means that the function is allowed to have removable discontinuities at points of such a set of measure zero that vanish by modifying the function values.

However, in general, $H^1(\Omega) \not\subset C(\bar{\Omega})$.

As an example for this, we consider a circular domain in dimension $d = 2$:

$$\Omega = B_R(0) = \left\{ x \in \mathbb{R}^2 \;\middle|\; |x| < R \right\}\,, \quad R < 1\,.$$

Then the function

$$v(x) := |\log |x||^\gamma \qquad \text{for some } \gamma < \frac{1}{2}$$

is in $H^1(\Omega)$, but not in $C(\bar{\Omega})$ (see Exercise 3.3).

The following problem now arises: In general, one cannot speak of a value $v(x)$ for some $x \in \Omega$ because a set of one point $\{x\}$ has (Lebesgue) measure zero. How do we then have to interpret the Dirichlet boundary conditions? A way out is to consider the boundary (pieces of the boundary, respectively) not as arbitrary points but as $(d-1)$-dimensional "spaces" (manifolds).

The above question can therefore be reformulated as follows: Is it possible to interpret v on $\partial\Omega$ as a function of $L^2(\partial\Omega)$ ($\partial\Omega$ "$\subset$" $\mathbb{R}^{d-1}$) ?

It is indeed possible if we have some minimal regularity of $\partial\Omega$ in the following sense: It has to be possible to choose locally, for some boundary point $x \in \partial\Omega$, a coordinate system in such a way that the boundary is locally a hyperplane in this coordinate system and the domain lies on one side. Depending on the smoothness of the parametrisation of the hyperplane we then speak of *Lipschitz*, C^k- (*for* $k \in \mathbb{N}$), and C^∞- *domains* (for an exact definition see Appendix A.5).

Examples:

(1) A circle $\Omega = \left\{x \in \mathbb{R}^d \;\middle|\; |x - x_0| < R\right\}$ is a C^k-domain for all $k \in \mathbb{N}$, and hence a C^∞-domain.

(2) A rectangle $\Omega = \left\{x \in \mathbb{R}^d \;\middle|\; 0 < x_i < a_i,\ i = 1, \ldots, d\right\}$ is a Lipschitz domain, but not a C^1-domain.

(3) A circle with a cut $\Omega = \left\{x \in \mathbb{R}^d \;\middle|\; |x - x_0| < R,\ x \neq x_0 + \lambda e_1 \text{ for } 0 \leq \lambda < R\right\}$ is not a Lipschitz domain, since Ω does not lie on one side of $\partial\Omega$ (see Figure 3.1).

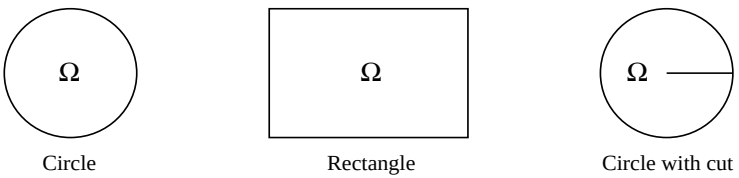


Figure 3.1. Domains of different smoothness.

Hence, suppose Ω is a Lipschitz domain. Since only a finite number of overlapping coordinate systems are sufficient for the description of $\partial\Omega$, using these, it is possible to introduce a $(d-1)$-dimensional measure on $\partial\Omega$ and define the space $L^2(\partial\Omega)$ of square integrable functions with respect to this measure (see Appendix A.5 or [37] for an extensive description). In the following, let $\partial\Omega$ be equipped with this $(d-1)$-dimensional measure $d\sigma$, and integrals over the boundary are to be interpreted accordingly. This also holds for Lipschitz subdomains of Ω, since they are given by the finite elements.

Theorem 3.5 (Trace Theorem) *Suppose Ω is a bounded Lipschitz domain. We define*

$$C^\infty(\mathbb{R}^d)|_\Omega \;:=\; \left\{v : \Omega \to \mathbb{R} \;\middle|\; v \text{ can be extended to } \tilde{v} : \mathbb{R}^d \to \mathbb{R} \text{ and } \tilde{v} \in C^\infty(\mathbb{R}^d)\right\}.$$

Then, $C^\infty(\mathbb{R}^d)|_\Omega$ is dense in $H^1(\Omega)$*; that is, with respect to $\|\cdot\|_1$ an arbitrary $w \in H^1(\Omega)$ can be approximated arbitrarily well by some $v \in C^\infty(\mathbb{R}^d)|_\Omega$.*

The mapping that restricts v to $\partial\Omega$,

$$\begin{aligned} \gamma_0 : \big(C^\infty(\mathbb{R}^d)|_\Omega, \|\cdot\|_1\big) &\to \big(L^2(\partial\Omega), \|\cdot\|_0\big)\,, \\ v &\mapsto v|_{\partial\Omega}\,, \end{aligned}$$

is continuous.

Thus there exists a unique, linear, and continuous extension

$$\gamma_0 : \big(H^1(\Omega), \|\cdot\|_1\big) \to \big(L^2(\partial\Omega), \|\cdot\|_0\big)\,.$$

Proof: See, for example, [37]. □

Therefore, in short form, $\gamma_0(v) \in L^2(\partial\Omega)$, and there exists some constant $C > 0$ such that

$$\|\gamma_0(v)\|_0 \le C\|v\|_1 \quad \text{for all } v \in H^1(\Omega)\,.$$

Here $\gamma_0(v) \in L^2(\partial\Omega)$ is called the *trace* of $v \in H^1(\Omega)$.

The mapping γ_0 is not surjective; that is, $\big\{\gamma_0(v) \;\big|\; v \in H^1(\Omega)\big\}$ is a real subset of $L^2(\partial\Omega)$. For all $v \in C^\infty(\mathbb{R}^d)|_\Omega$ we have

$$\gamma_0(v) = v|_{\partial\Omega}\,.$$

In the following we will use again $v|_{\partial\Omega}$ or "v on $\partial\Omega$" for $\gamma_0(v)$, but in the sense of Theorem 3.5. According to this theorem, definition (2.20) is well-defined with the interpretation of u on $\partial\Omega$ as the trace:

Definition 3.6 $H_0^1(\Omega) := \big\{v \in H^1(\Omega) \;\big|\; \gamma_0(v) = 0 \text{ (as a function on } \partial\Omega)\big\}$.

Theorem 3.7 *Suppose $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain. Then $C_0^\infty(\Omega)$ is dense in $H_0^1(\Omega)$.*

Proof: See [37]. □

The assertion of Theorem 3.5, that $C^\infty(\mathbb{R}^d)|_\Omega$ is dense in $H^1(\Omega)$, has severe consequences for the treatment of functions in $H^1(\Omega)$ which are in general not very smooth. It is possible to consider them as smooth functions if at the end only relations involving continuous expressions in $\|\cdot\|_1$ (and not requiring something like $\|\partial_i v\|_\infty$) arise. Then, by some "density argument" the result can be transferred to $H^1(\Omega)$ or, as for the trace term, new terms can be defined for functions in $H^1(\Omega)$. Thus, for the proof of Lemma 3.4 it is necessary simply to verify estimate (3.9), for example for $v \in C^1[a,b]$. By virtue of Theorem 3.7, analogous results hold for $H_0^1(\Omega)$.
Hence, for $v \in H^1(\Omega)$ integration by parts is possible:

Theorem 3.8 *Suppose $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain. The outer unit normal vector $\nu = (\nu_i)_{i=1,\dots,d} : \partial\Omega \to \mathbb{R}^d$ is defined almost everywhere and $\nu_i \in L^\infty(\partial\Omega)$.*

For $v, w \in H^1(\Omega)$ and $i = 1, \ldots, d$,

$$\int_\Omega \partial_i v \, w \, dx = -\int_\Omega v \, \partial_i w \, dx + \int_{\partial\Omega} v \, w \, \nu_i \, d\sigma \, .$$

Proof: See, for example, [14] or [37]. □

If $v \in H^2(\Omega)$, then due to the above theorem, $v|_{\partial\Omega} := \gamma_0(v) \in L^2(\partial\Omega)$ and $\partial_i v|_{\partial\Omega} := \gamma_0(\partial_i v) \in L^2(\partial\Omega)$, since also $\partial_i v \in H^1(\Omega)$. Hence, the *normal derivative*

$$\partial_\nu v|_{\partial\Omega} := \sum_{i=1}^{d} \partial_i v|_{\partial\Omega} \, \nu_i$$

is well-defined and belongs to $L^2(\partial\Omega)$.

Thus, the trace mapping

$$\begin{aligned} \gamma : H^2(\Omega) &\to L^2(\partial\Omega) \times L^2(\partial\Omega) \, , \\ v &\mapsto (v|_{\partial\Omega}, \partial_\nu v|_{\partial\Omega}) \, , \end{aligned}$$

is well-defined and continuous. The continuity of this mapping follows from the fact that it is a composition of continuous mappings:

$$\begin{aligned} v \in H^2(\Omega) \overset{\text{continuous}}{\mapsto} \partial_i v \in H^1(\Omega) &\overset{\text{continuous}}{\mapsto} \partial_i v|_{\partial\Omega} \in L^2(\partial\Omega) \\ &\overset{\text{continuous}}{\mapsto} \partial_i v|_{\partial\Omega} \, \nu_i \in L^2(\partial\Omega) \, . \end{aligned}$$

Corollary 3.9 *Suppose $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain.*

(1) Let $w \in H^1(\Omega)$, $q_i \in H^1(\Omega)$, $i = 1, \ldots, d$. Then

$$\int_\Omega q \cdot \nabla w \, dx = -\int_\Omega \nabla \cdot q \, w \, dx + \int_{\partial\Omega} q \cdot \nu \, w \, d\sigma \, . \tag{3.10}$$

(2) Let $v \in H^2(\Omega)$, $w \in H^1(\Omega)$. Then

$$\int_\Omega \nabla v \cdot \nabla w \, dx = -\int_\Omega \Delta v \, w \, dx + \int_{\partial\Omega} \partial_\nu v \, w \, d\sigma \, .$$

The integration by parts formulas also hold more generally if only it is ensured that the function whose trace has to be formed belongs to $H^1(\Omega)$. For example, if $K = (k_{ij})_{ij}$, where $k_{ij} \in W^1_\infty(\Omega)$ and $v \in H^2(\Omega)$, $w \in H^1(\Omega)$, it follows that

$$\int_\Omega K\nabla v \cdot \nabla w \, dx = -\int_\Omega \nabla \cdot (K\nabla v) \, w \, dx + \int_{\partial\Omega} K\nabla v \cdot \nu \, w \, d\sigma \tag{3.11}$$

with *conormal derivative* (see (0.41))

$$\partial_{\nu_K} v := K\nabla v \cdot \nu = \nabla v \cdot K^T \nu = \sum_{i,j=1}^{d} k_{ij} \partial_j v \, \nu_i .$$

Here it is important that the components of $K\nabla v$ belong to $H^1(\Omega)$, using the fact that for $v \in L^2(\Omega)$, $k \in L^\infty(\Omega)$,

$$kv \in L^2(\Omega) \quad \text{and} \quad \|kv\|_0 \le \|k\|_\infty \|v\|_0 \,.$$

Theorem 3.10 *Suppose $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain. If $k > d/2$, then*

$$H^k(\Omega) \subset C(\bar{\Omega}) \,,$$

and the embedding is continuous.

Proof: See, for example, [37]. □

For dimension $d = 2$ this requires $k > 1$, and for dimension $d = 3$ we need $k > \frac{3}{2}$. Therefore, in both cases $k = 2$ satisfies the assumption of the above theorem.

Exercises

3.1 Prove the Lax–Milgram Theorem in the following way:

(a) Show, by using the Riesz representation theorem, the equivalence of (3.5) with the operator equation

$$A\bar{u} = f$$

for $A \in L[V, V]$ and $f \in V$.

(b) Show, for $T_\varepsilon \in L[V, V]$, $T_\varepsilon v := v - \varepsilon(Av - f)$ and $\varepsilon > 0$, that for some $\varepsilon > 0$, the operator T_ε is a contraction on V. Then conclude the assertion by Banach's fixed-point theorem (in the Banach space setting, cf. Remark 8.5).

3.2 Prove estimate (3.9) by showing that even for $v \in H^1(a, b)$,

$$|v(x) - v(y)| \le |v|_1 |x - y|^{1/2} \quad \text{for } x, y \in (a, b) \,.$$

3.3 Suppose $\Omega \subset \mathbb{R}^2$ is the open disk with radius $\frac{1}{2}$ and centre 0. Prove that for the function $u(x) := \big|\ln|x|\big|^\alpha$, $x \in \Omega \setminus \{0\}$, $\alpha \in (0, \frac{1}{2})$ we have $u \in H^1(\Omega)$, but u cannot be extended continuously to $x = 0$.

3.4 Suppose $\Omega \subset \mathbb{R}^2$ is the open unit disk. Prove that each $u \in H^1(\Omega)$ has a trace $u|_{\partial\Omega} \in L_2(\partial\Omega)$ satisfying $\|u\|_{0,\partial\Omega} \le \sqrt[4]{8}\, \|u\|_{1,\Omega}$.

3.2 Elliptic Boundary Value Problems of Second Order

In this section we integrate boundary value problems for the linear, stationary case of the differential equation (0.33) into the general theory of Section 3.1.

Concerning the domain we will assume that Ω is a bounded Lipschitz domain.

We consider the equation

$$(Lu)(x) := -\nabla \cdot (K(x)\nabla u(x)) + c(x) \cdot \nabla u(x) + r(x)u(x) = f(x) \text{ for } x \in \Omega \tag{3.12}$$

with the data

$$K : \Omega \to \mathbb{R}^{d,d}, \quad c : \Omega \to \mathbb{R}^d, \quad r, f : \Omega \to \mathbb{R}.$$

Assumptions about the Coefficients and the Right-Hand Side
For an interpretation of (3.12) in the classical sense, we need

$$\partial_i k_{ij}\,, c_i\,, r\,, f \in C(\bar{\Omega})\,, \quad i, j \in \{1, \ldots, d\}\,, \tag{3.13}$$

and for an interpretation in the sense of $L^2(\Omega)$ with weak derivatives, and hence for a solution in $H^2(\Omega)$,

$$\partial_i k_{ij}\,, c_i\,, r \in L^\infty(\Omega)\,,\ f \in L^2(\Omega)\,, \quad i, j \in \{1, \ldots, d\}\,. \tag{3.14}$$

Once we have obtained the variational formulation, weaker assumptions about the smoothness of the coefficients will be sufficient for the verification of the properties (3.2)–(3.4), which are required by the Lax–Milgram, namely,

$$\begin{aligned} &k_{ij}\,, c_i\,, \nabla \cdot c\,, r \in L^\infty(\Omega)\,,\ f \in L^2(\Omega)\,, \quad i, j \in \{1, \ldots, d\}\,, \\ &\qquad \text{and if } |\Gamma_1 \cup \Gamma_2|_{d-1} > 0\,, \quad \nu \cdot c \in L^\infty(\Gamma_1 \cup \Gamma_2)\,. \end{aligned} \tag{3.15}$$

Here we refer to a definition of the boundary conditions as in (0.36)–(0.39) (see also below). Furthermore, the *uniform ellipticity* of L is assumed: There exists some constant $k_0 > 0$ such that for (almost) every $x \in \Omega$,

$$\sum_{i,j=1}^{d} k_{ij}(x)\xi_i\xi_j \geq k_0|\xi|^2 \quad \text{for all } \xi \in \mathbb{R}^d \tag{3.16}$$

(that is, the coefficient matrix K is positive definite uniformly in x). Moreover, K should be symmetric.

If K is a diagonal matrix, that is, $k_{ij}(x) = k_i(x)\delta_{ij}$ (this is in particular the case if $k_i(x) = k(x)$ with $k : \Omega \to \mathbb{R}$, $i \in \{1, \ldots, d\}$, where $K\nabla u$ becomes $k\nabla u$), this means that

$$(3.16) \quad \Leftrightarrow \quad k_i(x) \geq k_0 \text{ for (almost) every } x \in \Omega\,, \quad i \in \{1, \ldots, d\}\,.$$

Finally, there exists a constant $r_0 \geq 0$ such that

$$r(x) - \frac{1}{2}\nabla \cdot c(x) \geq r_0 \quad \text{for (almost) every } x \in \Omega\,. \tag{3.17}$$

Boundary Conditions
As in Section 0.5, suppose $\Gamma_1, \Gamma_2, \Gamma_3$ is a disjoint decomposition of the boundary $\partial\Omega$ (cf. (0.39)):

$$\partial\Omega = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3\,,$$

where Γ_3 is a closed subset of the boundary. For given functions $g_j : \Gamma_j \to \mathbb{R}$, $j = 1, 2, 3$, and $\alpha : \Gamma_2 \to \mathbb{R}$ we assume on $\partial\Omega$

- Neumann boundary condition (cf. (0.41) or (0.36))

$$K\nabla u \cdot \nu = \partial_{\nu_K} u = g_1 \quad \text{on } \Gamma_1\,, \tag{3.18}$$

- mixed boundary condition (cf. (0.37))

$$K\nabla u \cdot \nu + \alpha u = \partial_{\nu_K} u + \alpha u = g_2 \quad \text{on } \Gamma_2\,, \tag{3.19}$$

- Dirichlet boundary condition (cf. (0.38))

$$u = g_3 \quad \text{on } \Gamma_3\,. \tag{3.20}$$

Concerning the boundary data the following is assumed: For the classical approach we need

$$g_j \in C(\overline{\Gamma}_j)\,, \quad j = 1, 2, 3\,, \quad \alpha \in C(\overline{\Gamma}_2)\,, \tag{3.21}$$

whereas for the variational interpretation,

$$g_j \in L^2(\Gamma_j)\,, \quad j = 1, 2, 3\,, \quad \alpha \in L^\infty(\Gamma_2) \tag{3.22}$$

is sufficient.

3.2.1 Variational Formulation of Special Cases

The basic strategy for the derivation of the variational formulation of boundary value problems (3.12) has already been demonstrated in Section 2.1. Assuming the existence of a classical solution of (3.12) the following steps are performed in general:

Step 1: Multiplication of the differential equation by test functions that are chosen compatible with the type of boundary condition and subsequent integration over the domain Ω.

Step 2: Integration by parts under incorporation of the boundary conditions in order to derive a suitable bilinear form.

Step 3: Verification of the required properties like ellipticity and continuity.

In the following the above steps will be described for some important special cases.

(I) Homogeneous Dirichlet Boundary Condition
$\partial\Omega = \Gamma_3\,,\ g_3 \equiv 0\,,\ V := H_0^1(\Omega)$

Suppose u is a solution of (3.12), (3.20); that is, in the sense of classical solutions let $u \in C^2(\Omega) \cap C(\bar{\Omega})$ and the differential equation (3.12) be satisfied pointwise in Ω under the assumptions (3.13) as well as $u = 0$ pointwise on $\partial\Omega$. However, the weaker case in which $u \in H^2(\Omega) \cap V$ and the differential equation is satisfied in the sense of $L^2(\Omega)$, now under the assumptions (3.14), can also be considered.

Multiplying (3.12) by $v \in C_0^\infty(\Omega)$ (in the classical case) or by $v \in V$, respectively, then integrating by parts according to (3.11) and taking into account that $v = 0$ on $\partial\Omega$ by virtue of the definition of $C_0^\infty(\Omega)$ and $H_0^1(\Omega)$, respectively, we obtain

$$\begin{aligned} a(u,v) &:= \int_\Omega \{K\nabla u \cdot \nabla v + c \cdot \nabla u\, v + r\, uv\}\, dx \qquad (3.23)\\ &= b(v) := \int_\Omega f v\, dx \qquad \text{for all } v \in C_0^\infty(\Omega) \text{ or } v \in V\,. \end{aligned}$$

The bilinear form a is symmetric if c vanishes (almost everywhere).
For $f \in L^2(\Omega)$,

$$b \text{ is continuous on } (V, \|\cdot\|_1)\,. \qquad (3.24)$$

This follows directly from the Cauchy–Schwarz inequality, since

$$|b(v)| \le \int_\Omega |f|\,|v|\, dx \le \|f\|_0\, \|v\|_0 \le \|f\|_0\, \|v\|_1 \quad \text{for } v \in V\,.$$

Further, by (3.15),

$$a \text{ is continuous } (V, \|\cdot\|_1)\,. \qquad (3.25)$$

Proof: First, we obtain

$$|a(u,v)| \le \int_\Omega \{|K\nabla u|\,|\nabla v| + |c|\,|\nabla u||v| + |r|\,|u|\,|v|\}\, dx\,.$$

Here $|\cdot|$ denotes the absolute value of a real number or the Euclidean norm of a vector. Using also $\|\cdot\|_2$ for the (associated) spectral norm, and $\|\cdot\|_\infty$ for the $L^\infty(\Omega)$ norm of a function, we further introduce the following notations:

$$C_1 := \max\left\{\big\|\|K\|_2\big\|_\infty, \|r\|_\infty\right\} < \infty\,, \quad C_2 := \big\||c|\big\|_\infty < \infty\,.$$

By virtue of

$$|K(x)\nabla u(x)| \le \|K(x)\|_2\, |\nabla u(x)|,$$

we continue to estimate as follows:

$$|a(u,v)| \le \underbrace{C_1 \int_\Omega \{|\nabla u|\,|\nabla v| + |u|\,|v|\}\,dx}_{=:A_1} + \underbrace{C_2 \int_\Omega |\nabla u|\,|v|\,dx}_{=:A_2}\,.$$

The integrand of the first addend is estimated by the Cauchy–Schwarz inequality for $\mathbb{R}^2$, and then the Cauchy–Schwarz inequality for $L^2(\Omega)$ is applied:

$$\begin{aligned} A_1 &\le C_1 \int_\Omega \left\{|\nabla u|^2 + |u|^2\right\}^{1/2} \left\{|\nabla v|^2 + |v|^2\right\}^{1/2} dx \\ &\le C_1 \left\{\int_\Omega |u|^2 + |\nabla u|^2\,dx\right\}^{1/2} \left\{\int_\Omega |v|^2 + |\nabla v|^2\,dx\right\}^{1/2} = C_1 \|u\|_1\,\|v\|_1\,. \end{aligned}$$

Dealing with A_2, we can employ the Cauchy–Schwarz inequality for $L^2(\Omega)$ directly:

$$\begin{aligned} A_2 &\le C_2 \left\{\int_\Omega |\nabla u|^2\,dx\right\}^{1/2} \left\{\int_\Omega |v|^2\,dx\right\}^{1/2} \\ &\le C_2 \|u\|_1\,\|v\|_0 \le C_2 \|u\|_1\,\|v\|_1 \quad \text{for all } u, v \in V\,. \end{aligned}$$

Thus, the assertion follows. □

Remark 3.11 In the proof of the propositions (3.24) and (3.25) it has not been used that the functions u, v satisfy homogeneous Dirichlet boundary conditions. Therefore, under the assumptions (3.15) these properties hold for every subspace $V \subset H^1(\Omega)$.

Conditions for the V-Ellipticity of a

(A) a is symmetric; that is $c = 0$ (a.e.): Condition (3.17) then has the simple form $r(x) \ge r_0$ for almost all $x \in \Omega$.

(A1) $c = 0, \quad r_0 > 0$:

Because of (3.16) we directly get

$$a(u,u) \ge \int_\Omega \{k_0 |\nabla u|^2 + r_0 |u|^2\}\,dx \ge C_3 \|u\|_1^2 \quad \text{for all } u \in V\,,$$

where $C_3 := \min\{k_0, r_0\}$. This also holds for every subspace $V \subset H^1(\Omega)$.

(A2) $c = 0, \quad r_0 \ge 0$:

According to the Poincaré inequality (Theorem 2.18), there exists some constant $C_P > 0$, independent of u, such that for $u \in H_0^1(\Omega)$

$$\|u\|_0 \le C_P \left\{\int_\Omega |\nabla u|^2\,dx\right\}^{1/2}.$$

Taking into account (3.16) and using the simple decomposition $k_0 = \frac{k_0}{1+C_P^2} + \frac{C_P^2}{1+C_P^2}\, k_0$ we can further conclude that

$$\begin{aligned} a(u,u) &\geq \int_\Omega k_0 |\nabla u|^2 \, dx \\ &\geq \frac{k_0}{1+C_P^2} \int_\Omega |\nabla u|^2 \, dx + \frac{C_P^2}{1+C_P^2}\, k_0 \frac{1}{C_P^2} \int_\Omega |u|^2 \, dx = C_4 \|u\|_1^2 , \end{aligned} \tag{3.26}$$

where $C_4 := \dfrac{k_0}{1+C_P^2} > 0$.

For this estimate it is essential that u satisfies the homogeneous Dirichlet boundary condition.

(B) $\big\| |c| \big\|_\infty > 0$:

First of all, we consider a smooth function $u \in C_0^\infty(\Omega)$. From $u \nabla u = \frac{1}{2} \nabla u^2$ we get by integrating by parts

$$\int_\Omega c \cdot \nabla u \, u \, dx = \frac{1}{2} \int_\Omega c \cdot \nabla u^2 \, dx = -\frac{1}{2} \int_\Omega \nabla \cdot c \, u^2 \, dx .$$

Since according to Theorem 3.7 the space $C_0^\infty(\Omega)$ is dense in V, the above relation also holds for $u \in V$. Consequently, by virtue of (3.16) and (3.17) we obtain

$$\begin{aligned} a(u,u) &= \int_\Omega \left\{ K \nabla u \cdot \nabla u + \left(r - \frac{1}{2} \nabla \cdot c \right) u^2 \right\} dx \\ &\geq \int_\Omega \{ k_0 |\nabla u|^2 + r_0 |u|^2 \} \, dx \quad \text{for all } u \in V . \end{aligned} \tag{3.27}$$

Hence, a distinction concerning r_0 as in **(A)** with the same results (constants) is possible.

Summarizing, we have therefore proven the following application of the Lax–Milgram Theorem (Theorem 3.1):

Theorem 3.12 *Suppose $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain. Under the assumptions* (3.15)–(3.17) *the homogeneous Dirichlet problem has one and only one weak solution $u \in H_0^1(\Omega)$.*

(II) Mixed Boundary Conditions

$\partial\Omega = \Gamma_2$, $V = H^1(\Omega)$

Suppose u is a solution of (3.12), (3.19); that is, in the classical sense let $u \in C^2(\Omega) \cap C^1(\bar{\Omega})$ and the differential equation (3.12) be satisfied pointwise in Ω and (3.19) pointwise on $\partial\Omega$ under the assumptions (3.13), (3.21). However, the weaker case can again be considered, now under the assumptions (3.14), (3.22), that $u \in H^2(\Omega)$ and the differential equation is satisfied in the sense of $L^2(\Omega)$ as well as the boundary condition (3.19) in the sense of $L^2(\partial\Omega)$.

As in **(I)**, according to (3.11),

$$\begin{aligned} a(u,v) &:= \int_\Omega \{K\nabla u\cdot\nabla v + c\cdot\nabla u\, v + r\, uv\}\, dx + \int_{\partial\Omega} \alpha\, uv\, d\sigma \quad (3.28)\\ &= b(v) := \int_\Omega fv\, dx + \int_{\partial\Omega} g_2 v\, d\sigma \quad \text{for all } v\in V\,. \end{aligned}$$

Under the assumptions (3.15), (3.22) the continuity of b and a, respectively, ((3.24) and (3.25)) can easily be shown. The additional new terms can be estimated, for instance under the assumptions (3.15), (3.22), by the Cauchy–Schwarz inequality and the Trace Theorem (Theorem 3.4) as follows:

$$\left|\int_{\partial\Omega} g_2 v\, d\sigma\right| \le \|g_2\|_{0,\partial\Omega}\|v|_{\partial\Omega}\|_{0,\partial\Omega} \le C\|g_2\|_{0,\partial\Omega}\|v\|_1 \quad \text{for all } v\in V$$

and

$$\left|\int_{\partial\Omega} \alpha uv\, d\sigma\right| \le \|\alpha\|_{\infty,\partial\Omega}\|u|_{\partial\Omega}\|_{0,\partial\Omega}\|v|_{\partial\Omega}\|_{0,\partial\Omega} \le C^2\|\alpha\|_{\infty,\partial\Omega}\|u\|_1\|v\|_1\,,$$

respectively, for all $u, v\in V$, where $C>0$ denotes the constant appearing in the Trace Theorem.

Conditions for the V-Ellipticity of a

For the proof of the V-ellipticity we proceed similarly to **(I)(B)**, but now taking into account the mixed boundary conditions. For the convective term we have

$$\int_\Omega c\cdot\nabla u\, u\, dx = \frac{1}{2}\int_\Omega c\cdot\nabla u^2\, dx = -\frac{1}{2}\int_\Omega \nabla\cdot c\, u^2\, dx + \frac{1}{2}\int_{\partial\Omega} \nu\cdot c\, u^2\, d\sigma\,,$$

and thus

$$a(u,u) = \int_\Omega \left\{K\nabla u\cdot\nabla u + \left(r - \frac{1}{2}\nabla\cdot c\right)u^2\right\} dx + \int_{\partial\Omega}\left(\alpha + \frac{1}{2}\nu\cdot c\right)u^2\, d\sigma.$$

This shows that $\alpha + \frac{1}{2}\nu\cdot c \ge 0$ on $\partial\Omega$ should additionally be assumed. If $r_0>0$ in (3.17), then the V-ellipticity of a follows directly. However, if only $r_0\ge 0$ is valid, then the so-called *Friedrichs' inequality*, a refined version of the Poincaré inequality, helps (see [25, Theorem 1.9]).

Theorem 3.13 *Suppose $\Omega\subset\mathbb{R}^d$ is a bounded Lipschitz domain and let the set $\tilde\Gamma\subset\partial\Omega$ have a positive $(d-1)$-dimensional measure. Then there exists some constant $C_F>0$ such that for all $v\in H^1(\Omega)$,*

$$\|v\|_1 \le C_F\left\{\int_{\tilde\Gamma} v^2\, d\sigma + \int_\Omega |\nabla v|^2\, dx\right\}^{1/2}. \quad (3.29)$$

If $\alpha + \frac{1}{2}\nu\cdot c \ge \alpha_0 > 0$ for $x\in\tilde\Gamma\subset\Gamma_2$ and $\tilde\Gamma$ has a positive $(d-1)$-dimensional measure, then $r_0\ge 0$ is already sufficient for the V-ellipticity.

Indeed, using Theorem 3.13, we have

$$a(u,u) \geq k_0|u|_1^2 + \alpha_0 \int_{\tilde{\Gamma}} u^2\, d\sigma \geq \min\{k_0,\alpha_0\} \left\{ |u|_1^2 + \int_{\tilde{\Gamma}} u^2\, d\sigma \right\} \geq C_5 \|u\|_1^2$$

with $C_5 := C_F^{-2} \min\{k_0,\alpha_0\}$. Therefore, we obtain the existence and uniqueness of a solution analogously to Theorem 3.12.

(III) General Case

First, we consider the case of a **homogeneous Dirichlet boundary condition** on Γ_3 with $|\Gamma_3|_{d-1} > 0$. For this, we define

$$V := \left\{ v \in H^1(\Omega) : \gamma_0(v) = 0 \text{ on } \Gamma_3 \right\} . \tag{3.30}$$

Here V is a closed subspace of $H^1(\Omega)$, since the trace mapping $\gamma_0 : H^1(\Omega) \to L^2(\partial\Omega)$ and the restriction of a function from $L^2(\partial\Omega)$ to $L^2(\Gamma_3)$ are continuous.

Suppose u is a solution of (3.12), (3.18)–(3.20); that is, in the sense of classical solutions let $u \in C^2(\Omega) \cap C^1(\bar{\Omega})$ and the differential equation (3.12) be satisfied pointwise in Ω and the boundary conditions (3.18)–(3.20) pointwise on their respective parts of $\partial\Omega$ under the assumptions (3.13), (3.21). However, the weaker case that $u \in H^2(\Omega)$ and the differential equation is satisfied in the sense of $L^2(\Omega)$ and the boundary conditions (3.18)–(3.20) are satisfied in the sense of $L^2(\Gamma_j)$, $j = 1,2,3$, under the assumptions (3.14), (3.22) can also be considered here.

As in **(I)**, according to (3.11),

$$\begin{aligned} a(u,v) &:= \int_\Omega \{K\nabla u \cdot \nabla v + c \cdot \nabla u\, v + r\, uv\}\, dx + \int_{\Gamma_2} \alpha\, uv\, d\sigma \qquad (3.31) \\ &= b(v) := \int_\Omega fv\, dx + \int_{\Gamma_1} g_1 v\, d\sigma + \int_{\Gamma_2} g_2 v\, d\sigma \quad \text{for all } v \in V . \end{aligned}$$

Under the assumptions (3.15), (3.22) the continuity of a and b, (3.25)) and ((3.24) can be proven analogously to **(II)**.

Conditions for V-Ellipticity of a

For the verification of the V-ellipticity we again proceed similarly to **(II)**, but now the boundary conditions are more complicated. Here we have for the convective term

$$\int_\Omega c \cdot \nabla u\, u\, dx = -\frac{1}{2} \int_\Omega \nabla \cdot c\, u^2\, dx + \frac{1}{2} \int_{\Gamma_1 \cup \Gamma_2} \nu \cdot c u^2\, d\sigma ,$$

and therefore

$$\begin{aligned} a(u,u) &= \int_\Omega \left\{ K\nabla u \cdot \nabla u + \left(r - \frac{1}{2} \nabla \cdot c \right) u^2 \right\} dx \\ &\quad + \frac{1}{2} \int_{\Gamma_1} \nu \cdot c\, u^2\, d\sigma + \int_{\Gamma_2} \left(\alpha + \frac{1}{2} \nu \cdot c \right) u^2\, d\sigma . \end{aligned}$$

In order to ensure the V-ellipticity of a we need, besides the obvious conditions

$$\nu \cdot c \geq 0 \quad \text{on } \Gamma_1 \quad \text{and} \quad \alpha + \frac{1}{2}\nu \cdot c \geq 0 \quad \text{on } \Gamma_2\,, \tag{3.32}$$

the following corollary from Theorem 3.13.

Corollary 3.14 *Suppose $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain and $\tilde{\Gamma} \subset \partial\Omega$ has a positive $(d-1)$-dimensional measure. Then there exists some constant $C_F > 0$ such that for all $v \in H^1(\Omega)$ with $v|_{\tilde{\Gamma}} = 0$,*

$$\|v\|_0 \leq C_F \left\{ \int_\Omega |\nabla v|^2 \, dx \right\}^{1/2} = C_F |v|_1\,.$$

This corollary yields the same results as in the case of homogeneous Dirichlet boundary conditions on the whole of $\partial\Omega$.

If $|\Gamma_3|_{d-1} = 0$, then by tightening conditions (3.32) for c and α, the application of Theorem 3.13 as done in **(II)** may be successful.

Summary

We will now present a summary of our considerations for the case of homogeneous Dirichlet boundary conditions.

Theorem 3.15 *Suppose $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain. Under the assumptions* (3.15), (3.16), (3.22) *with $g_3 = 0$, the boundary value problem* (3.12), (3.18)–(3.20) *has one and only one* weak solution $u \in V$, *if*

(1) $r - \frac{1}{2}\nabla \cdot c \geq 0$ *in* Ω.

(2) $\nu \cdot c \geq 0$ *on* Γ_1.

(3) $\alpha + \frac{1}{2}\nu \cdot c \geq 0$ *on* Γ_2.

(4) Additionally, one of the following conditions is satisfied:

- *(a)* $|\Gamma_3|_{d-1} > 0$.
- *(b) There exists some* $\tilde{\Omega} \subset \Omega$ *with* $|\tilde{\Omega}|_d > 0$ *and* $r_0 > 0$ *such that* $r - \frac{1}{2}\nabla \cdot c \geq r_0$ *on* $\tilde{\Omega}$.
- *(c) There exists some* $\tilde{\Gamma}_1 \subset \Gamma_1$ *with* $|\tilde{\Gamma}_1|_{d-1} > 0$ *and* $c_0 > 0$ *such that* $\nu \cdot c \geq c_0$ *on* $\tilde{\Gamma}_1$.
- *(d) There exists some* $\tilde{\Gamma}_2 \subset \Gamma_2$ *with* $|\tilde{\Gamma}_2|_{d-1} > 0$ *and* $\alpha_0 > 0$ *such that* $\alpha + \frac{1}{2}\nu \cdot c \geq \alpha_0$ *on* $\tilde{\Gamma}_2$.

Remark 3.16 We point out that by using different techniques in the proof, it is possible to weaken conditions (4)(b)–(d) in such a way that only the following has to be assumed:

(*b*) $\left|\left\{x \in \Omega : r - \frac{1}{2}\nabla \cdot c > 0\right\}\right|_d > 0$,

(*c*) $\left|\left\{x \in \Gamma_1 : \nu \cdot c > 0\right\}\right|_{d-1} > 0$,

(*d*) $\left|\left\{x \in \Gamma_2 : \alpha + \frac{1}{2}\nu \cdot c > 0\right\}\right|_{d-1} > 0$.

However, we stress that the conditions of Theorem 3.15 are only sufficient, since concerning the V-ellipticity, it might also be possible to balance an indefinite addend by some "particular definite" addend. But this would require conditions in which the constants C_P and C_F are involved.

Note that the pure Neumann problem for the Poisson equation

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega\,, \\ \partial_\nu u &= g \quad \text{on } \partial\Omega \end{aligned} \tag{3.33}$$

is excluded by the conditions of Theorem 3.15. This is consistent with the fact that not always a solution of (3.33) exists, and if a solution exists, it obviously is not unique (see Exercise 3.8).

Before we investigate inhomogeneous Dirichlet boundary conditions, the application of the theorem will be illustrated by an example of a natural situation described in Chapter 0.

For the linear stationary case of the differential equation (0.33) in the form

$$\nabla \cdot (c\,u - K\nabla u) + \tilde{r}\,u = f$$

we obtain, by differentiating and rearranging the convective term,

$$-\nabla \cdot (K\nabla u) + c \cdot \nabla u + (\nabla \cdot c + \tilde{r})\,u = f\,,$$

which gives the form (3.12) with $r := \nabla \cdot c + \tilde{r}$. The boundary $\partial\Omega$ consists only of two parts Γ_1 and Γ_2. Therein, Γ_1 an *outflow boundary* and Γ_2 an *inflow boundary;* that is, the conditions

$$c \cdot \nu \geq 0 \quad \text{on } \Gamma_1 \quad \text{and} \quad c \cdot \nu \leq 0 \quad \text{on } \Gamma_2$$

hold. Frequently prescribed boundary conditions are

$$\begin{aligned} -(c\,u - K\nabla u) \cdot \nu &= -\nu \cdot c\,u \quad \text{on } \Gamma_1\,, \\ -(c\,u - K\nabla u) \cdot \nu &= g_2 \qquad\quad \text{on } \Gamma_2\,. \end{aligned}$$

They are based on the following assumptions: On the inflow boundary Γ_2 the normal component of the total (mass) flux is prescribed but on the outflow boundary Γ_1, on which in the extreme case $K = 0$ the boundary conditions would drop out, only the following is required:

- the normal component of the total (mass) flux is continuous over Γ_1,
- the ambient mass flux that is outside Ω consists only of a convective part,
- the extensive variable (for example, the concentration) is continuous over Γ_1, that is, the ambient concentration in x is also equal to $u(x)$.

Therefore, after an obvious reformulation we get, in accordance with the definitions of Γ_1 and Γ_2 due to (3.18), (3.19), the Neumann boundary

condition (3.18), and the mixed boundary condition (3.19),

$$\begin{aligned} K\nabla u \cdot \nu &= 0 \quad \text{on } \Gamma_1\,, \\ K\nabla u \cdot \nu + \alpha\, u &= g_2 \quad \text{on } \Gamma_2\,, \end{aligned}$$

where $\alpha := -\nu \cdot c$.

Now the conditions of Theorem 3.15 can be checked:

We have $r - \frac{1}{2}\nabla\cdot c = \tilde{r} + \frac{1}{2}\nabla\cdot c$; therefore, for the latter term the inequality in (1) and (4)(b) must be satisfied. Further, the condition $\nu \cdot c \geq 0$ on Γ_1 holds due to the characterization of the outflow boundary. Because of $\alpha + \frac{1}{2}\nu \cdot c = -\frac{1}{2}\nu \cdot c$, the condition (3) is satisfied due to the definition of the inflow boundary.

Now we address the case of **inhomogeneous Dirichlet boundary conditions** ($|\Gamma_3|_{d-1} > 0$).

This situation can be reduced to the case of homogeneous Dirichlet boundary conditions, if we are able to choose some (fixed) element $w \in H^1(\Omega)$ in such a way that (in the sense of trace) we have

$$\gamma_0(w) = g_3 \quad \text{on } \Gamma_3\,. \tag{3.34}$$

The existence of such an element w is a necessary assumption for the existence of a solution $\tilde{u} \in H^1(\Omega)$. On the other hand, such an element w can exist only if g_3 belongs to the range of the mapping

$$H^1(\Omega) \ni v \;\mapsto\; \gamma_0(v)|_{\Gamma_3} \in L^2(\Gamma_3).$$

However, this is not valid for all $g_3 \in L^2(\Gamma_3)$, since the range of the trace operator of $H^1(\Omega)$ is a proper subset of $L^2(\partial\Omega)$.

Therefore, we assume the existence of such an element w. Since only the homogeneity of the Dirichlet boundary conditions of the test functions plays a role in derivation (3.31) of the bilinear form a and the linear form b, we first obtain with the space V, defined in (3.30), and

$$\tilde{V} := \big\{v \in H^1(\Omega) :\, \gamma_0(v) = g_3 \text{ on } \Gamma_3\big\} = \big\{v \in H^1(\Omega) : v - w \in V\big\}$$

the following variational formulation:

Find $\tilde{u} \in \tilde{V}$ such that

$$a(\tilde{u}, v) = b(v) \quad \text{for all } v \in V\,.$$

However, this formulation does not fit into the theoretical concept of Section 3.1 since the space $\tilde{V}$ is not a linear one.

If we put $\tilde{u} := u + w$, then this is equivalent to the following:

Find $u \in V$ such that

$$a(u, v) = b(v) - a(w, v) =: \tilde{b}(v) \quad \text{for all } v \in V\,. \tag{3.35}$$

Now we have a variational formulation for the case of inhomogeneous Dirichlet boundary conditions that has the form required in the theory.

Remark 3.17 In the existence result of Theorem 3.1, the only assumption is that b has to be a continuous linear form in V.

For $d = 1$ and $\Omega = (a, b)$ this is also satisfied, for instance, for the special linear form

$$\delta_\gamma(v) := v(\gamma) \quad \text{for } v \in H^1(a, b),$$

where $\gamma \in (a, b)$ is arbitrary but fixed, since by Lemma 3.4 the space $H^1(a, b)$ is continuously embedded in the space $C[a, b]$. Thus, for $d = 1$ point sources ($b = \delta_\gamma$) are also allowed. However, for $d \geq 2$ this does not hold since $H^1(\Omega) \not\subset C(\bar{\Omega})$.

Finally, we will once again state the **general assumptions** under which the variational formulation of the boundary value problem (3.12), (3.18)–(3.20) in the space (3.30),

$$V = \left\{ v \in H^1(\Omega) : \gamma_0(v) = 0 \text{ on } \Gamma_3 \right\},$$

has properties that satisfy the conditions of the Lax–Milgram Theorem (Theorem 3.1):

- $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain.
- $k_{ij}, c_i, \nabla \cdot c, r \in L^\infty(\Omega)$, $f \in L^2(\Omega)$, $i, j \in \{1, \ldots, d\}$, and, if $|\Gamma_1 \cup \Gamma_2|_{d-1} > 0$, $\nu \cdot c \in L^\infty(\Gamma_1 \cup \Gamma_2)$ (i.e., (3.15)).
- There exists some constant $k_0 > 0$ such that in Ω, we have $\xi \cdot K(x)\xi \geq k_0|\xi|^2$ for all $\xi \in \mathbb{R}^d$ (i.e., (3.16)),
- $g_j \in L^2(\Gamma_j)$, $j = 1, 2, 3$, $\alpha \in L^\infty(\Gamma_2)$ (i.e., (3.22)).
- The following hold:
  (1) $r - \frac{1}{2}\nabla \cdot c \geq 0$ in Ω.
  (2) $\nu \cdot c \geq 0$ on Γ_1.
  (3) $\alpha + \frac{1}{2}\nu \cdot c \geq 0$ on Γ_2.
  (4) Additionally, one of the following conditions is satisfied:
    (a) $|\Gamma_3|_{d-1} > 0$.
    (b) There exists some $\tilde{\Omega} \subset \Omega$ with $|\tilde{\Omega}|_d > 0$ and $r_0 > 0$ such that $r - \frac{1}{2}\nabla \cdot c \geq r_0$ on $\tilde{\Omega}$.
    (c) There exists some $\tilde{\Gamma}_1 \subset \Gamma_1$ with $|\tilde{\Gamma}_1|_{d-1} > 0$ and $c_0 > 0$ such that $\nu \cdot c \geq c_0$ on $\tilde{\Gamma}_1$.
    (d) There exists some $\tilde{\Gamma}_2 \subset \Gamma_2$ with $|\tilde{\Gamma}_2|_{d-1} > 0$ and $\alpha_0 > 0$ such that $\alpha + \frac{1}{2}\nu \cdot c \geq \alpha_0$ on $\tilde{\Gamma}_2$.
- If $|\Gamma_3|_{d-1} > 0$, then there exists some $w \in H^1(\Omega)$ with $\gamma_0(w) = g_3$ on Γ_3 (i.e., (3.34)).

3.2.2 An Example of a Boundary Value Problem of Fourth Order

The Dirichlet problem for the *biharmonic equation* reads as follows:

Find $u \in C^4(\Omega) \cap C^1(\bar{\Omega})$ such that

$$\left\{ \begin{array}{rcl} \Delta^2 u & = & f \quad \text{in } \Omega\,, \\ \partial_\nu u & = & u = 0 \quad \text{on } \partial\Omega\,, \end{array} \right. \tag{3.36}$$

where

$$\Delta^2 u := \Delta\,(\Delta u) = \sum_{i,j=1}^{d} \partial_i^2\,\left(\partial_j^2 u\right).$$

In the case $d = 1$ this collapses to $\Delta^2 u = u^{(4)}$.

For $u, v \in H^2(\Omega)$ it follows from Corollary 3.9 that

$$\int_\Omega (u\,\Delta v - \Delta u\, v)\; dx = \int_{\partial\Omega} \{u\,\partial_\nu v - \partial_\nu u\, v\} d\sigma$$

and hence for $u \in H^4(\Omega)$, $v \in H^2(\Omega)$ (by replacing u with Δu in the above equation),

$$\int_\Omega \Delta u\,\Delta v\, dx = \int_\Omega \Delta^2 u\, v\, dx - \int_{\partial\Omega} \partial_\nu \Delta u\, v\, d\sigma + \int_{\partial\Omega} \Delta u\,\partial_\nu v\, d\sigma\,.$$

For a Lipschitz domain Ω we define

$$H_0^2(\Omega) := \left\{ v \in H^2(\Omega) \;\middle|\; v = \partial_\nu v = 0 \text{ on } \partial\Omega \right\}$$

and obtain the variational formulation of (3.36) in the space $V := H_0^2(\Omega)$:

Find $u \in V$, such that

$$a(u, v) := \int_\Omega \Delta u\,\Delta v\, dx = b(v) := \int_\Omega f v\, dx \quad \text{for all } v \in V\,.$$

More general, for a boundary value problem of order $2m$ in conservative form, we obtain a variational formulation in $H^m(\Omega)$ or $H_0^m(\Omega)$.

3.2.3 Regularity of Boundary Value Problems

In Section 3.2.1 we stated conditions under which the linear elliptic boundary value problem admits a unique solution u ($\tilde{u}$, respectively) in some subspace V of $H^1(\Omega)$. In many cases, for instance for the interpolation of the solution or in the context of error estimates (also in norms other than the $\|\cdot\|_V$ norm) it is not sufficient that u ($\tilde{u}$, respectively) have only first weak derivatives in $L^2(\Omega)$.

Therefore, within the framework of the so-called regularity theory, the question of the assumptions under which the weak solution belongs to $H^2(\Omega)$, for instance, has to be answered. These additional conditions contain conditions about

- the smoothness of the boundary of the domain,
- the shape of the domain,
- the smoothness of the coefficients and the right-hand side of the differential equation and the boundary conditions,
- the kind of the transition of boundary conditions in those points, where the type is changing,

which can be quite restrictive as a whole. Therefore, in what follows we often assume only the required smoothness. Here we cite as an example one regularity result ([13, Theorem 8.12]).

Theorem 3.18 *Suppose Ω is a bounded C^2-domain and $\Gamma_3 = \partial\Omega$. Further, assume that $k_{ij} \in C^1(\bar{\Omega})$, $c_i, r \in L^\infty(\Omega)$, $f \in L^2(\Omega)$, $i,j \in \{1,\dots,d\}$, as well as (3.16). Suppose there exists some function $w \in H^2(\Omega)$ with $\gamma_0(w) = g_3$ on Γ_3. Let $\tilde{u} = u + w$ and let u be a solution of (3.35). Then $\tilde{u} \in H^2(\Omega)$ and*

$$\|\tilde{u}\|_2 \le C\{\|u\|_0 + \|f\|_0 + \|w\|_2\}$$

with a constant $C > 0$ independent of u, f, and w.

One drawback of the above result is that it excludes polyhedral domains. If the convexity of Ω is additionally assumed, then it can be transferred to this case. Simple examples of boundary value problems in domains with reentrant corners show that one cannot avoid such additional assumptions (see Exercise 3.5).

Exercises

3.5 Consider the boundary value problem (1.1), (1.2) for $f = 0$ in the sector $\Omega := \{(x,y) \in \mathbb{R}^2 \mid x = r\cos\varphi,\ y = r\sin\varphi \text{ with } 0 < r < 1,\ 0 < \varphi < \alpha\}$ for some $0 < \alpha < 2\pi$, thus with the interior angle α. Derive as in (1.23), by using the ansatz $w(z) := z^{1/\alpha}$, a solution $u(x,y) = \Im w(x+iy)$ for an appropriate boundary function g. Then check the regularity of u, that is, $u \in H^k(\Omega)$, in dependence of α.

3.6 Consider the problem (1.29) with the transmission condition (1.30) and, for example, Dirichlet boundary conditions and derive a variational formulation for this.

3.7 Consider the variational formulation:
Find $u \in H^1(\Omega)$ such that

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx + \int_{\partial\Omega} g v \, d\sigma \quad \text{for all } v \in H^1(\Omega), \tag{3.37}$$

where Ω is a bounded Lipschitz domain, $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$.

(a) Let $u \in H^1(\Omega)$ be a solution of this problem. Show that $-\Delta u$ exists in the weak sense in $L^2(\Omega)$ and

$$-\Delta u = f\,.$$

(b) If additionally $u \in H^2(\Omega)$, then $\partial_\nu u|_{\partial\Omega}$ exists in the sense of trace in $L^2(\partial\Omega)$ and

$$\partial_\nu u = g$$

where this equality is to be understood as

$$\int_{\partial\Omega} (\partial_\nu u - g) v \, d\sigma = 0 \quad \text{for all } v \in H^1(\Omega)\,.$$

3.8 Consider the variational equation (3.37) for the Neumann problem for the Poisson equation as in Exercise 3.7.

(a) If a solution $u \in H^1(\Omega)$ exists, then the compatibility condition

$$\int_\Omega f \, dx + \int_{\partial\Omega} g \, d\sigma = 0 \tag{3.38}$$

has to be fulfilled.

(b) Consider the following bilinear form on $H^1(\Omega)$:

$$\tilde{a}(u,v) := \int_\Omega \nabla u \cdot \nabla v \, dx + \left(\int_\Omega u \, dx\right)\left(\int_\Omega v \, dx\right)\,.$$

Show that $\tilde{a}$ is V-elliptic on $H^1(\Omega)$.
Hint: Do it by contradiction using the fact that a bounded sequence in $H^1(\Omega)$ possesses a subsequence converging in $L^2(\Omega)$ (see, e.g., [37]).

(c) Consider the unique solution $\tilde{u} \in H^1(\Omega)$ of

$$\tilde{a}(u,v) = \int_\Omega f v \, dx + \int_{\partial\Omega} g v \, d\sigma \quad \text{for all } v \in H^1(\Omega)\,.$$

Then:

$$|\Omega| \int_\Omega \tilde{u} \, dx = \int_\Omega f \, dx + \int_{\partial\Omega} g \, d\sigma\,.$$

Furthermore, if (3.38) is valid, then $\tilde{u}$ is a solution of (3.37) (with $\int_\Omega \tilde{u} \, dx = 0$).

3.9 Show analogously to Exercise 3.7: A weak solution $u \in V \subset H^1(\Omega)$ of (3.31), where V is defined in (3.30), with data satisfying (3.14) and (3.22), fulfills a differential equation in $L^2(\Omega)$. The boundary conditions are fulfilled in the following sense:

$$\int_{\Gamma_1} \partial_{\nu_K} u \, v \, d\sigma + \int_{\Gamma_2} (\partial_{\nu_K} u + \alpha\, u) v \, d\sigma = \int_{\Gamma_1} g_1 v \, d\sigma + \int_{\Gamma_2} g_2 v \, d\sigma \ \text{ for all } v \in V\,.$$

3.3 Element Types and Affine Equivalent Triangulations

In order to be able to exploit the theory developed in Sections 3.1 and 3.2 we make the assumption that Ω is a Lipschitz domain.

The finite element discretization of the boundary value problem (3.12) with the boundary conditions (3.18)–(3.20) corresponds to performing a Galerkin approximation (cf. (2.23)) of the variational equation (3.35) with the bilinear form a and the linear form b, supposed to be defined as in (3.31), and some $w \in H^1(\Omega)$ with the property $w = g_3$ on Γ_3. The solution of the weak formulation of the boundary value problem is then given by $\tilde{u} := u + w$, if u denotes the solution of the variational equation (3.35).

Since the bilinear form a is in general not symmetric, (2.21) and (2.23), respectively (the variational equation), are no longer equivalent to (2.22) and (2.24), respectively (the minimization problem), so that in the following we pursue only the first, more general, ansatz.

The Galerkin approximation of the variational equation (3.35) reads as follows: Find some $u \in V_h$ such that

$$a(u_h, v) = b(v) - a(w, v) = \tilde{b}(v) \quad \text{for all } v \in V_h\,. \tag{3.39}$$

The space V_h that is to be defined has to satisfy $V_h \subset V$. Therefore, we speak of a *conforming* finite element discretization, whereas for a *nonconforming* discretization this property, for instance, can be violated. The ansatz space is defined piecewise with respect to a triangulation $\mathcal{T}_h$ of Ω with the goal of getting small supports for the basis functions. A triangulation in two space dimensions consisting of triangles has already been defined in definition (2.25). The generalization in d space dimensions reads as follows:

Definition 3.19 A *triangulation* $\mathcal{T}_h$ of a set $\Omega \subset \mathbb{R}^d$ consists of a finite number of subsets K of Ω with the following properties:

- **(T1)** Every $K \in \mathcal{T}_h$ is closed.
- **(T2)** For every $K \in \mathcal{T}_h$ its nonempty interior $\mathrm{int}\,(K)$ is a Lipschitz domain.
- **(T3)** $\overline{\Omega} = \cup_{K \in \mathcal{T}_h} K$.
- **(T4)** For different K_1 and K_2 of $\mathcal{T}_h$ the intersection of $\mathrm{int}\,(K_1)$ and $\mathrm{int}\,(K_2)$ is empty.

The sets $K \in \mathcal{T}_h$, which are called somewhat inaccurately *elements* in the following, form a nonoverlapping decomposition of $\overline{\Omega}$. Here the formulation is chosen in such a general way, since in Section 3.8 elements with curved boundaries will also be considered. In Definition 3.19 some condition, which corresponds to the property (3) of definition (2.25), is still missing. In the following this will be formulated specifically for each element type. The

parameter h is a measure for the size of all elements and mostly chosen as

$$h = \max \left\{ \operatorname{diam}(K) \;\middle|\; K \in \mathcal{T}_h \right\} ;$$

that is, for instance, for triangles h is the length of the triangle's largest edge.

For a given vector space V_h let

$$P_K := \{ v|_K \mid v \in V_h \} \quad \text{for } K \in \mathcal{T}_h \,, \tag{3.40}$$

that is,

$$V_h \subset \left\{ v : \Omega \to \mathbb{R} \;\middle|\; v|_K \in P_K \text{ for all } K \in \mathcal{T}_h \right\} .$$

In the example of "linear triangles" in (2.27) we have $P_K = \mathcal{P}_1$, the polynomials of first order. In the following definitions the space P_K will always consist of polynomials or of smooth "polynomial-like" functions, such that we can assume $P_K \subset H^1(K) \cap C(K)$. Here, $H^1(K)$ is an abbreviation for $H^1(\operatorname{int}(K))$. The same holds for similar notation.

As the following theorem shows, elements $v \in V_h$ of a conforming ansatz space $V_h \subset V$ have therefore to be continuous :

Theorem 3.20 *Suppose $P_K \subset H^1(K) \cap C(K)$ for all $K \in \mathcal{T}_h$. Then*

$$V_h \subset C(\bar{\Omega}) \iff V_h \subset H^1(\Omega)$$

and, respectively, for $V_{0h} := \left\{ v \in V_h \;\middle|\; v = 0 \text{ on } \partial\Omega \right\}$,

$$V_{0h} \subset C(\bar{\Omega}) \iff V_{0h} \subset H^1_0(\Omega) \,.$$

Proof: See, for example, [9, Theorem 5.1 (p. 62)] or also Exercise 3.10. □

If $V_h \subset C(\bar{\Omega})$, then we also speak of C^0-*elements*. Hence with this notion we do not mean only the $K \in \mathcal{T}_h$, but these provided with the local ansatz space P_K (and the degrees of freedom still to be introduced). For a boundary value problem of fourth order, $V_h \subset H^2(\Omega)$ and hence the requirement $V_h \subset C^1(\bar{\Omega})$ are necessary for a conforming finite element ansatz. Therefore, this requires, analogously to Theorem 3.20, so-called C^1-*elements*. By *degrees of freedom* we denote a finite number of values that are obtained for some $v \in P_K$ from evaluating linear functionals on P_K. The set of these functionals is denoted by Σ_K. In the following, these will basically be the function values in fixed points of the element K, as in the example of (2.27). We refer to these points as *nodes*. (Sometimes, this term is used only for the vertices of the elements, which at least in our examples are always nodes.) If the degrees of freedom are only function values, then we speak of *Lagrange elements* and specify Σ by the corresponding nodes of the element. Other possible degrees of freedom are values of derivatives in fixed nodes or also integrals. Values of derivatives are necessary if we want to obtain C^1-elements.

As in the example of (2.27) (cf. Lemma 2.10), V_h is defined by specifying P_K and the degrees of freedom on K for $K \in \mathcal{T}_h$. These have to be chosen such that, on the one hand, they enforce the continuity of $v \in V_h$ and, on the other hand, the satisfaction of the homogeneous Dirichlet boundary conditions at the nodes. For this purpose, compatibility between the Dirichlet boundary condition and the triangulation is necessary, since it will be required in (T6).

As can be seen from the proof of Lemma 2.10, it is essential

(F1) that the interpolation problem, locally defined on $K \in \mathcal{T}_h$ by the degrees of freedom, is uniquely solvable in P_K, (3.41)

(F2) that this also holds on the $(d-1)$-dimensional boundary surfaces F of $K \in \mathcal{T}_h$ for the degrees of freedom from F and the functions $v|_F$ where $v \in P_K$; this then ensures the continuity of $v \in V_h$, if P_K and $P_{K'}$ match in the sense of $P_K|_F = P_{K'}|_F$ for $K, K' \in \mathcal{T}_h$ intersecting in F (see Figure 3.2). (3.42)

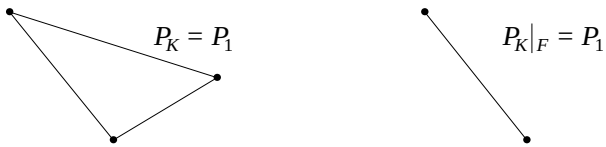


Figure 3.2. Compatibility of the ansatz space on the boundary surface and the degrees of freedom there.

The following *finite elements* defined by their basic domain $K(\in \mathcal{T}_h)$, the local ansatz space P_K, and the degrees of freedom Σ_K satisfy these properties.

For this, let $\mathcal{P}_k(K)$ be the set of mappings $p : K \to \mathbb{R}$ of the following form:

$$p(x) = p(x_1, \ldots, x_d) = \sum_{|\alpha| \le k} \gamma_{\alpha_1 \ldots \alpha_d} x_1^{\alpha_1} \cdots x_d^{\alpha_d} = \sum_{|\alpha| \le k} \gamma_\alpha x^\alpha \,, \tag{3.43}$$

hence the polynomials of order k in d variables. The set $\mathcal{P}_k(K)$ forms a vector space, and since $p \in \mathcal{P}_k(K)$ is differentiable arbitrarily often, $\mathcal{P}_k(K)$ is a subset of all function spaces introduced so far (provided that the boundary conditions do not belong to their definition).

For both, $K \in \mathcal{T}_h$ and $K = \mathbb{R}^d$ we have

$$\dim \mathcal{P}_k(K) = \dim \mathcal{P}_k(\mathbb{R}^d) = \binom{d+k}{k} \,, \tag{3.44}$$

as even $\mathcal{P}_k(\mathbb{R}^d)|_K = \mathcal{P}_k(K)$ (see Exercise 3.12). Therefore, for short we will use the notation $\mathcal{P}_1 = \mathcal{P}_1(K)$ if the dimension of the basic space is fixed.

We start with *simplicial finite elements*, that is, elements whose basic domain is a regular d-simplex of $\mathbb{R}^d$. By this we mean the following:

Definition 3.21 A set $K \subset \mathbb{R}^d$ is called a *regular d-simplex* if there exist $d+1$ distinct points $a_1, \ldots, a_{d+1} \in \mathbb{R}^d$, the vertices of K, such that

$$a_2 - a_1, \ldots, a_{d+1} - a_1 \qquad \text{are linearly independent} \tag{3.45}$$

(that is, $a_1, \ldots, a_{d+1}$ do not lie in a hyperplane) and

$$\begin{aligned} K &= \operatorname{conv}\{a_1, \ldots, a_{d+1}\} \\ &:= \left\{ x = \sum_{i=1}^{d+1} \lambda_i a_i \;\middle|\; 0 \le \lambda_i (\le 1)\,,\ \sum_{i=1}^{d+1} \lambda_i = 1 \right\} \\ &= \left\{ x = a_1 + \sum_{i=2}^{d+1} \lambda_i (a_i - a_1) \;\middle|\; \lambda_i \ge 0\,,\ \sum_{i=2}^{d+1} \lambda_i \le 1 \right\}. \end{aligned} \tag{3.46}$$

A *face* of K is a $(d-1)$-simplex defined by d points of $\{a_1, \ldots, a_{d+1}\}$. The particular d-simplex

$$\hat{K} := \operatorname{conv}\{\hat{a}_1, \ldots, \hat{a}_{d+1}\} \text{ with } \hat{a}_1 = 0\,,\ \hat{a}_{i+1} = e_i\,,\ i = 1, \ldots, d\,, \tag{3.47}$$

is called the *standard simplicial reference element*.

In the case $d = 2$ we get a triangle with $\dim \mathcal{P}_1 = 3$ (cf. Lemma 2.10). The faces are the 3 edges of the triangle. In the case $d = 3$ we get a tetrahedron with $\dim \mathcal{P}_1 = 4$, the faces are the 4 triangle surfaces, and finally, in the case $d = 1$ it is a line segment with $\dim \mathcal{P}_1 = 2$ and the two boundary points as faces.

More precisely, a face is not interpreted as a subset of $\mathbb{R}^d$, but of a $(d-1)$-dimensional space that, for instance, is spanned by the vectors $a_2 - a_1, \ldots, a_d - a_1$ in the case of the defining points $a_1, \ldots, a_d$.

Sometimes, we also consider *degenerate d-simplices*, where the assumption (3.45) of linear independence is dropped. We consider, for instance, a line segment in the two-dimensional space as it arises as an edge of a triangular element. In the one-dimensional parametrisation it is a regular 1-simplex, but in $\mathbb{R}^2$ a degenerate 2-simplex.

The unique coefficients $\lambda_i = \lambda_i(x)$, $i = 1, \ldots, d+1$, in (3.46), are called *barycentric coordinates* of x. This defines mappings $\lambda_i : K \to \mathbb{R}$, $i = 1, \ldots, d+1$.

We consider a_j as a column of a matrix; that is, for $j = 1, \ldots, d$, $a_j = (a_{ij})_{i=1,\ldots,d}$. The defining conditions for $\lambda_i = \lambda_i(x)$ can be written as a $(d+1) \times (d+1)$ system of equations:

$$\left.\begin{array}{rcl} \sum\limits_{j=1}^{d+1} a_{ij}\lambda_j &=& x_i \\ \sum\limits_{j=1}^{d+1} \lambda_j &=& 1 \end{array}\right\} \Leftrightarrow B\lambda = \begin{pmatrix} x \\ 1 \end{pmatrix} \tag{3.48}$$

for

$$B = \begin{pmatrix} a_{11} & \cdots & a_{1,d+1} \\ \vdots & \ddots & \vdots \\ a_{d1} & \cdots & a_{d,d+1} \\ 1 & \cdots & 1 \end{pmatrix}. \tag{3.49}$$

The matrix B is nonsingular due to assumption (3.45); that is, $\lambda(x) = B^{-1}\binom{x}{1}$, and hence

$$\lambda_i(x) = \sum_{j=1}^{d} c_{ij}x_j + c_{i,d+1} \quad \text{for all} \quad i = 1, \ldots, d+1\,,$$

where $C = (c_{ij})_{ij} := B^{-1}$.

Consequently, the λ_i are affine-linear, and hence $\lambda_i \in \mathcal{P}_1$. The level surfaces $\{x \in K \mid \lambda_i(x) = \mu\}$ correspond to intersections of hyperplanes with the simplex K (see Figure 3.3). The level surfaces for distinct μ_1 and μ_2 are parallel to each other, that is, in particlular, to the level surface for $\mu = 0$, which corresponds to the triangle face spanned by all the vertices apart of a_i.

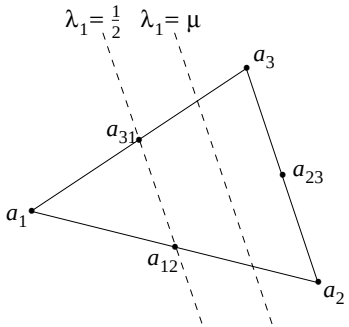


Figure 3.3. Barycentric coordinates and hyperplanes.

By (3.48), the barycentric coordinates can be defined for arbitrary $x \in \mathbb{R}^d$ (with respect to some fixed d-simplex K). Then

$$x \in K \iff 0 \le \lambda_i(x) \le 1 \quad \text{for all } i = 1, \ldots, d+1\,.$$

Applying Cramer's rule to the system $B\lambda = \binom{x}{1}$, we get for the ith barycentric coordinate

$$\lambda_i(x) = \frac{1}{\det(B)} \det \begin{pmatrix} a_{11} & \cdots & x_1 & \cdots & a_{1,d+1} \\ \vdots & & \vdots & & \vdots \\ a_{d1} & \cdots & x_d & \cdots & a_{d,d+1} \\ 1 & \cdots & 1 & \cdots & 1 \end{pmatrix}.$$

Here, in the ith column a_i has been replaced with x. Since in general,

$$\mathrm{vol}\,(K) = \mathrm{vol}\,(\hat{K})\,|\det(B)| \tag{3.50}$$

for the reference simplex $\hat{K}$ defined by (3.47) (cf. (2.50)), we have for the volume of the d-simplex $K = \mathrm{conv}\,\{a_1, \ldots, a_{d+1}\}$,

$$\mathrm{vol}\,(K) = \frac{1}{d!}\left|\det\begin{pmatrix} a_{11} & \cdots & a_{1,d+1} \\ \vdots & \ddots & \vdots \\ a_{d1} & \cdots & a_{d,d+1} \\ 1 & \cdots & 1 \end{pmatrix}\right|,$$

and from this,

$$\lambda_i(x) = \pm\frac{\mathrm{vol}\,(\mathrm{conv}\,\{a_1, \ldots, x, \ldots, a_{d+1}\})}{\mathrm{vol}\,(\mathrm{conv}\,\{a_1, \ldots, a_i, \ldots, a_{d+1}\})}\,. \tag{3.51}$$

The sign is determined by the arrangement of the coordinates.

In the case $d = 2$ for example, we have

$$\begin{aligned} & \mathrm{vol}\,(K) = \det(B)/2 \\ \iff \quad & a_1, a_2, a_3 \text{ are ordered positively (that is, counterclockwise).} \end{aligned}$$

Here, conv $\{a_1, \ldots, x, \ldots, a_{d+1}\}$ is the d-simplex that is generated by replacing a_i with x and is possibly degenerate if x lies on a face of K (then $\lambda_i(x) = 0$). Hence, in the case $d = 2$ we have for $x \in K$ that the barycentric coordinates $\lambda_i(x)$ are the relative areas of the triangles that are spanned by x and the vertices other than a_i. Therefore, we also speak of *surface coordinates* (see Figure 3.4). Analogous interpretations hold for $d = 3$. Using the barycentric coordinates, we can now easily specify points that admit a geometric characterization. The midpoint $a_{ij} := \frac{1}{2}(a_i + a_j)$ of a line segment that is given by a_i and a_j satisfies, for instance,

$$\lambda_i(x) = \lambda_j(x) = \frac{1}{2}\,.$$

By the *barycentre* of a d-simplex we mean

$$a_S := \frac{1}{d+1}\sum_{i=1}^{d+1} a_i\,; \quad \text{thus} \quad \lambda_i(a_S) = \frac{1}{d+1} \quad \text{for all } i = 1, \ldots, d+1\,. \tag{3.52}$$

A geometric interpretation follows directly from the above considerations.

In the following suppose conv $\{a_1, \ldots, a_{d+1}\}$ to be a regular d-simplex. We make the following definition:

Finite Element: Linear Ansatz on the Simplex

$$\begin{aligned} K &= \mathrm{conv}\,\{a_1, \ldots, a_{d+1}\}\,, \\ P &= \mathcal{P}_1(K)\,, \\ \Sigma &= \{p\,(a_i)\,,\, i = 1, \ldots, d+1\}\,. \end{aligned} \tag{3.53}$$

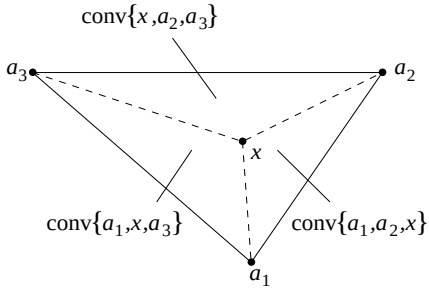


Figure 3.4. Barycentric coordinates as surface coordinates.

The *local interpolation problem in* P, *given by the degrees of freedom* Σ, namely,

$$\begin{aligned} \text{find some} \quad & p \in P \text{ for } u_1, \ldots, u_{d+1} \in \mathbb{R} \text{ such that} \\ & p(a_i) = u_i \quad \text{for all} \quad i = 1, \ldots, d+1\,, \end{aligned}$$

can be interpreted as the question of finding the inverse image of a linear mapping from P to $\mathbb{R}^{|\Sigma|}$. By virtue of (3.44),

$$|\Sigma| = d + 1 = \dim P\,.$$

Since both vector spaces have the same dimension, the solvability of the interpolation problem is equivalent to the uniqueness of the solution. This consideration holds independently of the type of the degrees of freedom (as far as they are linear functionals on P). Therefore, we need only to ensure the solvability of the interpolation problem. This is obtained by specifying

$$N_1, \ldots, N_{d+1} \in P \quad \text{with } N_i\,(a_j) = \delta_{ij} \quad \text{for all } i, j = 1, \ldots, d+1\,,$$

the so-called *shape functions* (see (2.29) for $d = 2$). Then the solution of the interpolation problem is given by

$$p(x) = \sum_{i=1}^{d+1} u_i N_i(x) \tag{3.54}$$

and analogously in the following; that is, the shape functions form a basis of P and the coefficients in the representation of the interpolating function are exactly the degrees of freedom $u_1, \ldots, u_{d+1}$.

Due to the above considerations, the specification of the shape functions can easily be done by choosing

$$N_i = \lambda_i\,.$$

Finite Element: Quadratic Ansatz on the Simplex
Here, we have

$$K \quad = \quad \text{conv } \{a_1, \ldots, a_{d+1}\}\,,$$

$$\begin{aligned} P &= \mathcal{P}_2(K)\,, \\ \Sigma &= \{p\,(a_i)\,, p\,(a_{ij})\,, \quad i = 1, \ldots, d+1,\ i < j \le d+1\}\,, \end{aligned} \tag{3.55}$$

where the a_{ij} denote the midpoints of the edges (see Figure 3.5).

Since here we have

$$|\Sigma| = \frac{(d+1)(d+2)}{2} = \dim P\,,$$

it also suffices to specify the shape functions. They are given by

$$\begin{aligned} &\lambda_i\,(2\lambda_i - 1)\,, && i = 1, \ldots, d+1\,, \\ &4\lambda_i\lambda_j, && i, j = 1, \ldots, d+1\,,\ i < j\,. \end{aligned}$$

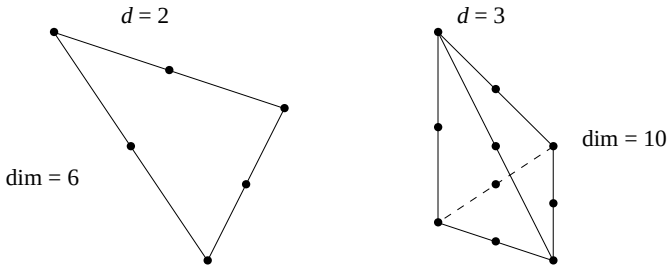


Figure 3.5. Quadratic simplicial elements.

If we want to have polynomials of higher degree as local ansatz functions, but still Lagrange elements, then degrees of freedom also arise in the interior of K:

Finite Element: Cubic Ansatz on the Simplex

$$\begin{aligned} K &= \operatorname{conv}\{a_1, \ldots, a_{d+1}\}\,, \\ P &= \mathcal{P}_3(K)\,, \\ \Sigma &= \{p(a_i), p(a_{i,i,j}), p(a_{i,j,k})\}\,, \end{aligned} \tag{3.56}$$

where

$$a_{i,i,j} := \frac{2}{3}a_i + \frac{1}{3}a_j \qquad \text{for} \quad i, j = 1, \ldots, d+1\,,\ i \ne j\,,$$

$$a_{i,j,k} := \frac{1}{3}(a_i + a_j + a_k) \qquad \text{for} \quad i, j, k = 1, \ldots, d+1\,,\ i < j < k\,.$$

Since here $|\Sigma| = \dim P$ also holds, it is sufficient to specify the shape functions, which is possible by

$$\begin{aligned} &\frac{1}{2}\lambda_i(3\lambda_i - 1)(3\lambda_i - 2), && i = 1, \ldots, d+1\,, \\ &\frac{9}{2}\lambda_i\lambda_j(3\lambda_i - 1), && i, j = 1, \ldots, d+1\,,\ i \ne j\,, \end{aligned}$$

$$27\lambda_i\lambda_j\lambda_k, \qquad i,j,k = 1,\dots,d+1\,,\ i<j<k\,.$$

Thus for $d = 2$ the value at the barycentre arises as a degree of freedom. This, and in general the $a_{i,j,k}, i < j < k$, can be dropped if the ansatz space P is reduced (see [9, p. 70]).

All finite elements discussed so far have degrees of freedom that are defined in convex combinations of the vertices. On the other hand, two regular d-simplices can be mapped bijectively onto each other by a unique affine-linear F, that is, $F \in \mathcal{P}_1$ such that as defining condition, the vertices of the simplices should be mapped onto each other. If we choose, besides the general simplex K, the standard reference element $\hat{K}$ defined by (3.47), then $F = F_K : \hat{K} \to K$ is defined by

$$F(\hat{x}) = B\hat{x} + a_1\,, \tag{3.57}$$

where $B = (a_2 - a_1, \dots, a_{d+1} - a_1)$.

Since for F we have

$$F\left(\sum_{i=1}^{d+1}\lambda_i\hat{a}_i\right) = \sum_{i=1}^{d+1}\lambda_i F(\hat{a}_i) \quad \text{for} \quad \lambda_i \geq 0\,,\ \sum_{i=1}^{d+1}\lambda_i = 1\,,$$

F is indeed a bijection that maps the degrees of freedom onto each other as well as the faces of the simplices. Since the ansatz spaces P and $\hat{P}$ remain invariant under the transformation F_K, the finite elements introduced so far are (in their respective classes) *affine equivalent* to each other and to the *reference element*.

Definition 3.22 Two Lagrange elements $(K, P, \Sigma), (\hat{K}, \hat{P}, \hat{\Sigma})$ are called *equivalent* if there exists a bijective $F : \hat{K} \to K$ such that

$$\begin{aligned} &\left\{F(\hat{a}) \;\middle|\; \hat{a} \in \hat{K} \text{ generates a degree of freedom on } \hat{K}\right\} \\ &\qquad = \left\{a \;\middle|\; a \in K \text{ generates a degree of freedom on } K\right\} \\ &\text{and} \\ &P \;=\; \left\{p : K \to \mathbb{R} \;\middle|\; p \circ F \in \hat{P}\right\}. \end{aligned} \tag{3.58}$$

They are called *affine equivalent* if F is affine-linear.

Here we have formulated the definition in a more general way, since in Section 3.8 elements with more general F will be introduced: For *isoparametric* elements the same functions F as in the ansatz space are admissible for the transformation. From the elements discussed so far only the simplex with linear ansatz is thus isoparametric. Hence, in the (affine) equivalent case a transformation not only of the points is defined by

$$\hat{x} = F^{-1}(x)\,,$$

but also of the mappings, defined on K and $\hat{K}$, (not only of P and $\hat{P}$) is given by

$$\hat{v} : \hat{K} \to \mathbb{R}\,, \quad \hat{v}(\hat{x}) := v(F(\hat{x}))$$

for $v : K \to \mathbb{R}$ and vice versa.

We can also use the techniques developed so far in such a way that only the reference element is defined, and then a general element is obtained from this by an affine-linear transformation. As an example of this, we consider elements on a cube.

Suppose $\hat{K} := [0,1]^d = \left\{x \in \mathbb{R}^d \;\middle|\; 0 \le x_i \le 1,\, i = 1,\dots,d\right\}$ is the unit cube. The *faces* of $\hat{K}$ are defined by setting a coordinate to 0 or 1; thus for instance,

$$\prod_{i=1}^{j-1}[0,1] \times \{0\} \times \prod_{j+1}^{d}[0,1]\,.$$

Let $Q_k(K)$ denote the set of polynomials on K that are of the form

$$p(x) = \sum_{\substack{0 \le \alpha_i \le k \\ i=1,\dots,d}} \gamma_{\alpha_1,\dots,\alpha_d} x_1^{\alpha_1} \cdots x_d^{\alpha_d}\,.$$

Hence, we have $\mathcal{P}_k \subset Q_k \subset \mathcal{P}_{dk}$.

Therefore, we define a reference element generally for $k \in \mathbb{N}$ as follows:

Finite Element: d-polynomial Ansatz on the Cuboid

$$\begin{aligned} \hat{K} &= [0,1]^d\,, \\ \hat{P} &= Q_k(\hat{K})\,, \\ \hat{\Sigma} &= \left\{p(\hat{x}) \;\middle|\; \hat{x} = \left(\frac{i_1}{k},\dots,\frac{i_d}{k}\right),\; i_j \in \{0,\dots,k\},\; j = 1,\dots,d\right\}, \end{aligned} \tag{3.59}$$

which is depicted in Figure 3.6. Again, we have $|\hat{\Sigma}| = \dim \hat{P}$, such that for the unique solvability of the local interpolation problem we have only to specify the shape functions. They are obtained on $\hat{K}$ as the product of the corresponding shape functions for the case $d = 1$, thus of the *Lagrange basis polynomials*

$$p_{i_1,\dots,i_d}(\hat{x}) := \prod_{j=1}^{d}\left(\prod_{\substack{i'_j=0 \\ i'_j \neq i_j}}^{k} \frac{k\hat{x}_j - i'_j}{i_j - i'_j}\right).$$

Interior degrees of freedom arise from $k = 2$ onward. Hence the ansatz space on the general element K is, according to the definition above,

$$P = \left\{\hat{p} \circ F_K^{-1} \;\middle|\; \hat{p} \in Q_k(\hat{K})\right\}.$$

In the case of a general rectangular cuboid, that is, if B in (3.57) is a diagonal matrix, then $P = Q_k(K)$ holds, analogously to the simplices. However, for a general B additional polynomial terms arise that do not belong to Q_k (see Exercise 3.14).

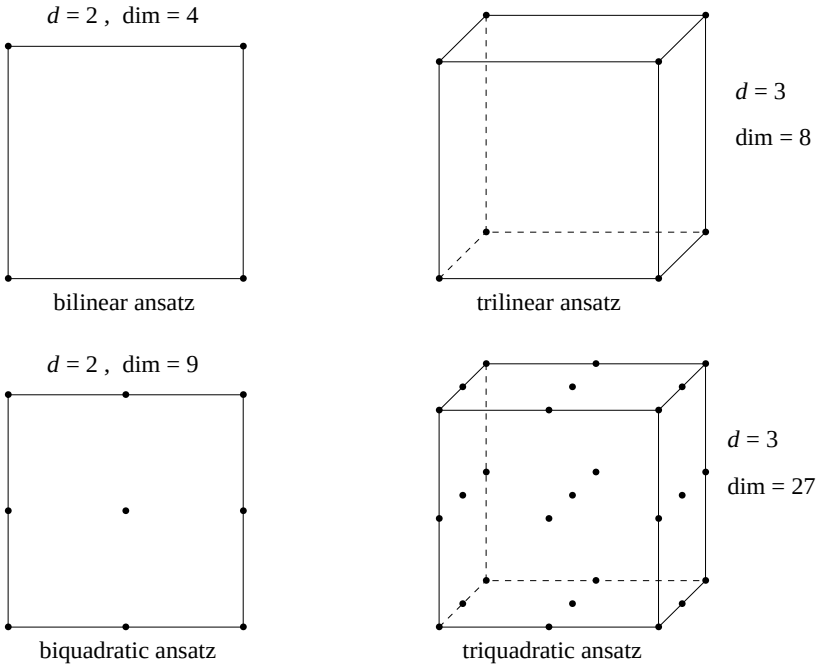


Figure 3.6. Quadratic and cubic elements on the cube.

An affine-linear transformation does not generate general cuboids but only d-epipeds, thus for $d = 3$ parallelepipeds and for $d = 2$ only parallelograms. To map the unit square to an arbitrary general convex quadrilateral, we need some transformation of Q_1, that is, isoparametric elements (see (3.142)).

Let $\mathcal{T}_h$ be a triangulation of d-simplices or of affinely transformed d-unit cubes. In particular, $\Omega = \mathrm{int}(\cup_{K\in\mathcal{T}_h} K)$ is polygonally bounded. The condition (F1) in (3.41) is always satisfied. In order to be able to satisfy the condition (F2) in (3.42) as well, a further assumption in addition to (T1)–(T4) has to be made about the triangulation:

(T5) Every face of some $K \in \mathcal{T}_h$ is either a subset of the boundary Γ of Ω or identical to a face of another $\tilde{K} \in \mathcal{T}_h$.

In order to ensure the validity of the homogeneous Dirichlet boundary condition on Γ_3 for the $v_h \in V_h$ that have to be defined, we additionally assume the following:

(T6) The boundary sets $\overline{\Gamma}_1$, $\overline{\Gamma}_2$, Γ_3 decompose into faces of elements $K \in \mathcal{T}_h$.

A face F of $K \in \mathcal{T}_h$ that is lying on $\partial\Omega$ is therefore only allowed to contain a point from the intersection $\overline{\Gamma}_i \cap \overline{\Gamma}_j$ for $i \neq j$, if and only if the point is a

boundary point of F. We recall that the set Γ_3 has been defined as being closed in $\partial\Omega$.

In the following, we suppose that these conditions are always satisfied. A triangulation that also satisfies (T5) and (T6) is called *conforming*.

Then, for all of the above finite elements,

- If $K, K' \in \mathcal{T}_h$ have a common face F, then the degrees of freedom of K and K' coincide on F. (3.60)
- F itself becomes a finite element (that is, the local interpolation problem is uniquely solvable) with the ansatz space $P_K|_F$ and the degrees of freedom on F. (3.61)

We now choose V_h as follows:

$$V_h := \left\{ v : \Omega \to \mathbb{R} \;\middle|\; v|_K \in P_K \text{ for } K \in \mathcal{T}_h \text{ and } v \text{ is uniquely given in the degrees of freedom} \right\}. \tag{3.62}$$

Analogously to the proof of Lemma 2.10, we can see that $v \in V_h$ is continuous over the face of an element; thus $V_h \subset C(\bar{\Omega})$, that is, $V_h \subset H^1(\Omega)$ according to Theorem 3.20.

Further, $u|_F = 0$ if F is a face of $K \in \mathcal{T}_h$ with $F \subset \partial\Omega$ and the specifications in the degrees of freedom of F are zero (Dirichlet boundary conditions only in the nodes); that is, the homogeneous Dirichlet boundary conditions are satisfied by enforcing them in the degrees of freedom. Due to the assumption (T6), the boundary set Γ_3 is fully taken into account in this way.

Consequently, we the following theorem:

Theorem 3.23 *Suppose $\mathcal{T}_h$ is a conforming triangulation of d-simplices or d-epipeds of a domain $\Omega \subset \mathbb{R}^d$. The elements are defined as in one of the examples* (3.53), (3.55), (3.56), (3.59).

Let the degrees of freedom be given in the nodes $a_1, \ldots, a_M$. Suppose they are numbered in such a way that $a_1, \ldots, a_{M_1} \in \Omega \cup \Gamma_1 \cup \Gamma_2$ and $a_{M_1+1}, \ldots, a_M \in \Gamma_3$. If the ansatz space V_h is defined by (3.62), *then an element $v \in V_h$ is determined uniquely by specifying $v(a_i), i = 1, \ldots, M$, and*

$$v \in H^1(\Omega).$$

If $v(a_i) = 0$ for $i = M_1 + 1, \ldots, M$, then we also have

$$v = 0 \quad \text{on } \Gamma_3.$$

Exactly as in Section 2.2 (see (2.32)), functions $\varphi_i \in V_h$ are uniquely determined by the interpolation condition

$$\varphi_i(a_j) = \delta_{ij}, \quad i, j = 1, \ldots, M.$$

By the same consideration as there and as for the shape functions (see (3.54)) we observe that the φ_i form a basis of V_h, the *nodal basis*, since

each $v \in V_h$ has a unique representation

$$v(x) = \sum_{i=1}^{M} v(a_i)\varphi_i(x)\,. \tag{3.63}$$

If for Dirichlet boundary conditions, the values in the boundary nodes $a_i, i = M_1 + 1, \ldots, M$, are given as zero, then the index has to run only up to M_1.

The support $\operatorname{supp}\varphi_i$ of the basis functions thus consists of all elements that contain the node a_i, since in all other elements φ_i assumes the value 0 in the degrees of freedom and hence vanishes identically. In particular, for an interior degree of freedom, that is, for some a_i with $a_i \in \operatorname{int}(K)$ for an element $K \in \mathcal{T}_h$, we have $\operatorname{supp}\varphi_i = K$.

Different element types can also be combined (see Figure 3.7) if only (3.60) is satisfied, thus, for instance for $d = 2$ (3.59), $k = 1$, can be combined with (3.53) or (3.59), $k = 2$, with (3.55).

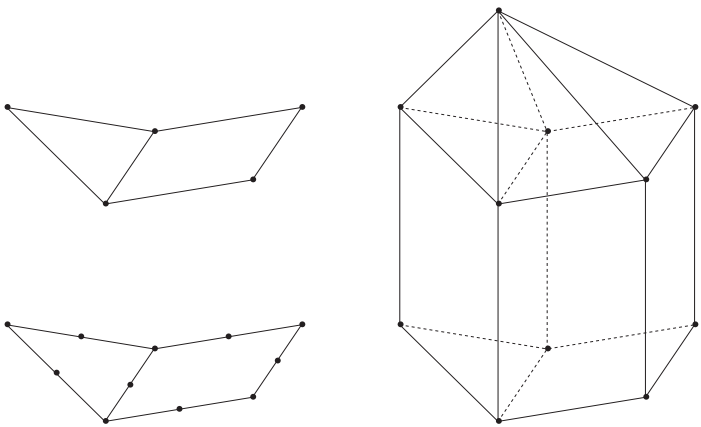


Figure 3.7. Conforming combination of different element types.

For $d = 3$ a combination of simplices and parallelepipeds is not possible, since they have different types of faces. Tetrahedra can be combined with prisms at their two triangular surfaces, whereas their three quadrilateral surfaces (see Exercise 3.17) allow for a combination of prisms with parallelepipeds. Possibly also pyramids are necessary as transition elements (see [57]).

So far, the degrees of freedom have always been function values (*Lagrange elements*). If, additionally, derivative values are specified, then we speak of *Hermite elements*. As an example, we present the following:

Finite Element: Cubic Hermite Ansatz on the Simplex

$K = \operatorname{conv}\{a_1, \ldots, a_{d+1}\}$,

$$
\begin{aligned}
P &= \mathcal{P}_3(K)\,, \qquad (3.64)\\
\Sigma &= \big\{p(a_i)\,,\ i=1,\dots,d+1\,,\ p(a_{i,j,k})\,,\ i,j,k=1,\dots,d+1\,,\ i<j<k\,,\\
&\qquad \nabla p(a_i)\cdot(a_j-a_i)\,,\ i,j=1,\dots,d+1\,,\ i\neq j\big\}\,.
\end{aligned}
$$

Instead of the directional derivatives we could also have chosen the partial derivatives as degrees of freedom, but would not have generated affine equivalent elements in that way. In order to ensure that directional derivatives in the directions ξ and $\hat{\xi}$ are mapped onto each other by the transformation, the directions have to satisfy

$$\xi = B\hat{\xi}\,,$$

where B is the linear part of the transformation F according to (3.57). This is satisfied for (3.64), but would be violated for the partial derivatives, that is, $\xi = \hat{\xi} = e_i$. This has also to be taken into account for the question of which degrees of freedom have to be chosen for Dirichlet boundary conditions (see Exercise 3.19). Thus, the desired property that the degrees of freedom be defined "globally" is lost here. Nevertheless, we do not have a C^1-element: The ansatz (3.64) ensures only the continuity of the tangential, not of the normal derivative over a face.

Finite Element: Bogner–Fox–Schmit Rectangle
The simplest C^1-*element* is for $d = 2$:

$$
\begin{aligned}
\hat{K} &= [0,1]^2\,,\\
\hat{P} &= Q_3(\hat{K})\,, \qquad (3.65)\\
\hat{\Sigma} &= \{p(a),\ \partial_1 p(a),\ \partial_2 p(a),\ \partial_{12} p(a) \quad \text{for all vertices } a\}\ ;
\end{aligned}
$$

that is, the element has 16 degrees of freedom.

In the case of Hermite elements, the above propositions concerning the nodal basis hold analogously with an appropriate extension of the identity (3.63).

Further, all considerations of Section 2.2 concerning the determination of the Galerkin approximation as a solution of a system of equations (2.34) also hold, since there only the (bi)linearity of the forms is supposed. Therefore using the nodal basis, the quantity $a(\varphi_j, \varphi_i)$ has to be computed as the (i,j)th matrix entry of the system of equations that has to be set up for the bilinear form a. The form of the bilinear form (3.31) shows that the consideration of Section 2.2, concerning that there is at most a nonzero entry at position (i,j) if,

$$\operatorname{supp}\varphi_i \cap \operatorname{supp}\varphi_j \neq \emptyset\,, \qquad (3.66)$$

still holds.

Since in the examples discussed, $\operatorname{supp}\varphi_i$ consists of at most of those elements containing the node a_i (see Figure 3.10), the nodes have to be *adjacent*, for the validity of (3.66); that is, they should belong to some

common element. In particular, an interior degree of freedom of some element is connected only with the nodes of the same element: This can be used to eliminate such nodes from the beginning (*static condensation*).

The following consideration can be helpful for the choice of the element type: An increase in the size of polynomial ansatz spaces increases the (computational) cost by an increase in the number of nodes and an increase in the population of the matrix.

As an example for $d = 2$ we consider triangles with linear (a) and quadratic (b) ansatz (see Figure 3.8).

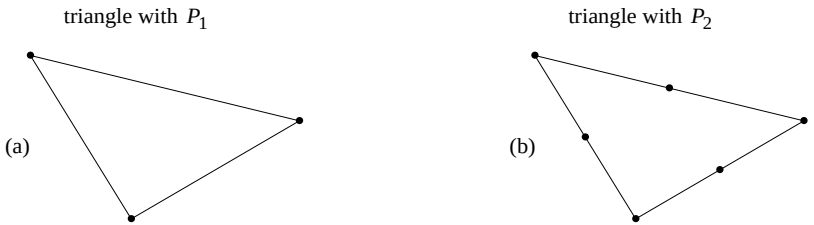


Figure 3.8. Comparison between linear and quadratic triangles.

In order to have the same number of nodes we compare (b) with the discretization parameter h with (a) with the discretization parameter $h/2$ (one step of "red refinement") (see Figure 3.9).

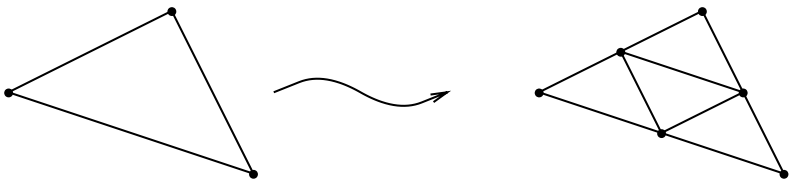


Figure 3.9. Generation of the same number of nodes.

However, this shows that we have a denser population in (b) than in (a).

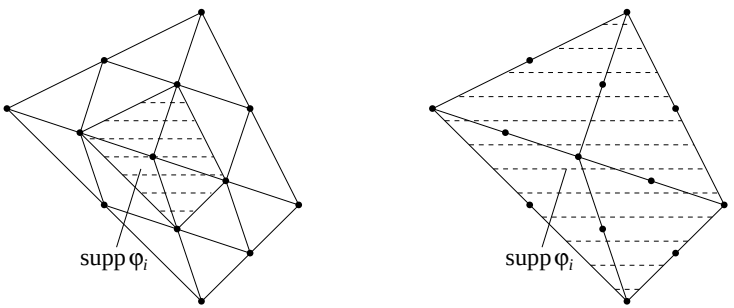


Figure 3.10. Supports of the basis functions.

To have still an advantage by using the higher polynomial order, the ansatz (b) has to have a higher convergence rate. In Theorem 3.29 we will prove the following estimate for a regular family of triangulations $\mathcal{T}_h$ (see Definition 3.28):

- If $u \in H^2(\Omega)$, then for (a) and (b) we have the estimate

$$\|u - u_h\|_1 \leq C_1 h\,. \tag{3.67}$$

- If $u \in H^3(\Omega)$, then for (b) but not for (a) we have the estimate

$$\|u - u_h\|_1 \leq C_2 h^2\,. \tag{3.68}$$

For the constants we may in general expect $C_2 > C_1$.

In order to be able to make a comparison between the variants (a) and (b), we consider in the following the case of a rectangle $\Omega = (0, a) \times (0, b)$. The number of the nodes is then proportional to $1/h^2$ if the elements are all "essentially" of the same size.

However, if we consider the number of nodes M as given, then h is proportional to $1/\sqrt{M}$.

Using this in the estimate (3.67), we get for a solution $u \in H^2(\Omega)$,

$$\begin{aligned} &\text{in the case (a) for } h/2\text{:} && \|u - u_{h/2}\|_1 &&\leq\ C_1 \frac{1}{2\sqrt{M}}\,,\\ &\text{in the case (b) for } h\text{:} && \|u - u_h\|_1 &&\leq\ \bar{C}_1 \frac{1}{\sqrt{M}}\,. \end{aligned}$$

If both constants are the same, this means an advantage for the variant (a).

On the other hand, if the solution is smoother and satisfies $u \in H^3(\Omega)$, then the estimate (3.68), which can be applied only to the variant (b), yields

$$\begin{aligned} &\text{in the case (a) for } h/2\text{:} && \|u - u_{h/2}\|_1 &&\leq\ C_1 \frac{1}{2\sqrt{M}}\,,\\ &\text{in the case (b) for } h\text{:} && \|u - u_h\|_1 &&\leq\ C_2 \frac{1}{M}\,. \end{aligned}$$

By an elementary reformulation, we get

$$C_2 \frac{1}{M} < (<) C_1 \frac{1}{2\sqrt{M}} \iff M > (>)\, 4 \frac{C_2^2}{C_1^2}\,,$$

which gives an advantage for (b) if the number of variables M is chosen, depending on C_2/C_1, sufficiently large. However, the denser population of the matrix in (b) has to be confronted with this.

Hence, a higher-order polynomial ansatz has an advantage only if the smoothness of the solution leads to a higher convergence rate. Especially for nonlinear problems with less-smooth solutions, a possible advantage of the higher-order ansatz has to be examined critically.

Exercises

3.10 Prove the implication "⇒" in Theorem 3.20.

Hint: For $v \in V_h$ define a function w_i by $w_i|_{\mathrm{int}(K)} := \partial_i v$, $i = 1, \ldots, d$, and show that w_i is the ith partial derivative of v.

3.11 Construct the element stiffness matrix for the Poisson equation on a rectangle with quadratic bilinear rectangular elements. Verify that this finite element discretization of the Laplace operator can be interpreted as a finite difference method with the difference stencil according to (1.22).

3.12 Prove that:

(a) $\dim \mathcal{P}_k(\mathbb{R}^d) = \binom{d+k}{k}$.

(b) $\mathcal{P}_k(\mathbb{R}^d)|_K = \mathcal{P}_k(K)$ if $\mathrm{int}\,(K) \neq \emptyset$.

3.13 Prove for given vectors $a_1, \ldots, a_{d+1} \in \mathbb{R}^d$ that $a_2 - a_1, \ldots, a_{d+1} - a_1$ are linear independent if and only if $a_1 - a_i, \ldots, a_{i-1} - a_i, a_{i+1} - a_i, \ldots, a_{d+1} - a_i$ are linearly independent for some $i \in \{2, \ldots, d\}$.

3.14 Determine for the polynomial ansatz on the cuboid as reference element (3.59) the ansatz space P that is obtained by an affine-linear transformation to a d-epiped.

3.15 Suppose K is a rectangle with the (counterclockwise numbered) vertices $a_1, \ldots, a_4$ and the corresponding edge midpoints a_{12}, a_{23}, a_{34}, a_{41}. Show that the elements f of $Q_1(K)$ are not determined uniquely by the degrees of freedom $f(a_{12}), f(a_{23}), f(a_{34}), f(a_{41})$.

3.16 Check the given shape functions for (3.55) and (3.56).

3.17 Define a reference element in $\mathbb{R}^3$ by

$$\begin{aligned}
\hat{K} &= \mathrm{conv}\,\{\hat{a}_1, \hat{a}_2, \hat{a}_3\} \times [0,1] \text{ with } \hat{a}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix},\ \hat{a}_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix},\ \hat{a}_3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \\
\hat{P} &= \left\{ p_1(x_1, x_2)\, p_2(x_3) \;\middle|\; p_1 \in \mathcal{P}_1(\mathbb{R}^2)\,,\ p_2 \in \mathcal{P}_1(\mathbb{R}) \right\}, \\
\hat{\Sigma} &= \left\{ p(\hat{x}) \;\middle|\; \hat{x} = (\hat{a}_i, j)\,,\ i = 0, 1, 2\,,\ j = 0, 1 \right\}.
\end{aligned}$$

Show the unique solvability of the local interpolation problem and describe the elements obtained by affine-linear transformation.

3.18 Suppose $d+1$ points a_j, $j = 1, \ldots, d+1$, in $\mathbb{R}^d$ are given with the property as in Exercise 3.13. Additionally, we define as in (3.48), (3.49) the barycentric coordinates $\lambda_j = \lambda_j(x; S)$ of x with respect to the d-simplex S generated by the points a_j. Show that for each bijective affine-linear

mapping $\ell : \mathbb{R}^d \to \mathbb{R}^d$, $\lambda_j(x; S) = \lambda_j(\ell(x); \ell(S))$, which means that the barycentric coordinates are invariant under such transformations.

3.19 Discuss for the cubic Hermite ansatz (3.64) and Dirichlet boundary conditions the choice of the degrees of freedom with regard to the angle between two edges of boundary elements that is either $\alpha \neq 2\pi$ or $\alpha = 2\pi$.

3.20 Construct a nodal basis for the Bogner–Fox–Schmit element in (3.65).

3.4 Convergence Rate Estimates

In this section we consider further a finite element approximation in the framework described in the previous section: The bounded basic domain $\Omega \subset \mathbb{R}^d$ of the boundary value problem is decomposed into conforming triangulations $\mathcal{T}_h$, which may also consist of different types of elements. Here, by an element we mean not only the set $K \in \mathcal{T}_h$, but this equipped with some ansatz space P_K and degrees of freedom Σ_K. However, the elements are supposed to decompose into a fixed number of subsets, independent of h, each consisting of elements that are affine equivalent to each other. Different elements have to be compatible with each other such that the ansatz space V_h, introduced in (3.62), is well-defined. The smoothness of the functions arising in this way has to be consistent with the boundary value problem, in so far as $V_h \subset V$ is guaranteed. In the following we consider only one element type; the generalization to the more general situation will be obvious. The goal is to prove *a priori estimates* of the form

$$\|u - u_h\| \leq C|u|h^\alpha \tag{3.69}$$

with constants $C > 0$, $\alpha > 0$ and norms and seminorms $\|\cdot\|$ and $|\cdot|$, respectively.

We do not attempt to give the constant C explicitly, although in principle, this is possible (with other techniques of proof). In particular, in the following C has to be understood generically; that is, by C we denote at different places different values, which, however, are independent of h. Therefore, the estimate (3.69) does not serve only to estimate numerically the error for a fixed triangulation $\mathcal{T}_h$. It is rather useful for estimating what gain in accuracy can be expected by increasing the effort, which then corresponds to the reduction of h by some refinement (see the discussion around (3.67)). Independently of the *convergence rate* α, (3.69) provides the certainty that an arbitrary accuracy in the desired norm $\|\cdot\|$ can be obtained at all. In the following, we will impose some geometric conditions on the family $(\mathcal{T}_h)_h$, which have always to be understood uniformly in h. For a fixed triangulation these conditions are always trivially satisfied, since here

we have a finite number of elements. For a family $(\mathcal{T}_h)_h$ with $h \to 0$, thus for increasing refinement, this number becomes unbounded. In the following estimates we have therefore to distinguish between "variable" values like the number of nodes $M = M(h)$ of $\mathcal{T}_h$, and "fixed" values like the dimension d or the dimension of P_K or equivalence constants in the renorming of P_K, which can all be included in the generic constant C.

3.4.1 Energy Norm Estimates

If we want to derive estimates in the norm of the Hilbert space V underlying the variational equation for the boundary value problem, concretely, in the norm of Sobolev spaces, then Céa's lemma (Theorem 2.17) shows that for this purpose it is necessary only to specify a comparison element $v_h \in V_h$ for which the inequality

$$\|u - v_h\| \leq C|u|h^\alpha \tag{3.70}$$

holds. For $\|\cdot\| = \|\cdot\|_1$, these estimates are called *energy norm estimates* due to the equivalence of $\|\cdot\|_1$ and $\|\cdot\|_a$ (cf. (2.46)) in the symmetric case. Therefore, the comparison element v_h has to approximate u as well as possible, and in genera,l it is specified as the image of a linear operator I_h:

$$v_h = I_h(u)\,.$$

The classical approach consists in choosing for I_h the *interpolation operator* with respect to the degrees of freedom. To simplify the notation, we restrict ourselves in the following to Lagrange elements, the generalization to Hermite elements is also easily possible.

We suppose that the triangulation $\mathcal{T}_h$ has its degrees of freedom in the nodes $a_1, \ldots, a_M$ with the corresponding nodal basis $\varphi_1, \ldots, \varphi_M$. Then let

$$I_h(u) := \sum_{i=1}^{M} u(a_i)\varphi_i \ \in V_h\,. \tag{3.71}$$

For the sake of $I_h(u)$ being well-defined, $u \in C(\bar{\Omega})$ has to be assumed in order to ensure that u can be evaluated in the nodes. This requires a certain smoothness assumption about the solution u, which we formulate as

$$u \in H^{k+1}(\Omega)\,.$$

Thus, if we assume again $d \leq 3$ for the sake of simplicity, the embedding theorem (Theorem 3.10) ensures that I_h is well-defined on $H^{k+1}(\Omega)$ for $k \geq 1$. For the considered C^0-elements, we have $I_h(u) \in H^1(\Omega)$ by virtue of Theorem 3.20. Therefore, we can substantiate the desired estimate (3.70) to

$$\|u - I_h(u)\|_1 \leq Ch^\alpha |u|_{k+1}\,. \tag{3.72}$$

Sobolev (semi) norms can be decomposed into expressions over subsets of Ω, thus, for instance, the elements of $\mathcal{T}_h$,

$$|u|_l^2 = \int_\Omega \sum_{|\alpha|=l} |\partial^\alpha u|^2 \, dx = \sum_{K\in\mathcal{T}_h} \int_K \sum_{|\alpha|=l} |\partial^\alpha u|^2 \, dx = \sum_{K\in\mathcal{T}_h} |u|_{l,K}^2 \,,$$

and, correspondingly,

$$\|u\|_l^2 = \sum_{K\in\mathcal{T}_h} \|u\|_{l,K}^2 \,,$$

where, if Ω is not basic domain, this will be included in the indices of the norm. Since the elements K are considered as being closed, K should more precisely be replaced by $\operatorname{int}(K)$. By virtue of this decomposition, it is sufficient to prove the estimate (3.72) for the elements K. This has some analogy to the (elementwise) assembling described in Section 2.4.2, which is also to be seen in the following. On K, the operator I_h reduces to the analogously defined local interpolation operator. Suppose the nodes of the degrees of freedom on K are $a_{i_1}, \ldots, a_{i_L}$, where $L \in \mathbb{N}$ is the same for all $K \in \mathcal{T}_h$ due to the equivalence of elements. Then

$$I_h(u)|_K = I_K(u|_K) \quad \text{for } u \in C(\bar{\Omega}) \,,$$

where

$$I_K(u) := \sum_{j=1}^{L} u(a_{i_j})\varphi_{i_j} \quad \text{for } u \in C(K) \,,$$

since both functions of P_K solve the same interpolation problem on K (cf. Lemma 2.10). Since we have an (affine) equivalent triangulation, the proof of the local estimate

$$\|u - I_K(u)\|_{m,K} \le Ch^\alpha |u|_{k+1,K} \tag{3.73}$$

is generally done in three steps:

- Transformation to some reference element $\hat{K}$,
- Proof of (3.73) on $\hat{K}$,
- Back transformation to the element K.

To be precise, the estimate (3.73) will even be proved with h_K instead of h, where

$$h_K := \operatorname{diam}(K) \quad \text{for } K \in \mathcal{T}_h \,,$$

and in the second step, the fixed value $h_{\hat{K}}$ is incorporated in the constant. The powers of h_K are due to the transformation steps.

Therefore, let some reference element $\hat{K}$ with the nodes $\hat{a}_1, \ldots, \hat{a}_L$ be chosen as fixed. By assumption, there exists some bijective, affine-linear

mapping

$$\begin{aligned} F &= F_K : \hat{K} \to K\,, \\ F(\hat{x}) &= B\hat{x} + d\,, \end{aligned} \tag{3.74}$$

(cf. (2.30) and (3.57)). By this transformation, functions $v : K \to \mathbb{R}$ are mapped to functions $\hat{v} : \hat{K} \to \mathbb{R}$ by

$$\hat{v}(\hat{x}) := v(F(\hat{x}))\,. \tag{3.75}$$

This transformation is also *compatible* with the local interpolation operator in the following sense:

$$\widehat{I_K(v)} = I_{\hat{K}}(\hat{v}) \quad \text{for } v \in C(K)\,. \tag{3.76}$$

This follows from the fact that the nodes of the elements as well as the shape functions are mapped onto each other by F.

For a classically differentiable function the chain rule (see (2.49)) implies

$$\nabla_x v(F(\hat{x})) = B^{-T} \nabla_{\hat{x}} \hat{v}(\hat{x})\,, \tag{3.77}$$

and corresponding formulas for higher-order derivatives, for instance,

$$D_x^2 v(F(\hat{x})) = B^{-T} D_{\hat{x}}^2 \hat{v}(\hat{x}) B^{-1}\,,$$

where $D_x^2 v(x)$ denotes the matrix of the second-order derivatives. These chain rules hold also for corresponding $v \in H^l(K)$ (Exercise 3.22).

The situation becomes particularly simple in one space dimension ($d = 1$). The considered elements reduce to a polynomial ansatz on simplices, which here are intervals. Thus

$$\begin{aligned} F : \hat{K} = [0,1] &\to K = [a_{i_1}, a_{i_2}]\,, \\ \hat{x} &\mapsto h_K \hat{x} + a_{i_1}\,, \end{aligned}$$

where $h_K := a_{i_2} - a_{i_1}$ denotes the length of the element. Hence, for $l \in \mathbb{N}$,

$$\partial_x^l v(F(\hat{x})) = h_K^{-l} \partial_{\hat{x}}^l \hat{v}(\hat{x})\,.$$

By the substitution rule for integrals (cf. (2.50)) an additional factor $|\det(B)| = h_K$ arises such that, for $v \in H^l(K)$, we have

$$|v|_{l,K}^2 = \left(\frac{1}{h_K}\right)^{2l-1} |\hat{v}|_{l,\hat{K}}^2\,.$$

Hence, for $0 \le m \le k+1$ it follows by (3.76) that

$$|v - I_K(v)|_{m,K}^2 = \left(\frac{1}{h_K}\right)^{2m-1} \left|\hat{v} - I_{\hat{K}}(\hat{v})\right|_{m,\hat{K}}^2\,.$$

Thus, what is missing, is an estimate of the type

$$\left|\hat{v} - I_{\hat{K}}(\hat{v})\right|_{m,\hat{K}} \le C |\hat{v}|_{k+1,\hat{K}} \tag{3.78}$$

for $\hat{v} \in H^{k+1}(\hat{K})$. In specific cases this can partly be proven directly but in the following a general proof, which is also independent of $d = 1$, will be sketched. For this, the mapping

$$\begin{aligned} G : H^{k+1}(\hat{K}) &\to H^m(\hat{K}) , \\ \hat{v} &\mapsto \hat{v} - I_{\hat{K}}(\hat{v}) , \end{aligned} \tag{3.79}$$

is considered. The mapping is linear but also continuous, since

$$\begin{aligned} \left\| I_{\hat{K}}(\hat{v}) \right\|_{m,\hat{K}} &\leq \left\| \sum_{i=1}^{L} \hat{v}(\hat{a}_i) \hat{\varphi}_i \right\|_{k+1,\hat{K}} \\ &\leq \sum_{i=1}^{L} \|\hat{\varphi}_i\|_{k+1,\hat{K}} \, \|\hat{v}\|_{\infty,\hat{K}} \leq C \|\hat{v}\|_{k+1,\hat{K}} , \end{aligned} \tag{3.80}$$

where the continuity of the embedding of $H^{k+1}(\hat{K})$ in $H^m(\hat{K})$ (see (3.8)) and of $H^{k+1}(\hat{K})$ in $C(\hat{K})$ (Theorem 3.10) is used, and the norm contribution from the fixed basis functions $\hat{\varphi}_i$ is included in the constant.

If the ansatz space $\hat{P}$ is chosen in such a way that $\mathcal{P}_k \subset \hat{P}$, then G has the additional property

$$G(p) = 0 \quad \text{for } p \in \mathcal{P}_k ,$$

since these polynomials are interpolated then exactly. Such mappings satisfy the Bramble–Hilbert lemma, which will directly be formulated, for further use, in a more general way.

Theorem 3.24 (Bramble–Hilbert lemma)
Suppose $K \subset \mathbb{R}^d$ is open, $k \in \mathbb{N}_0$, $1 \leq p \leq \infty$, and $G : W_p^{k+1}(K) \to \mathbb{R}$ is a continuous linear functional that satisfies

$$G(q) = 0 \quad \textit{for all } q \in \mathcal{P}_k . \tag{3.81}$$

Then there exists some constant $C > 0$ independent of G such that for all $v \in W_p^{k+1}(K)$

$$|G(v)| \leq C \, \|G\| \, |v|_{k+1,p,K} .$$

Proof: See [9, Theorem 28.1]. □

Here $\|G\|$ denotes the operator norm of G (see (A4.25)). The estimate with the full norm $\|\cdot\|_{k+1,p,K}$ on the right-hand side (and $C = 1$) would hence only be the operator norm's definition. The condition (3.81) allows the reduction to the highest seminorm.

For the application of the Bramble–Hilbert lemma (Theorem 3.24), which was formulated only for functionals, to the operator G according to (3.79) an additional argument is required (alternatively, Theorem 3.24 could be generalized):

Generally, for $\hat{w} \in H^m(\hat{K})$ (as in every normed space) we have

$$\|\hat{w}\|_{m,\hat{K}} = \sup_{\substack{\varphi \in (H^m(\hat{K}))' \\ \|\varphi\| \le 1}} \varphi(\hat{w}) \,, \tag{3.82}$$

where the norm applying to φ is the operator norm defined in (A4.25).

For any fixed $\varphi \in (H^m(\hat{K}))'$ the linear functional on $H^{k+1}(\hat{K})$ is defined by

$$\tilde{G}(\hat{v}) := \varphi(G(\hat{v})) \quad \text{for} \quad \hat{v} \in H^{k+1}(\hat{K}) \,. \tag{3.83}$$

According to (3.80), $\tilde{G}$ is continuous and it follows that

$$\|\tilde{G}\| \le \|\varphi\| \, \|G\| \,.$$

Theorem 3.24 is applicable to $\tilde{G}$ and yields

$$|\tilde{G}(\hat{v})| \le C \, \|\varphi\| \, \|G\| \, |\hat{v}|_{k+1,\hat{K}} \,.$$

By means of (3.82) it follows that

$$\|G(\hat{v})\|_{m,\hat{K}} \le C \, \|G\| \, |\hat{v}|_{k+1,\hat{K}} \,.$$

The same proof can also be used in the proof of Theorem 3.31 (3.94).

Applied to G defined in (3.79), the estimate (3.80) shows that the operator norm $\left\|\mathrm{Id} - I_{\hat{K}}\right\|$ can be estimated independently from m (but dependent on k and the $\hat{\varphi}_i$) and can be incorporated in the constant that gives (3.78) in general, independent of the one-dimensional case.

Therefore, in the one-dimensional case we can continue with the estimation and get

$$|v - I_K(v)|^2_{m,K} \le \left(\frac{1}{h_K}\right)^{2m-1} C|\hat{v}|^2_{k+1,\hat{K}} \le C(h_K)^{1-2m+2(k+1)-1}|v|^2_{k+1,K} \,.$$

Since due to $I_h(v) \in H^1(\Omega)$ we have for $m = 0, 1$

$$\sum_{K \in \mathcal{T}_h} |v - I_K(v)|^2_{m,K} = |v - I_h(v)|^2_m \,,$$

we have proven the following Theorem:

Theorem 3.25 *Consider in one space dimension $\Omega = (a, b)$ the polynomial Lagrange ansatz on elements with maximum length h and suppose that for the respective local ansatz spaces P, the inclusion $\mathcal{P}_k \subset P$ is satisfied for some $k \in \mathbb{N}$. Then there exists some constant $C > 0$ such that for all $v \in H^{k+1}(\Omega)$ and $0 \le m \le k+1$,*

$$\left(\sum_{K \in \mathcal{T}_h} |v - I_K(v)|^2_{m,K}\right)^{1/2} \le C h^{k+1-m}|v|_{k+1} \,.$$

If the solution u of the boundary value problem (3.12), (3.18)–(3.20) *belongs to $H^{k+1}(\Omega)$, then we have for the finite element approximation u_h according*

to (3.39)*,*

$$\|u - u_h\|_1 \le Ch^k |u|_{k+1} .$$

Note that for $d = 1$ a direct proof is also possible (see Exercise 3.21).

Now we address to the general d-dimensional situation: The seminorm $|\cdot|_1$ is transformed, for instance, as follows (cf. (2.49)):

$$|v|_{1,K}^2 = \int_K |\nabla_x v|^2 \, dx = \int_{\hat{K}} B^{-T} \nabla_{\hat{x}} \hat{v} \cdot B^{-T} \nabla_{\hat{x}} \hat{v} \, |\det(B)| \, d\hat{x} . \tag{3.84}$$

From this, it follows for $\hat{v} \in H^1(\hat{K})$ that

$$|v|_{1,K} \le C \, \|B^{-1}\| \, |\det(B)|^{1/2} \, |\hat{v}|_{1,\hat{K}} .$$

Since d is one of the mentioned "fixed" quantities and all norms on $\mathbb{R}^{d,d}$ are equivalent, the matrix norm $\|\cdot\|$ can be chosen arbitrarily, and it is also possible to change between such norms. In the above considerations K and $\hat{K}$ had equal rights; thus similarly for $v \in H^1(K)$, we have

$$|\hat{v}|_{1,\hat{K}} \le C \, \|B\| \, |\det(B)|^{-1/2} \, |v|_{1,K} .$$

In general, we have the following theorem:

Theorem 3.26 *Suppose K and $\hat{K}$ are bounded domains in $\mathbb{R}^d$ that are mapped onto each other by an affine bijective linear mapping F, defined in* (3.74)*. If $v \in W_p^l(K)$ for $l \in \mathbb{N}$ and $p \in [1, \infty]$, then we have for $\hat{v}$ (defined in* (3.75)*), $\hat{v} \in W_p^l(\hat{K})$, and for some constant $C > 0$ independent of v,*

$$|\hat{v}|_{l,p,\hat{K}} \le C \, \|B\|^l \, |\det(B)|^{-1/p} \, |v|_{l,p,K} , \tag{3.85}$$

$$|v|_{l,p,K} \le C \, \|B^{-1}\|^l \, |\det(B)|^{1/p} \, |\hat{v}|_{l,p,\hat{K}} . \tag{3.86}$$

Proof: See [9, Theorem 15.1]. □

For further use, also this theorem has been formulated in a more general way than would be necessary here. Here, only the case $p = 2$ is relevant.

Hence, if we use the estimate of Theorem 3.24, then the value $\|B\|$ (for some matrix norm) has to be related to the geometry of K. For this, let for $K \in \mathcal{T}_h$,

$$\varrho_K := \sup \left\{ \operatorname{diam}(S) \;\middle|\; S \text{ is a ball in } \mathbb{R}^d \text{ and } S \subset K \right\} .$$

Hence, in the case of a triangle, h_K denotes the longest edge and ϱ_K the diameter of the inscribed circle. Similarly, the reference element has its (fixed) parameters $\hat{h}$ and $\hat{\varrho}$. For example, for the reference triangle with the vertices $\hat{a}_1 = (0,0)$, $\hat{a}_2 = (1,0)$, $\hat{a}_3 = (0,1)$ we have that $\hat{h} = 2^{1/2}$ and $\hat{\varrho} = 2 - 2^{1/2}$.

Theorem 3.27 *For $F = F_K$ according to* (3.74)*, in the spectral norm $\|\cdot\|_2$, we have*

$$\|B\|_2 \le \frac{h_K}{\hat{\varrho}} \quad \textit{and} \quad \|B^{-1}\|_2 \le \frac{\hat{h}}{\varrho_K}\,.$$

Proof: Since K and $\hat{K}$ have equal rights in the assertion, it suffices to prove one of the statements: We have (cf. (A4.25))

$$\|B\|_2 = \sup_{|\xi|_2=\hat{\varrho}} \left| B\left(\frac{1}{\hat{\varrho}}\xi\right)\right|_2 = \frac{1}{\hat{\varrho}} \sup_{|\xi|_2=\hat{\varrho}} |B\xi|_2\,.$$

For every $\xi \in \mathbb{R}^d$ with $|\xi|_2 = \hat{\varrho}$ there exist some points $\hat{y}, \hat{z} \in \hat{K}$ such that $\hat{y}-\hat{z} = \xi$. Since $B\xi = F(\hat{y})-F(\hat{z})$ and $F(\hat{y}), F(\hat{z}) \in K$, we have $|B\xi|_2 \le h_K$. Consequently, by the above identity we get the first inequality. □

If we combine the local estimates of (3.78), Theorem 3.26, and Theorem 3.27, we obtain for $v \in H^{k+1}(K)$ and $0 \le m \le k+1$,

$$|v - I_K(v)|_{m,K} \le C \left(\frac{h_K}{\varrho_K}\right)^m h_K^{k+1-m}\, |v|_{k+1,K}\,, \tag{3.87}$$

where $\hat{\varrho}$ and $\hat{h}$ are included in the constant C. In order to obtain some convergence rate result, we have to control the term h_K/ϱ_K. If this term is bounded (uniformly for all triangulations), we get the same estimate as in the one-dimensional case (where even $h_K/\varrho_K = 1$). Conditions of the form

$$\varrho_K \ge \sigma h_K^{1+\alpha}$$

for some $\sigma > 0$ and $0 \le \alpha < \frac{k+1}{m} - 1$ for $m \ge 1$ would also lead to convergence rate results. Here we pursue only the case $\alpha = 0$.

Definition 3.28 A family of triangulations $(\mathcal{T}_h)_h$ is called *regular* if there exists some $\sigma > 0$ such that for all $h > 0$ and all $K \in \mathcal{T}_h$,

$$\varrho_K \ge \sigma h_K\,.$$

From estimate (3.87) we conclude directly the following theorem:

Theorem 3.29 *Consider a family of Lagrange finite element discretizations in $\mathbb{R}^d$ for $d \le 3$ on a regular family of triangulations $(\mathcal{T}_h)_h$ in the generality described at the very beginning. For the respective local ansatz spaces P suppose $\mathcal{P}_k \subset P$ for some $k \in \mathbb{N}$.*

Then there exists some constant $C > 0$ such that for all $v \in H^{k+1}(\Omega)$ and $0 \le m \le k+1$,

$$\left(\sum_{K\in\mathcal{T}_h} |v - I_K(v)|_{m,K}^2\right)^{1/2} \le C h^{k+1-m} |v|_{k+1}\,. \tag{3.88}$$

If the solution u of the boundary value problem (3.12), (3.18)–(3.20) *belongs to $H^{k+1}(\Omega)$, then for the finite element approximation u_h defined in* (3.39), *it follows that*

$$\|u - u_h\|_1 \leq Ch^k |u|_{k+1} \,. \tag{3.89}$$

Remark 3.30 Indeed, here and also in Theorem 3.25 a sharper estimate has been shown, which, for instance for (3.89), has the following form:

$$\|u - u_h\|_1 \leq C \left(\sum_{K \in \mathcal{T}_h} h_K^{2k} |u|_{k+1,K}^2 \right)^{1/2} . \tag{3.90}$$

In the following we will discuss what the regularity assumption means in the two simplest cases:

For a rectangle and the cuboid K, whose edge lengths can be assumed, without any loss of generality, to be of order $h_1 \leq h_2 [\leq h_3]$, we have

$$\frac{h_K}{\varrho_K} = \left(1 + \left(\frac{h_2}{h_1} \right)^2 \left[+ \left(\frac{h_3}{h_1} \right)^2 \right] \right)^{1/2} .$$

This term is uniformly bounded if and only if there exists some constant $\alpha (\geq 1)$ such that

$$\begin{array}{ccccc} h_1 & \leq & h_2 & \leq & \alpha h_1 \,, \\ h_1 & \leq & h_3 & \leq & \alpha h_1 \,. \end{array} \tag{3.91}$$

In order to satisfy this condition, a refinement in one space direction has to imply a corresponding one in the other directions, although in certain *anisotropic* situations only the refinement in one space direction is recommendable. If, for instance, the boundary value problem (3.12), (3.18)–(3.20) with $c = r = 0$, but space-dependent conductivity K, is interpreted as the simplest ground water model (see (0.18)), then it is typical that K varies discontinuously due to some *layering* or more complex geological structures (see Figure 3.11).

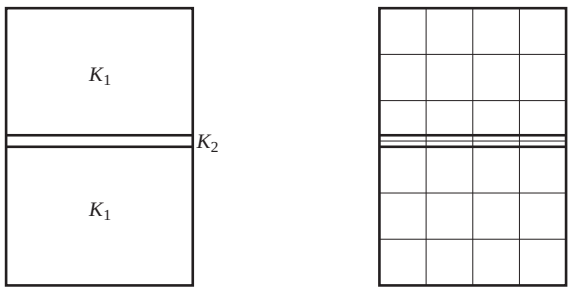


Figure 3.11. Layering and anisotropic triangulation.

If thin layers arise in such a case, on the one hand they have to be *resolved*; that is, the triangulation has to be compatible with the layering and there

have to be sufficiently many elements in this layer. On the other hand, the solution often changes less strongly in the direction of the layering than over the boundaries of the layer, which suggests an *anisotropic* triangulation, that is, a strongly varying dimensioning of the elements. The restriction (3.91) is not compatible with this, but in the case of rectangles this is due only to the techniques of proof. In this simple situation, the local interpolation error estimate can be performed directly, at least for $P = Q_1(K)$, without any transformation such that the estimate (3.89) (for $k = 1$) is obtained without any restrictions like (3.91).

The next simple example is a triangle K: The smallest angle $\alpha_{\min} = \alpha_{\min}(K)$ includes the longest edge h_K, and without loss of generality, the situation is as illustrated in Figure 3.12.

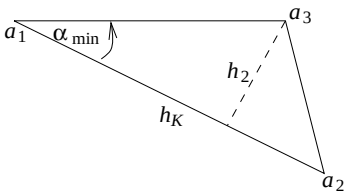


Figure 3.12. Triangle with the longest edge and the height as parameters.

For the 2×2 matrix $B = (a_2 - a_1\,,\ a_3 - a_1)$, in the Frobenius norm $\|\cdot\|_F$ (see (A3.5)) we have

$$\|B^{-1}\|_F = \frac{1}{|\det(B)|}\|B\|_F\,,$$

and further, with the height h_2 over h_K,

$$\det(B) = h_K h_2\,, \tag{3.92}$$

since $\det(B)/2$ is the area of the triangle, as well as

$$\|B\|_F^2 = |a_2 - a_1|_2^2 + |a_3 - a_1|_2^2 \geq h_K^2\,,$$

such that

$$\|B\|_F\|B^{-1}\|_F \geq h_K/h_2\,,$$

and thus by virtue of $\cot\alpha_{\min} < h_K/h_2$,

$$\|B\|_F\|B^{-1}\|_F > \cot\alpha_{\min}\,.$$

Since we get by analogous estimates

$$\|B\|_F\|B^{-1}\|_F \leq 4\cot\alpha_{\min}\,,$$

it follows that $\cot\alpha_{\min}$ describes the asymptotic behavior of $\|B\|\|B^{-1}\|$ for a fixed chosen arbitrary matrix norm. Therefore, from Theorem 3.27 we

get the existence of some constant $C > 0$ independent of h such that for all $K \in \mathcal{T}_h$,

$$\frac{h_K}{\varrho_K} \geq C \cot \alpha_{\min}(K)\,. \tag{3.93}$$

Consequently, a family of triangulations $(\mathcal{T}_h)_h$ of triangles can only be regular if all angles of the triangles are uniformly bounded from below by some positive constant. This condition sometimes is called the *minimum angle condition*. In the situation of Figure 3.11 it would thus not be allowed to decompose the flat rectangles in the thin layer by means of a Friedrichs–Keller triangulation. Obviously, using directly the estimates of Theorem 3.26 we see that the minimum angle condition is sufficient for the estimates of Theorem 3.29. This still leaves the possibility open that less severe conditions are also sufficient.

3.4.2 The Maximum Angle Condition on Triangles

In what follows we show that the condition (3.93) is due only to the techniques of proof, and at least in the case of the linear ansatz, it has indeed only to be enssured that the largest angle is uniformly bounded away from π. Therefore, this allows the application of the described approach in the layer example of Figure 3.11.

The estimate (3.87) shows that for $m = 0$ the crucial part does not arise; hence only for $m = k = 1$ do the estimates have to be investigated. It turns out to be useful to prove the following sharper form of the estimate (3.78):

Theorem 3.31 *For the reference triangle $\hat{K}$ with linear ansatz functions there exists some constant $C > 0$ such that for all $\hat{v} \in H^2(\hat{K})$ and $j = 1, 2$,*

$$\left\| \frac{\partial}{\partial \hat{x}_j} \big(\hat{v} - I_{\hat{K}}(\hat{v})\big) \right\|_{0,\hat{K}} \leq C \left| \frac{\partial}{\partial \hat{x}_j} \hat{v} \right|_{1,\hat{K}}\,.$$

Proof: In order to simplify the notation, we drop the hat $\hat{}$ in the notation of the reference situation in the proof. Hence, we have $K = \operatorname{conv}\{a_1, a_2, a_3\}$ with $a_1 = (0,0)^T$, $a_2 = (1,0)^T$, and $a_3 = (0,1)^T$. We consider the following linear mappings: $F_1 : H^1(K) \to L^2(K)$ is defined by

$$F_1(w) := \int_0^1 w(s,0)\, ds\,,$$

and, analogously, F_2 as the integral over the boundary part $\operatorname{conv}\{a_1, a_3\}$. The image is taken as constant function on K. By virtue of the Trace Theorem (Theorem 3.5), and the continuous embedding of $L^2(0,1)$ in $L^1(0,1)$, the F_i are well-defined and continuous. Since we have for $w \in \mathcal{P}_0(K)$,

$$F_i(w) = w\,,$$

the Bramble–Hilbert lemma (Theorem 3.24) implies the existence of some constant $C > 0$ such that for $w \in H^1(K)$,

$$\|F_i(w) - w\|_{0,K} \le C|w|_{1,K} \,. \tag{3.94}$$

This can be seen in the following way: Let $v \in H^1(K)$ be arbitrary but fixed, and for this, consider on $H^1(K)$ the functional

$$G(w) := \langle F_i(w) - w, F_i(v) - v\rangle \quad \text{for } w \in H^1(K) \,.$$

We have $G(w) = 0$ for $w \in \mathcal{P}_0(K)$ and

$$|G(w)| \le \|F_i(w) - w\|_{0,K}\|F_i(v) - v\|_{0,K} \le C\|F_i(v) - v\|_{0,K}\|w\|_{1,K}$$

by the above consideration. Thus by Theorem 3.24,

$$|G(w)| \le C\,\|F_i(v) - v\|_{0,K}\,|w|_{1,K} \,.$$

For $v = w$ this implies (3.94). On the other hand, for $w := \partial_1 v$ it follows that

$$\begin{aligned} F_1(\partial_1 v) &= v(1,0) - v(0,0) = (I_K(v))(1,0) - (I_K(v))(0,0) = \\ &= \partial_1(I_K(v))(x_1,x_2) \end{aligned}$$

for $(x_1, x_2) \in K$ and, analogously, $F_2(\partial_2 v) = \partial_2(I_K(v))(x_1, x_2)$. This, substituted into (3.94), gives the assertion. □

Compared with estimate (3.78), for example in the case $j = 1$ the term $\frac{\partial^2}{\partial \hat{x}_2^2}\hat{v}$ does not arise on the right-hand side: The derivatives and thus the space directions are therefore treated "more separately."

Next, the effect of the transformation will be estimated more precisely. For this, let $\alpha_{\max} = \alpha_{\max}(K)$ be the largest angle arising in $K \in \mathcal{T}_h$, supposed to include the vertex a_1, and let $h_1 = h_{1K} := |a_2 - a_1|_2$, $h_2 = h_{2K} := |a_3 - a_1|$ (see Figure 3.13).

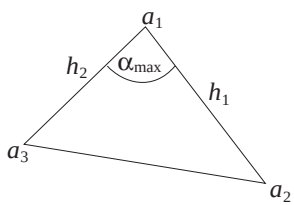


Figure 3.13. A general triangle.

As a variant of (3.86) (for $l = 1$) we have the following:

Theorem 3.32 *Suppose K is a general triangle. With the above notation for $v \in H^1(K)$ and the transformed $\hat{v} \in H^1(\hat{K})$,*

$$|v|_{1,K} \le \sqrt{2}\,|\det(B)|^{-1/2} \left(h_2^2 \left\| \frac{\partial}{\partial \hat{x}_1} \hat{v} \right\|_{0,\hat{K}}^2 + h_1^2 \left\| \frac{\partial}{\partial \hat{x}_2} \hat{v} \right\|_{0,\hat{K}}^2 \right)^{1/2} .$$

Proof: We have

$$B = (a_2 - a_1, a_3 - a_1) =: \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$$

and hence

$$\left| \begin{pmatrix} b_{11} \\ b_{21} \end{pmatrix} \right| = h_1 , \quad \left| \begin{pmatrix} b_{12} \\ b_{22} \end{pmatrix} \right| = h_2 . \tag{3.95}$$

From

$$B^{-T} = \frac{1}{\det(B)} \begin{pmatrix} b_{22} & -b_{21} \\ -b_{12} & b_{11} \end{pmatrix}$$

and (3.84) it thus follows that

$$|v|_{1,K}^2 = \frac{1}{|\det(B)|} \int_{\hat{K}} \left| \begin{pmatrix} b_{22} \\ -b_{12} \end{pmatrix} \frac{\partial}{\partial \hat{x}_1} \hat{v} + \begin{pmatrix} -b_{21} \\ b_{11} \end{pmatrix} \frac{\partial}{\partial \hat{x}_2} \hat{v} \right|^2 d\hat{x}$$

and from this the assertion. □

In modification of the estimate (3.85) (for $l = 2$) we prove the following result:

Theorem 3.33 *Suppose K is a general triangle with diameter $h_K = \operatorname{diam}(K)$. With the above notation for $\hat{v} \in H^2(\hat{K})$ and the transformed $v \in H^2(K)$,*

$$\left| \frac{\partial}{\partial \hat{x}_i} \hat{v} \right|_{1,\hat{K}} \le 4 |\det(B)|^{-1/2} h_i h_K |v|_{2,K} \quad \textit{for } i = 1, 2 .$$

Proof: According to (3.84) we get by exchanging K and $\hat{K}$,

$$|\hat{w}|_{1,\hat{K}}^2 = \int_K B^T \nabla_x w \cdot B^T \nabla_x w \, dx \, |\det(B)|^{-1}$$

and, consequently, for $\hat{w} = \frac{\partial}{\partial \hat{x}_i} \hat{v}$, thus by (3.77) for $w = (B^T \nabla_x v)_i$,

$$\left| \frac{\partial}{\partial \hat{x}_i} \hat{v} \right|_{1,\hat{K}}^2 = \int_K \left| B^T \nabla_x \left(\left(B^T \nabla_x v \right)_i \right) \right|^2 dx \, |\det(B)|^{-1} .$$

According to (3.95), the norm of the ith row vector of B^T is equal to h_i, which implies the assertion. □

Instead of the regularity of the family of triangulations and hence the uniform bound for $\cot \alpha_{\min}(K)$ (see (3.93)) we require the following definition:

Definition 3.34 A family of triangulations $(\mathcal{T}_h)_h$ of triangles satisfies the *maximum angle condition* if there exists some constant $\overline{\alpha} < \pi$ such that for all $h > 0$ and $K \in \mathcal{T}_h$ the maximum angle $\alpha_{\max}(K)$ of K satisfies

$$\alpha_{\max}(K) \leq \overline{\alpha}\,.$$

Since $\alpha_{\max}(K) \geq \pi/3$ is always satisfied, the maximum angle condition is equivalent to the existence of some constant $\tilde{s} > 0$, such that

$$\sin(\alpha_{\max}(K)) \geq \tilde{s} \quad \text{for all } K \in \mathcal{T}_h \text{ and } h > 0\,. \tag{3.96}$$

The relation of this condition to the above estimates is given by (cf. (3.92))

$$\det(B) = h_1 h_2 \sin \alpha_{\max}\,. \tag{3.97}$$

Inserting the estimates of Theorem 3.32 (for $v - I_K(v)$), Theorem 3.31, and Theorem 3.33 into each other and recalling (3.96), (3.97), the following theorem follows from Céa's lemma (Theorem 2.17):

Theorem 3.35 *Consider the linear ansatz* (3.53) *on a family of triangulations $(\mathcal{T}_h)_h$ of triangles that satisfies the maximum angle condition. Then there exists some constant $C > 0$ such that for $v \in H^2(\Omega)$,*

$$\|v - I_h(v)\|_1 \leq C\, h\, |v|_2\,.$$

If the solution u of the boundary value problem (3.12), (3.18)–(3.20) *belongs to $H^2(\Omega)$, then for the finite element approximation u_h defined in* (3.39) *we have the estimate*

$$\|u - u_h\|_1 \leq Ch|u|_2\,. \tag{3.98}$$

Exercise 3.26 shows the necessity of the maximum angle condition. Again, a remark analogous to Remark 3.30 holds. For an analogous investigation of tetrahedra we refer to [58].

With a modification of the above considerations and an additional condition *anisotropic error estimates* of the form

$$|v - I_h(v)|_1 \leq C \sum_{i=1}^{d} h_i\, |\partial_i v|_1$$

can be proven for $v \in H^2(\Omega)$, where the h_i denote length parameter depending on the element type. In the case of triangles, these are the longest edge ($h_1 = h_K$) and the height on it as shown in Figure 3.12 (see [41]).

3.4.3 L^2 Error Estimates

The error estimate (3.89) also contains a result about the approximation of the gradient (and hence of the flux), but it is linear only for $k = 1$, in

contrast to the error estimate of Chapter 1 (Theorem 1.6). The question is whether an improvement of the convergence rate is possible if we strive only for an estimate of the function values. The *duality argument* of Aubin and Nitsche shows that this is correct, if the adjoint boundary value problem is regular, where we have the following definition:

Definition 3.36 The *adjoint boundary value problem* for (3.12), (3.18)–(3.20) is defined by the bilinear form

$$(u, v) \mapsto a(v, u) \quad \text{for } u, v \in V$$

with V from (3.30). It is called *regular* if for every $f \in L^2(\Omega)$ there exists a unique solution $u = u_f \in V$ of the adjoint boundary value problem

$$a(v, u) = \langle f, v \rangle_0 \quad \text{for all } v \in V$$

and even $u_f \in H^2(\Omega)$ is satisfied, and for some constant $C > 0$ a stability estimate of the form

$$|u_f|_2 \le C \|f\|_0 \quad \text{for given } f \in L^2(\Omega)$$

is satisfied.

The V-ellipticity and the continuity of the bilinear form (3.2), (3.3) directly carry over from (3.31) to the adjoint boundary value problem, so that in this case the unique existence of $u_f \in V$ is ensured. More precisely, the adjoint boundary value problem is obtained by an exchange of the arguments in the bilinear form, which does not effect any change in its symmetric parts. The nonsymmetric part of (3.31) is $\int_\Omega c \cdot \nabla u \, v \, dx$, which becomes $\int_\Omega c \cdot \nabla v \, u \, dx$. By virtue of

$$\int_\Omega c \cdot \nabla v \, u \, dx = -\int_\Omega \nabla \cdot (cu) \, v \, dx + \int_{\partial\Omega} c \cdot \nu \, uv \, d\sigma$$

the transition to the adjoint boundary value problem therefore means the exchange of the convective part $c \cdot \nabla u$ by a convective part, now in divergence form and in the opposite direction $-c$, namely $\nabla \cdot (-cu)$, with the correponding modification of the boundary condition. Hence, in general we may expect a similar regularity behavior to that in the original boundary value problem, which was discussed in Section 3.2.3. For a regular adjoint problem we get an improvement of the convergence rate in $\| \cdot \|_0$:

Theorem 3.37 (Aubin and Nitsche)
Consider the situation of Theorem 3.29 *or Theorem* 3.35 *and suppose the adjoint boundary value problem is regular. Then there exists some constant* $C > 0$ *such that for the solution* u *of the boundary value problem* (3.12), (3.18)–(3.20) *and its finite element approximation* u_h *defined by* (3.39),

(1) $$\|u - u_h\|_0 \le Ch \|u - u_h\|_1 ,$$

(2) $$\|u - u_h\|_0 \le Ch \|u\|_1 ,$$

(3) $$\|u - u_h\|_0 \leq Ch^{k+1}|u|_{k+1}\,, \quad \text{if } u \in H^{k+1}(\Omega).$$

Proof: The assertions (2) and (3) follow directly from (1). On the one hand, by using $\|u - u_h\|_1 \leq \|u\|_1 + \|u_h\|_1$ and the stability estimate (2.44), on the other hand directly from (3.89) and (3.98), respectively.

For the proof of (1), we consider the solution u_f of the adjoint problem with the right-hand side $f = u - u_h \in V \subset L^2(\Omega)$. Choosing the test function $u - u_h$ and using the error equation (2.39) gives

$$\|u - u_h\|_0^2 = \langle u - u_h, u - u_h\rangle_0 = a(u - u_h, u_f) = a(u - u_h, u_f - v_h)$$

for all $v_h \in V_h$. If we choose specifically $v_h = I_h(u_f)$, then from the continuity of the bilinear form, Theorem 3.29, and Theorem 3.35, and the regularity assumption it follows that

$$\begin{aligned} \|u - u_h\|_0^2 &\leq C\|u - u_h\|_1 \|u_f - I_h(u_f)\|_1 \\ &\leq C\|u - u_h\|_1 h|u_f|_2 \leq C\|u - u_h\|_1 h\|u - u_h\|_0\,. \end{aligned}$$

Division by $\|u - u_h\|_0$ gives the assertion, which is trivial in the case $\|u - u_h\|_0 = 0$. □

Thus, if a rough right-hand side in (3.12) prevents convergence from being ensured by Theorem 3.29 or Theorem 3.35, then the estimate (2) can still be used to get a convergence estimate (of lower order).

In the light of the considerations from Section 1.2, the result of Theorem 3.37 is surprising, since we have only (pointwise) consistency of first order. On the other hand, Theorem 1.6 also raises the question of convergence rate results in $\|\cdot\|_\infty$ which then would give a result stronger, in many respects, than Theorem 1.6. Although the considerations described here (as in Section 3.9) can be the starting point of such L^∞ estimates, we get the most far-reaching results with the weighted norm technique (see [9, pp. 155 ff.]), whose description is not presented here.

The above theorems contain convergence rate results under regularity assumptions that may often, even though only locally, be violated. In fact, there also exist (weaker) results with less regularity assumptions. However, the following observation seems to be meaningful: Estimate (3.90) indicates that on subdomains, where the solution has less regularity, on which the (semi) norms of the solutions thus become large, local refinement is advantageous (without improving the convergence rate by this). Adaptive mesh refinement strategies on the basis of a posteriori error estimates described in Chapter 4 provide a systematical approach in this direction.

Exercises

3.21 Prove for the linear finite element ansatz (3.53) in one space dimension that for $K \in \mathcal{T}_h$ and $v \in H^2(K)$, the following estimate

holds:

$$|v - I_K(v)|_{1,K} \leq h_K |v|_{2,K} .$$

Hint: Rolle's theorem and Exercise 2.5 (b) (Poincaré inequality).

Generalize the considerations to an arbitrary polynomial ansatz $P = \mathcal{P}_k$ in one space dimension by proving

$$|v - I_K(v)|_{1,K} \leq h_K^k |v|_{k+1,K} \quad \text{for } v \in H^{k+1}(K) .$$

3.22 Prove the chain rule (3.77) for $v \in H^1(K)$.

3.23 Derive analogously to Theorem 3.29 a convergence rate result for the Hermite elements (3.64) and (3.65) (Bogner–Fox–Schmit element) and the boundary value problem (3.12) with Dirichlet boundary conditions.

3.24 Derive analogously to Theorem 3.29 a convergence rate result for the Bogner–Fox–Schmit element (3.65) and the boundary value problem (3.36).

3.25 Let a triangle K with the vertices a_1, a_2, a_3 and a function $u \in C^2(K)$ be given. Show that if u is interpolated by a linear polynomial $I_K(u)$ with $(I_K(u))(a_i) = u(a_i)$, $i = 1, 2, 3$, then, for the error the estimate

$$\sup_{x \in K} |u(x) - (I_K(u))(x)| + h \sup_{x \in K} |\nabla(u - I_K(u))(x)| \leq 2M \frac{h^2}{\cos(\alpha/2)}$$

holds, where h denotes the diameter, α the size of the largest interior angle of K and M an upper bound for the maximum of the norm of the Hessian matrix of u on K.

3.26 Consider a triangle K with the vertices $a_1 := (-h, 0)$, $a_2 := (h, 0)$, $a_3 := (0, \varepsilon)$, and $h, \varepsilon > 0$. Suppose that the function $u(x) := x_1^2$ is linearly interpolated on K such that $(I_h(u))(a_i) = u(a_i)$ for $i = 1, 2, 3$.
Determine $\|\partial_2(I_h(u) - u)\|_{2,K}$ as well as $\|\partial_2(I_h(u) - u)\|_{\infty,K}$ and discuss the consequences for of different orders of magnitude of h and ε.

3.27 Suppose that no further regularity properties are known for the solution $u \in V$ of the boundary value problem (3.12). Show under the assumptions of Section 3.4 that for the finite element approximation $u_h \in V_h$

$$\|u - u_h\|_1 \to 0 \quad \text{for } h \to 0 .$$

3.5 The Implementation of the Finite Element Method: Part 2

3.5.1 Incorporation of Dirichlet Boundary Conditions: Part 2

In the theoretical analysis of boundary value problems with inhomogeneous Dirichlet boundary conditions $u = g_3$ on Γ_3, the existence of a function $w \in H^1(\Omega)$ with $w = g_3$ on Γ_3 has been assumed so far. The solution $u \in V$ (with homogeneous Dirichlet boundary conditions) is then defined according to (3.31) such that $\tilde{u} = u + w$ satisfies the variational equation with test functions in V:

$$a(u + w, v) = b(v) \quad \text{for all } v \in V\,. \tag{3.99}$$

For the Galerkin approximation u_h, which has been analyzed in Section 3.4, this means that the parts $-a(w, \varphi_i)$ with nodal basis functions φ_i, $i = 1, \ldots, M_1$, go into the right-hand side of the system of equations (2.34), and then $\tilde{u}_h := u_h + w$ has to be considered as the solution of the inhomogeneous problem

$$a(u_h + w, v) = b(v) \quad \text{for all } v \in V_h\,. \tag{3.100}$$

If we complete the basis of V_h by the basis functions $\varphi_{M_1+1}, \ldots, \varphi_M$ for the Dirichlet boundary nodes $a_{M_1+1}, \ldots, a_M$ and denote the generated space by X_h,

$$X_h = \text{span}\,\{\varphi_1, \ldots, \varphi_{M_1}, \varphi_{M_1+1}, \ldots, \varphi_M\}\,, \tag{3.101}$$

that is the ansatz space without taking into account boundary conditions, then in particular, $\tilde{u}_h \in X_h$ does not hold in general. This approach does not correspond to the practice described in Section 2.4.3. That practice, applied to a general variational equation, reads as follows:

For all degrees of freedom $1, \ldots, M_1, M_1 + 1, \ldots, M$ the system of equations is built with the components

$$a(\varphi_j, \varphi_i)\,, \quad i, j = 1, \ldots, M\,, \tag{3.102}$$

for the stiffness matrix and

$$b(\varphi_i)\,, \quad i = 1, \ldots, M\,, \tag{3.103}$$

for the load vector. The vector of unknowns is therefore

$$\tilde{\boldsymbol{\xi}} = \begin{pmatrix} \boldsymbol{\xi} \\ \hat{\boldsymbol{\xi}} \end{pmatrix} \quad \text{with} \quad \boldsymbol{\xi} \in \mathbb{R}^{M_1}\,,\ \hat{\boldsymbol{\xi}} \in \mathbb{R}^{M_2}\,.$$

For Dirichlet boundary conditions the equations $M_1+1, \ldots, M$ are replaced by

$$\tilde{\xi}_i = g_3(a_i)\,, \quad i = M_1 + 1, \ldots, M\,,$$

and the concerned variables are eliminated in equations $1, \ldots, M_1$. Of course, it is assumed here that $g_3 \in C(\Gamma_3)$. This procedure can also be interpreted in the following way: If we set

$$A_h := (a(\varphi_j, \varphi_i))_{i,j=1,\ldots,M_1}\,,\quad \hat{A}_h := (a(\varphi_j, \varphi_i))_{i=1,\ldots,M_1,\, j=M_1+1,\ldots,M}\,,$$

then the first M_1 equations of the generated system of equations are

$$A_h \boldsymbol{\xi} + \hat{A}_h \hat{\boldsymbol{\xi}} = \boldsymbol{q}_h\,,$$

where $\boldsymbol{q}_h \in \mathbb{R}^{M_1}$ consists of the first M_1 components according to (3.103). Hence the elimination leads to

$$A_h \boldsymbol{\xi} = \boldsymbol{q}_h - \hat{A}_h \hat{\boldsymbol{\xi}} \tag{3.104}$$

with $\hat{\boldsymbol{\xi}} = (g_3(a_i))_{i=M_1+1,\ldots,M_2}$. Suppose

$$w_h := \sum_{i=M_1+1}^{M} g_3(a_i)\,\varphi_i \in X_h \tag{3.105}$$

is the ansatz function that satisfies the boundary conditions in the Dirichlet nodes and assumes the value 0 in all other nodes. The system of equations (3.104) is then equivalent to

$$a(\breve{u}_h + w_h, v) = b(v) \quad \text{for all } v \in V_h \tag{3.106}$$

for $\breve{u}_h = \sum_{i=1}^{M_1} \xi_i \varphi_i \in V_h$ (that is, the "real" solution), in contrast to the variational equation (3.100) was used in the analysis. This consideration also holds if another h-dependent bilinear form a_h and analogously a linear form b_h instead of the linear form b is used for assembling. In the following we assume that there exists some function $w \in C(\bar{\Omega})$ that satisfies the boundary condition on Γ_3. Instead of (3.106), we consider the finite-dimensional auxiliary problem of finding some $\breve{\breve{u}}_h \in V_h$, such that

$$a(\breve{\breve{u}}_h + \bar{I}_h(w), v) = b(v) \quad \text{for all } v \in V_h\,. \tag{3.107}$$

Here $\bar{I}_h : C(\bar{\Omega}) \to X_h$ is the interpolation operator with respect to all degrees of freedom,

$$\bar{I}_h(v) := \sum_{i=1}^{M_1+M_2} v(a_i)\varphi_i\,,$$

whereas in Section 3.4 we considered the interpolation operator I_h for functions that vanish on Γ_3. In the following, when analyzing the effect of quadrature, we will show that — also for some approximation of a and b —

$$\tilde{u}_h := \breve{\breve{u}}_h + \bar{I}_h(w) \in X_h \tag{3.108}$$

is an approximation of $u + w$ of the quality established in Theorem 3.29 (see Theorem 3.42). We have $w_h - \bar{I}_h(w) \in V_h$ and hence also $\breve{u}_h + w_h -$

$\bar{I}_h(w) \in V_h$. If (3.107) is uniquely solvable, which follows from the general assumption of the V-ellipticity of a (3.3), we have

$$\breve{u}_h + w_h - \bar{I}_h(w) = \breve{\breve{u}}_h$$

and hence for $\tilde{u}_h$, according to (3.108),

$$\tilde{u}_h = \breve{u}_h + w_h \,. \tag{3.109}$$

In this way the described implementation practice for Dirichlet boundary conditions is justified.

3.5.2 Numerical Quadrature

We consider again a boundary value problem in the variational formulation (3.31) and a finite element discretization in the general form described in Sections 3.3 and 3.4. If we step through Section 2.4.2 describing the assembling within a finite element code, we notice that the general element-to-element approach with transformation to the reference element is here also possible, with the exception that due to the general coefficient functions K, c, r and f, the arising integrals can not be evaluated exactly in general. If K_m is a general element with degrees of freedom in $a_{r_1}, \ldots, a_{r_L}$, then the components of the element stiffness matrix for $i, j = 1, \ldots, L$ are

$$\begin{aligned} A_{ij}^{(m)} &= \int_{K_m} K\nabla\varphi_{r_j} \cdot \nabla\varphi_{r_i} + c \cdot \nabla\varphi_{r_j}\varphi_{r_i} + r\varphi_{r_j}\varphi_{r_i}\, dx \\ &\quad + \int_{K_m \cap \Gamma_2} \alpha\varphi_{r_j}\varphi_{r_i} d\sigma \qquad (3.110) \\ &=: \int_{K_m} v_{ij}(x)\, dx + \int_{K_m \cap \Gamma_2} w_{ij}(\sigma)\, d\sigma \\ &= \int_{\hat{K}} \hat{v}_{ij}(\hat{x})\, d\hat{x}\, |\det(B)| + \int_{\hat{K}'} \hat{w}_{ij}(\hat{\sigma})\, d\hat{\sigma}\, |\det(\tilde{B})| \,. \end{aligned}$$

Here, K_m is affine equivalent to the reference element $\hat{K}$ by the mapping $F(\hat{x}) = B\hat{x} + d$. By virtue of the conformity of the triangulation (T6), the boundary part $K_m \cap \bar{\Gamma}_2$ consists of none, one, or more complete faces of K_m. For simplicity, we restrict ourselves to the case of one face that is affine equivalent to the reference element $\hat{K}'$ by some mapping $\tilde{F}(\hat{\sigma}) = \tilde{B}\hat{\sigma} + \tilde{d}$ (cf. (3.42)). The generalization to the other cases is obvious. The functions $\hat{v}_{ij}$ and analogously $\hat{w}_{ij}$ are the transformed functions defined in (3.75).

Correspondingly, we get as components for the right-hand side of the system of equations, that is, for the load vector,

$$\begin{aligned} \left(\boldsymbol{q}^{(m)}\right)_i &= \int_{\hat{K}} \hat{f}(\hat{x}) N_i(\hat{x})\, d\hat{x}\, |\det(B)| \qquad (3.111) \\ &\quad + \int_{\hat{K}_1'} \hat{g}_1(\hat{\sigma}) N_i(\hat{\sigma})\, d\hat{\sigma}\, |\det(\tilde{B}_1)| + \int_{\hat{K}_2'} \hat{g}_2(\hat{\sigma}) N_i(\hat{\sigma})\, d\hat{\sigma}\, |\det(\tilde{B}_2)| \,. \end{aligned}$$

$i = 1, \ldots, L$. Here, the N_i, $i = 1, \ldots, L$, are the shape functions; that is, the local nodal basis functions on $\hat{K}$.

If the transformed integrands contain derivatives with respect to x, they can be transformed into derivatives with respect to $\hat{x}$. For instance, for the first addend in $A_{ij}^{(m)}$ we get, as an extension of (2.50),

$$\int_{\hat{K}} K(F(\hat{x}))B^{-T}\nabla_{\hat{x}}N_j(\hat{x}) \cdot B^{-T}\nabla_{\hat{x}}N_i(\hat{x})\, d\hat{x}\, |\det(B)|\,.$$

The shape functions, their derivatives, and their integrals over $\hat{K}$ are known which has been used in (2.52) for the exact integration. Since general coefficient functions arise, this is in general, but also in the remaining special cases no longer possible, for example for polynomial $K(x)$ it is also not recommendable due to the corresponding effort. Instead, one should approximate these integrals (and, analogously, also the boundary integrals) by using some *quadrature formula*.

A quadrature formula on $\hat{K}$ for the approximation of $\int_{\hat{K}} \hat{v}(\hat{x})\, d\hat{x}$ has the form

$$\sum_{i=1}^{R} \hat{\omega}_i\, \hat{v}(\hat{b}_i) \tag{3.112}$$

with *weights* $\hat{\omega}_i$ and *quadrature* or *integration points* $\hat{b}_i \in \hat{K}$. Hence, applying (3.112) assumes the evaluability of $\hat{v}$ in $\hat{b}_i$, which is in the following ensured by the continuity of $\hat{v}$. This implies the same assumption for the coefficients, since the shape functions N_i and their derivatives are continuous. In order to ensure the numerical stability of a quadrature formula, it is usually required that

$$\hat{\omega}_i > 0 \quad \text{for all} \quad i = 1, \ldots, R\,, \tag{3.113}$$

which we will also do. Since all the considered finite elements are such that their faces with the enclosed degrees of freedom represent again a finite element (in $\mathbb{R}^{d-1}$) (see (3.42)), the boundary integrals are included in a general discussion. In principle, different quadrature formulas can be applied for each of the above integrals, but here we will disregard this possibility (with the exception of distinguishing between volume and boundary integrals because of their different dimensions).

A quadrature formula on $\hat{K}$ generates a quadrature formula on a general element K, recalling

$$\int_K v(x)\, dx = \int_{\hat{K}} \hat{v}(\hat{x})\, d\hat{x}\, |\det(B)|$$

by

$$\sum_{i=1}^{R} \omega_{i,K}\, v(b_{i,K})\,,$$

where $\omega_i = \omega_{i,K} = \hat{\omega}_i |\det(B)|$ and $b_i = b_{i,K} := F(\hat{b}_i)$ are dependent on K. The positivity of the weights is preserved. Here, again $F(\hat{x}) = B\hat{x} + d$ denotes the affine-linear transformation from $\hat{K}$ to K. The errors of the quadrature formulas

$$\begin{aligned} \hat{E}(\hat{v}) &:= \int_{\hat{K}} \hat{v}(\hat{x})\, d\hat{x} - \sum_{i=1}^{R} \hat{\omega}_i\, \hat{v}(\hat{b}_i)\,, \\ E_K(v) &:= \int_K v(x)\, dx - \sum_{i=1}^{R} \omega_i\, v(b_i) \end{aligned} \tag{3.114}$$

are related to each other by

$$E_K(v) = |\det(B)| \hat{E}(\hat{v})\,. \tag{3.115}$$

The *accuracy* of a quadrature formula will be defined by the requirement that for l as large as possible,

$$\hat{E}(\hat{p}) = 0 \quad \text{for } \hat{p} \in \mathcal{P}_l(\hat{K})$$

is satisfied, which transfers directly to the integration over K. A quadrature formula should further provide the desired accuracy by using quadrature nodes as less as possible, since the evaluation of the coefficient functions is often expensive. In contrast, for the shape functions and their derivatives a single evaluation is sufficient. In the following we discuss some examples of quadrature formulas for the elements that have been introduced in Section 3.3.

The most obvious approach consists in using *nodal quadrature formulas*, which have the nodes $\hat{a}_1, \ldots, \hat{a}_L$ of the reference element $(\hat{K}, \hat{P}, \hat{\Sigma})$ as quadrature nodes. The requirement of exactness in $\hat{P}$ is then equivalent to

$$\hat{\omega}_i = \int_{\hat{K}} N_i(\hat{x})\, d\hat{x}\,, \tag{3.116}$$

so that the question of the validity of (3.113) remains.

We start with the **unit simplex $\hat{K}$** defined in (3.47). Here, the weights of the quadrature formulas can be given directly on a general simplex K: If the shape functions are expressed by their barycentric coordinates λ_i, the integrals can be computed by

$$\int_K \lambda_1^{\alpha_1} \lambda_2^{\alpha_2} \cdots \lambda_{d+1}^{\alpha_{d+1}}(x)\, dx = \frac{\alpha_1! \alpha_2! \cdots \alpha_{d+1}!}{(\alpha_1 + \alpha_2 + \cdots + \alpha_{d+1} + d)!} \frac{\operatorname{vol}(K)}{\operatorname{vol}(\hat{K})} \tag{3.117}$$

(see Exercise 3.28).

If $\boldsymbol{P = \mathcal{P}_1(K)}$ and thus the quadrature nodes are the vertices, it follows that

$$\omega_i = \int_K \lambda_i(x)\, dx \; = \frac{1}{d+1} \operatorname{vol}(K) \quad \text{for all } i = 1, \ldots, d+1\,. \tag{3.118}$$

For $\boldsymbol{P = \mathcal{P}_2(K)}$ and $d = 2$ we get, by the shape functions $\lambda_i(2\lambda_i - 1)$, the weights 0 for the nodes a_i and, by the shape functions $4\lambda_i\lambda_j$, the weights

$$\omega_i = \frac{1}{3}\mathrm{vol}\,(K) \quad \text{for} \quad b_i = a_{ij}\,,\ i,j = 1,\ldots,3\,,\ i > j\,,$$

so that we have obtained here a quadrature formula that is superior to (3.118) (for $d = 2$). However, for $d \geq 3$ this ansatz leads to negative weights and is thus useless. We can also get the exactness in $\boldsymbol{\mathcal{P}_1(K)}$ by a single quadrature node, by the barycentre (see (3.52)):

$$\omega_1 = \mathrm{vol}\,(K) \quad \text{and} \quad b_1 = a_S = \frac{1}{d+1}\sum_{i=1}^{d+1} a_i\,,$$

which is obvious due to (3.117).

As a formula that is exact for $\boldsymbol{\mathcal{P}_2(K)}$ and $d = 3$ (see [53]) we present $R = 4$, $\omega_i = \frac{1}{4}\,\mathrm{vol}\,(K)$, and the b_i are obtained by cyclic exchange of the barycentric coordinates:

$$\left(\frac{5-\sqrt{5}}{20}\,,\frac{5-\sqrt{5}}{20}\,,\frac{5-\sqrt{5}}{20}\,,\frac{5+3\sqrt{5}}{20}\right).$$

On the **unit cuboid** $\boldsymbol{\hat{K}}$ we obtain nodal quadrature formulas, which are exact for $\boldsymbol{Q_k(\hat{K})}$, from the Newton–Côtes formulas in the one-dimensional situation by

$$\begin{aligned}\hat{\omega}_{i_1\ldots i_d} \;&=\; \hat{\omega}_{i_1}\cdots\hat{\omega}_{i_d} \quad \text{for} \quad \hat{b}_{i_1\ldots i_d} = \left(\frac{i_1}{k},\ldots,\frac{i_d}{k}\right) \qquad (3.119)\\ &\qquad \text{for } i_j \in \{0,\ldots,k\} \text{ and } j = 1,\ldots,d\,.\end{aligned}$$

Here the $\hat{\omega}_{i_j}$ are the weights of the Newton–Côtes formula for $\int_0^1 f(x)dx$ (see [30, p. 128]). As in (3.118), for $k = 1$ we have here a generalization of the *trapezoidal rule* (cf. (2.38), (8.31)) with the weights 2^{-d} in the 2^d vertices. From $k = 8$ on, negative weights arise. This can be avoided and the accuracy for a given number of points increased if the Newton–Côtes integration is replaced by the *Gauss–(Legendre) integration*: In (3.119), i_j/k has to be replaced by the jth node of the kth Gauss–Legendre formula (see [30, p. 156] there on $[-1,1]$) and analogously $\hat{\omega}_{i_j}$. In this way, by $(k+1)^d$ quadrature nodes the exactness in $\boldsymbol{Q_{2k+1}(\hat{K})}$, not only in $Q_k(\hat{K})$, is obtained.

Now the question as to which quadrature formula should be chosen arises. For this, different criteria can be considered (see also (8.29)). Here, we require that the convergence rate result that was proved in Theorem 3.29 should not be deteriorated. In order to investigate this question we have to clarify which problem is solved by the approximation $\bar{u}_h \in V_h$ based on quadrature. To simplify the notation, from now on we do not consider boundary integrals, that is, only Dirichlet and homogeneous Neumann

boundary conditions are allowed. However, the generalization should be clear. Replacing the integrals in (3.111) and (3.111) by quadrature formulas $\sum_{i=1}^{R} \hat{\omega}_i \hat{v}(\hat{b}_i)$ leads to some approximation $\bar{A}_h$ of the stiffness matrix and $\bar{\boldsymbol{q}}_h$ of the load vector in the form

$$\bar{A}_h = (a_h(\varphi_j, \varphi_i))_{i,j} \ , \quad \bar{\boldsymbol{q}}_h = (b_h(\varphi_i))_i \ ,$$

for $i, j = 1, \ldots, M$. Here the φ_i are the basis functions of X_h (see (3.101)) without taking into account the Dirichlet boundary condition and

$$\begin{aligned} a_h(v,w) \quad &:= \quad \sum_{K \in \mathcal{T}_h} \sum_{l=1}^{R} \omega_{l,K} (K \nabla v \cdot \nabla w)(b_{l,K}) \\ &\quad + \sum_{K \in \mathcal{T}_h} \sum_{l=1}^{R} \omega_{l,K} (c \cdot \nabla v w)(b_{l,K}) + \sum_{K \in \mathcal{T}_h} \sum_{l=1} \omega_{l,K} (r v w)(b_{l,K}) \\ &\quad \text{for } v, w \in X_h \, , \qquad (3.120) \\ b_h(v) \quad &:= \quad \sum_{K \in \mathcal{T}_h} \sum_{l=1} \omega_{l,K} (f v)(b_{l,K}) \qquad \text{for } v \in X_h \, . \end{aligned}$$

The above-given mappings a_h and b_h are well-defined on $X_h \times X_h$ and X_h, respectively, if the coefficient functions can be evaluated in the quadrature nodes. Here we take into account that for some element K, ∇v for $v \in X_h$ can have jump discontinuities on ∂K. Thus, for the quadrature nodes $b_{l,K} \in \partial K$ in $\nabla v(b_{l,K})$ we have to choose the value "belonging to $b_{l,K}$" that corresponds to the limit of sequences in the interior of K. We recall that in general a_h and b_h are not defined for functions of V. Obviously, a_h is bilinear and b_h is linear. If we take into account the analysis of incorporating the Dirichlet boundary conditions in (3.99)–(3.106), we get a system of equations for the degrees of freedom $\bar{\boldsymbol{\xi}} = (\xi_1, \ldots, \xi_{M_1})^T$, which is equivalent to the variational equation on V_h for $\bar{u}_h = \sum_{i=1}^{M_1} \bar{\xi}_i \varphi_i \in V_h$:

$$a_h(\bar{u}_h, v) = b_h(v) - a_h(w_h, v) \quad \text{for all } v \in V_h \qquad (3.121)$$

with w_h according to (3.105). As has been shown in (3.109), (3.121) is equivalent, in the sense of the total approximation $\bar{u}_h + w_h$ of $u + w$, to the variational equation for $\bar{\bar{u}}_h \in V_h$,

$$a_h(\bar{\bar{u}}_h, v) = \bar{b}_h(v) := b_h(v) - a_h(\bar{I}_h(w), v) \quad \text{for all } v \in V_h \, , \qquad (3.122)$$

if this system of equations is uniquely solvable.

Exercises

3.28 Prove equation (3.117) by first proving the equation for $K = \hat{K}$ and then deducing from this the assertion for the general simplex by Exercise 3.18.

3.29 Let K be a triangle with vertices a_1, a_2, a_3. Further, let a_{12}, a_{13}, a_{23} denote the corresponding edge midpoints, a_{123} the barycenter and $|K|$ the area of K. Check that the quadrature formula

$$Q_h(u) := \frac{|K|}{60}\left[3\sum_{i=1}^{3} u(a_i) + 8\sum_{i<j} u(a_{ij}) + 27u(a_{123})\right]$$

computes the integral $Q(u) := \int_K u\,dx$ exactly for polynomials of third degree.

3.6 Convergence Rate Results in the Case of Quadrature and Interpolation

The purpose of this section is to analyze the approximation quality of a solution $\bar{\bar{u}}_h + \bar{I}_h(w)$ according to (3.122) and thus of $\bar{u}_h + w_h$ according to (3.121) of the boundary value problem (3.12), (3.18)–(3.20).

Hence, we have left the field of Galerkin methods, and we have to investigate the influence of the errors

$$a - a_h\,, \quad b - a(w,\cdot) - b_h + a_h(\bar{I}_h(w),\cdot).$$

To this end, we consider in general the variational equation in a normed space $(V, \|\cdot\|)$

$$u \in V \text{ satisfies} \qquad a(u,v) = l(v) \qquad \text{for all } v \in V\,, \tag{3.123}$$

and the approximation in subspaces $V_h \subset V$ for $h > 0$,

$$u_h \in V_h \text{ satisfies} \quad a_h(u_h, v) = l_h(v) \quad \text{for all } v \in V_h\,. \tag{3.124}$$

Here a and a_h are bilinear forms on $V \times V$ and $V_h \times V_h$, respectively, and l, l_h are linear forms on V and V_h, respectively. Then we have the following theorem

Theorem 3.38 (First Lemma of Strang)
Suppose there exists some $\alpha > 0$ such that for all $h > 0$ and $v \in V_h$,

$$\alpha\|v\|^2 \le a_h(v,v)\,, \tag{3.125}$$

and let a be continuous in $V \times V$.

Then, there exists some constant C independent of V_h such that

$$\begin{aligned}\|u - u_h\| \quad \le \quad & C\left\{\inf_{v\in V_h}\left\{\|u-v\| + \sup_{w\in V_h}\frac{|a(v,w) - a_h(v,w)|}{\|w\|}\right\}\right. \\ & \left. + \sup_{w\in V_h}\frac{|l(w) - l_h(w)|}{\|w\|}\right\}.\end{aligned} \tag{3.126}$$

Proof: Let $v \in V_h$ be arbitrary. Then it follows from (3.123)–(3.125) that

$$\begin{aligned}\alpha\|u_h - v\|^2 &\leq a_h(u_h - v, u_h - v)\\ &= a(u - v, u_h - v) + \big(a(v, u_h - v) - a_h(v, u_h - v)\big)\\ &\quad + \big(l_h(u_h - v) - l(u_h - v)\big)\end{aligned}$$

and moreover, by the continuity of a (cf. (3.2)),

$$\begin{aligned}\alpha\|u_h - v\| \leq\ & M\|u - v\| + \sup_{w\in V_h} \frac{|a(v,w) - a_h(v,w)|}{\|w\|}\\ &+ \sup_{w\in V_h} \frac{|l_h(w) - l(w)|}{\|w\|} \quad \text{for } v \in V_h\,.\end{aligned}$$

By means of $\|u - u_h\| \leq \|u - v\| + \|u_h - v\|$ and taking the infimum over all $v \in V_h$, the assertion follows. □

For $a_h = a$ and $l_h = l$ the assertion reduces to Céa's lemma (Theorem 2.17), which was the initial point for the analysis of the convergence rate in Section 3.4. Here we can proceed analogously. For that purpose, the following conditions must be fulfilled additionally:

- The *uniform V_h-ellipticity* of a_h according to (3.125) must be ensured.
- For the *consistency errors*

$$A_h(v) := \sup_{w\in V_h} \frac{|a(v,w) - a_h(v,w)|}{\|w\|} \tag{3.127}$$

  for an arbitrarily chosen comparison function $v \in V_h$ and for

$$\sup_{w\in V_h} \frac{|l(w) - l_h(w)|}{\|w\|}$$

  the behavior in h must be analyzed.

The first requirement is not crucial if only a itself is V-elliptic and A_h tends suitably to 0 for $h \to 0$:

Lemma 3.39 *Suppose the bilinear form a is V-elliptic and there exists some function $C(h)$ with $C(h) \to 0$ for $h \to 0$ such that*

$$A_h(v) \leq C(h)\,\|v\| \quad \textit{for } v \in V_h\,.$$

Then there exists some $\bar{h} > 0$ such that a_h is uniformly V_h-elliptic for $h \leq \bar{h}$.

Proof: By assumption, there exists some $\alpha > 0$ such that for $v \in V_h$,

$$\alpha\|v\|^2 \leq a_h(v,v) + a(v,v) - a_h(v,v)$$

and

$$|a(v,v) - a_h(v,v)| \leq A_h(v)\|v\| \leq C(h)\|v\|^2\,.$$

Therefore, for instance, choose $\bar{h}$ such that $C(h) \leq \alpha/2$ for $h \leq \bar{h}$. □

We concretely address the analysis of the influence of numerical quadrature, that is, a_h is defined as in (3.120) and l_h corresponds to $\bar{b}_h$ in (3.122) with the approximate linear form b_h according to (3.120). Since this is an extension of the convergence results (in $\|\cdot\|_1$) given in Section 3.4, the assumptions about the finite element discretization are as summarized there at the beginning. In particular, the triangulations $\mathcal{T}_h$ consist of elements that are affine equivalent to each other. Furthermore, for a simplification of the notation, let again $d \leq 3$ and only Lagrange elements are considered. In particular, let the general assumptions about the boundary value problems which are specified at the end of Section 3.2.1 be satisfied.

According to Theorem 3.38, the uniform V_h-ellipticity of a_h must be ensured and the consistency errors (for an appropriate comparison element $v \in V_h$) must have the correct convergence behavior. If the step size h is small enough, the first proposition is implied by the second proposition by virtue of Lemma 3.39. Now, simple criteria that are independent of this restriction will be presented. The quadrature formulas satisfy the properties (3.112), (3.113) introduced in Section 3.5; in particular, the weights are positive.

Lemma 3.40 *Suppose the coefficient function K satisfies* (3.16) *and let $c = 0$ in Ω, let $|\Gamma_3|_{d-1} > 0$, and let $r \geq 0$ in Ω. If $P \subset \mathcal{P}_k(K)$ for the ansatz space and if the quadrature formula is exact for $\mathcal{P}_{2k-2}(K)$, then a_h is uniformly V_h-elliptic.*

Proof: Let $\alpha > 0$ be the constant of the uniform positive definiteness of $K(x)$. Then we have for $v \in V_h$:

$$a_h(v,v) \geq \alpha \sum_{K \in \mathcal{T}_h} \sum_{l=1}^{R} \omega_{l,K} \, |\nabla v|^2(b_{l,K}) = \alpha \int_\Omega |\nabla v|^2(x) \, dx = \alpha |v|_1^2 \,,$$

since $|\nabla v|^2\big|_K \in \mathcal{P}_{2k-2}(K)$. The assertion follows from Corollary 3.14. □

Further results of this type can be found in [9, pp. 194]. To investigate the consistency error we can proceed similarly to the estimation of the interpolation error in Section 3.4: The error is split into the sum of the errors over the elements $K \in \mathcal{T}_h$ and there transformed by means of (3.115) into the error over the reference element $\hat{K}$. The derivatives (in $\hat{x}$) arising in the error estimation over $\hat{K}$ are backtransformed by using Theorem 3.26 and Theorem 3.27, which leads to the desired h_K-factors. But note that powers of $\|B^{-1}\|$ or similar terms do not arise. If the powers of $\det(B)$ arising in both transformation steps cancel each other (which will happen), in this way no condition about the geometric quality of the family of triangulations arises. Of course, these results must be combined with estimates for the

approximation error of V_h, for which, in particular, both approaches of Section 3.4 (either regularity or maximum angle condition) are admissible.

For the sake of simplicity, we restrict our attention in the following to the case of the polynomial ansatz space $P = \mathcal{P}_k(K)$. More general results of similar type, in particular for triangulations with the cuboid element and $\hat{P} = Q_k(\hat{K})$ as reference element, are summarized in [9, p. 207].

We recall the notation and the relations introduced in (3.114), (3.115) for the local errors. In the following theorems we make use of the Sobolev spaces W^l_∞ on Ω and on K with the norms $\|\cdot\|_{l,\infty}$ and $\|\cdot\|_{l,\infty,K}$, respectively, and the seminorms $|\cdot|_{l,\infty}$ and $|\cdot|_{l,\infty,K}$, respectively. The essential local assertion is the following:

Theorem 3.41 *Suppose $k \in \mathbb{N}$ and $\hat{P} = \mathcal{P}_k(\hat{K})$ and the quadrature formula is exact for $\mathcal{P}_{2k-2}(\hat{K})$:*

$$\hat{E}(\hat{v}) = 0 \quad \text{for all } \hat{v} \in \mathcal{P}_{2k-2}(\hat{K})\,. \tag{3.128}$$

Then there exist some constant $C > 0$ independent of $h > 0$ and $K \in \mathcal{T}_h$ such that for $l \in \{1, k\}$ the following estimates are given:

(1) $$|E_K(apq)| \leq Ch_K^l \|a\|_{k,\infty,K} \|p\|_{l-1,K} \|q\|_{0,K}$$
for $a \in W^k_\infty(K)\,,\ p, q \in \mathcal{P}_{k-1}(K)\,,$

(2) $$|E_K(cpq)| \leq Ch_K^l \|c\|_{k,\infty,K} \|p\|_{l-1,K} \|q\|_{1,K}$$
for $c \in W^k_\infty(K)\,,\ p \in \mathcal{P}_{k-1}(K)\,,\ q \in \mathcal{P}_k(K)\,,$

(3) $$|E_K(rpq)| \leq Ch_K^l \|r\|_{k,\infty,K} \|p\|_{l,K} \|q\|_{1,K}$$
for $r \in W^k_\infty(K)\,,\ p, q \in \mathcal{P}_k(K)\,,$

(4) $$|E_K(fq)| \leq Ch_K^k \|f\|_{k,\infty,K} \mathrm{vol}\,(K)^{1/2} \|q\|_{1,K}$$
for $f \in W^k_\infty(K)\,,\ q \in \mathcal{P}_k(K)\,.$

The (unnecessarily varied) notation of the coefficients already indicates the field of application of the respective estimate. The smoothness assumption concerning the coefficients in (1)–(3) can be weakened to some extent. We prove only assertion (1). However, a direct application of this proof to assertions (2)–(4) leads to a loss of convergence rate (or higher exactness conditions for the quadrature). Here, quite technical considerations including the insertion of projections are necessary, which can be found to some extent in [9, pp. 201–203]. In the following proof we intensively make use of the fact that all norms are equivalent on the "fixed" finite-dimensional ansatz space $\mathcal{P}_k(\hat{K})$. The assumption (3.128) is equivalent to the same condition on a general element. However, the formulation already indicates an assumption that is also sufficient in more general cases.

Proof of Theorem 3.41, (1): We consider a general element $K \in \mathcal{T}_h$ and mappings $a \in W_\infty^k(K)$, $p, q \in \mathcal{P}_{k-1}(K)$ on it and, moreover, mappings $\hat{a} \in W_\infty^k(\hat{K})$, $\hat{p}, \hat{q} \in \mathcal{P}_{k-1}(\hat{K})$ defined according to (3.75). First, the proof is done for $l = k$. On the reference element $\hat{K}$, for $\hat{v} \in W_\infty^k(\hat{K})$ and $\hat{q} \in \mathcal{P}_{k-1}(\hat{K})$), we have

$$\left|\hat{E}(\hat{v}\hat{q})\right| = \left| \int_{\hat{K}} \hat{v}\hat{q}\, d\hat{x} - \sum_{l=1}^{R} \hat{\omega}_l\, (\hat{v}\hat{q})(\hat{b}_l) \right| \le C\, \|\hat{v}\hat{q}\|_{\infty,\hat{K}} \le C\, \|\hat{v}\|_{\infty,\hat{K}}\, \|\hat{q}\|_{\infty,\hat{K}}\,,$$

where the continuity of the embedding of $W_\infty^k(\hat{K})$ in $C(\hat{K})$ is used (see [8, p. 181]). Therefore, by the equivalence of $\|\cdot\|_{\infty,\hat{K}}$ and $\|\cdot\|_{0,\hat{K}}$ on $\mathcal{P}_{k-1}(\hat{K})$, it follows that

$$\left|\hat{E}(\hat{v}\hat{q})\right| \le C\, \|\hat{v}\|_{k,\infty,\hat{K}}\, \|\hat{q}\|_{0,\hat{K}}\,.$$

If a fixed $\hat{q} \in \mathcal{P}_{k-1}(\hat{K})$ is chosen, then a linear continuous functional G is defined on $W_\infty^k(\hat{K})$ by $\hat{v} \mapsto \hat{E}(\hat{v}\hat{q})$ that has the following properties:

$$\|G\| \le C\|\hat{q}\|_{0,\hat{K}} \quad \text{and} \quad G(\hat{v}) = 0 \quad \text{for} \quad \hat{v} \in \mathcal{P}_{k-1}(\hat{K})$$

by virtue of (3.128).

The Bramble–Hilbert lemma (Theorem 3.24) implies

$$\left|\hat{E}(\hat{v}\hat{q})\right| \le C\, |\hat{v}|_{k,\infty,\hat{K}}\, \|\hat{q}\|_{0,\hat{K}}\,.$$

According to the assertion we now choose

$$\hat{v} = \hat{a}\hat{p} \quad \text{for} \quad \hat{a} \in W^{k,\infty}(\hat{K})\,,\ \hat{p} \in \mathcal{P}_{k-1}(\hat{K})\,,$$

and we have to estimate $|\hat{a}\hat{p}|_{k,\infty\hat{K}}$ (thanks to the Bramble–Hilbert lemma not $\|\hat{a}\hat{p}\|_{k,\infty,\hat{K}}$). The Leibniz rule for the differentiation of products implies the estimate

$$|\hat{a}\hat{p}|_{k,\infty,\hat{K}} \le C \sum_{j=0}^{k} |\hat{a}|_{k-j,\infty,\hat{K}}\, |\hat{p}|_{j,\infty,\hat{K}}\,. \tag{3.129}$$

Here the constant C depends only on k, but not on the domain $\hat{K}$.

Since $\hat{p} \in \mathcal{P}_{k-1}(\hat{K})$, the last term of the sum in (3.129) can be omitted. Therefore, we have obtained the following estimate holding for $\hat{a} \in W_\infty^k(\hat{K})$, $\hat{p}, \hat{q} \in \mathcal{P}_{k-1}(\hat{K})$:

$$\begin{aligned} \left|\hat{E}(\hat{a}\hat{p}\hat{q})\right| &\le C \left\{ \sum_{j=0}^{k-1} |\hat{a}|_{k-j,\infty,\hat{K}}\, |\hat{p}|_{j,\infty,\hat{K}} \right\} \|\hat{q}\|_{0,\hat{K}} \\ &\le C \left\{ \sum_{j=0}^{k-1} |\hat{a}|_{k-j,\infty,\hat{K}}\, |\hat{p}|_{j,\hat{K}} \right\} \|\hat{q}\|_{0,\hat{K}}\,. \end{aligned} \tag{3.130}$$

The last estimate uses the equivalence of $\|\cdot\|_\infty$ and $\|\cdot\|_0$ on $\mathcal{P}_{k-1}(\hat{K})$.

We suppose that the transformation F of $\hat{K}$ to the general element K has, as usual, the linear part B. The first transformation step yields the

factor $|\det(B)|$ according to (3.115), and for the backtransformation it follows from Theorem 3.26 and Theorem 3.27 that

$$\begin{aligned} |\hat{a}|_{k-j,\infty,\hat{K}} &\leq C\, h_K^{k-j}\, |a|_{k-j,\infty,K}\,, \\ |\hat{p}|_{j,\hat{K}} &\leq C\, h_K^{j}\, |\det(B)|^{-1/2}\, |p|_{j,K}\,, \\ \|\hat{q}\|_{0,\hat{K}} &\leq C\, |\det(B)|^{-1/2}\, \|q\|_{0,K} \end{aligned} \tag{3.131}$$

for $0 \leq j \leq k-1$. Here a, p, q are the mappings $\hat{a}, \hat{p}, \hat{q}$ (back)transformed according to (3.75). Substituting these estimates into (3.130) therefore yields

$$\left|E_K(apq)\right| \leq C\, h_K^k \left\{ \sum_{j=0}^{k-1} |a|_{k-j,\infty,K}\, |p|_{j,K} \right\} \|q\|_{0,K}$$

and from this, assertion (1) follows for $l = k$.

If $l = 1$, we modify the proof as follows. Again, in (3.130) we estimate by using the equivalence of norms:

$$\begin{aligned} \left|\hat{E}(\hat{a}\hat{p}\hat{q})\right| &\leq C \left\{ \sum_{j=0}^{k-1} |\hat{a}|_{k-j,\infty,\hat{K}}\, \|\hat{p}\|_{j,\infty,\hat{K}} \right\} \|\hat{q}\|_{0,\hat{K}} \\ &\leq C \left\{ \sum_{j=0}^{k-1} |\hat{a}|_{k-j,\infty,\hat{K}} \right\} \|\hat{p}\|_{0,\hat{K}}\, \|\hat{q}\|_{0,\hat{K}}\,. \end{aligned}$$

The first and the third estimates of (3.131) remain applicable; the second estimate is replaced with the third such that we have

$$\left|E_K(apq)\right| \leq C\, h_K \left\{ \sum_{j=0}^{k-1} |a|_{k-j,\infty,K} \right\}, \|p\|_{0,K}\, \|q\|_{0,K}$$

since the lowest h_K-power arises for $j = k-1$. This estimate yields the assertion (1) for $l = 1$. □

Finally, we can now verify the assumptions of Theorem 3.38 with the following result:

Theorem 3.42 *Consider a family of affine equivalent Lagrange finite element discretizations in $\mathbb{R}^d$, $d \leq 3$, with $P = \mathcal{P}_k$ for some $k \in \mathbb{N}$ as local ansatz space. Suppose that the family of triangulations is regular or satisfies the maximum angle condition in the case of triangles with $k = 1$. Suppose that the applied quadrature formulas are exact for $\mathcal{P}_{2k-2}$. Let the function w satisfying the Dirichlet boundary condition and let the solution u of the boundary value problem* (3.12), (3.18)–(3.20) *(with $g_3 = 0$) belong to $H^{k+1}(\Omega)$.*

Then there exist some constants $C > 0$, $\bar{h} > 0$ independent of u and w such that for the finite element approximation $\bar{u}_h + w_h$ according to (3.105),

(3.121), *it follows for* $h \le \bar{h}$ *that*

$$\left\|u + w - (\bar{u}_h + w_h)\right\|_1 \le C\, h^k \left\{ |u|_{k+1} + |w|_{k+1} + \left(\sum_{i,j=1}^{d} \|k_{ij}\|_{k,\infty} + \sum_{i=1}^{d} \|c_i\|_{k,\infty} + \|r\|_{k,\infty} \right) \left(\|u\|_{k+1} + \|w\|_{k+1} \right) + \|f\|_{k,\infty} \right\}.$$

Proof: According to (3.108), we aim at estimating $\left\|u+w-(\bar{\bar{u}}_h+\bar{I}_h(w))\right\|_1$, where $\bar{\bar{u}}_h$ satisfies (3.122).

By virtue of Theorem 3.29 or Theorem 3.35 (set formally $\Gamma_3 = \emptyset$) we have

$$\|w - \bar{I}_h(w)\|_1 \le Ch^k |w|_{k+1}\,. \tag{3.132}$$

For the bilinear form a_h defined in (3.120), it follows from Theorem 3.41 for $v, w \in V_h$ and $l \in \{0, k\}$ that

$$\begin{aligned} \left|a(v,w) - a_h(v,w)\right| &\le \sum_{K\in\mathcal{T}_h} \left\{ \sum_{i,j=1}^{d} \left|E_K(k_{ij}\partial_j(v|_K)\partial_i(w|_K))\right| \right. \\ &\qquad \left. + \sum_{i=1}^{d} \left|E_K(c_i\partial_i(v|_K)w)\right| + \left|E_K(rvw)\right| \right\} \\ &\le C \sum_{K\in\mathcal{T}_h} h_K^l \left\{ \sum_{i,j=1}^{d} \|k_{ij}\|_{k,\infty,K} + \sum_{i=1}^{d} \|c_i\|_{k,\infty,K} + \|r\|_{k,\infty,K} \right\} \\ &\qquad \times \|v\|_{l,K} \|w\|_{1,K} \\ &\le C\, h^l \left\{ \sum_{i,j=1}^{d} \|k_{ij}\|_{k,\infty} + \sum_{i=1}^{d} \|c_i\|_{k,\infty} + \|r\|_{k,\infty} \right\} \\ &\qquad \times \left(\sum_{K\in\mathcal{T}_h} \|v\|_{l,K}^2 \right)^{1/2} \|w\|_1, \end{aligned} \tag{3.133}$$

by estimating the $\|\cdot\|_{k,\infty,K}$-norms in terms of norms on the domain Ω and then applying the Cauchy–Schwarz inequality with "index" $K \in \mathcal{T}_h$.

From this we obtain for $l = 1$ an estimate of the form

$$|a(v,w) - a_h(v,w)| \le Ch\|v\|_1\|w\|_1$$

such that the estimate required in Lemma 3.39 holds (with $C(h) = C \cdot h$). Therefore, there exists some $\bar{h} > 0$ such that a_h is uniformly V_h-elliptic for $h \le \bar{h}$. Hence, the estimate (3.126) is applicable, and the first addend, the approximation error, behaves as asserted according to Theorem 3.29 or Theorem 3.35 (again, choose $v = I_h(u)$ for the comparison element).

In order to estimate the consistency error of a_h, a comparison element $v \in V_h$ has to be found for which the corresponding part of the norm in

(3.133) is uniformly bounded. This is satisfied for $v = I_h(u)$, since

$$\begin{aligned}\left(\sum_{K\in\mathcal{T}_h} \|I_h(u)\|_{k,K}^2 \right)^{1/2} &\leq \|u\|_k + \left\{ \sum_{K\in\mathcal{T}_h} \|u - I_h(u)\|_{k,K}^2 \right\}^{1/2} \\ &\leq \|u\|_k + Ch|u|_{k+1} \leq \|u\|_{k+1}\end{aligned}$$

due to Theorem 3.29 or Theorem 3.35.

Hence, the consistency error in a behaves as asserted according to (3.133), so that only the consistency error of l has to be investigated: We have

$$l - l_h = b - b_h - a(w,\cdot) + a_h(\bar{I}_h(w),\cdot)\,,$$

where b_h is defined in (3.120).

If $v \in V_h$, then

$$\big|a(w,v) - a_h(\bar{I}_h(w),v)\big| \leq \big|a(w,v) - a(\bar{I}_h(w),v)\big| + \big|a(\bar{I}_h(w),v) - a_h(\bar{I}_h(w),v)\big|.$$

For the first addend the continuity of a implies

$$\big|a(w,v) - a(\bar{I}_h(w),v)\big| \leq C \left\|w - \bar{I}_h(w)\right\|_1 \|v\|_1\,,$$

so that the corresponding consistency error part behaves like $\left\|w - \bar{I}_h(w)\right\|_1$, which has already been estimated in (3.132). The second addend just corresponds to the estimate used for the consistency error in a (here, the difference between I_h and $\bar{I}_h$ is irrelevant), so that the same contribution to the convergence rate, now with $\|u\|_{k+1}$ replaced by $\|w\|_{k+1}$, arises. Finally, Theorem 3.41, (4) yields for $v \in V_h$,

$$\begin{aligned}|b(v) - b(v_h)| &\leq \sum_{K\in\mathcal{T}_h} |E_K(fv)| \leq C \sum_{K\in\mathcal{T}_h} h_K^k \,\mathrm{vol}\,(K)^{1/2} \|f\|_{k,\infty,K} \|v\|_{1,K} \\ &\leq C\, h^k\, |\Omega|^{1/2} \|f\|_{k,\infty} \|v\|_1\end{aligned}$$

by proceeding as in (3.133). This implies the last part of the asserted estimate. □

If the uniform V_h-ellipticity of a_h is ensured in a different way (perhaps by Lemma 3.40), one can dispense with the smallness assumption about h. If estimates as given in Theorem 3.41 are also available for other types of elements, then triangulations consisting of combinations of various elements can also be considered.

3.7 The Condition Number of Finite Element Matrices

The stability of solution algorithms for linear systems of equations as described in Section 2.5 depends on the condition number of the system matrix (see [28, Chapter 1]). The condition number also plays an important role for the convergence behavior of iterative methods, which will be discussed in Chapter 5. Therefore, in this section we shall estimate the spectral condition number (see Appendix A.3) of the stiffness matrix

$$A = \big(a(\varphi_j, \varphi_i)\big)_{i,j=1,\dots,M} \tag{3.134}$$

and also of the mass matrix (see (7.45))

$$B = \big(\langle \varphi_j, \varphi_i \rangle_0\big)_{i,j=1,\dots,M}\,, \tag{3.135}$$

which is of importance for time-dependent problems. Again, we consider a finite element discretization in the general form of Section 3.4 restricted to Lagrange elements. In order to simplify the notation, we assume the affine equivalence of all elements. Further we suppose that

- the family $(\mathcal{T}_h)_h$ of triangulations is regular.

We assume that the variational formulation of the boundary value problem leads to a bilinear form a that is V-elliptic and continuous on $V \subset H^1(\Omega)$.

As a modification of definition (1.18), let the following norm (which is also induced by a scalar product) be defined in the ansatz space $V_h = \operatorname{span}\{\varphi_1, \dots, \varphi_M\}$:

$$\|v\|_{0,h} := \left(\sum_{K \in \mathcal{T}_h} h_K^d \sum_{a_i \in K} |v(a_i)|^2 \right)^{1/2}. \tag{3.136}$$

Here, $a_1, \dots, a_M$ denote the nodes of the degrees of freedom, where in order to simplify the notation, M instead of M_1 is used for the number of degrees of freedom. The norm properties follow directly from the corresponding properties of $|\cdot|_2$ except for the definiteness. But the definiteness follows from the uniqueness of the interpolation problem in V_h with respect to degrees of freedom a_i.

Theorem 3.43 *(1) There exist constants $C_1, C_2 > 0$ independent of h such that for $v \in V_h$:*

$$C_1 \|v\|_0 \le \|v\|_{0,h} \le C_2 \|v\|_0\,.$$

(2) There exists a constant $C > 0$ independent of h such that for $v \in V_h$,

$$\|v\|_1 \le C \left(\min_{K \in \mathcal{T}_h} h_K \right)^{-1} \|v\|_0\,.$$

Proof: As already known from Sections 3.4 and 3.6, the proof is done locally in $K \in \mathcal{T}_h$ and there transformed to the reference element $\hat{K}$ by means of $F(\hat{x}) = B\hat{x} + d$.

Ad (1): All norms are equivalent on the local ansatz space $\hat{P}$, thus also $\|\cdot\|_{0,\hat{K}}$ and the Euclidean norm in the degrees of freedom. Hence, there exist some $\hat{C}_1, \hat{C}_2 > 0$ such that for $\hat{v} \in \hat{P}$,

$$\hat{C}_1 \|\hat{v}\|_{0,\hat{K}} \leq \left(\sum_{i=1}^{L} |\hat{v}(\hat{a}_i)|^2 \right)^{1/2} \leq \hat{C}_2 \|\hat{v}\|_{0,\hat{K}} .$$

Here, $\hat{a}_1, \ldots, \hat{a}_L$ are the degrees of freedom in $\hat{K}$. Due to (3.50) we have

$$\mathrm{vol}\,(K) = \mathrm{vol}\,(\hat{K})\,|\det(B)|\,,$$

and according to the definition of h_K and the regularity of the family $(\mathcal{T}_h)_h$, there exist constants $\tilde{C}_i > 0$ independent of h such that

$$\tilde{C}_1\, h_K^d \leq \tilde{C}_3\, \varrho_K^d \leq |\det(B)| \leq \tilde{C}_2\, h_K^d\,.$$

By the transformation rule we thus obtain for $v \in P_K$, the ansatz space on K, that

$$\begin{aligned}
\hat{C}_1 \|v\|_{0,K} &= \hat{C}_1\,|\det(B)|^{1/2}\,\|\hat{v}\|_{0,\hat{K}} \leq \left(\tilde{C}_2\, h_K^d\right)^{1/2} \left(\sum_{i=1}^{L} |\hat{v}(\hat{a}_i)|^2 \right)^{1/2} \\
&= \tilde{C}_2^{1/2} \left(\sum_{a_i \in K} h_K^d |v(a_i)|^2 \right)^{1/2} = \left(\tilde{C}_2\, h_K^d\right)^{1/2} \left(\sum_{i=1}^{L} |\hat{v}(\hat{a}_i)|^2 \right)^{1/2} \\
&\leq \left(\tilde{C}_2\, h_K^d\right)^{1/2} \hat{C}_2\, \|\hat{v}\|_{0,\hat{K}} = \left(\tilde{C}_2\, h_K^d\right)^{1/2} \hat{C}_2\, |\det(B)|^{-1/2}\, \|v\|_{0,K} \\
&\leq \tilde{C}_2^{1/2}\, \hat{C}_2\, \tilde{C}_1^{-1/2}\, \|v\|_{0,K}.
\end{aligned}$$

This implies assertion (1).

Ad (2): Arguing as before, now using the equivalence of $\|\cdot\|_{1,\hat{K}}$ and $\|\cdot\|_{0,\hat{K}}$ in $\hat{P}$, it follows by virtue of (3.86) for $v \in P_K$ (with the generic constant C) that

$$\|v\|_{1,K} \leq C\,|\det(B)|^{1/2}\,\|B^{-1}\|_2\,\|\hat{v}\|_{0,\hat{K}} \leq C\,\|B^{-1}\|_2\,\|v\|_{0,K} \leq C\,h_K^{-1}\,\|v\|_{0,K}$$

by Theorem 3.27 and the regularity of $(\mathcal{T}_h)_h$, and from this, the assertion (2). □

In order to make the norm $\|\cdot\|_{0,h}$ comparable with the (weighted) Euclidean norm we assume in the following:

- There exists a constant $C_A > 0$ independent of h such that for every node of $\mathcal{T}_h$, the number of elements to which this node belongs is bounded by C_A. (3.137)

This condition is (partly) redundant: For $d = 2$ and triangular elements, the condition follows from the uniform lower bound (3.93) for the smallest angle as an implication of the regularity. Note that the condition need not be satisfied if only the maximum angle condition is required.

In general, if $C \in \mathbb{R}^{M,M}$ is a matrix with real eigenvalues $\lambda_1 \leq \cdots \leq \lambda_M$ and an orthonormal basis of eigenvectors $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_M$, for instance a symmetric matrix, then it follows for $\boldsymbol{\xi} \in \mathbb{R}^M \setminus \{\mathbf{0}\}$ that

$$\lambda_1 \leq \frac{\boldsymbol{\xi}^T C \boldsymbol{\xi}}{\boldsymbol{\xi}^T \boldsymbol{\xi}} \leq \lambda_M \,, \tag{3.138}$$

and the bounds are assumed for $\boldsymbol{\xi} = \boldsymbol{\xi}_1$ and $\boldsymbol{\xi} = \boldsymbol{\xi}_M$.

Theorem 3.44 *There exists a constant $C > 0$ independent of h such that we have*

$$\kappa(B) \leq C \left(\frac{h}{\min\limits_{K \in \mathcal{T}_h} h_K} \right)^d$$

for the spectral condition number of the mass matrix B according to (3.135).

Proof: $\kappa(B) = \lambda_M / \lambda_1$ must be determined. For arbitrary $\boldsymbol{\xi} \in \mathbb{R}^M \setminus \{\mathbf{0}\}$ we have

$$\frac{\boldsymbol{\xi}^T B \boldsymbol{\xi}}{\boldsymbol{\xi}^T \boldsymbol{\xi}} = \frac{\boldsymbol{\xi}^T B \boldsymbol{\xi}}{\|v\|_{0,h}^2} \, \frac{\|v\|_{0,h}^2}{\boldsymbol{\xi}^T \boldsymbol{\xi}} \,,$$

where $v := \sum_{i=1}^M \xi_i \varphi_i \in V_h$. By virtue of $\boldsymbol{\xi}^T B \boldsymbol{\xi} = \langle v, v \rangle_0$, the first factor on the right-hand side is uniformly bounded from above and below according to Theorem 3.43. Further, by (3.137) and $\boldsymbol{\xi} = (v(a_1), \ldots, v(a_M))^T$ it follows that

$$\min_{K \in \mathcal{T}_h} h_K^d \, |\boldsymbol{\xi}|^2 \leq \|v\|_{0,h}^2 \leq C_A \, h^d \, |\boldsymbol{\xi}|^2 \,,$$

and, thus the second factor is estimated from above and below. This leads to estimates of the type

$$\lambda_1 \geq C_1 \min_{K \in \mathcal{T}_h} h_K^d \,, \quad \lambda_M \leq C_2 \, h^d \,,$$

and from this, the assertion follows. □

Therefore, if the family of triangulations $(\mathcal{T}_h)_h$ is *quasi-uniform* in the sense that there exists a constant $C > 0$ independent of h such that

$$h \leq C \, h_K \quad \text{for all } K \in \mathcal{T}_h \,, \tag{3.139}$$

then $\kappa(B)$ is uniformly bounded.

In order to be able to argue analogously for the stiffness matrix, we assume that we stay close to the symmetric case:

Theorem 3.45 *Suppose the stiffness matrix A (3.134) admits real eigenvalues and a basis of eigenvectors. Then there exists a constant $C > 0$ independent of h such that the following estimates for the spectral condition number κ hold:*

$$\begin{aligned} \kappa(B^{-1}A) &\leq C \left(\min_{K \in \mathcal{T}_h} h_K \right)^{-2}, \\ \kappa(A) &\leq C \left(\min_{K \in \mathcal{T}_h} h_K \right)^{-2} \kappa(B). \end{aligned}$$

Proof: With the notation of (3.138), we proceed analogously to the proof of Theorem 3.44. Since

$$\frac{\boldsymbol{\xi}^T A \boldsymbol{\xi}}{\boldsymbol{\xi}^T \boldsymbol{\xi}} = \frac{\boldsymbol{\xi}^T A \boldsymbol{\xi}}{\boldsymbol{\xi}^T B \boldsymbol{\xi}} \frac{\boldsymbol{\xi}^T B \boldsymbol{\xi}}{\boldsymbol{\xi}^T \boldsymbol{\xi}},$$

it suffices to bound the first factor on the right-hand side from above and below. This also yields a result for the eigenvalues of $B^{-1}A$, since we have for the variable $\boldsymbol{\eta} := B^{1/2}\boldsymbol{\xi}$,

$$\frac{\boldsymbol{\xi}^T A \boldsymbol{\xi}}{\boldsymbol{\xi}^T B \boldsymbol{\xi}} = \frac{\boldsymbol{\eta}^T B^{-1/2} A B^{-1/2} \boldsymbol{\eta}}{\boldsymbol{\eta}^T \boldsymbol{\eta}},$$

and the matrix $B^{-1/2}AB^{-1/2}$ possesses the same eigenvalues as $B^{-1}A$ by virtue of $B^{-1/2}(B^{-1/2}AB^{-1/2})B^{1/2} = B^{-1}A$. Here, $B^{1/2}$ is the symmetric positive definite matrix that satisfies $B^{1/2}B^{1/2} = B$, and $B^{-1/2}$ is its inverse.

Since $\boldsymbol{\xi}^T A \boldsymbol{\xi} / \boldsymbol{\xi}^T B \boldsymbol{\xi} = a(v,v)/\langle v, v \rangle_0$ and

$$\begin{aligned} a(v,v) &\geq \alpha \|v\|_1^2 \geq \alpha \|v\|_0^2, \\ a(v,v) &\leq C\|v\|_1^2 \leq C \left(\min_{K \in \mathcal{T}_h} h_K \right)^{-2} \|v\|_0^2, \end{aligned} \tag{3.140}$$

with a generic constant $C > 0$ (the last estimate is due to Theorem 3.43, 2), it follows that

$$\alpha \leq \frac{a(v,v)}{\langle v, v \rangle_0} = \frac{\boldsymbol{\xi}^T A \boldsymbol{\xi}}{\boldsymbol{\xi}^T B \boldsymbol{\xi}} = \frac{a(v,v)}{\langle v, v \rangle_0} \leq C \left(\min_{K \in \mathcal{T}_h} h_K \right)^{-2}, \tag{3.141}$$

and from this the assertion. □

The analysis of the eigenvalues of the model problem in Example 2.12 shows that the above-given estimates are not too pessimistic.

3.8 General Domains and Isoparametric Elements

All elements considered so far are bounded by straight lines or plane surfaces. Therefore, only polyhedral domains can be decomposed exactly by means of a triangulation. Depending on the application, domains with a curved boundary may appear. With the available elements the obvious way of dealing with such domains is the following (in the two-dimensional case): for elements K that are close to the boundary put only the nodes of one edge on the boundary $\partial\Omega$. This implies an approximation error for the domain, for $\Omega_h := \bigcup_{K\in\mathcal{T}_h} K$, there holds in general neither $\Omega \subset \Omega_h$ nor $\Omega_h \subset \Omega$ (see Figure 3.14).

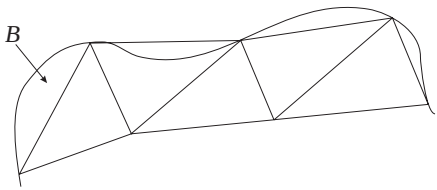


Figure 3.14. Ω and Ω_h.

As the simplest example, we consider homogeneous Dirichlet boundary conditions, thus $V = H_0^1(\Omega)$, on a convex domain for which therefore $\Omega_h \subset \Omega$ is satisfied. If an ansatz space V_h is introduced as in Section 3.3, then functions defined on Ω_h are generated. Therefore, these functions must be extended to Ω in such a way that they vanish on $\partial\Omega$, and consequently, for the generated function space $\tilde{V}_h$, $\tilde{V}_h \subset V$. This is supposed to be done by adding the domains B whose boundary consists of a boundary part of some element $K \in \mathcal{T}_h$ close to the boundary and a subset of $\partial\Omega$ to the set of elements with the ansatz space $P(B) = \{0\}$. Céa's lemma (Theorem 2.17) can still be applied, so that for an error estimate in $\|\cdot\|_1$ the question of how to choose a comparison element $v \in \tilde{V}_h$ arises. The ansatz $v = \tilde{I}_h(u)$, where $\tilde{I}_h(u)$ denotes the interpolation on Ω_h extended by 0 on the domains B, is admissible only for the (multi-)linear ansatz: Only in this case are all nodes of an edge "close to the boundary" located on $\partial\Omega$ and therefore have homogeneous degrees of freedom, so that the continuity on these edges is ensured. For the present, let us restrict our attention to this case, so that $\|u - \tilde{I}_h(u)\|_1$ has to be estimated where u is the solution of the boundary value problem.

The techniques of Section 3.4 can be applied to all $K \in \mathcal{T}_h$, and by the conditions assumed there about the triangulation, this yields

$$\begin{aligned} \|u - u_h\|_1 &\leq C\big(\|u - I_h(u)\|_{1,\Omega_h} + \|u\|_{1,\Omega\setminus\Omega_h}\big) \\ &\leq C\big(h|u|_{2,\Omega_h} + \|u\|_{1,\Omega\setminus\Omega_h}\big)\,. \end{aligned}$$

If $\partial\Omega \in C^2$, then we have the estimate

$$\|u\|_{1,\Omega\setminus\Omega_h} \le Ch\|u\|_{2,\Omega}$$

for the new error part due to the approximation of the domain, and thus the convergence rate is preserved. Already for a quadratic ansatz this is no longer satisfied, where only

$$\|u - u_h\|_1 \le Ch^{3/2}\|u\|_3$$

holds instead of the order $O(h^2)$ of Theorem 3.29 (see [31, pp. 194 ff]). One may expect that this decrease of the approximation quality arises only locally close to the boundary, however, one may also try to obtain a better approximation of the domain by using curved elements. Such elements can be defined on the basis of the reference elements $(\hat{K}, \hat{P}, \hat{\Sigma})$ of Lagrange type introduced in Section 3.3 if a general element is obtained from this one by an *isoparametric transformation*; that is, choose an

$$F \in (\hat{P})^d \tag{3.142}$$

that is injective and then

$$K := F(\hat{K})\,, \quad P := \left\{\hat{p} \circ F^{-1} \;\middle|\; \hat{p} \in \hat{P}\right\}, \quad \Sigma := \left\{F(\hat{a}) \;\middle|\; \hat{a} \in \hat{\Sigma}\right\}.$$

Since the bijectivity of $F : \hat{K} \to K$ is ensured by requirement, a finite element is thus defined in terms of (3.58). By virtue of the unique solvability of the interpolation problem, F can be defined by prescribing $a_1, \ldots, a_L$, $L = |\hat{\Sigma}|$, and requiring

$$F(\hat{a}_i) = a_i\,, \quad i = 1, \ldots, L\,.$$

However, this does not in general ensure the injectivity. Since, on the other hand, in the grid generation process elements are created by defining the nodes (see Section 4.1), geometric conditions about their positions that characterize the injectivity of F are desirable. A typical curved element that can be used for the approximation of the boundary can be generated on the basis of the unit simplex with $\hat{P} = \mathcal{P}_2(\hat{K})$ (see Figure 3.15).

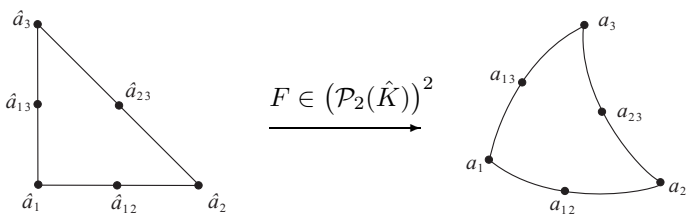


Figure 3.15. Isoparametric element: quadratic ansatz on triangle.

Elements with, in general, one curved edge and otherwise straight edges thus are suggested for the problem of boundary approximation. They are

combined with affine "quadratic triangles" in the interior of the domain. *Subparametric elements* can be generated analogously to the *isoparametric elements* if (the components of) the transformations in (3.142) are restricted to some subspace $\hat{P}_T \subset \hat{P}$. If $\hat{P}_T = \mathcal{P}_1(\hat{K})$, we again obtain the affine equivalent elements.

However, isoparametric elements are also important if, for instance, the unit square or cube is supposed to be the reference element. Only the isoparametric transformation allows for "general" quadrilaterals and hexahedra, respectively, which are preferable in anisotropic cases (for instance in generalization of Figure 3.11) to simplices due to their adaptability to local coordinates. In what follows, let $\hat{K} = [0,1]^d$, $\hat{P} = Q_1(\hat{K})$.

In general, since also a finite element (in $\mathbb{R}^{d-1}$) is defined for every face $\hat{S}$ of $\hat{K}$ with $\hat{P}|_{\hat{S}}$ and $\hat{\Sigma}|_{\hat{S}}$, the "faces" of K, that is, $F[\hat{S}]$, are already uniquely defined by the related nodes.

Consequently, if $d = 2$, the edges of the general quadrilateral are straight lines (see Figure 3.16), but if $d = 3$, we have to expect curved surfaces (hyperbolic paraboloids) for a general hexahedron.

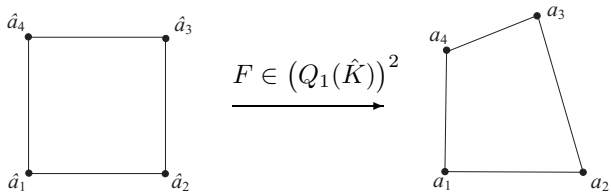


Figure 3.16. Isoparametric element: bilinear ansatz on rectangle.

A geometric characterization of the injectivity of F is still unknown (to our knowledge) for $d = 3$, but it can be easily derived for $d = 2$: Let the nodes a_1, a_2, a_3, a_4 be numbered counterclockwise and suppose that they are not on a straight line, and thus (by rearranging) $T = \operatorname{conv}(a_1, a_2, a_4)$ forms a triangle such that

$$2 \operatorname{vol}(T) = \det(B) > 0 .$$

Here $F_T(\hat{x}) = B\hat{x} + d$ is the affine-linear mapping that maps the reference triangle $\operatorname{conv}(\hat{a}_1, \hat{a}_2, \hat{a}_4)$ bijectively to T. If $\tilde{a}_3 := F_T^{-1}(a_3)$, then the quadrilateral $\tilde{K}$ with the vertices $\hat{a}_1, \hat{a}_2, \tilde{a}_3, \hat{a}_4$ is mapped bijectively to K by F_T.

The transformation F can be decomposed into

$$F = F_T \circ F_Q ,$$

where $F_Q \in \left(Q_1(\hat{K})\right)^2$ denotes the mapping defined by

$$F_Q(\hat{a}_i) = \hat{a}_i , \quad i = 1, 2, 4 , \quad F_Q(\hat{a}_3) = \tilde{a}_3$$

(see Figure 3.17).

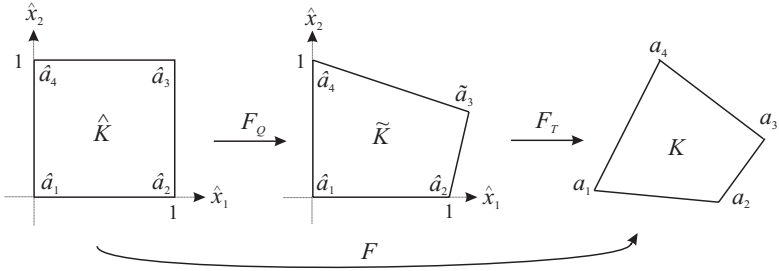


Figure 3.17. Decomposition of the bilinear isoparametric mapping.

Therefore, the bijectivity of F is equivalent to the bijectivity of F_Q.

We characterize a "uniform" bijectivity which is defined by $\det(DF(\hat{x}_1, \hat{x}_2)) \neq 0$ for the functional matrix $DF(\hat{x}_1, \hat{x}_2)$:

Theorem 3.46 *Suppose Q is a quadrilateral with the vertices $a_1,\ldots,a_4$ (numbered counterclockwise). Then,*

$$\det(DF(\hat{x}_1, \hat{x}_2)) \neq 0 \quad \textit{for all } (\hat{x}_1, \hat{x}_2) \in [0,1]^2 \quad \Longleftrightarrow$$
$$\det(DF(\hat{x}_1, \hat{x}_2)) > 0 \quad \textit{for all } (\hat{x}_1, \hat{x}_2) \in [0,1]^2 \quad \Longleftrightarrow$$

Q is convex and does not degenerate into a triangle or straight line.

Proof: By virtue of

$$\det(DF(\hat{x}_1, \hat{x}_2)) = \det(B) \det(DF_Q(\hat{x}_1, \hat{x}_2))$$

and $\det(B) > 0$, F can be replaced with F_Q in the assertion. Since

$$F_Q(\hat{x}_1, \hat{x}_2) = \begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \end{pmatrix} + \begin{pmatrix} \tilde{a}_{3,1} - 1 \\ \tilde{a}_{3,2} - 1 \end{pmatrix} \hat{x}_1 \hat{x}_2 \,,$$

it follows by some simple calculations that

$$\det(DF_Q(\hat{x}_1, \hat{x}_2)) = 1 + (\tilde{a}_{3,2} - 1)\hat{x}_1 + (\tilde{a}_{3,1} - 1)\hat{x}_2$$

is an affine-linear mapping because the quadratic parts just cancel each other. This mapping assumes its extrema on $[0,1]^2$ at the 4 vertices, where we have the following values:

$$(0,0) : 1\,, \quad (1,0) : \tilde{a}_{3,2}\,, \quad (0,1) : \tilde{a}_{3,1}\,, \quad (1,1) : \tilde{a}_{3,1} + \tilde{a}_{3,2} - 1\,.$$

A uniform sign is thus obtained if and only if the function is everywhere positive. This is the case if and only if

$$\tilde{a}_{3,1}\,,\ \tilde{a}_{3,2}\,,\ \tilde{a}_{3,1} + \tilde{a}_{3,2} - 1 \ > \ 0\,,$$

which just characterizes the convexity and the nondegeneration of $\tilde{K}$. By the transformation F_T this also holds for K. □

According to this theorem it is not allowed that a quadrilateral degenerates into a triangle (now with linear ansatz). But a more careful analysis [55] shows that this does not affect negatively the quality of the approximation.

In general, for isoparametric elements we have the following:

From the point of view of implementation, only slight modifications have to be made: In the integrals (3.111), (3.111) transformed to the reference element or their approximation by quadrature (3.120), $|\det B|$ has to be replaced with $|\det(DF(\hat{x}))|$ (in the integrand).

The analysis of the order of convergence can be done along the same lines as in Section 3.4 (and 3.6), however, the transformation rules for the integrals become more complex (see [9, pp. 237 ff.]).

3.9 The Maximum Principle for Finite Element Methods

In this section maximum and comparision principles that have been introduced for the finite difference method are outlined for the finite element method.

In the case of two-dimensional domains Ω the situation has been well investigated for linear elliptic boundary value problems of second order and linear elements. For higher-dimensional problems ($d > 2$) as well as other types of elements, the corresponding assumptions are much more complex, or there does not necessarily exist any maximum principle.

From now on, let $\Omega \subset \mathbb{R}^2$ be a polygonally bounded domain and let X_h denote the finite element space of continuous, piecewise linear functions for a conforming triangulation $\mathcal{T}_h$ of Ω where the function values in the nodes on the Dirichlet boundary Γ_3 are included in the degrees of freedom. First, we consider the discretization developed for the Poisson equation $-\Delta u = f$ with $f \in L^2(\Omega)$. The algebraization of the method is done according to the scheme described in Section 2.4.3. According to this, first all nodes inside Ω and on Γ_1 and Γ_2 are numbered consecutively from 1 to a number M_1. The nodal values $u_h(a_r)$ for $r = 1, \ldots, M_1$ are arranged in the vector $\boldsymbol{u}_h$. Then, the nodes that belong to the Dirichlet boundary are numbered from $M_1 + 1$ to some number $M_1 + M_2$, the corresponding nodal values generate the vector $\hat{\boldsymbol{u}}_h$. The combination of $\boldsymbol{u}_h$ and $\hat{\boldsymbol{u}}_h$ gives the vector of all nodal values $\tilde{\boldsymbol{u}}_h = \binom{\boldsymbol{u}_h}{\hat{\boldsymbol{u}}_h} \in \mathbb{R}^M$, $M = M_1 + M_2$.

This leads to a linear system of equations of the form (1.31) described in Section 1.4:

$$A_h \boldsymbol{u}_h = -\hat{A}_h \hat{\boldsymbol{u}}_h + \boldsymbol{f}$$

with $A_h \in \mathbb{R}^{M_1, M_1}$, $\hat{A}_h \in \mathbb{R}^{M_1, M_2}$, $\boldsymbol{u}_h, \boldsymbol{f} \in \mathbb{R}^{M_1}$ and $\hat{\boldsymbol{u}}_h \in \mathbb{R}^{M_2}$.

Recalling the support properties of the basis functions $\varphi_i, \varphi_j \in X_h$, we obtain for a general element of the (extended) stiffness matrix $\tilde{A}_h :=$

$\left(A_h \mid \hat{A}_h\right) \in \mathbb{R}^{M_1,M}$ following the relation

$$(\tilde{A}_h)_{ij} = \int_\Omega \nabla\varphi_j \cdot \nabla\varphi_i \, dx = \int_{\operatorname{supp}\varphi_i \cap \operatorname{supp}\varphi_j} \nabla\varphi_j \cdot \nabla\varphi_i \, dx \,.$$

Therefore, if $i \neq j$, the actual domain of integration consists of at most two triangles. Hence, for the present it is reasonable to consider only one triangle as the domain of integration .

Lemma 3.47 *Suppose $\mathcal{T}_h$ is a conforming triangulation of Ω. Then for an arbitrary triangle $K \in \mathcal{T}_h$ with the vertices a_i, a_j $(i \neq j)$, the following relation holds:*

$$\int_K \nabla\varphi_j \cdot \nabla\varphi_i \, dx = -\frac{1}{2}\cot\alpha_{ij}^K \,,$$

where α_{ij}^K denotes the interior angle of K that is opposite to the edge with the boundary points a_i, a_j.

Proof: Suppose the triangle K has the vertices a_i, a_j, a_k (see Figure 3.18). On the edge opposite to the point a_j, we have

$$\varphi_j \equiv 0 \,.$$

Therefore, $\nabla\varphi_j$ has the direction of a normal vector to this edge and — by considering in which direction φ_j increases — the orientation opposite to the outward normal vector ν_{ki}, that is,

$$\nabla\varphi_j = -\,|\nabla\varphi_j|\,\nu_{ki} \quad \text{with} \quad |\nu_{ki}| = 1 \,. \tag{3.143}$$

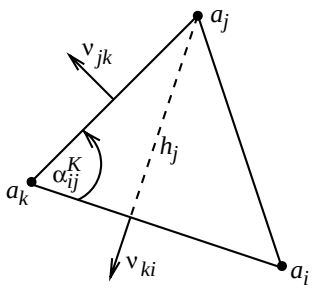


Figure 3.18. Notation for the proof of Lemma 3.47.

In order to calculate $|\nabla\varphi_j|$ we use the following: From (3.143) we obtain

$$|\nabla\varphi_j| = -\nabla\varphi_j \cdot \nu_{ki} \,;$$

that is, we have to compute a directional derivative. By virtue of $\varphi_j(a_j) = 1$, we have

$$\nabla\varphi_j \cdot \nu_{ki} = \frac{0-1}{h_j} = -\frac{1}{h_j} \,,$$

where h_j denotes the height of K with respect to the edge opposite a_j. Thus we have obtained the relation

$$\nabla\varphi_j = -\frac{1}{h_j}\nu_{ki}\,.$$

Hence we have

$$\nabla\varphi_j \cdot \nabla\varphi_i = \frac{\nu_{ki}\cdot\nu_{jk}}{h_j\,h_i} = -\frac{\cos\alpha_{ij}^K}{h_j\,h_i}\,.$$

Since

$$2\,|K| = h_j\,|a_k - a_i| = h_i\,|a_j - a_k| = |a_k - a_i|\,|a_j - a_k|\,\sin\alpha_{ij}^K\,,$$

we obtain

$$\nabla\varphi_j \cdot \nabla\varphi_i = -\frac{\cos\alpha_{ij}^K}{4\,|K|^2}\,|a_k - a_i|\,|a_j - a_k| = -\frac{1}{2}\cot\alpha_{ij}^K\,\frac{1}{|K|}\,,$$

so that the assertion follows by integration. □

Corollary 3.48 *If K and K' are two triangles of $\mathcal{T}_h$ which have a common edge spanned by the nodes a_i, a_j, then*

$$(\tilde{A}_h)_{ij} = \int_{K\cup K'} \nabla\varphi_j\cdot\nabla\varphi_i\,dx = -\frac{1}{2}\,\frac{\sin(\alpha_{ij}^K + \alpha_{ij}^{K'})}{(\sin\alpha_{ij}^K)(\sin\alpha_{ij}^{K'})}\,.$$

Proof: The formula follows from the addition theorem for the cotangent function. □

Lemma 3.47 and Corollary 3.48 are the basis for the proof of the assumption (1.32)* in the case of the extended system matrix $\tilde{A}_h$. Indeed, additional assumptions about the triangulation $\mathcal{T}_h$ are necessary:

Angle condition: For any two triangles of $\mathcal{T}_h$ with a common edge, the sum of the interior angles opposite to this edge does not exceed the value π. If a triangle has an edge on the boundary part Γ_1 or Γ_2, then the angle opposite this edge must not be obtuse.

Connectivity condition: For every pair of nodes both belonging to $\Omega\cup\Gamma_1\cup\Gamma_2$ there exists a polygonal line between these two nodes such that the polygonal line consists only of triangle edges whose boundary points also belong to $\Omega\cup\Gamma_1\cup\Gamma_2$ (see Figure 3.19).

Discussion of assumption (1.32)*: The proof of (1), (2), (5), (6)* is rather elementary. For the "diagonal elements,"

$$(A_h)_{rr} = \int_\Omega |\nabla\varphi_r|^2\,dx = \sum_{K\subset\operatorname{supp}\varphi_r}\int_K |\nabla\varphi_r|^2\,dx > 0\,,\quad r = 1,\ldots,M_1\,,$$

which already is (1). Checking the sign conditions (2) and (5) for the "nondiagonal elements" of $\tilde{A}_h$ requires the analysis of two cases:

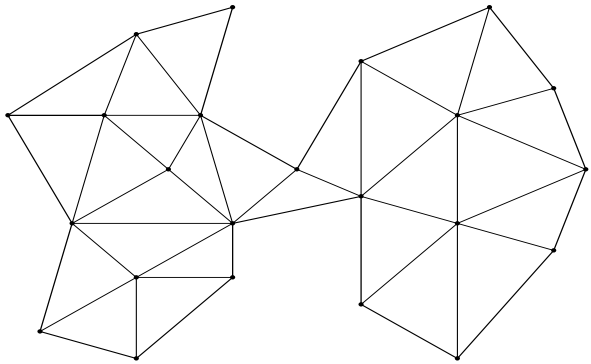


Figure 3.19. Example of a nonconnected triangulation ($\Gamma_3 = \partial\Omega$).

(i) For $r = 1, \ldots, M_1$ and $s = 1, \ldots, M$ with $r \neq s$, there exist two triangles that have the common vertices a_r, a_s.

(ii) There exists only one triangle that has a_r as well as a_s as vertices.

In case (i), Corollary 3.48 can be applied, since if K, K' just denote the two triangles with a common edge spanned by a_r, a_s, then $0 < \alpha_{rs}^K + \alpha_{rs}^{K'} \leq \pi$ and thus $(\tilde{A}_h)_{rs} \leq 0$, $r \neq s$. In case (ii), Lemma 3.47, due to the part of the angle condition that refers to the boundary triangles, can be applied directly yielding the assertion.

Further, since $\sum_{s=1}^M \varphi_s = 1$ in Ω, we obtain

$$\sum_{s=1}^{M} (\tilde{A}_h)_{rs} = \sum_{s=1}^{M} \int_\Omega \nabla\varphi_s \cdot \nabla\varphi_r \, dx = \int_\Omega \nabla \left(\sum_{s=1}^{M} \varphi_s \right) \cdot \nabla\varphi_r \, dx = 0 \,.$$

This is (6)*.

The sign condition in (3) now follows from (6)* and (5), since we have

$$\sum_{s=1}^{M_1} (A_h)_{rs} = \underbrace{\sum_{s=1}^{M} (\tilde{A}_h)_{rs}}_{=0} - \sum_{s=M_1+1}^{M} (\hat{A}_h)_{rs} \geq 0 \,. \tag{3.144}$$

The difficult part of the proof of (3) consists in showing that at least one of these inequalities (3.144) is satisfied strictly. This is equivalent to the fact that at least one element $(\hat{A}_h)_{rs}$, $r = 1, \ldots, M_1$ and $s = M_1 + 1, \ldots, M$, is negative, which can be shown in terms of an indirect proof by using Lemma 3.47 and Corollary 3.48, but is not done here in order to save space. Simultaneously, this also proves the condition (7).

The remaining condition (4)* is proved similarly. First, due to the connectivity condition, the existence of geometric connections between pairs of nodes by polygonal lines consisting of edges is obvious. It is more difficult to prove that under all possible connections there exists one along which

the corresponding matrix elements do not vanish. This can be done by the same technique of proof as used in the second part of (3), which, however, is not presented here.

If the angle condition given above is replaced with a stronger angle condition in which stretched and right angles are excluded, then the proof of (3) and (4)* becomes trivial.

Recalling the relations

$$\max_{x\in\overline{\Omega}} u_h(x) = \max_{r\in\{1,\dots,M\}} (\tilde{\boldsymbol{u}}_h)_r$$

and

$$\max_{x\in\Gamma_3} u_h(x) = \max_{r\in\{M_1+1,\dots,M\}} (\hat{\boldsymbol{u}}_h)_r\,,$$

which hold for linear elements, the following result can be derived from Theorem 1.10.

Theorem 3.49 *If the triangulation $\mathcal{T}_h$ satisfies the angle condition and the connectivity condition, then we have the following estimate for the finite element solution u_h of the Poisson equation in the space of linear elements for a nonpositive right-hand side $f \in L^2(\Omega)$:*

$$\max_{x\in\overline{\Omega}} u_h(x) \le \max_{x\in\Gamma_3} u_h(x)\,.$$

Finally, we make two remarks concerning the case of more general differential equations.

If an equation with a variable scalar diffusion coefficient $k : \Omega \to \mathbb{R}$ is considered instead of the Poisson equation, then the relation in Corollary 3.48 loses its purely geometric character. Even if the diffusion coefficient is supposed to be elementwise constant, the data-dependent relation

$$(\tilde{A}_h)_{ij} = -\frac{1}{2}\left\{k_K \cot\alpha_{ij}^K + k_{K'} \cot\alpha_{ij}^{K'}\right\}$$

would arise, where k_K and $k_{K'}$ denote the constant restriction of k to the triangles K and K', respectively. The case of matrix-valued coefficients $K : \Omega \to \mathbb{R}^{d,d}$ is even more problematic.

The second remark concerns differential expressions that also contain lower-order terms, that is, convective and reactive parts. If the diffusive term $-\nabla\cdot(K\nabla u)$ can be discretized in such a way that a maximum principle holds, then this maximum principle is preserved if the discretization of the other terms leads to matrices whose "diagonal elements" are nonnegative and whose "nondiagonal elements" are nonpositive. These matrix properties are much simpler than the conditions (1.32) and (1.32)*. However, satisfying these properties causes difficulties in special cases, e.g., for convection-dominated equations (see Chapter 9), unless additional restrictive assumptions are made or special discretization schemes are used.

4

Grid Generation and A Posteriori Error Estimation

4.1 Grid Generation

As one of the first steps, the implementation of the finite element method (and also of the finite volume method as described in Chapter 6) requires a "geometric discretization" of the domain Ω.

This part of a finite element program is usually included in the so-called *preprocessor* (see also Section 2.4.1). In general, a finite element program consists further of the intrinsic *kernel* (*assembling* of the finite-dimensional system of algebraic equations, rearrangement of data (if necessary), solution of the algebraic problem) and the *postprocessor* (editing of the results, extraction of intermediate results, preparation for graphic output, a posteriori error estimation).

4.1.1 Classification of Grids

Grids can be grouped according to different criteria: One criterion considers the geometric shape of the elements (triangles, quadrilaterals, tetrahedra, hexahedra, prisms, pyramids; possibly with curved boundaries). A further criterion distinguishes the logical structure of the grid (structured or unstructured grids). Beside these rough classes, in practice one can find a large number of variants combining grids of different classes (combined grids).

A *structured grid in the strict sense* is characterized by a regular arrangement of the grid points (nodes), that is, the connectivity pattern between neighbouring nodes is identical everywhere in the interior of the grid. The

only exceptions of that pattern may occur near the boundary of the domain Ω.

Typical examples of structured grids are rectangular Cartesian two- or three-dimensional grids as they are also used within the framework of the finite difference methods described in Chapter 1 (see, e.g., Figure 1.1).

A *structured grid in the wider sense* is obtained by the application of a piecewise smooth bijective transformation to some "reference grid", which is a structured grid in the strict sense. Grids of this type are also called logically structured, because only the logical structure of the connectivity pattern is fixed in the interior of the grid. However, the edges or faces of the geometric elements of a logically structured grid are not necessarily straight or even.

Logically structured grids have the advantage of simple implementation, because the pattern already defines the neighbours of a given node. Furthermore, there exist efficient methods for the solution of the algebraic system resulting from the discretization, including parallelized resolution algorithms.

In contrast to structured grids, unstructured grids do not have a self-repeating node pattern. Moreover, elements of different geometric type can be combined in unstructured grids.

Unstructured grids are suitable tools for the modelling of complex geometries of Ω and for the adjustment of the grid to the numerical solution (local grid adaptation).

In the subsequent sections, a survey of a few methods for generating unstructured grids will be given. Methods to produce structured grids can be found, for instance, in the books [23] or [33].

4.1.2 Generation of Simplicial Grids

A simplicial grid consists of triangles (in two dimensions) or tetrahedra (in three dimensions). To generate simplicial grids, the following three types of methods are widely used:

- overlay methods,
- Delaunay triangulations,
- advancing front methods.

Overlay Methods

The methods of this type start with a structured grid (the overlay grid) that covers the whole domain. After that, this basic grid is modified near the boundary to fit to the domain geometry. The so-called *quadtree* (in two dimensions) or *octree technique* (in three dimensions) forms a typical example of an overlay method, where the overlay grid is a relatively coarse rectangular Cartesian two- or three-dimensional grid. The substantial part

of the algorithm consists of fitting routines for those parts of the starting grid that are located near the boundary and of simplicial subdivisions of the obtained geometric elements. The fitting procedures perform recursive subdivisions of the boundary rectangles or rectangular parallelepipeds in such a way that at the end every geometric element contains at most one geometry defining point (i.e., a vertex of Ω or a point of $\partial\Omega$, where the type of boundary conditions changes). Finally, the so-called *smoothing step*, which optimizes the grid with respect to a certain regularity criterion, can be supplemented; see Section 4.1.4.

Typically, grids generated by overlay methods are close to structured grids in the interior of the domain. Near the boundary, they lose the structure. Further details can be found in the references [68] and [72].

Delaunay Triangulations

The core algorithm of these methods generates, for a given cloud of isolated points (nodes), a triangulation of their convex hull. Therefore, a grid generator based on this principle has to include a procedure for the generation of this point set (for example, the points resulting from an overlay method) as well as certain fitting procedures (to cover, for example, nonconvex domains, too).

The Delaunay triangulation of the convex hull of a given point set in $\mathbb{R}^d$ is characterized by the following property (*empty sphere criterion*, Figure 4.1): Any open d-ball, the boundary of which contains $d+1$ points from the given set, does not contain any other points from that set. The triangulation can be generated from the so-called *Voronoi tesselation* of $\mathbb{R}^d$ for the given point set. In two dimensions, this procedure is described in Chapter 6, which deals with finite volume methods (Section 6.2.1). How-

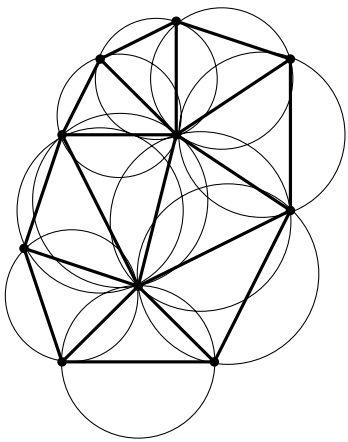


Figure 4.1. Empty sphere criterion in two dimensions ($d = 2$).

ever, practical algorithms ([48] or [71]) apply the empty sphere criterion more directly.

The interesting theoretical properties of Delaunay triangulations are one of the reasons for the "popularity" of this method. In two dimensions, the so-called *max-min-angle property* is valid: Among all triangulations of the convex hull $\overline{G}$ of a given point set, the Delaunay triangulation maximizes the minimal interior angle over all triangles. In the case $d = 3$, this nice property does not remain true. In contrast, even badly shaped elements (the so-called *sliver elements*) may occur. A further important property of a two-dimensional Delaunay triangulation is that the sum of two angles opposite an interior edge is not more than π. For example, such a requirement is a part of the angle condition formulated in Section 3.9.

Advancing Front Methods

The idea of these methods, which are also known in the literature (see, e.g., [50], [56], [60], [62]) as *moving front methods*, is to generate a triangulation recursively from a discretization of the current boundary. The methods start with a partition of the boundary of $G_0 := \Omega$. For $d = 2$, this "initial front" is a polygonal line, whereas in $d = 3$ it is a triangulation of a curved surface (the so-called "2.5-dimensional triangulation"). The method consists of an iteration of the following general step (Figure 4.2): An element of the current front (i.e., a straight-line segment or a triangle) is taken and then, either generating a new inner point or taking an already existing point, a new simplex K_j that belongs to $\overline{G}_{j-1}$ is defined. After the data of the new simplex are saved, the simplex is deleted from $\overline{G}_{j-1}$. In this way, a smaller domain G_j with a new boundary ∂G_j (a new "current front") results. The general step is repeated until the current front is empty. Often, the grid generation process is supplemented by the so-called smoothing step; see Section 4.1.4.

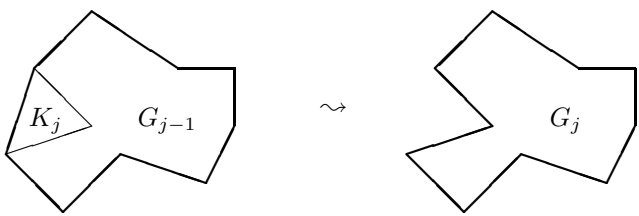


Figure 4.2. Step j of the advancing front method: The new simplex K_j is deleted from the domain G_{j-1}.

4.1.3 Generation of Quadrilateral and Hexahedral Grids

Grids consisting of quadrilaterals or hexahedra can also be generated by means of overlay methods (e.g., [66]) or *advancing front methods* (e.g., [46], [47]). An interesting application of simplicial *advancing front methods* in the two-dimensional case is given in the paper [73]. The method is based on the simple fact that any two triangles sharing a common edge form a quadrilateral. Obviously, a necessary condition for the success of the method is that the triangulation should consist of an even number of triangles. Unfortunately, the generalization of the method to the three-dimensional situation is difficult, because a comparatively large number of adjacent tetrahedra should be united to form a hexahedron.

Multiblock Methods
The basic idea of these methods is to partition the domain into a small number of subdomains ("blocks") of simple shape (quadrilaterals, hexahedra, as well as triangles, tetrahedra, prisms, pyramids, etc.) and then generate structured or logically structured grids in the individual subdomains (see, e.g., [23], [33]).

In multiblock grids, special attention has to be devoted to the treatment of common boundaries of adjacent blocks. Unless special discretization methods such as, for example, the so-called *mortar finite element method* (cf. [45]) are used in this situation, there may be a conflict between certain compatibility conditions at the common block interfaces (to ensure, e.g., the continuity of the finite element functions across the interfaces) on the one hand and the output directives of an error estimation procedure that may advise to refine a block-internal grid locally on the other hand.

Hierarchically Structured Grids
These grids are a further, hybrid variant of structured and unstructured grids, though not yet very widespread. Starting with a logically structured grid, hierarchically structured grids are generated by a further logically structured refinement of certain subdomains. As in multiblock methods, the interfaces between blocks of different refinement degrees have to be treated carefully.

Combined Grids
Especially in three-dimensional situations, the generation of "purely" hexahedral grids may be very difficult for complicated geometries of the domain. Therefore, the so-called combined grids that consist of hexahedral grids in geometrically simple subdomains and tetrahedral, prismatic, pyramidal, etc. grids in more critical subregions are used.

Chimera Grids
These grids are also called *overset grids* (see, e.g., [51]). In contrast to the multiblock grids described above, here the domain is covered by a compar-

atively small number of domains of simple shape, and then structured or logically structured grids are generated on the individual domains. That is, a certain overlapping of the blocks and thus of the subgrids is admitted.

4.1.4 Grid Optimization

Many grid generation codes include "smoothing algorithms" that optimize the grid with respect to certain regularity criteria. In the so-called *r-method* (relocation method) the nodes are slightly moved, keeping the logical structure (connectivities) of the grid fixed. Another approach is to improve the grid connectivities themselves.

A typical example for r-methods is given by the so-called *Laplacian smoothing* (or *barycentric smoothing*), where any inner grid point is moved into the barycentre of its neighbours (see [50]). A local weighting of selected neighbours can also be used (*weighted barycentric smoothing*). From a formal point of view, the application of the Laplacian smoothing corresponds to the solution of a system of linear algebraic equations that is obtained from the equations of the arithmetic (or weighted) average of the nodes. The matrix of this system is large but sparse. The structure of this matrix is very similar to the one that results from a finite volume discretization of the Poisson equation as described in Section 6.2 (see the corresponding special case of (6.9)). In general, there is no need to solve this system exactly. Typically, only one to three steps of a simple iterative solver (as presented in Section 5.1) are performed. When the domain is almost convex, Laplacian smoothing will produce good results. It is also clear that for strongly nonconvex domains or other special situations, the method may produce invalid grids.

Among the methods to optimize the grid connectivities, the so-called *2:1-rule* and, in the two-dimensional case, the *edge swap* (or *diagonal swap*, [59]) are well known. The 2:1-rule is used within the quadtree or octree method to reduce the difference of the refinement levels between neighbouring quadrilaterals or hexahedra to one by means of additional refinement steps; see Figure 4.3.

In the edge swap method, a triangular grid is improved. Since any two triangles sharing an edge form a convex quadrilateral, the method decides which of the two diagonals of the quadrilateral optimizes a given criterion. If the optimal diagonal does not coincide with the common edge, the other configuration will be taken; i.e., the edge will be swapped.

Finally, it should be mentioned that there exist grid optimization methods that delete nodes or even complete elements from the grid.

4.1.5 Grid Refinement

A typical grid refinement algorithm for a triangular grid, the so-called *red/green refinement*, has previously been introduced in Section 2.4.1. A

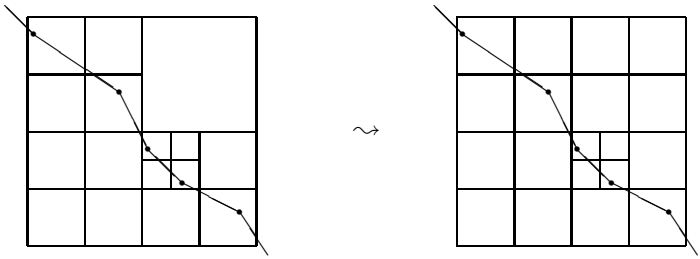


Figure 4.3. 2:1-rule.

further class of methods is based on bisection, that is, a triangle is divided by the median of an edge. A *method of bisection* is characterized by the number of bisections used within one refinement step (*stage number of the method of bisection*) and by the criterion of how to select the edge where the new node is to be located. A popular strategy is to take the longest of the three edges. The general (recursive) refinement step for some triangle K is of the following form:

(i) Find the longest edge of K and insert the median connecting the midpoint of that edge with the opposite vertex.

(ii) If the resulting new node is not a vertex of an already existing triangle or is not a boundary point of the domain Ω, then the adjacent triangle that shares the refined edge has to be divided, too.

Since the longest edge of the adjacent triangle need not coincide with the common edge, the application of this scheme leads to a nonconforming triangulation, in general. To obtain a conforming triangulation, all new nodes resulting from substep (i) have to be detected, and then certain closure rules have to be applied.

The red/green refinement as well as the method of bisection can be generalized to the three-dimensional case. However, since the number of different configurations is significantly larger than in the case $d = 2$, only a few illustrative examples will be given.

The red/green refinement of a tetrahedron K (see Figure 4.4) yields a partition of K into eight subtetrahedra with the following properties: All vertices of the subtetrahedra coincide either with vertices or with edge midpoints of K. At all the faces of K, the two-dimensional red/green refinement scheme can be observed.

In addition to the difficulties arising in the two-dimensional situation, the (one-stage) bisection applied to the longest edge of a tetrahedron also yields faces that violate the conformity conditions. Therefore, the closure rules are rather complicated, and in practice, multistage (often three-stage)

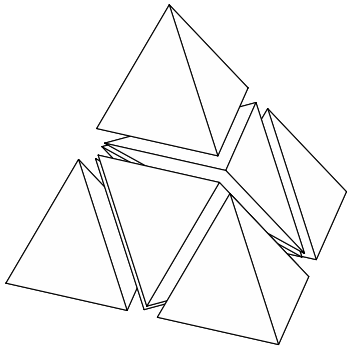


Figure 4.4. Representation of the red/green refinement of a tetrahedron.

methods of bisection are used to circumvent these difficulties (see Figure 4.5).

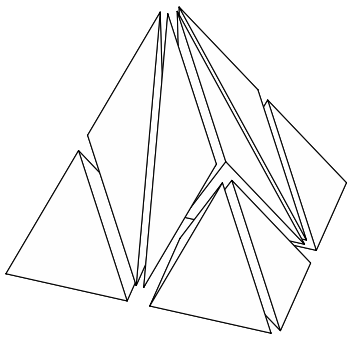


Figure 4.5. Representation of the bisection of a tetrahedron.

Grid refinement may be necessary in those parts of the domain where the weak solution of the variational equation has low regularity. The figure of the front cover (taken from [70]) shows the domain for a density-driven flow problem, where the inflow and the outflow pass through very small, nearly point-sized surfaces. The refinement is the result of a grid adaptation strategy based on a posteriori error estimators (see Section 4.2). In time-dependent problems, where those parts of the grid in which a refinement is needed may also vary, *grid coarsening* is necessary to limit the expense. A simple grid coarsening can be achieved, for example, by cancelling former refinement steps in a conforming way.

Exercises

4.1 For a given triangle K, the circumcentre can be computed by finding the intersection of the perpendicular bisectors associated with two edges

of K. This can be achieved by solving a linear system of equations with respect to the coordinates of the circumcentre.

(a) Give such a system.
(b) How can the radius of the circumcircle be obtained from this solution?

4.2 Given a triangle K, denote by h_i the length of edge i, $i \in \{1, 2, 3\}$. Prove that the following expression equals the radius of the circumcircle (without using the circumcentre!):

$$\frac{h_1 h_2 h_3}{4|K|}.$$

4.3 Let K_1, K_2 be two triangles sharing an edge.

(a) Show the equivalence of the following edge swap criteria:
Angle criterion: Select the diagonal of the so-formed quadrilateral that maximizes the minimum of the six interior angles among the two configurations.
Circle criterion: Choose the diagonal of the quadrilateral for which the open circumcircle disks to the resulting triangles do not contain any of the remaining vertices.
(b) If α_1, α_2 denote the two interior angles that are located opposite the common edge of the triangles K_1 and K_2, respectively, then the circle criterion states that an edge swap is to be performed if

$$\alpha_1 + \alpha_2 > \pi.$$

Prove this assertion.
(c) The criterion in (b) is numerically expensive. Show that the following test is equivalent:

$$\begin{aligned} & [(a_{1,1} - a_{3,1})(a_{2,1} - a_{3,1}) + (a_{1,2} - a_{3,2})(a_{2,2} - a_{3,2})] \\ & \quad * [(a_{2,1} - a_{4,1})(a_{1,2} - a_{4,2}) - (a_{1,1} - a_{4,1})(a_{2,2} - a_{4,2})] \\ < \; & [(a_{2,1} - a_{4,1})(a_{1,1} - a_{4,1}) + (a_{2,2} - a_{4,2})(a_{1,2} - a_{4,2})] \\ & \quad * [(a_{2,1} - a_{3,1})(a_{1,2} - a_{3,2}) - (a_{1,1} - a_{3,1})(a_{2,2} - a_{3,2})]. \end{aligned}$$

Here $a_i = (a_{i,1}, a_{i,2})^T$, $i \in \{1, 2, 3\}$, denote the vertices of a triangle ordered clockwise, and $a_4 = (a_{4,1}, a_{4,2})^T$ is the remaining vertex of the quadrilateral, the position of which is tested in relation to the circumcircle defined by a_1, a_2, a_3.
Hint: Addition theorems for the sin function.

4.2 A Posteriori Error Estimates and Grid Adaptation

In the practical application of discretization methods to partial differential equations, an important question is how much the computed approximative solution u_h deviates from the weak solution u of the given problem.

Typically, a certain norm of the error $u - u_h$ is taken as a measure of this deviation. For elliptic or parabolic differential equations of order two, a common norm to quantify the error is the energy norm (respectively an equivalent norm) or the L^2-norm. Some practically important problems involve the approximation of the so-called derived quantities which can be mathematically interpreted in terms of values of certain linear functionals of the solution u. In such a case, an estimate of the corresponding error is also of interest.

Example 4.1

$J(u) = \int_{\Gamma_0} \nu \cdot \nabla u \, d\sigma$: flux of u through a part of the boundary $\Gamma_0 \subset \partial\Omega$,
$J(u) = \int_{\Omega_0} u \, dx$: integral mean of u on some subdomain $\Omega_0 \subset \Omega$.

In the following we will consider some estimates for a norm $\|\cdot\|$ of the error $u - u_h$ and explain the corresponding terminology. Similar statements remain true if $\|u - u_h\|$ is replaced by $|J(u) - J(u_h)|$.

The error estimates given in the previous chapters are characterized by the fact that no information about the computed solution u_h is needed. Estimates of this type are called *a priori* error estimates.

For example, consider a variational equation with a bilinear form that satisfies (for some space V such that $H_0^1(\Omega) \subset V \subset H^1(\Omega)$ and $\|\cdot\| := \|\cdot\|_1$) the assumptions (2.42), (2.43) and use numerically piecewise linear, continuous finite elements. Then Céa's lemma (Theorem 2.17) together with the interpolation error estimate from Theorem 3.29 implies the estimate

$$\|u - u_h\|_1 \leq \frac{M}{\alpha} \|u - I_h(u)\|_1 \leq \frac{M}{\alpha} Ch \,, \tag{4.1}$$

where the constant C depends on the weak solution u of the variational equality.

Here C has the special form

$$C = \bar{C} \left\{ \int_\Omega \sum_{|\alpha|=2} |\partial^\alpha u|^2 \, dx \right\}^{1/2} \tag{4.2}$$

with $\bar{C} > 0$ independent of u. Unfortunately, the structure of the bound (4.2) does not allow an immediate numerical application of (4.1).

But even if the constant C could be estimated and (4.1) could be used to determine the discretization parameter h (maximum diameter of the triangles in $\mathcal{T}_h$) for a prescribed tolerance, in general this would lead to

a grid that is too fine. This corresponds to an algebraic problem that is too large. The reason is that the described approach determines a global parameter, whereas the true error measure may have different magnitudes in different regions of the domain Ω.

So we should aim at error estimates of type

$$\|u - u_h\| \leq D\eta \tag{4.3}$$

or

$$D_1\eta \leq \|u - u_h\| \leq D_2\eta\,, \tag{4.4}$$

where the constants $D, D_1, D_2 > 0$ do not depend on the discretization parameters and

$$\eta = \left\{ \sum_{K \in \mathcal{T}_h} \eta_K^2 \right\}^{1/2} . \tag{4.5}$$

Here the quantities η_K should be computable using only the data — including possibly $u_h|_K$ — which are known on the particular element K.

If the bounds η (or the terms η_K, respectively) in (4.3) (respectively (4.4)) depend on u_h, i.e., they can be evaluated only if u_h is known, then they are called (local) *a posteriori error estimators* in the wider sense.

Often the bounds also depend on the weak solution u of the variational equality, so in fact, they cannot be evaluated immediately. In such a case they should be replaced by computable quantities that do not depend on u in a direct way. So, if the bounds can be evaluated without knowing u but using possibly u_h, then they are called (local) a posteriori error estimators in the strict sense.

Inequalities of the form (4.3) guarantee, for a given tolerance $\varepsilon > 0$, that the inequality $\eta \leq \varepsilon$ implies that the error measure does not exceed ε up to a multiplicative constant. In this sense the error estimator η is called *reliable*. Now, if the computed approximative solution u_h is sufficiently precise in the described sense, then the computation can be finished. If u_h is such that $\eta > \varepsilon$, then the question of how to modify the discretization in order to achieve the tolerance or, if the computer resources are nearly exhausted, how to minimize the overshooting of η, arises. That is, the information given by the evaluation of the bounds has to be used to adapt the discretization and then to perform a new run of the solution process. A typical modification is to refine or to coarsen the grid.

Error estimators may overestimate the real error measure significantly; thus a grid adaptation procedure based on such an error estimate generates a grid that is too fine, and consequently, the corresponding algebraic problem is too large.

This effect can be reduced or even avoided if the error estimator satisfies a two-sided inequality like (4.4). Then the ratio D_2/D_1 is a measure of the *efficiency* of the error estimator.

An error estimator η is called *asymptotically exact* if for an arbitrary convergent sequence of approximations $\{u_h\}$ with $\|u - u_h\| \to 0$ the following limit is valid:

$$\frac{\eta}{\|u - u_h\|} \to 1 \,.$$

Usually, a posteriori error estimators are designed for a well-defined class of boundary or initial-boundary value problems. Within a given class of problems, the question regarding the sensitivity of the constants D in (4.3) or D_1, D_2 in (4.4), with respect to the particular data of the problem (e.g., coefficients, inhomogeneities, geometry of the domain, grid geometry, ...), arises. If this dependence of the data is not crucial, then the error estimator is called *robust* within this class.

Grid Adaptation

Let us assume that the local error estimators η_K composing an efficient error estimator η for an approximate solution u_h on some grid $\mathcal{T}_h$ really reflect the error on the element K and that this local error can be improved by a refinement of K (e.g., following the principles of Section 4.1.5). Then the following grid adaptation strategies can be applied until the given tolerance ε is reached or the computer resources are exhausted.

Equidistribution strategy: The objective of the grid adaptation (refinement or coarsening of elements) is to get a new grid $\mathcal{T}_h^{\text{new}}$ such that the local error estimators η_K^{new} for this new grid take one and the same value for all elements $K \in \mathcal{T}_h^{\text{new}}$; that is (cf. (4.5))

$$\eta_K^{\text{new}} \approx \frac{\varepsilon}{\sqrt{|\mathcal{T}_h^{\text{new}}|}} \quad \text{for all} \quad K \in \mathcal{T}_h^{\text{new}} \,.$$

Since the number of elements of the new grid enters the right-hand side of this criterion, the strategy is an implicit method. In practical use, it is approximated iteratively.

Cut-off strategy: Given a parameter $\kappa \in (0,1)$, a threshold value $\kappa\eta$ is defined. Then the elements K with $\eta_K > \kappa\eta$ will be refined.

Reduction strategy: Given a parameter $\kappa \in (0,1)$, an auxiliary tolerance $\varepsilon_\eta := \kappa\eta$ is defined. Then a couple of steps following the equidistribution strategy with the tolerance ε_η are performed.

In practice, the equidistribution strategy may perform comparatively slowly and thus may reduce the efficiency of the complete solution process. The cut-off method does not allow grid coarsening. It is rather sensitive to the choice of the parameter κ. Among all three strategies, the reduction method represents the best compromise.

Design of A Posteriori Error Estimators

In the following, three basic principles of the design of a posteriori error estimators will be described. In order to illustrate the underlying ideas and to avoid unnecessary technical difficulties, a model problem will be treated: Consider a diffusion-reaction equation on a polygonally bounded domain $\Omega \subset \mathbb{R}^2$ with homogeneous Dirichlet boundary conditions

$$\begin{aligned} -\Delta u + ru &= f \qquad \text{in } \Omega\,, \\ u &= 0 \qquad \text{on } \partial\Omega\,, \end{aligned}$$

where $f \in L^2(\Omega)$ and $r \in C(\overline{\Omega})$ with $r(x) \geq 0$ for all $x \in \Omega$. The problem is discretized using piecewise linear, continuous finite element functions as described in Section 2.2.

Setting $a(u,v) := \int_\Omega (\nabla u \cdot \nabla v + ruv)\,dx$ for $u, v \in V := H_0^1(\Omega)$, we have the following variational (weak) formulation:

Find $u \in V$ such that $a(u,v) = \langle f, v\rangle_0$ for all $v \in V$.

The corresponding finite element method reads as follows:

Find $u_h \in V_h$ such that $a(u_h, v_h) = \langle f, v_h\rangle_0$ for all $v_h \in V_h$.

Residual Error Estimators

Similar to the derivation of the a priori error estimate in the proof of Céa's lemma (Theorem 2.17), the V-ellipticity of a (2.43) implies that

$$\alpha \|u - u_h\|_1^2 \leq a(u - u_h, u - u_h)\,.$$

Without loss of generality we may suppose $u - u_h \in V \setminus \{0\}$, hence

$$\|u - u_h\|_1 \leq \frac{1}{\alpha} \frac{a(u - u_h, u - u_h)}{\|u - u_h\|_1} \leq \frac{1}{\alpha} \sup_{v \in V} \frac{a(u - u_h, v)}{\|v\|_1}\,. \tag{4.6}$$

We observe that the term

$$a(u - u_h, v) = a(u, v) - a(u_h, v) = \langle f, v\rangle_0 - a(u_h, v) \tag{4.7}$$

is the *residual* of the variational equation; i.e., the right-hand side of inequality (4.6) can be interpreted as a certain norm of the variational residual.

In a next step, the variational residual will be split into local terms according to the given grid, and these terms are transformed by means of integration by parts. For arbitrary $v \in V$, from (4.7) it follows that

$$\begin{aligned} a(u - u_h, v) &= \sum_{K \in \mathcal{T}_h} \left\{ \int_K f v\,dx - \int_K (\nabla u_h \cdot \nabla v + r u_h v)\,dx \right\} \\ &= \sum_{K \in \mathcal{T}_h} \left\{ \int_K [f - (-\Delta u_h + r u_h)] v\,dx - \int_{\partial K} \nu \cdot \nabla u_h v\,d\sigma \right\}. \end{aligned}$$

The first factor in the integrals over the elements K is the classical elementwise residual of the differential equation:

$$r_K(u_h) := [f - (-\Delta u_h + r u_h)]\big|_K$$

All quantities entering $r_K(u_h)$ are known. In the case considered here we even have $-\Delta u_h = 0$ on K, hence $r_K(u_h) = [f - r u_h]\big|_K$.

The integrals over the boundary of the elements K are further split into a sum over the integrals along the element edges $E \subset \partial K$:

$$\int_{\partial K} \nu \cdot \nabla u_h v \, d\sigma = \sum_{E \subset \partial K} \int_E \nu \cdot \nabla u_h v \, d\sigma \,.$$

Since $v = 0$ on $\partial\Omega$, only the integrals along edges lying in Ω contribute to the sum. Denoting by $\mathcal{E}_h$ the set of all interior edges of all elements $K \in \mathcal{T}_h$ and assigning a fixed unit normal ν_E to any of those edges, we see that in the summation of the split boundary integrals over all $K \in \mathcal{T}_h$ there occur exactly two integrals along one and the same edge $E \in \mathcal{E}_h$. This observation results in the relation

$$\sum_{K \in \mathcal{T}_h} \int_{\partial K} \nu \cdot \nabla u_h v \, d\sigma = \sum_{E \in \mathcal{E}_h} \int_E [\nu_E \cdot \nabla u_h]_E \, v \, d\sigma \,,$$

where, for a piecewise continuous function $w : \Omega \to \mathbb{R}$, the term

$$[w]_E(x) := \lim_{\varrho \to +0} w(x + \varrho \nu_E) - \lim_{\varrho \to +0} w(x - \varrho \nu_E) \,, \quad x \in E \,,$$

denotes the *jump* of the function w across the edge E. If w is the normal derivative of u_h in the fixed direction ν_E, i.e., $w = \nu_E \cdot \nabla u_h$, then its jump does not depend on the particular orientation of ν_E (see Exercise 4.6).

In summary, we have the following relation:

$$a(u - u_h, v) = \sum_{K \in \mathcal{T}_h} \int_K r_K(u_h) v \, dx - \sum_{E \in \mathcal{E}_h} \int_E [\nu_E \cdot \nabla u_h]_E \, v \, d\sigma \,.$$

Using the error equation (2.39), we obtain for an arbitrary element $v_h \in V_h$ the fundamental identity

$$\begin{aligned} a(u - u_h, v) &= a(u - u_h, v - v_h) \\ &= \sum_{K \in \mathcal{T}_h} \int_K r_K(u_h)(v - v_h) \, dx \\ &\quad - \sum_{E \in \mathcal{E}_h} \int_E [\nu_E \cdot \nabla u_h]_E \, (v - v_h) \, d\sigma \,, \end{aligned}$$

which is the starting point for the construction of further estimates.

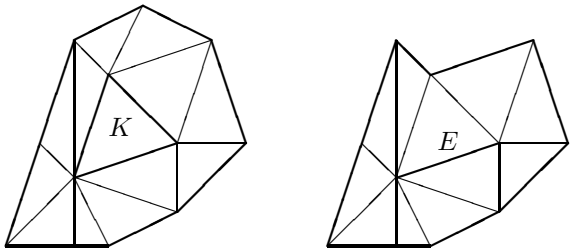


Figure 4.6. The triangular neighbourhoods $\Delta(K)$ (left) and $\Delta(E)$ (right).

First we see that the Cauchy–Schwarz inequality immediately implies

$$\begin{aligned} a(u-u_h, v-v_h) \quad &\leq \quad \sum_{K\in\mathcal{T}_h} \|r_K(u_h)\|_{0,K}\|v-v_h\|_{0,K} \\ &\quad + \sum_{E\in\mathcal{E}_h} \left\|[\nu_E\cdot\nabla u_h]_E\right\|_{0,E}\|v-v_h\|_{0,E}\,. \end{aligned} \tag{4.8}$$

To get this bound as small as possible, the function $v_h \in V_h$ is chosen such that the element $v \in V$ is approximated adequately in both spaces $L^2(K)$ and $L^2(E)$. One suggestion is the use of an interpolating function according to (2.47). However, since $V \notin C(\overline{\Omega})$, this interpolant is not defined. Therefore other approximation procedures have to be applied. Roughly speaking, suitable approximation principles, due to Clément [52] or Scott and Zhang [67], are based on taking certain local integral means. However, at this place we cannot go further into these details and refer to the cited literature. In fact, for our purposes it is important only that such approximations exist. Their particular design is of minor interest.

We will formulate the relevant facts as a lemma. To do so, we need some additional notation (see Figure 4.6):

triangular neighbourhood of a triangle K : $\quad \Delta(K) := \bigcup_{K':K'\cap K\neq\emptyset} K'$,

triangular neighbourhood of an edge E : $\quad \Delta(E) := \bigcup_{K':K'\cap E\neq\emptyset} K'$.

Thus $\Delta(K)$ consists of the union of the supports of those nodal basis functions that are associated with the vertices of K, whereas $\Delta(E)$ is formed by the union of those nodal basis functions that are associated with the boundary points of E. Furthermore, the length of the edge E is denoted by $h_E := |E|$.

Lemma 4.2 *Let a regular family $(\mathcal{T}_h)$ of triangulations of the domain Ω be given. Then for any $v \in V$ there exists an element $Q_h v \in V_h$ such that for all triangles $K \in \mathcal{T}_h$ and all edges $E \in \mathcal{E}_h$ the following estimates are valid:*

$$\|v - Q_h v\|_{0,K} \quad \leq \quad Ch_K|v|_{1,\Delta(K)}\,,$$

$$\|v - Q_h v\|_{0,E} \quad \leq \quad C\sqrt{h_E}|v|_{1,\Delta(E)}\,,$$

where the constant $C > 0$ depends only on the family of triangulations.

Now, setting $v_h = Q_h v$ in (4.8), the discrete Cauchy-Schwarz inequality yields

$$\begin{aligned} a(u-u_h,v) \;\leq\; & C \sum_{K\in\mathcal{T}_h} h_K \|r_K(u_h)\|_{0,K}|v|_{1,\Delta(K)} \\ & + C \sum_{E\in\mathcal{E}_h} \sqrt{h_E}\big\|[\nu_E\cdot\nabla u_h]_E\big\|_{0,E}|v|_{1,\Delta(E)} \\ \leq\; & C\left\{\sum_{K\in\mathcal{T}_h} h_K^2\|r_K(u_h)\|_{0,K}^2\right\}^{1/2}\left\{\sum_{K\in\mathcal{T}_h}|v|_{1,\Delta(K)}^2\right\}^{1/2} \\ & + C\left\{\sum_{E\in\mathcal{E}_h} h_E\big\|[\nu_E\cdot\nabla u_h]_E\big\|_{0,E}^2\right\}^{1/2}\left\{\sum_{E\in\mathcal{E}_h}|v|_{1,\Delta(E)}^2\right\}^{1/2}. \end{aligned}$$

A detailed investigation of the two second factors shows that we can decompose the integrals over $\Delta(K)$, $\Delta(E)$, according to

$$\int_{\Delta(K)}\ldots = \sum_{K'\subset\Delta(K)}\int_{K'}\ldots\,, \qquad \int_{\Delta(E)}\ldots = \sum_{K'\subset\Delta(E)}\int_{K'}\ldots\,.$$

This leads to a repeated summation of the integrals over the single elements K. However, due to the regularity of the family of triangulations, the multiplicity of these summations is bounded independently of the particular triangulation (see (3.93)). So we arrive at the estimates

$$\sum_{K\in\mathcal{T}_h}|v|_{1,\Delta(K)}^2 \leq C|v|_1^2, \qquad \sum_{E\in\mathcal{E}_h}|v|_{1,\Delta(E)}^2 \leq C|v|_1^2\,.$$

Using the inequality $a+b \leq \sqrt{2(a^2+b^2)}$ for $a,b\in\mathbb{R}$, we get

$$\begin{aligned} & a(u-u_h,v) \\ & \leq \quad C\left\{\sum_{K\in\mathcal{T}_h} h_K^2\|r_K(u_h)\|_{0,K}^2 + \sum_{E\in\mathcal{E}_h} h_E\|[\nu_E\cdot\nabla u_h]_E\|_{0,E}^2\right\}^{1/2}|v|_1\,. \end{aligned}$$

Finally, (4.6) yields

$$\|u-u_h\|_1 \leq D\eta \qquad \text{with} \quad \eta^2 := \sum_{K\in\mathcal{T}_h}\eta_K^2$$

and

$$\eta_K^2 := h_K^2\|f-ru_h\|_{0,K}^2 + \frac{1}{2}\sum_{E\subset\partial K\setminus\partial\Omega} h_E\big\|[\nu_E\cdot\nabla u_h]_E\big\|_{0,E}^2\,. \tag{4.9}$$

Here we have taken into account that in the transformation of the edge sum

$$\sum_{E\in\mathcal{E}_h} \dots \qquad \text{into the double sum} \quad \sum_{K\in\mathcal{T}_h} \sum_{E\subset\partial K\setminus\partial\Omega} \dots$$

the latter sums up every interior edge twice.

In summary, we have obtained an a posteriori error estimate of the form (4.3). By means of refined arguments it is also possible to derive lower bounds for $\|u-u_h\|_1$. For details, we refer to the literature, for example [35].

Error Estimation by Gradient Recovery

If we are interested in an estimate of the error $u-u_h \in V = H_0^1(\Omega)$ measured in the H^1- or energy norm $\|\cdot\|$, this problem can be simplified by means of the fact that both norms are equivalent on V to the H^1-seminorm

$$|u-u_h|_1 = \left\{ \int_\Omega |\nabla u - \nabla u_h|^2 \, dx \right\}^{1/2} =: \|\nabla u - \nabla u_h\|_0 \,.$$

This is a simple consequence of the definitions and the Poincaré inequality (see Theorem 2.18). That is, there exist constants $C_1, C_2 > 0$ independent of h such that

$$C_1|u-u_h|_1 \le \|u-u_h\| \le C_2|u-u_h|_1 \tag{4.10}$$

(cf. Exercise 4.8). Consequently, ∇u remains the only unknown quantity in the error bound.

The idea of error estimation by means of gradient recovery is to replace the unknown gradient of the weak solution u by a suitable quantity $R_h u_h$ that is computable from the approximative solution u_h at moderate expense. A popular example of such a technique is the so-called *Z^2 estimate*. Here we will describe a simple version of it. Further applications can be found in the original papers by Zienkiewicz and Zhu, e.g., [74].

Similar to the notation introduced in the preceding subsection, for a given node a the set

$$\Delta(a) := \bigcup_{K':a\in\partial K'} K'$$

denotes the triangular neighbourhood of a (see Figure 4.7). This set coincides with the support of the piecewise linear, continuous basis function associated with that node.

The gradient ∇u_h of a piecewise linear, continuous finite element function u_h is constant on every triangle K. This suggests that at any node a of the triangulation $\mathcal{T}_h$ we define the average $R_h u_h(a)$ of the values of the gradients on those triangles sharing the vertex a:

$$R_h u_h(a) := \frac{1}{|\Delta(a)|} \sum_{K\subset\Delta(a)} \nabla u_h|_K \, |K| \,.$$

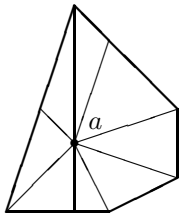


Figure 4.7. The triangular neighbourhood $\Delta(a)$.

Interpolating the two components of these nodal values of $R_h u_h$ separately in V_h, we get a *recovery operator* $R_h : V_h \to V_h \times V_h$.

Now a local error estimator can be defined by the simple restriction of the quantity $\overline{\eta} := \|R_h u_h - \nabla u_h\|_0$ onto a single element K:

$$\overline{\eta}_K := \|R_h u_h - \nabla u_h\|_{0,K} .$$

A nice insight into the properties of this local estimator was given by Rodríguez ([64], see also [35]), who compared it with the corresponding residual estimator (4.9). Namely, neglecting in the residual estimator just the residual part, i.e., setting

$$\tilde{\eta}_K^2 := \frac{1}{2} \sum_{E \subset \partial K \setminus \partial\Omega} h_E \big\| [\nu_E \cdot \nabla u_h]_E \big\|_{0,E}^2 \qquad \text{and} \quad \tilde{\eta}^2 := \sum_{K \in \mathcal{T}_h} \tilde{\eta}_K^2 ,$$

then the following result is true:

Theorem 4.3 *There exist two constants $c_1, c_2 > 0$ depending only on the family of triangulations such that*

$$c_1 \tilde{\eta} \le \overline{\eta} \le c_2 \tilde{\eta} .$$

The motivation for the method of gradient recovery is to be seen in the fact that $R_h u_h$ possesses special convergence properties. Namely, under certain assumptions the recovered gradient $R_h u_h$ converges asymptotically to ∇u faster than ∇u_h does. In such a case $R_h u_h$ is said to be a *superconvergent approximation* to ∇u.

If superconvergence holds, the simple decomposition

$$\nabla u - \nabla u_h = R_h u_h - \nabla u_h + \nabla u - R_h u_h$$

demonstrates that the first difference on the right-hand side represents the asymptotically dominating, computable part of the gradient error $\nabla u - \nabla u_h$. In other words, if we could define, for the class of problems under consideration, a superconvergent gradient recovery $R_h u_h$ that is computable with moderate expense, then the quantities $\overline{\eta}_K$ and $\overline{\eta}$ defined above may serve as a tool for a posteriori error estimation.

Unfortunately, such superconvergence properties are valid only under rather restrictive assumptions (especially with respect to the grid and to the regularity of the weak solution). Thus it is difficult to obtain a full mathematical foundation in practice. Nevertheless, gradient recovery is often applied and yields satisfactory results in many situations.

The following example, which is due to Repin [63], shows that a recovered gradient does not have to reflect the real behaviour of the error.

Example 4.4 Consider the following boundary value problem for $d = 1$ and $\Omega = (0, 1)$:

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

If f is constant, the exact solution reads $u(x) = x(2 + (1-x)f)/2$. Suppose we have found the function $v_h = x$ as an approximate solution. For an arbitrary partition of Ω into subintervals, this function is piecewise linear and it satisfies the boundary conditions formulated above. Now let R_h be an arbitrary gradient recovery operator that is able to reproduce at least constants. Since $v_h' = 1$, we have $v_h' - R_h v_h = 0$, whereas the real error is $v_h' - u' = (x - \frac{1}{2})f$.

An interpretation of this effect is that the function v_h does not solve the corresponding discrete (Galerkin) equations. But this property of u_h is used for the proof of superconvergence. This property also plays an important role in the derivation of the residual error estimates, because the error equation is used therein.

Dual-Weighted Residual Error Estimators

The aforementioned a posteriori error estimates have two disadvantages: On the one hand, certain global constants, which are not known in general, are part of the bounds. Typical examples are α^{-1} in (4.6) and the constants C_1, C_2 in the equivalence relation (4.10). On the other hand, we obtained scaling factors like h_K and $\sqrt{h_E}$ simply by using a particular approximation operator.

In the following, we will outline a method that attempts to circumvent these drawbacks. It is especially appropriate for the estimation of errors of functionals depending linearly on the solution.

So let $J : V \to \mathbb{R}$ denote a linear, continuous functional. We are interested in an estimate of $|J(u) - J(u_h)|$. Therefore, the following auxiliary *dual* problem is considered:

$$\text{Find } w \in V \text{ such that } a(v, w) = J(v) \text{ for all } v \in V.$$

Taking $v = u - u_h$, we get immediately

$$J(u) - J(u_h) = J(u - u_h) = a(u - u_h, w)\,.$$

If $w_h \in V_h$ is an arbitrary element, the error equation (2.39) yields

$$J(u) - J(u_h) = a(u - u_h, w - w_h)\,.$$

Obviously, the right-hand side is of the same structure as in the derivation of the estimate (4.8). Consequently, by using the same arguments it follows that

$$\begin{aligned} |J(u) - J(u_h)| \quad \leq \quad & \sum_{K \in \mathcal{T}_h} \|r_K(u_h)\|_{0,K} \|w - w_h\|_{0,K} \\ & + \sum_{E \in \mathcal{E}_h} \left\| [\nu_E \cdot \nabla u_h]_E \right\|_{0,E} \|w - w_h\|_{0,E} \,. \end{aligned}$$

In contrast to the previous approaches, here the norms of $w - w_h$ will not be theoretically analyzed but numerically approximated. This can be done by an approximation of the dual solution w. There are several (more or less heuristic) ways to do this.

(1) Estimation of the approximation error: Here, the norms of $w - w_h$ are estimated as in the case of residual error estimators. Since the result depends on the unknown H^1-seminorm of w, which is equivalent to the L^2-norm of ∇w, the finite element solution $w_h \in V_h$ of the auxiliary problem is used to approximate ∇w. It is a great disadvantage of this approach that again global constants enter in the final estimate through the estimation of the approximation error. Furthermore, the discrete auxiliary problem is of similar complexity to that of the original discrete problem.

(2) Higher-order discretizations of the auxiliary problem: The auxiliary problem is solved numerically by using a method that is more accurate than the original method to determine a solution in V_h. Then w is replaced by that solution and $w_h \in V_h$ by an interpolant of that solution. Unfortunately, since the discrete auxiliary problem is of comparatively large dimension, this approach is rather expensive.

(3) Approximation by means of higher-order recovery: This method works similarly to the approach described in the previous subsection; w is replaced by an element that is recovered from the finite element solution $w_h \in V_h$ of the auxiliary problem. The recovered element approximates w with higher order in both norms than w_h does. This method exhibits two problems: On the one hand, the auxiliary problem has to be solved numerically, and on the other hand, ensuring the corresponding superconvergence properties may be difficult.

At the end of this section we want to mention how the method could be used to estimate certain norms of the error. In the case where the norms are induced by particular scalar products, there is a simple, formal way. For example, for the L^2-norm we have

$$\|u - u_h\|_0 = \frac{\langle u - u_h, u - u_h \rangle_0}{\|u - u_h\|_0} \,.$$

Keeping u and u_h fixed, we get with the definition

$$J(v) := \frac{\langle v, u - u_h \rangle_0}{\|u - u_h\|_0}$$

a linear, continuous functional $J : H^1(\Omega) \to \mathbb{R}$ such that $J(u) - J(u_h) = \|u - u_h\|_0$.

The practical difficulty of this approach consists in the fact that to be able to find the solution w of the auxiliary problem we have to know the values of J, but they depend on the unknown element $u - u_h$. The idea of approximating these values immediately implies two problems: There is additional expense, and the influence of the approximation quality on the accuracy of the obtained bounds has to be analyzed.

Exercises

4.4 Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with a polygonal, Lipschitz continuous boundary and $V := H_0^1(\Omega)$. Now consider a V-elliptic, continuous bilinear form a and a continuous linear form b. The problem

$$u \in V : \quad a(u, v) = b(v) \quad \text{for all } v \in V$$

is discretized using piecewise linear, continuous finite elements. If E_i denotes the support of the nodal basis functions of V_h associated with the vertex a_i, show that the abstract local error indicators

$$\eta_i := \sup_{v \in H_0^1(E_i)} \frac{a(e, v)}{\|v\|}$$

can be estimated by means of the solutions $e_i \in H_0^1(E_i)$ of the local boundary value problems

$$e_i \in H_0^1(E_i) : \quad a(e_i, v) = b(v) - a(u_h, v) \quad \text{for all } v \in H_0^1(E_i)$$

as follows (M and α denote the constants appearing in the continuity and ellipticity conditions on a):

$$\alpha \|e_i\| \le \eta_i \le M \|e_i\| \,.$$

If necessary, the elements of $H_0^1(E_i)$ are extended by zero to the whole domain Ω.

4.5 A linear polynomial on some triangle is uniquely defined either by its values at the vertices or by its values at the edge midpoints. For a fixed triangulation of a polygonally bounded, simply connected domain $\Omega \subset \mathbb{R}^2$, there can be defined two finite element spaces by identifying common degrees of freedom of adjacent triangles.

(a) Show that the dimension of the space defined by the degrees of freedom located at the vertices is less than the dimension of the other space (provided that the triangulation consists of more than one triangle).

(b) How can one explain this "loss of degrees of freedom"?

4.6 Denote by $\mathcal{T}_h$ a triangulation of the domain $\Omega \subset \mathbb{R}^d$. Show that for a function $v : \Omega \to \mathbb{R}$ that is continuously differentiable on each element the jump $[\nu_E \cdot \nabla v]_E$ of the normal derivative of v across an element edge E does not depend on the orientation of the normal ν_E.

4.7 Let a regular family of triangulations $(\mathcal{T}_h)$ of a domain $\Omega \subset \mathbb{R}^2$ be given. Show that there exist constants $C > 0$ that depend only on the family $(\mathcal{T}_h)$ such that

$$\begin{aligned} \sum_{K \in \mathcal{T}_h} |v|^2_{0,\Delta(K)} &\leq C\|v\|^2_0 \qquad \text{for all } v \in L^2(\Omega)\,, \\ \sum_{E \in \mathcal{E}_h} |v|^2_{0,\Delta(E)} &\leq C\|v\|^2_0 \qquad \text{for all } v \in L^2(\Omega)\,. \end{aligned}$$

4.8 Let $\Omega \subset \mathbb{R}^d$ be a bounded domain. Show that there are constants $C_1, C_2 > 0$ such that for all $v \in H^1_0(\Omega)$,

$$C_1|v|_1 \leq \|v\|_1 \leq C_2|v|_1\,.$$

5

Iterative Methods for Systems of Linear Equations

We consider again the system of linear equations

$$Ax = b \tag{5.1}$$

with nonsingular matrix $A \in \mathbb{R}^{m,m}$, right-hand side $b \in \mathbb{R}^m$, and solution $x \in \mathbb{R}^m$. As shown in Chapters 2 and 3, such systems of equations arise from finite element discretizations of elliptic boundary value problems. The matrix A is the stiffness matrix and thus sparse, as can be seen from (2.37). A *sparse* matrix is vaguely a matrix with so many vanishing elements that using this structure in the solution of (5.1) is advantageous. Taking advantage of a band or hull structure was discussed in Section 2.5. More precisely, if (5.1) represents a finite element discretization, then it is not sufficient to know the properties of the solution method for a fixed m. It is on the contrary necessary to study a sequence of problems with growing dimension m, as it appears by the refinement of a triangulation. In the strict sense we understand by the notion *sparse matrices* a sequence of matrices in which the number of nonzero elements per row is bounded independently of the dimension. This is the case for the stiffness matrices due to (2.37) if the underlying sequence of triangulations is regular in the sense of Definition 3.28, for example. In finite element discretizations of time-dependent problems (Chapter 7) as well as in finite volume discretizations (Chapter 6) systems of equations of equal properties arise, so that the following considerations can be also applied there.

The described matrix structure is best applied in iterative methods that have the operation matrix $\times$ vector as an essential module, where either the system matrix A or a matrix of similar structure derived from it is

concerned. If the matrix is sparse in the strict sense, then $O(m)$ elementary operations are necessary. In particular, list-oriented storage schemes can be of use, as pointed out in Section 2.5.

The effort for the approximative solution of (5.1) by an iterative method is determined by the number of elementary operations per iteration step and the number of iterations k that are necessary in order to reach the desired *relative error level* $\varepsilon > 0$, i.e., to meet the demand

$$\left\|x^{(k)} - x\right\| \leq \varepsilon \left\|x^{(0)} - x\right\| . \tag{5.2}$$

Here $\left(x^{(k)}\right)_k$ is the sequence of iterates for the initial value $x^{(0)}$, $\|\cdot\|$ a fixed norm in $\mathbb{R}^m$, and $x = A^{-1}b$ the exact solution of (5.1).

For all methods to be discussed we will have *linear convergence* of the kind

$$\left\|x^{(k)} - x\right\| \leq \varrho^k \left\|x^{(0)} - x\right\| \tag{5.3}$$

with a *contraction number* ϱ with $0 < \varrho < 1$, which in general depends on the dimension m. To satisfy (5.2), k iterations are thus sufficient, with

$$k \geq \left(\ln \frac{1}{\varepsilon}\right) \Big/ \left(\ln \frac{1}{\varrho}\right) . \tag{5.4}$$

The computational effort of a method obviously depends on the size of ε, although this will be seen as fixed and only the dependence on the dimension m is considered: often ε will be omitted in the corresponding Landau's symbols. The methods differ therefore by their convergence behaviour, described by the contraction number ϱ and especially by its dependence on m (for specific classes of matrices and boundary value problems). A method is *(asymptotically) optimal* if the contraction numbers are bounded independently of m:

$$\varrho(m) \leq \overline{\varrho} < 1 . \tag{5.5}$$

In this case the total effort for a sparse matrix is $O(m)$ elementary operations, as for a matrix $\times$ vector step. Of course, for a more exact comparison, the corresponding constants, which also reflect the effort of an iteration step, have to be exactly estimated.

While direct methods solve the system of equations (5.1) with machine precision, provided it is solvable in a stable manner, one can freely choose the accuracy with iterative methods. If (5.1) is generated by the discretization of a boundary value problem, it is recommended to solve it only with that accuracy with which (5.1) approximates the boundary value problem. Asymptotic statements hereto have, among others, been developed in (3.89), (7.129) and give an estimation of the approximation error by Ch^α, with constants $C, \alpha > 0$, whereby h is the mesh size of the corresponding triangulation. Since the constants in these estimates are usually unknown, the error level can be adapted only asymptotically in m, in order to gain

an *algorithmic error* of equal asymptotics compared to the error of approximation. Although this contradicts the above-described point of view of a constant error level, it does not alter anything in the comparison of the methods: The respective effort always has to be multiplied by a factor $O(\ln m)$ if in d space dimensions $m \sim h^{-d}$ is valid, and the relations between the methods compared remain the same.

Furthermore, the choice of the error level ε will be influenced by the quality of the initial iterate. Generally, statements about the initial iterate are only possible for special situations: For parabolic initial boundary value problems (Chapter 7) and a one-step time discretization it is recommended to use the approximation of the old time level as initial iterate. In the case of a hierarchy of space discretizations, a *nested iteration* is possible (Section 5.6), where the initial iterates will naturally result.

5.1 Linear Stationary Iterative Methods

5.1.1 General Theory

We begin with the study of the following class of affine-linear iteration functions,

$$\Phi(x) := Mx + Nb\,, \tag{5.6}$$

with matrices $M, N \in \mathbb{R}^{m,m}$ to be specified later. By means of Φ an iteration sequence $x^{(0)}, x^{(1)}, x^{(2)}, \ldots$ is defined through a *fixed-point iteration*

$$x^{(k+1)} := \Phi\big(x^{(k)}\big)\,, \quad k = 0, 1, \ldots\,, \tag{5.7}$$

from an initial approximation $x^{(0)}$. Methods of this kind are called *linear stationary*, because of their form (5.6) with a fixed *iteration matrix* M. The function $\Phi : \mathbb{R}^m \to \mathbb{R}^m$ is continuous, so that in case of convergence of $x^{(k)}$ for $k \to \infty$, for the limit x we have

$$x = \Phi(x) = Mx + Nb\,.$$

In order to achieve that the fixed-point iteration defined by (5.6) is *consistent* with $Ax = b$, i.e., each solution of (5.1) is also a fixed point, we must require

$$A^{-1}b = MA^{-1}b + Nb \;\text{ for arbitrary }\; b \in \mathbb{R}^m\,,$$

i.e., $A^{-1} = MA^{-1} + N$, and thus

$$I = M + NA\,. \tag{5.8}$$

On the other hand, if N is nonsingular, which will always be the case in the following, then (5.8) also implies that a fixed point of (5.6) solves the system of equations.

Assuming the validity of (5.8), the fixed-point iteration for (5.6) can also be written as

$$x^{(k+1)} = x^{(k)} - N\big(Ax^{(k)} - b\big)\,, \tag{5.9}$$

because

$$Mx^{(k)} + Nb = (I - NA)\,x^{(k)} + Nb\,.$$

If N is nonsingular, we have additionally an equivalent form given by

$$W\big(x^{(k+1)} - x^{(k)}\big) = -\big(Ax^{(k)} - b\big) \tag{5.10}$$

with $W := N^{-1}$. The *correction* $x^{(k+1)} - x^{(k)}$ for $x^{(k)}$ is given by the *residual*

$$g^{(k)} := Ax^{(k)} - b$$

through (5.9) or (5.10), possibly by solving a system of equations. In order to compete with the direct method, the solution of (5.10) should require one order in m fewer elementary operations. For dense matrices no more operations than $O(m^2)$ should be necessary as are already necessary for the calculation of $g^{(k)}$. The same holds for sparse matrices, for example band matrices. On the other side the method should converge, and that as quickly as possible.

In the form (5.6) Φ is Lipschitz continuous for a given norm $\|\cdot\|$ on $\mathbb{R}^m$ with Lipschitz constant $\|M\|$, where $\|\cdot\|$ is a norm on $\mathbb{R}^{m,m}$ that is consistent with the vector norm (see (A3.9)).

More precisely, for a consistent iteration the *error*

$$e^{(k)} := x^{(k)} - x\,,$$

with $x = A^{-1}b$ still denoting the exact solution, even satisfies

$$e^{(k+1)} = Me^{(k)}\,,$$

because (5.7) and (5.8) imply

$$e^{(k+1)} = x^{(k+1)} - x = Mx^{(k)} + Nb - Mx - NAx = Me^{(k)}\,. \tag{5.11}$$

The *spectral radius of* M, that is, the maximum of the absolute values of the (complex) eigenvalues of M, will be denoted by $\varrho(M)$.

The following general convergence theorem holds:

Theorem 5.1 *A fixed-point iteration given by* (5.6) *to solve* $Ax = b$ *is globally and linearly convergent if*

$$\varrho(M) < 1\,. \tag{5.12}$$

This is satisfied if for a matrix norm $\|\cdot\|$ *on* $\mathbb{R}^{m,m}$ *induced by a norm* $\|\cdot\|$ *on* $\mathbb{R}^m$ *we have*

$$\|M\| < 1\,. \tag{5.13}$$

If the consistency condition (5.8) *holds and the matrix and vector norms applied are consistent, then the convergence is monotone in the following sense:*

$$\left\|e^{(k+1)}\right\| \leq \left\|M\right\| \left\|e^{(k)}\right\| . \tag{5.14}$$

Proof: Assuming (5.12), then for $\varepsilon = (1 - \varrho(M))/2 > 0$ there is a norm $\|\cdot\|_S$ on $\mathbb{R}^m$ such that the induced norm $\|\cdot\|_S$ on $\mathbb{R}^{m,m}$ satisfies

$$\|M\|_S \leq \varrho(M) + \varepsilon < 1$$

(see [16, p. 34]). The function Φ is a contraction with respect to this special norm on $\mathbb{R}^m$. Therefore, Banach's fixed-point theorem (Theorem 8.4) can be applied on $X = (\mathbb{R}^m, \|\cdot\|_S)$, which ensures the global convergence of the sequence $\left(x^{(k)}\right)_k$ to a fixed point $\bar{x}$ of Φ.

If (5.13) holds, Φ is a contraction even with respect to the norm $\|\cdot\|$ on $\mathbb{R}^m$, and $\|M\|$ is the Lipschitz constant. Finally relation (5.14) follows from (5.11). □

In any case, we have convergence in any norm on $\mathbb{R}^m$, since they are all equivalent. Linear convergence for (5.12) holds only in the generally not available norm $\|\cdot\|_S$ with $\|M\|_S$ as contraction number.

As termination criterion for the concrete iteration methods to be introduced, often

$$\left\|g^{(k)}\right\| \leq \delta \left\|g^{(0)}\right\| \tag{5.15}$$

is used with a control parameter $\delta > 0$, abbreviated as $\|g^{(k)}\| = 0$. The connection to the desired reduction of the relative error according to (5.2) is given by

$$\frac{\left\|e^{(k)}\right\|}{\left\|e^{(0)}\right\|} \leq \kappa(A) \frac{\left\|g^{(k)}\right\|}{\left\|g^{(0)}\right\|} , \tag{5.16}$$

where the condition number $\kappa(A) = \|A\|\|A^{-1}\|$ is to be computed with respect to a matrix norm that is consistent with the chosen vector norm. Relation (5.16) follows from

$$\begin{aligned} \left\|e^{(k)}\right\| &= \left\|A^{-1}g^{(k)}\right\| \leq \left\|A^{-1}\right\| \left\|g^{(k)}\right\| , \\ \left\|g^{(0)}\right\| &= \left\|Ae^{(0)}\right\| \leq \left\|A\right\| \left\|e^{(0)}\right\| . \end{aligned}$$

Therefore, for the selection of δ in (5.15) we have to take into account the behaviour of the condition number.

For the iteration matrix M, according to (5.8), we have

$$M = I - NA ,$$

or according to (5.10) with nonsingular W,

$$M = I - W^{-1}A .$$

To improve the convergence, i.e. to reduce $\varrho(M)$ (or $\|M\|$), we need

$$N \approx A^{-1} \text{ and } W \approx A\,,$$

which is in contradiction to the fast solvability of (5.10).

5.1.2 Classical Methods

The fast solvability of (5.10) (in $O(m)$ operations) is ensured by choosing

$$W := D\,, \tag{5.17}$$

where $A = L + D + R$ is the unique partition of A, with a strictly lower triangular matrix L, a strictly upper triangular matrix R, and the diagonal matrix D:

$$L := \begin{pmatrix} 0 & \cdots & \cdots & 0 \\ a_{2,1} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ a_{m,1} & \cdots & a_{m,m-1} & 0 \end{pmatrix}, \quad R := \begin{pmatrix} 0 & a_{1,2} & \cdots & a_{1,m} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{m-1,m} \\ 0 & \cdots & \cdots & 0 \end{pmatrix},$$

$$D := \begin{pmatrix} a_{11} & & & \\ & a_{22} & & \mathbf{0} \\ & & \ddots & \\ & \mathbf{0} & & a_{mm} \end{pmatrix}. \tag{5.18}$$

Assume $a_{ii} \neq 0$ for all $i = 1, \ldots, m$, or equivalently that D is nonsingular, which can be achieved by row and column permutation.

The choice of (5.17) is called the *method of simultaneous displacements* or *Jacobi's method*. In the formulation form (5.6) we have

$$\begin{aligned} N &= D^{-1}\,, \\ M_{\mathrm{J}} &= I - NA = I - D^{-1}A = -D^{-1}(L+R)\,. \end{aligned}$$

Therefore, the iteration can be written as

$$D\big(x^{(k+1)} - x^{(k)}\big) = -\big(Ax^{(k)} - b\big)$$

or

$$x^{(k+1)} = D^{-1}\big(-Lx^{(k)} - Rx^{(k)} + b\big) \tag{5.19}$$

or

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(-\sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{m} a_{ij}x_j^{(k)} + b_i\right) \text{ for all } i = 1, \ldots, m\,.$$

On the right side in the first sum it is reasonable to use the new iterate $x^{(k+1)}$ where it is already calculated. This leads us to the iteration

$$x^{(k+1)} = D^{-1}\big(-Lx^{(k+1)} - Rx^{(k)} + b\big) \tag{5.20}$$

or

$$(D + L)\, x^{(k+1)} = -Rx^{(k)} + b$$

or

$$(D + L)\left(x^{(k+1)} - x^{(k)}\right) = -\left(Ax^{(k)} - b\right), \tag{5.21}$$

the so-called *method of successive displacements* or *Gauss–Seidel method*. According to (5.21) we have here a consistent iteration with

$$W = D + L\,.$$

Since D is nonsingular, W is nonsingular. Written in the form (5.6) the method is defined by

$$\begin{aligned} N &= W^{-1} = (D + L)^{-1}\,, \\ M_{\mathrm{GS}} &= I - NA = I - (D + L)^{-1} A = -(D + L)^{-1} R\,. \end{aligned}$$

In contrast to the Jacobi iteration, the Gauss–Seidel iteration depends on the order of the equations. However, the derivation (5.20) shows that the number of operations per iteration step is equal,

Jacobi becomes Gauss–Seidel,

if $x^{(k+1)}$ is stored on the same vector as $x^{(k)}$.

A sufficient convergence condition is given by the following theorem:

Theorem 5.2 *Jacobi's method and the Gauss–Seidel method converge globally and monotonically with respect to* $\|\cdot\|_\infty$ *if the* strict row sum criterion

$$\sum_{\substack{j=1 \\ j\neq i}}^{m} |a_{ij}| < |a_{ii}| \quad \textit{for all } i = 1, \ldots, m \tag{5.22}$$

is satisfied.

Proof : The proof here is given only for the Jacobi iteration. For the other method see, for example, [16].

The inequality (5.22) is equivalent to $\|M_{\mathrm{J}}\|_\infty < 1$ because of $M_{\mathrm{J}} = -D^{-1}(L + R)$ if $\|\cdot\|_\infty$ is the matrix norm that is induced by $\|\cdot\|_\infty$, which means the maximum-row-sum norm (see (A3.6)). □

It can be shown that the Gauss–Seidel method converges "better" than Jacobi's method, as expected: Under the assumption of (5.22) for the respective iteration matrices,

$$\|M_{\mathrm{GS}}\|_\infty \leq \|M_{\mathrm{J}}\|_\infty < 1$$

(see, for example, [16]).

Theorem 5.3 *If A is symmetric and positive definite, then the Gauss–Seidel method converges globally. The convergence is monotone in the energy norm $\|\cdot\|_A$, where $\|x\|_A := \left(x^T Ax\right)^{1/2}$ for $x \in \mathbb{R}^m$.*

Proof: See [16, p. 90]. □

If the differential operator, and therefore the bilinear form, is symmetric, that is, if (3.12) holds with $c = 0$, then Theorem 5.3 can be applied. Concerning the applicability of Theorem 5.2, even for the Poisson equation with Dirichlet boundary conditions (1.1), (1.2) requirements for the finite element discretization are necessary in order to satisfy at least a weaker version of (5.22). This example then satisfies the *weak* row sum criterion only in the following sense:

$$\begin{aligned} &\sum_{\substack{j=1 \\ j \neq i}}^{m} |a_{ij}| \leq |a_{ii}| \quad \text{for all } i = 1, \ldots, m\,; \\ &\text{“}<\text{”} \text{ holds for at least one } i \in \{1, \ldots, m\}\,. \end{aligned} \tag{5.23}$$

In the case of the finite difference method (1.7) for the rectangular domain or the finite element method from Section 2.2, which leads to the same discretization matrix, (5.23) is satisfied. For a general triangulation with linear ansatz functions, conditions for the angles of the elements must be required (see the angle condition in Section 3.9). The condition (5.23) is also sufficient, if A is irreducible (see Appendix A.3).

Theorem 5.4 *If A satisfies the condition* (5.23) *and is irreducible, then Jacobi's method converges globally.*

Proof: See [28, p. 111]. □

The qualitative statement of convergence does not say anything about the usefulness of Jacobi's and the Gauss–Seidel method for finite element discretizations. As an example we consider the Dirichlet problem for the Poisson equation on a rectangular domain as in (1.5), with the five-point stencil discretization introduced in Section 1.2. We restrict ourselves to an equal number of nodes in both space directions for simplificity of the notation. This number is denoted by $n + 1$, differently than in Chapter 1. Therefore, $A \in \mathbb{R}^{m,m}$ according to (1.14), with $m = (n-1)^2$ being the number of interior nodes. The factor h^{-2} can be omitted by multiplying the equation by h^2.

In the above example the eigenvalues and therefore the spectral radius can be calculated explicitly. Due to $D = 4I$ we have for Jacobi's method

$$M = -\frac{1}{4}(A - 4I) = I - \frac{1}{4}A\,,$$

and therefore A and M have the same eigenvectors, namely,

$$\left(z^{k,l}\right)_{ij} = \sin\frac{ik\pi}{n}\,\sin\frac{jl\pi}{n}\,, \quad 1 \le i,j,k,l \le n-1\,,$$

with the eigenvalues

$$2\left(2-\cos\frac{k\pi}{n}-\cos\frac{l\pi}{n}\right) \tag{5.24}$$

for A and

$$\frac{1}{2}\cos\frac{k\pi}{n}+\frac{1}{2}\cos\frac{l\pi}{n} \tag{5.25}$$

for M with $1 \le k,\, l \le n-1$. This can be proven directly with the help of trigonometric identities (see, for example, [15, p. 53]). Thus we have

$$\varrho(M) = -\cos\frac{(n-1)\pi}{n} = \cos\frac{\pi}{n} = 1-\frac{\pi^2}{2n^2}+O\left(n^{-4}\right)\,. \tag{5.26}$$

With growing n the rate of convergence becomes worse. The effort to gain an approximative solution, which means to reduce the error level below a given threshold ε, is proportional to the number of iterations $\times$ operations for an iteration, as we discussed at the beginning of this chapter. Due to (5.4) and (5.12) the number of necessary operations is calculated as follows:

$$\frac{\ln(1/\varepsilon)}{-\ln(\varrho(M))}\cdot O(m) = \ln\frac{1}{\varepsilon}\cdot O\left(n^2\right)\cdot O(m) = \ln\frac{1}{\varepsilon}\,O(m^2)\,.$$

Here the well-known expansion $\ln(1+x) = x + O(x^2)$ is employed in the determination of the leading term of $-1/(\ln(\varrho(M))$. An analogous result with better constants holds for the Gauss–Seidel method.

In comparison to this, the elimination or the Cholesky method requires

$$O\left(\text{band-width}^2\cdot m\right) = O(m^2)$$

operations; i.e., they are of the same complexity. Therefore, both methods are of use for only moderately large m.

An iterative method has a superior complexity to that of the Cholesky method if

$$\varrho(M) = 1 - O(n^{-\alpha}) \tag{5.27}$$

with $\alpha < 2$. In the ideal case (5.5) holds; then the method needs $O(m)$ operations, which is asymptotically optimal.

In the following we will present a sequence of methods with increasingly better convergence properties for systems of equations that arise from finite element discretizations.

The simplest iteration is the *Richardson method*, defined by

$$M = I - A\,, \quad \text{i.e.,} \quad N = W = I\,. \tag{5.28}$$

For this method we have

$$\varrho(M) = \max\{|1 - \lambda_{\max}(A)|, |1 - \lambda_{\min}(A)|\} ,$$

with $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ being the largest and smallest eigenvalues of A, respectively. Therefore, this method is convergent for special matrices only. In the case of a nonsingular D, the Richardson method for the transformed system of equations

$$D^{-1}Ax = D^{-1}b$$

is equivalent to Jacobi's method.

More generally, the following can be shown: If a consistent method is defined by M, N with $I = M + NA$, and N nonsingular, then it is equivalent to the Richardson method applied to

$$NAx = Nb . \tag{5.29}$$

The Richardson method for (5.29) has the form

$$x^{(k+1)} - x^{(k)} = -\tilde{N}\left(NAx^{(k)} - Nb\right)$$

with $\tilde{N} = I$, which means the form (5.9), and vice versa.

Equation (5.29) can also be interpreted as a *preconditioning* of the system of equations (5.1), with the aim to reduce the spectral condition number $\kappa(A)$ of the system matrix, since this is essential for the convergence behaviour. This will be further specified in the following considerations (5.33), (5.73). As already seen in the aforementioned examples, the matrix NA will not be constructed explicitly, since N is in general densely occupied, even if N^{-1} is sparse. The evaluation of $y = NAx$ therefore means solving the auxiliary system of equations

$$N^{-1}y = Ax .$$

Obviously, we have the following:

Lemma 5.5 *If the matrix A is symmetric and positive definite, then for the Richardson method all eigenvalues of M are real and smaller than* 1.

5.1.3 Relaxation

We continue to assume that A is symmetric and positive definite. Therefore, divergence of the procedure can be caused only by negative eigenvalues of $I - A$ less than or equal to -1. In general, bad or nonconvergent iterative methods can be improved in their convergence behaviour by *relaxation* if they meet certain conditions.

For an iteration method, given in the form (5.6), (5.7), the corresponding *relaxation method* with relaxation parameter $\omega > 0$ is defined by

$$x^{(k+1)} := \omega\big(Mx^{(k)} + Nb\big) + (1 - \omega)x^{(k)} , \tag{5.30}$$

which means

$$M_\omega \;:=\; \omega M + (1-\omega)I\,, \quad N_\omega := \omega N\,, \tag{5.31}$$

or if the condition of consistency $M = I - NA$ holds,

$$\begin{aligned} x^{(k+1)} &= \omega\left(x^{(k)} - N\left(Ax^{(k)} - b\right)\right) + (1-\omega)x^{(k)} \\ &= x^{(k)} - \omega N\left(Ax^{(k)} - b\right)\,. \end{aligned}$$

Let us assume for the procedure (5.6) that all eigenvalues of M are real. For the smallest one $\lambda_{\min}$ and the largest one $\lambda_{\max}$ we assume

$$\lambda_{\min} \le \lambda_{\max} < 1\,;$$

this is, for example, the case for the Richardson method. Then also the eigenvalues of M_ω are real, and we conclude that

$$\lambda_i(M_\omega) = \omega\lambda_i(M) + 1 - \omega = 1 - \omega\big(1 - \lambda_i(M)\big)$$

if the $\lambda_i(B)$ are the eigenvalues of B in an arbitrary ordering. Hence

$$\varrho(M_\omega) = \max\left\{\,|1 - \omega\,(1 - \lambda_{\min}(M))|\,, |1 - \omega\,(1 - \lambda_{\max}(M))|\,\right\},$$

since $f(\lambda) := 1 - \omega(1-\lambda)$ is a straight line for a fixed ω (with $f(1) = 1$ and $f(0) = 1 - \omega$).

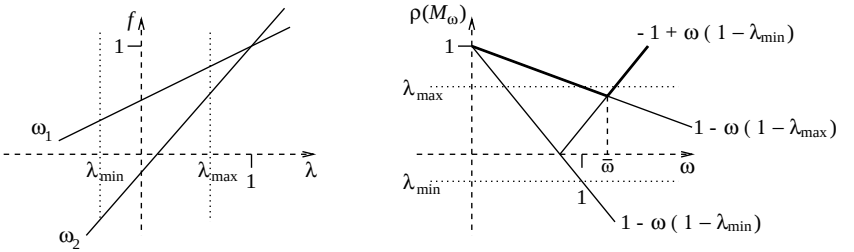


Figure 5.1. Calculation of $\bar\omega$.

For the *optimal* $\bar\omega$, i.e., $\bar\omega$ with

$$\varrho(M_{\bar\omega}) = \min_{\omega>0} \varrho(M_\omega)\,,$$

we therefore have, as can be seen from Figure 5.1,

$$\begin{aligned} & 1 - \bar\omega\,(1 - \lambda_{\max}(M)) = -1 + \bar\omega\,(1 - \lambda_{\min}(M)) \\ \iff \qquad & \bar\omega = \frac{2}{2 - \lambda_{\max}(M) - \lambda_{\min}(M)}\,. \end{aligned}$$

Hence $\bar\omega > 0$ and

$$\varrho(M_{\bar\omega}) = 1 - \bar\omega(1 - \lambda_{\max}(M)) < 1\,;$$

consequently, the method converges with optimal ω even in cases where it would not converge for $\omega = 1$. But keep in mind that one needs the eigenvalues of M to determine $\bar{\omega}$.

Moreover, we have

$$\bar{\omega} < 1 \quad \Leftrightarrow \quad \lambda_{\max}(M) + \lambda_{\min}(M) < 0\,.$$

If $\lambda_{\min}(M) \neq -\lambda_{\max}(M)$, that is, $\bar{\omega} \neq 1$, we will achieve an improvement by relaxation:

$$\varrho(M_{\bar{\omega}}) < \varrho(M)\,.$$

The case of $\omega < 1$ is called *underrelaxation*, whereas in the case of $\omega > 1$ we speak of an *overerrelaxation*.

In particular, for the Richardson method with the iteration matrix $M = I - A$, due to $\lambda_{\min}(M) = 1 - \lambda_{\max}(A)$ and $\lambda_{\max}(M) = 1 - \lambda_{\min}(A)$, the optimal $\bar{\omega}$ is given by

$$\bar{\omega} = \frac{2}{\lambda_{\min}(A) + \lambda_{\max}(A)}\,. \tag{5.32}$$

Hence

$$\varrho(M_{\bar{\omega}}) = 1 - \bar{\omega}\lambda_{\min}(A) = \frac{\lambda_{\max}(A) - \lambda_{\min}(A)}{\lambda_{\min}(A) + \lambda_{\max}(A)} = \frac{\kappa(A) - 1}{\kappa(A) + 1} < 1\,, \tag{5.33}$$

with the spectral *condition number* of A

$$\kappa(A) := \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$$

(see Appendix A.3).

For large $\kappa(A)$ we have

$$\varrho(M_{\bar{\omega}}) = \frac{\kappa(A) - 1}{\kappa(A) + 1} \approx 1 - \frac{2}{\kappa(A)}\,,$$

the variable of the proportionality being $\kappa(A)$. For the example of the five-point stencil discretization, due to (5.24),

$$\lambda_{\min}(A) + \lambda_{\max}(A) = 4\left(2 - \cos\frac{n-1}{n}\pi - \cos\frac{\pi}{n}\right) = 8\,,$$

and thus due to (5.32),

$$\bar{\omega} = \frac{1}{4}\,.$$

Hence the iteration matrix $M_{\bar{\omega}} = I - \frac{1}{4}A$ is identical to the Jacobi iteration: We have rediscovered Jacobi's method.

By means of (5.33) we can estimate the contraction number, since we know from (5.24) that

$$\kappa(A) = \frac{4\left(1 - \cos\frac{n-1}{n}\,\pi\right)}{4\left(1 - \cos\frac{\pi}{n}\right)} = \frac{1 + \cos\frac{\pi}{n}}{1 - \cos\frac{\pi}{n}} \approx \frac{4n^2}{\pi^2}\,. \tag{5.34}$$

This shows the stringency of Theorem 3.45, and again we can conclude that

$$\varrho(M_{\bar{\omega}}) = \cos\frac{\pi}{n} \approx 1 - \frac{\pi^2}{2n^2}\,. \tag{5.35}$$

Due to Theorem 3.45 the convergence behaviour seen for the model problem is also valid in general for quasi-uniform triangulations.

5.1.4 SOR and Block-Iteration Methods

We assume again that A is a general nonsingular matrix. For the relaxation of the Gauss–Seidel method we use it in the form

$$Dx^{(k+1)} = -Lx^{(k+1)} - Rx^{(k)} + b\,,$$

instead of the resolved form (5.20).

The relaxed method is then

$$Dx^{(k+1)} = \omega\big(-Lx^{(k+1)} - Rx^{(k)} + b\big) + (1-\omega)Dx^{(k)} \tag{5.36}$$

with a relaxation parameter $\omega > 0$. This is equivalent to

$$(D+\omega L)\,x^{(k+1)} = (-\omega R + (1-\omega)D)\,x^{(k)} + \omega b\,. \tag{5.37}$$

Hence

$$\begin{aligned} M_\omega &:= (D+\omega L)^{-1}\,(-\omega R + (1-\omega)D)\,, \\ N_\omega &:= (D+\omega L)^{-1}\,\omega\,. \end{aligned}$$

In the application to discretizations of boundary value problems, normally we choose $\omega > 1$, which means overrelaxation. This explains the name of the *SOR method* as an abbreviation of *successive overrelaxation*. The effort to execute an iteration step is hardly higher than for the Gauss–Seidel method. Although we have to add $3m$ operations to the evaluation of the right side of (5.36), the forward substitution to solve the auxiliary system of equations in (5.37) is already part of the form (5.36).

The calculation of the optimal $\bar{\omega}$ here is more difficult, because M_ω depends nonlinearly on ω. Only for special classes of matrices can the optimal $\bar{\omega}$ minimizing $\varrho(M_\omega)$ be calculated explicitly in dependence on $\varrho(M_1)$, the convergence rate of the (nonrelaxed) Gauss–Seidel method. Before we sketch this, we will look at some further variants of this procedure:

The matrix N_ω is nonsymmetric even for symmetric A. One gets a symmetric N_ω if after one SOR step another one is performed in which the indices are run through in reverse order $m, m-1, \ldots, 2, 1$, which means that L and R are exchanged. The two half steps

$$\begin{aligned} Dx^{(k+\frac{1}{2})} &= \omega\big(-Lx^{(k+\frac{1}{2})} - Rx^{(k)} + b\big) + (1-\omega)Dx^{(k)}\,, \\ Dx^{(k+1)} &= \omega\big(-Lx^{(k+\frac{1}{2})} - Rx^{(k+1)} + b\big) + (1-\omega)Dx^{(k+\frac{1}{2})}\,, \end{aligned}$$

make up one step of the *symmetric SOR*, the *SSOR method* for short. A special case is the *symmetric Gauss–Seidel method* for $\omega = 1$.

We write down the procedure for symmetric A, i.e., $R = L^T$ in the form (5.6), in which the symmetry of N becomes obvious:

$$\begin{aligned} M &= \left(D + \omega L^T\right)^{-1} \left[(1-\omega)D - \omega L\right] (D + \omega L)^{-1} \left[(1-\omega)D - \omega L^T\right] , \\ N &= \omega(2-\omega) \left(D + \omega L^T\right)^{-1} D \, (D + \omega L)^{-1} . \end{aligned} \tag{5.38}$$

The effort for SSOR is only slightly higher than for SOR if the vectors already calculated in the half steps are stored and used again, as for example $Lx^{(k+1/2)}$.

Other variants of these procedures are created if the procedures are not applied to the matrix itself but to a block partitioning

$$A = (A_{ij})_{i,j} \quad \text{with } A_{ij} \in \mathbb{R}^{m_i, m_j} , \quad i, j = 1, \ldots, p , \tag{5.39}$$

with $\sum_{i=1}^{p} m_i = m$. As an example we get the *block-Jacobi method*, which is analogous to (5.19) and has the form

$$\xi_i^{(k+1)} = A_{ii}^{-1} \left(- \sum_{j=1}^{i-1} A_{ij} \xi_j^{(k)} - \sum_{j=i+1}^{p} A_{ij} \xi_j^{(k)} + \beta_i \right) \text{ for all } i = 1, \ldots, p . \tag{5.40}$$

Here $x = (\xi_1, \ldots, \xi_p)^T$ and $b = (\beta_1, \ldots, \beta_p)^T$, respectively, are corresponding partitions of the vectors. By exchanging $\xi_j^{(k)}$ with $\xi_j^{(k+1)}$ in the first sum one gets the *block-Gauss–Seidel method* and then in the same way the relaxed variants. The iteration (5.40) includes p vector equations. For each of them we have to solve a system of equations with system matrix A_{ii}. To get an advantage compared to the pointwise method a much lower effort should be necessary than for the solution of the total system. This can require — if at all possible — a rearranging of the variables and equations. The necessary permutations will not be noted explicitly here. Such methods are applied in finite difference methods or other methods with structured grids (see Section 4.1) if an ordering of nodes is possible such that the matrices A_{ii} are diagonal or tridiagonal and therefore the systems of equations are solvable with $O(m_i)$ operations.

As an example we again discuss the five-point stencil discretization of the Poisson equation on a square with $n + 1$ nodes per space dimension. The matrix A then has the form (1.14) with $l = m = n$. If the nodes are numbered rowwise and we choose one block for each line, which means $p = n - 1$ and $m_i = n - 1$ for all $i = 1, \ldots, p$, then the matrices A_{ii} are tridiagonal. On the other hand, if one chooses a partition of the indices of the nodes in subsets S_i such that a node with index in S_i has neighbours only in other index sets, then for such a selection and arbitrary ordering within the index sets the matrices A_{ii} become diagonal. *Neighbours* here denote the nodes within a difference stencil or more generally, those nodes

$$\begin{pmatrix}
4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\
-1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\
0 & -1 & 4 & 0 & 0 & -1 & 0 & 0 & 0 \\
-1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 & 0 \\
0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 \\
0 & 0 & -1 & 0 & -1 & 4 & 0 & 0 & -1 \\
0 & 0 & 0 & -1 & 0 & 0 & 4 & -1 & 0 \\
0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1 \\
0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4
\end{pmatrix}$$

$m = 3 \times 3$: rowwise ordering.

$$\left(\begin{array}{ccccc|cccc}
4 & 0 & 0 & 0 & 0 & -1 & -1 & 0 & 0 \\
0 & 4 & 0 & 0 & 0 & -1 & 0 & -1 & 0 \\
0 & 0 & 4 & 0 & 0 & -1 & -1 & -1 & -1 \\
0 & 0 & 0 & 4 & 0 & 0 & -1 & 0 & -1 \\
0 & 0 & 0 & 0 & 4 & 0 & 0 & -1 & -1 \\
\hline
-1 & -1 & -1 & 0 & 0 & 4 & 0 & 0 & 0 \\
-1 & 0 & -1 & -1 & 0 & 0 & 4 & 0 & 0 \\
0 & -1 & -1 & 0 & -1 & 0 & 0 & 4 & 0 \\
0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 & 4
\end{array}\right)$$

red-black ordering:
red: node 1, 3, 5, 7, 9 from rowwise ordering
black: node 2, 4, 6, 8 from rowwise ordering

Figure 5.2. Comparison of orderings.

that contribute to the corresponding row of the discretization matrix. In the example of the five-point stencil, starting with rowwise numbering, one can combine all odd indices to a block S_1 (the "red nodes") and all even indices to a block S_2 (the "black" nodes). Here we have $p = 2$. We call this a *red-black ordering* (see Figure 5.2). If two "colours" are not sufficient, one can choose $p > 2$.

We return to the SOR method and its convergence: In the following the iteration matrix will be denoted by $M_{\mathrm{SOR}(\omega)}$ with the relaxation parameter ω. Likewise, M_{J} and M_{GS} are the iteration matrices of Jacobi's and the Gauss–Seidel method, respectively. General propositions are summarized in the following theorem:

Theorem 5.6 (of Kahan; Ostrowski and Reich)

(1) $\varrho\left(M_{\mathrm{SOR}(\omega)}\right) \geq |1-\omega|$ *for* $\omega \neq 0$.

(2) *If* A *is symmetric and positive definite, then*

$$\varrho\left(M_{\mathrm{SOR}(\omega)}\right) < 1 \quad \text{for } \omega \in (0,2)\,.$$

Proof: See [16, pp. 91 f.]. □

Therefore, we use only $\omega \in (0,2)$. For a useful procedure we need more information about the optimal relaxation parameter ω_{opt}, given by

$$\varrho\left(M_{\mathrm{SOR}(\omega_{\mathrm{opt}})}\right) = \min_{0<\omega<2} \varrho\left(M_{\mathrm{SOR}(\omega)}\right)\,,$$

and about the size of the contraction number. This is possible only if the ordering of equations and unknowns has certain properties:

Definition 5.7 A matrix $A \in \mathbb{R}^{m,m}$ is *consistently ordered* if for the partition (5.18), D is nonsingular and

$$C(\alpha) := \alpha^{-1} D^{-1} L + \alpha D^{-1} R$$

has eigenvalues independent of α for $\alpha \in \mathbb{C}\backslash\{0\}$.

There is a connection to the possibility of a multi-colour ordering, because a matrix in the block form (5.39) is consistently ordered if it is block-tridiagonal (i.e., $A_{ij} = 0$ for $|i-j| > 1$) and the diagonal blocks A_{ii} are nonsingular diagonal matrices (see [28, pp. 114 f.]).

In the case of a consistently ordered matrix one can prove a relation between the eigenvalues of M_{J}, M_{GS}, and $M_{\mathrm{SOR}(\omega)}$. From this we can see how much faster the Gauss–Seidel method converges than Jacobi's method:

Theorem 5.8 *If* A *is consistently ordered, then*

$$\varrho(M_{\mathrm{J}})^2 = \varrho(M_{\mathrm{GS}})\,.$$

Proof: For a special case see Remark 5.5.2 in [16]. □

Due to (5.4) we can expect a halving of the number of iteration steps, but this does not change the asymptotic statement (5.27).

Finally, in the case that Jacobi's method converges the following theorem holds:

Theorem 5.9 *Let* A *be consistently ordered with nonsingular diagonal matrix* D*, the eigenvalues of* M_{J} *being real and* $\beta := \varrho(M_{\mathrm{J}}) < 1$*. Then we have for the SOR method:*

(1) $\omega_{\mathrm{opt}} = \dfrac{2}{1+(1-\beta^2)^{1/2}}\,,$

$$(2)\quad \varrho(M_{\mathrm{SOR}(\omega)}) = \begin{cases} 1-\omega+\dfrac{1}{2}\omega^2\beta^2+\omega\beta\left(1-\omega+\dfrac{\omega^2\beta^2}{4}\right)^{1/2} \\ \qquad\qquad\qquad\qquad \textit{for } 0<\omega<\omega_{\mathrm{opt}} \\ \omega-1 \qquad\qquad\qquad\qquad \textit{for } \omega_{\mathrm{opt}}\le\omega<2\,, \end{cases}$$

$$(3)\quad \varrho\left(M_{\mathrm{SOR}(\omega_{\mathrm{opt}})}\right) = \frac{\beta^2}{(1+(1-\beta^2)^{1/2})^2}\,.$$

Proof: See [18, p. 216]. □

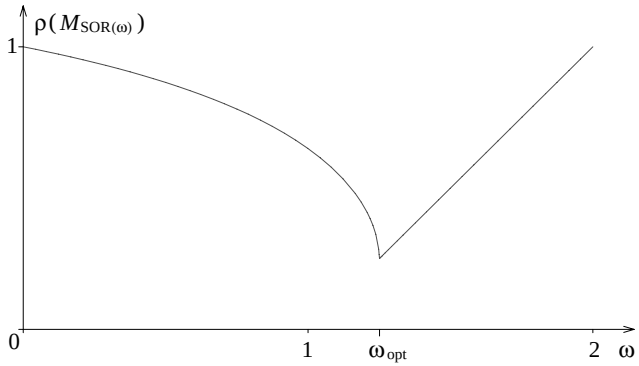


Figure 5.3. Dependence of $\varrho\left(M_{\mathrm{SOR}(\omega)}\right)$ on ω.

If $\varrho(M_{\mathrm{J}})$ is known for Jacobi's method, then ω_{opt} can be calculated. This is the case in the example of the five-point stencil discretization on a square: From (5.26) and Theorem 5.9 it follows that

$$\varrho(M_{\mathrm{GS}}) = \left(\cos\frac{\pi}{n}\right)^2 = 1-\frac{\pi^2}{n^2}+O(n^{-4})\,;$$

hence

$$\begin{aligned} \omega_{\mathrm{opt}} &= 2/\left(1+\sin\tfrac{\pi}{n}\right)\,, \\ \varrho\left(M_{\mathrm{SOR}(\omega_{\mathrm{opt}})}\right) &= \omega_{\mathrm{opt}}-1 = 1-2\tfrac{\pi}{n}+O(n^{-2})\,. \end{aligned}$$

Therefore, the optimal SOR method has a lower complexity than all methods described up to now.

Correspondingly, the number of operations to reach the relative error level $\varepsilon > 0$ is reduced to $\ln\frac{1}{\varepsilon}\,O(m^{3/2})$ operations in comparison to $\ln\frac{1}{\varepsilon}\,O(m^2)$ operations for the previous procedures.

Table 5.1 gives an impression of the convergence for the model problem. It displays the theoretically to be expected values for the numbers of iterations of the Gauss–Seidel method (m_{GS}), as well as for the SOR method

n	m_{GS}	m_{SOR}
8	43	8
16	178	17
32	715	35
64	2865	70
128	11466	140
256	45867	281

Table 5.1. Gauss–Seidel and optimal SOR method for the model problem.

with optimal relaxation parameter (m_{SOR}). Here we use the very moderate termination criterion $\varepsilon = 10^{-3}$ measured in the Euclidean norm.

The optimal SOR method is superior, even if we take into account the almost doubled effort per iteration step. But generally, ω_{opt} is not known explicitly. Figure 5.3 shows that it is probably better to overestimate ω_{opt} instead of underestimating. More generally, one can try to improve the relaxation parameter during the iteration:

If $\varrho(M_{\mathrm{J}})$ is a simple eigenvalue, then this also holds true for the spectral radius $\varrho(M_{\mathrm{SOR}(\omega)})$. The spectral radius can thus be approximated by the power method on the basis of the iterates. By Theorem 5.9 (3) one can approximate $\varrho(M_{\mathrm{J}})$, and by Theorem 5.9 (1) then also ω_{opt}.

This basic principle can be extended to an algorithm (see, for example, [18, Section 9.5]), but the upcoming overall procedure is no longer a linear stationary method.

5.1.5 Extrapolation Methods

Another possibility for an extension of the linear stationary methods, related to the adaption of the relaxation parameter, is the following: Starting with a linear stationary basic iteration $\tilde{x}^{k+1} := \Phi\left(\tilde{x}^k\right)$ we define a new iteration by

$$x^{(k+1)} := \omega_k \Phi\left(x^{(k)}\right) + (1 - \omega_k) x^{(k)} , \tag{5.41}$$

with *extrapolation factors* ω_k to be chosen. A generalization of this definition is to start with the iterates of the basic iteration $\tilde{x}^{(0)}, \tilde{x}^{(1)}, \ldots$. The iterates of the new method are to be determined by

$$x^{(k)} := \sum_{j=0}^{k} \alpha_{k_j} \tilde{x}^{(j)} ,$$

with α_{k_j} defined by a polynomial $p_k \in \mathcal{P}_k$, with the property $p_k(t) = \sum_{j=0}^{k} \alpha_{k_j} t^j$ and $p_k(1) = 1$. For an appropriate definition of such *extrapolation* or *semi-iterative methods* we need to know the spectrum of the basic iteration matrix M, since the error $e^{(k)} = x^{(k)} - x$ satisfies

$$e^{(k)} = p_k(M) e^{(0)} ,$$

where M is the iteration matrix of the basic iteration. This matrix should be normal, for example, such that

$$\|p_k(M)\|_2 = \varrho(p_k(M))$$

holds. Then we have the obvious estimation

$$\left|e^{(k)}\right|_2 \leq \left|p_k(M)e^{(0)}\right|_2 \leq \left\|p_k(M)\right\|_2 \left|e^{(0)}\right|_2 \leq \varrho(p_k(M))\left|e^{(0)}\right|_2 . \qquad (5.42)$$

If the method is to be defined in such a way that

$$\varrho(p_k(M)) = \max\left\{|p_k(\lambda)| \;\middle|\; \lambda \in \sigma(M)\right\}$$

is minimized by choosing p_k, then the knowledge of the spectrum $\sigma(M)$ is necessary. Generally, instead of this, we assume that suitable supersets are known: If $\sigma(M)$ is real and

$$a \leq \lambda \leq b \quad \text{for all } \lambda \in \sigma(M) ,$$

then, due to

$$\left|e^{(k)}\right|_2 \leq \max_{\lambda\in[a,b]} |p_k(\lambda)| \left|e^{(0)}\right|_2 ,$$

it makes sense to determine the polynomials p_k as a solution of the minimization problem on $[a, b]$,

$$\max_{\lambda\in[a,b]} |p_k(\lambda)| \to \min \quad \text{for all} \quad p \in \mathcal{P}_k \quad \text{with } p(1) = 1 . \qquad (5.43)$$

In the following sections we will introduce methods with an analogous convergence behaviour, without control parameters necessary for their definition.

For further information on semi-iterative methods see, for example, [16, Chapter 7].

Exercises

5.1 Investigate Jacobi's method and the Gauss–Seidel method for solving the linear system of equations $Ax = b$ with respect to their convergence if we have the following system matrices:

$$\text{(a)} \quad A = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix} , \qquad \text{(b)} \quad A = \frac{1}{2}\begin{pmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{pmatrix} .$$

5.2 Prove the consistency of the SOR method.

5.3 Prove Theorem 5.6, (1).

5.2 Gradient and Conjugate Gradient Methods

In this section let $A \in \mathbb{R}^{m,m}$ be symmetric and positive definite. Then the system of equations $Ax = b$ is equivalent to the problem

$$\text{Minimize} \quad f(x) := \frac{1}{2}x^T Ax - b^T x \quad \text{for } x \in \mathbb{R}^m\,, \tag{5.44}$$

since for such a functional the minima and stationary points coincide, where a *stationary point* is an x satisfying

$$0 = \nabla f(x) = Ax - b\,. \tag{5.45}$$

In contrast to the notation $x \cdot y$ for the "short" space vectors $x, y \in \mathbb{R}^d$ we write here the Euclidean scalar product as matrix product $x^T y$.

For the finite element discretization this corresponds to the equivalence of the Galerkin method (2.23) with the Ritz method (2.24) if A is the stiffness matrix and b the load vector (see (2.34) and (2.35)). More generally, Lemma 2.3 implies the equivalence of (5.44) and (5.45), if as bilinear form the so-called *energy scalar product*

$$\langle x, y\rangle_A := x^T Ay \tag{5.46}$$

is chosen.

A general iterative method to solve (5.44) has the following structure:

Define a search direction $d^{(k)}$.

$$\text{Minimize} \quad \alpha \mapsto \tilde{f}(\alpha) := f\big(x^{(k)} + \alpha d^{(k)}\big) \tag{5.47}$$

exactly or approximately, with the solution α_k.

$$\text{Define} \qquad x^{(k+1)} := x^{(k)} + \alpha_k d^{(k)}\,. \tag{5.48}$$

If f is defined as in (5.44), the exact α_k can be computed from the condition $\tilde{f}'(\alpha) = 0$ and

$$\tilde{f}'(\alpha) = \nabla f\big(x^{(k)} + \alpha d^{(k)}\big)^T d^{(k)}$$

as

$$\alpha_k = -\frac{{g^{(k)}}^T d^{(k)}}{{d^{(k)}}^T A d^{(k)}}\,, \tag{5.49}$$

where

$$g^{(k)} := Ax^{(k)} - b = \nabla f\big(x^{(k)}\big)\,. \tag{5.50}$$

The error of the kth iterate is denoted by $e^{(k)}$:

$$e^{(k)} := x^{(k)} - x\,.$$

Some relations that are valid in this general fromework are the following: Due to the one-dimensional minimization of f, we have

$${g^{(k+1)}}^T d^{(k)} = 0\,, \tag{5.51}$$

and from (5.50) we can conclude immediately that

$$Ae^{(k)} = g^{(k)}, \qquad e^{(k+1)} = e^{(k)} + \alpha_k d^{(k)}, \tag{5.52}$$

$$g^{(k+1)} = g^{(k)} + \alpha_k A d^{(k)}. \tag{5.53}$$

We consider the *energy norm*

$$\|x\|_A := \left(x^T A x\right)^{1/2} \tag{5.54}$$

induced by the energy scalar product. For a finite element stiffness matrix A with a bilinear form a we have the correspondence

$$\|x\|_A = a(u,u)^{1/2} = \|u\|_a$$

for $u = \sum_{i=1}^m x_i \varphi_i$ if the φ_i are the underlying basis functions. Comparing the solution $x = A^{-1}b$ with an arbitrary $y \in \mathbb{R}^m$ leads to

$$f(y) = f(x) + \frac{1}{2}\|y - x\|_A^2 \,, \tag{5.55}$$

so that condition (5.44) also minimizes the distance to x in $\|\cdot\|_A$. The energy norm will therefore have a special importance. Measured in the energy norm we have, due to (5.52),

$$\left\|e^{(k)}\right\|_A^2 = {e^{(k)}}^T g^{(k)} = {g^{(k)}}^T A^{-1} g^{(k)} \,,$$

and therefore due to (5.52) and (5.51),

$$\left\|e^{(k+1)}\right\|_A^2 = {g^{(k+1)}}^T e^{(k)} \,.$$

The vector $-\nabla f\left(x^{(k)}\right)$ in $x^{(k)}$ points in the direction of the locally steepest descent, which motivates the *gradient method*, i.e.,

$$d^{(k)} := -g^{(k)} \,, \tag{5.56}$$

and thus

$$\alpha_k = \frac{{d^{(k)}}^T d^{(k)}}{{d^{(k)}}^T A d^{(k)}} \,. \tag{5.57}$$

The above identities imply for the gradient method

$$\left\|e^{(k+1)}\right\|^2 = \left(g^{(k)} + \alpha_k A d^{(k)}\right)^T e^{(k)} = \|e^{(k)}\|_A^2 \left(1 - \alpha_k \frac{{d^{(k)}}^T d^{(k)}}{{d^{(k)}}^T A^{-1} d^{(k)}}\right)$$

and thus by means of the definition of α_k from (5.57)

$$\left\|x^{(k+1)} - x\right\|_A^2 = \left\|x^{(k)} - x\right\|_A^2 \left\{1 - \frac{\left({d^{(k)}}^T d^{(k)}\right)^2}{{d^{(k)}}^T A d^{(k)}\; {d^{(k)}}^T A^{-1} d^{(k)}}\right\}.$$

With the *inequality of Kantorovich* (see, for example, [28, p. 132]),

$$\frac{x^T A x \, x^T A^{-1} x}{\left(x^T x\right)^2} \le \left(\frac{1}{2}\kappa^{1/2} + \frac{1}{2}\kappa^{-1/2}\right)^2 \,,$$

where $\kappa := \kappa(A)$ is the spectral condition number, and the relation

$$1 - \frac{4}{\left(a^{1/2} + a^{-1/2}\right)^2} = \frac{(a-1)^2}{(a+1)^2} \quad \text{for } a > 0\,,$$

we obtain the following theorem:

Theorem 5.10 *For the gradient method we have*

$$\left\|x^{(k)} - x\right\|_A \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \left\|x^{(0)} - x\right\|_A . \tag{5.58}$$

This is the same estimate as for the optimally relaxed Richardson method (with the sharper estimate $\|M\|_A \leq \frac{\kappa-1}{\kappa+1}$ instead of $\varrho(M) \leq \frac{\kappa-1}{\kappa+1}$). The essential difference lies in the fact that this is possible without knowledge of the spectrum of A.

Nevertheless, for finite element discretizations we have the same poor convergence rate as for Jacobi's or similar methods. The reason for this deficiency lies in the fact that due to (5.51), we have ${g^{(k+1)}}^T g^{(k)} = 0$, but in general not ${g^{(k+2)}}^T g^{(k)} = 0$. On the contrary, these search directions are very often almost parallel, as can be seen from Figure 5.4.

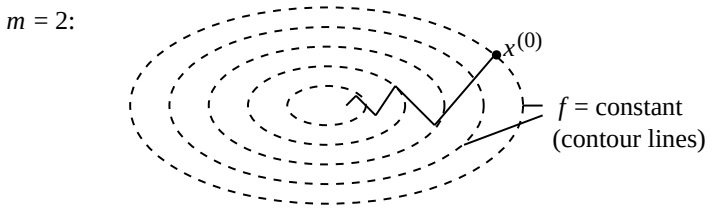


Figure 5.4. Zigzag behaviour of the gradient method.

The reason for this problem is the fact that for large κ the search directions $g^{(k)}$ and $g^{(k+1)}$ can be almost parallel with respect to the scalar products $\langle \cdot, \cdot \rangle_A$ (see Exercise 5.4), but with respect to $\| \cdot \|_A$ the distance to the solution will be minimized (see (5.55)).

The search directions $d^{(k)}$ should be orthogonal with respect to $\langle \cdot, \cdot \rangle_A$, which we call *conjugate*.

Definition 5.11 Vectors $d^{(0)}, \ldots, d^{(l)} \in \mathbb{R}^m$ are *conjugate* if they satisfy

$$\left\langle d^{(i)}, d^{(j)} \right\rangle_A = 0 \quad \text{for } i, j = 0, \ldots, l\,,\ i \neq j\,.$$

If the search directions of a method defined according to (5.48), (5.49) are chosen as conjugate, it is called a *method of conjugate directions*.

Let $d^{(0)}, \ldots, d^{(m-1)}$ be conjugate directions. Then they are also linearly independent and thus form a basis in which the solution x of (5.1) can be

represented, say by the coefficients γ_k:

$$x = \sum_{k=0}^{m-1} \gamma_k d^{(k)} .$$

Since the $d^{(k)}$ are conjugate and $Ax = b$ holds, we have

$$\gamma_k = \frac{{d^{(k)}}^T b}{{d^{(k)}}^T A d^{(k)}} , \tag{5.59}$$

and the γ_k can be calculated without knowledge of x. If the $d^{(k)}$ would by given a priori, for example by orthogonalization of a basis with respect to $\langle \cdot, \cdot \rangle_A$, then x would be determined by (5.59).

If we apply (5.59) to determine the coefficients for $x - x^{(0)}$ in the form

$$x - x^{(0)} = \sum_{k=0}^{m-1} \gamma_k d^{(k)} ,$$

which means replacing b with $b - Ax^{(0)}$ in (5.59), then we get

$$\gamma_k = -\frac{{g^{(0)}}^T d^{(k)}}{{d^{(k)}}^T A d^{(k)}} .$$

For the kth iterate we have, according to (5.48);

$$x^{(k)} = x^{(0)} + \sum_{i=0}^{k-1} \alpha_i d^{(i)}$$

and therefore (see (5.50))

$$g^{(k)} = g^{(0)} + \sum_{i=0}^{k-1} \alpha_i A d^{(i)} .$$

For a method of conjugate directions this implies

$${g^{(k)}}^T d^{(k)} = {g^{(0)}}^T d^{(k)}$$

and therefore

$$\gamma_k = -\frac{{g^{(k)}}^T d^{(k)}}{{d^{(k)}}^T A d^{(k)}} = \alpha_k ,$$

which means that $x = x^{(m)}$. A method of conjugate directions therefore is exact after at most m steps. Under certain conditions such a method may terminate before reaching this step number with $g^{(k)} = 0$ and the final iterate $x^{(k)} = x$. If m is very large, this exactness of a method of conjugate directions is less important than the fact that the iterates can be interpreted as the solution of a minimization problem approximating (5.44):

Theorem 5.12 *The iterates $x^{(k)}$ that are determined by a method of conjugate directions minimize the functional f from* (5.44) *as well as the error $\left\|x^{(k)} - x\right\|_A$ on $x^{(0)} + K_k(A; g^{(0)})$, where*

$$K_k(A; g^{(0)}) := \operatorname{span}\left\{d^{(0)}, \ldots, d^{(k-1)}\right\}.$$

This is due to

$$g^{(k)^T} d^{(i)} = 0 \quad \text{for } i = 0, \ldots, k-1\,. \tag{5.60}$$

Proof: It is sufficent to prove (5.60). Due to the one-dimensional minimization this holds for $k = 1$ and for $i = k-1$ (see (5.51) applied to $k-1$). To conclude the assertion for k from its knowledge for $k-1$, we note that (5.53) implies, for $0 \le i < k-1$,

$$d^{(i)^T}\left(g^{(k)} - g^{(k-1)}\right) = \alpha_{k-1} d^{(i)^T} A d^{(k-1)} = 0\,.$$

□

In the *method of conjugate gradients*, or *CG method*, the $d^{(k)}$ are determined during the iteration by the ansatz

$$d^{(k+1)} := -g^{(k+1)} + \beta_k d^{(k)}\,. \tag{5.61}$$

Then we have to clarify whether

$$\left\langle d^{(k)}, d^{(i)} \right\rangle_A = 0 \quad \text{for } k > i$$

can be obtained. The necessary requirement $\left\langle d^{(k+1)}, d^{(k)} \right\rangle_A = 0$ leads to

$$\begin{aligned} -\left\langle g^{(k+1)}, d^{(k)} \right\rangle_A + \beta_k \left\langle d^{(k)}, d^{(k)} \right\rangle_A = 0 \qquad &\Longleftrightarrow \\ \beta_k = \frac{g^{(k+1)^T} A d^{(k)}}{d^{(k)^T} A d^{(k)}}\,. \end{aligned} \tag{5.62}$$

In applying the method it is recommended not to calculate $g^{(k+1)}$ directly but to use (5.53) instead, because $Ad^{(k)}$ is already necessary to determine α_k and β_k.

The following equivalences hold:

Theorem 5.13 *In case the CG method does not terminate prematurely with $x^{(k-1)}$ being the solution of* (5.1)*, then we have for $1 \le k \le m$*

$$\begin{aligned} K_k(A; g^{(0)}) &= \operatorname{span}\left\{g^{(0)}, Ag^{(0)}, \ldots, A^{k-1} g^{(0)}\right\} \\ &= \operatorname{span}\left\{g^{(0)}, \ldots, g^{(k-1)}\right\}. \end{aligned} \tag{5.63}$$

Furthermore,

$$\begin{aligned} g^{(k)^T} g^{(i)} &= 0 \quad \text{for} \quad i = 0, \ldots, k-1, \text{ and} \\ \dim K_k(A; g^{(0)}) &= k\,. \end{aligned} \tag{5.64}$$

The space $K_k(A; g^{(0)}) = \text{span}\{g^{(0)}, Ag^{(0)}, \ldots, A^{k-1}g^{(0)}\}$ is called the *Krylov (sub)space* of dimension k of A with respect to $g^{(0)}$.

Proof: The identities (5.64) are immediate consequences of (5.63) and Theorem 5.12. The proof of (5.63) is given by induction:
For $k = 1$ the assertion is trivial. Let us assume that for $k \geq 1$ the identity (5.63) holds and therefore also (5.64) does. Due to (5.53) (applied to $k-1$) it follows that

$$g^{(k)} \in A\big[K_k\big(A; g^{(0)}\big)\big] \subset \text{span}\,\{g^{(0)}, \ldots, A^k g^{(0)}\}$$

and thus

$$\text{span}\,\{g^{(0)}, \ldots, g^{(k)}\} = \text{span}\,\{g^{(0)}, \ldots, A^k g^{(0)}\},$$

because the left space is contained in the right one and the dimension of the left subspace is maximal ($= k+1$) due to (5.64) and $g^{(i)} \neq 0$ for all $i = 0, \ldots, k$. The identity

$$\text{span}\,\{d^{(0)}, \ldots, d^{(k)}\} = \text{span}\,\{g^{(0)}, \ldots, g^{(k)}\}$$

follows from the induction hypothesis and (5.61). □

The number of operations per iteration can be reduced to one matrix vector, two scalar products, and three SAXPY operations, if the following equivalent terms are used:

$$\alpha_k = \frac{{g^{(k)}}^T g^{(k)}}{{d^{(k)}}^T A d^{(k)}}, \quad \beta_k = \frac{{g^{(k+1)}}^T g^{(k+1)}}{{g^{(k)}}^T g^{(k)}}. \tag{5.65}$$

Here a SAXPY operation is of the form

$$z := x + \alpha y$$

for vectors x, y, z and a scalar α.

The identities (5.65) can be seen as follows: Concerning α_k we note that because of (5.51) and (5.61),

$$-{g^{(k)}}^T d^{(k)} = -{g^{(k)}}^T\big(-g^{(k)} + \beta_{k-1} d^{(k-1)}\big) = {g^{(k)}}^T g^{(k)},$$

and concerning β_k, because of (5.53), (5.64), (5.62), and the identity (5.49) for α_k, we have

$${g^{(k+1)}}^T g^{(k+1)} = {g^{(k+1)}}^T\big(g^{(k)} + \alpha_k A d^{(k)}\big) = \alpha_k {g^{(k+1)}}^T A d^{(k)} = \beta_k {g^{(k)}}^T g^{(k)}$$

and hence the assumption. The algorithm is summarized in Table 5.2. Indeed, the algorithm defines conjugate directions:

Theorem 5.14 *If $g^{(k-1)} \neq 0$, then $d^{(k-1)} \neq 0$ and the $d^{(0)}, \ldots, d^{(k-1)}$ are conjugate.*

Choose any $x^{(0)} \in \mathbb{R}^m$ and calculate

$$d^{(0)} := -g^{(0)} = b - Ax^{(0)}\,.$$

For $k = 0, 1, \ldots$ put

$$\begin{aligned}
\alpha_k &= \frac{{g^{(k)}}^T g^{(k)}}{{d^{(k)}}^T A d^{(k)}}\,,\\
x^{(k+1)} &= x^{(k)} + \alpha_k d^{(k)}\,,\\
g^{(k+1)} &= g^{(k)} + \alpha_k A d^{(k)}\,,\\
\beta_k &= \frac{{g^{(k+1)}}^T g^{(k+1)}}{{g^{(k)}}^T g^{(k)}}\,,\\
d^{(k+1)} &= -g^{(k+1)} + \beta_k d^{(k)}\,,
\end{aligned}$$

until the termination criterion ("$|g^{(k+1)}|_2 = 0$") is fulfilled.

Table 5.2. CG method.

Proof: The proof is done by induction:
The case $k = 1$ is clear. Assume that $d^{(0)}, \ldots, d^{(k-1)}$ are all nonzero and conjugate. Thus according to Theorem 5.12 and Theorem 5.13 the identities (5.60)–(5.64) hold up to index k. Let us first prove that $d^{(k)} \neq 0$:

Due to $g^{(k)} + d^{(k)} = \beta_{k-1} d^{(k-1)} \in K_k(A; g^{(0)})$ the assertion $d^{(k)} = 0$ would imply directly $g^{(k)} \in K_k(A; g^{(0)})$. But relations (5.63) and (5.64) imply for the index k,

$${g^{(k)}}^T x = 0 \quad \text{for all } x \in K_k(A; g^{(0)})\,,$$

which contradicts $g^{(k)} \neq 0$.

In order to prove ${d^{(k)}}^T A d^{(i)} = 0$ for $i = 0, \ldots, k-1$, according to (5.62) we have to prove only the case $i \leq k-2$. We have

$${d^{(i)}}^T A d^{(k)} = -{d^{(i)}}^T A g^{(k)} + \beta_{k-1} {d^{(i)}}^T A d^{(k-1)}\,.$$

The first term disappears due to $Ad^{(i)} \in A\left(K_{k-1}\left(A; g^{(0)}\right)\right) \subset K_k\left(A; g^{(0)}\right)$, which means that $Ad^{(i)} \in \operatorname{span}\left\{d^{(0)}, \ldots, d^{(k-1)}\right\}$, and (5.60). The second term disappears because of the induction hypothesis. □

Methods that aim at minimizing the error or residual on $K_k\left(A; g^{(0)}\right)$ with respect to a norm $\|\cdot\|$ are called *Krylov subspace methods*. Here the error will be minimized in the energy norm $\|\cdot\| = \|\cdot\|_A$ according to (5.55) and Theorem 5.12.

Due to the representation of the Krylov space in Theorem 5.13 the elements $y \in x^{(0)} + K_k\left(A; g^{(0)}\right)$ are exactly the vectors of the form $y = x^{(0)} + q(A)g^{(0)}$, for any $q \in \mathcal{P}_{k-1}$ (for the notation $q(A)$ see Appendix

A.3). Hence it follows that

$$y - x = x^{(0)} - x + q(A)A\big(x^{(0)} - x\big) = p(A)\big(x^{(0)} - x\big)\,,$$

with $p(z) = 1 + q(z)z$, i.e., $p \in \mathcal{P}_k$ and $p(0) = 1$. On the other hand, any such polynomial can be represented in the given form (define q by $q(z) = (p(z) - 1)/z$). Thus Theorem 5.12 implies

$$\big\|x^{(k)} - x\big\|_A \le \|y - x\|_A = \big\|p(A)\big(x^{(0)} - x\big)\big\|_A \tag{5.66}$$

for any $p \in \mathcal{P}_k$ with $p(0) = 1$.

Let $z_1, \ldots, z_m$ be an orthonormal basis of eigenvectors, that is,

$$Az_j = \lambda_j z_j \quad \text{and} \quad z_i^T z_j = \delta_{ij} \quad \text{for } i, j = 1, \ldots, m\,. \tag{5.67}$$

Then we have $x^{(0)} - x = \sum_{j=1}^m c_j z_j$ for certain $c_j \in \mathbb{R}$, and hence

$$p(A)\big(x^{(0)} - x\big) = \sum_{j=1}^m p\left(\lambda_j\right) c_j z_j$$

and therefore

$$\big\|x^{(0)} - x\big\|_A^2 = \big(x^{(0)} - x\big)^T A\big(x^{(0)} - x\big) = \sum_{i,j=1}^m c_i c_j z_i^T A z_j = \sum_{j=1}^m \lambda_j \left|c_j\right|^2$$

and analogously

$$\big\|p(A)\left(x^{(0)} - x\right)\big\|_A^2 = \sum_{j=1}^m \lambda_j \left|c_j p(\lambda_j)\right|^2 \le \left(\max_{i=1,\ldots,m} |p(\lambda_i)|\right)^2 \big\|x^{(0)} - x\big\|_A^2\,. \tag{5.68}$$

Relations (5.66), (5.68) imply the following theorem:

Theorem 5.15 *For the CG method and any $p \in \mathcal{P}_k$ satisfying $p(0) = 1$, we have*

$$\big\|x^{(k)} - x\big\|_A \le \max_{i=1,\ldots,m} |p(\lambda_i)|\ \big\|x^{(0)} - x\big\|_A\,,$$

with the eigenvalues $\lambda_1, \ldots, \lambda_m$ of A.

If the eigenvalues of A are not known, but their location is, i.e., if one knows $a, b \in \mathbb{R}$ such that

$$a \le \lambda_1, \ldots, \lambda_m \le b\,, \tag{5.69}$$

then only the following weaker estimate can be used:

$$\big\|x^{(k)} - x\big\|_A \le \max_{\lambda \in [a,b]} |p(\lambda)|\ \big\|x^{(0)} - x\big\|_A\,. \tag{5.70}$$

Therefore, we have to find $p \in \mathcal{P}_m$ with $p(0) = 1$ that minimizes $\max\big\{|p(\lambda)| \;\big|\; \lambda \in [a, b]\big\}$.

This approximation problem in the maximum norm appeared already in (5.43), because there is a bijection between the sets $\{p \in P_k \mid p(1) = 1\}$ and $\{p \in P_k \mid p(0) = 1\}$ through

$$p \mapsto \tilde{p} \quad , \quad \tilde{p}(t) := p(1-t)\,. \tag{5.71}$$

Its solution can represented by using the Chebyshev polynomials of the first kind (see, for example, [38, p. 302]). They are recursively defined by

$$T_0(x) := 1\,, \quad T_1(x) := x\,, \quad T_{k+1}(x) := 2xT_k(x) - T_{k-1}(x) \quad \text{for } x \in \mathbb{R}$$

and have the representation

$$T_k(x) = \cos(k \arccos(x))$$

for $|x| \leq 1$. This immediately implies

$$|T_k(x)| \leq 1 \quad \text{for } |x| \leq 1\,.$$

A further representation, valid for $x \in \mathbb{R}$, is

$$T_k(x) = \frac{1}{2}\left(\left(x + \left(x^2-1\right)^{1/2}\right)^k + \left(x - \left(x^2-1\right)^{1/2}\right)^k\right)\,. \tag{5.72}$$

The optimal polynomial in (5.70) is then defined by

$$p(z) := \frac{T_k\left((b+a-2z)/(b-a)\right)}{T_k\left((b+a)/(b-a)\right)} \quad \text{for } z \in \mathbb{R}\,.$$

This implies the following result:

Theorem 5.16 *Let κ be the spectral condition number of A and assume $\kappa > 1$. Then*

$$\left\|x^{(k)} - x\right\|_A \leq \frac{1}{T_k\left(\frac{\kappa+1}{\kappa-1}\right)}\left\|x^{(0)} - x\right\|_A \leq 2\left(\frac{\kappa^{1/2}-1}{\kappa^{1/2}+1}\right)^k \left\|x^{(0)} - x\right\|_A\,. \tag{5.73}$$

Proof: Choose a as the smallest eigenvalue $\lambda_{\min}$ and b as the largest one $\lambda_{\max}$.

The first inequality follows immediately from (5.70) and $\kappa = b/a$. For the second inequality note that due to $(\kappa+1)/(\kappa-1) = 1 + 2/(\kappa-1) =: 1 + 2\eta \geq 1$, (5.72) implies

$$\begin{aligned} T_k\left(\frac{\kappa+1}{\kappa-1}\right) &\geq \frac{1}{2}\left(1 + 2\eta + \left((1+2\eta)^2 - 1\right)^{1/2}\right)^k \\ &= \frac{1}{2}\left(1 + 2\eta + 2\left(\eta(\eta+1)\right)^{1/2}\right)^k\,. \end{aligned}$$

Finally,

$$1 + 2\eta + 2\left(\eta(\eta+1)\right)^{1/2} = \left(\eta^{1/2} + (\eta+1)^{1/2}\right)^2 = \frac{(\eta+1)^{1/2} + \eta^{1/2}}{(\eta+1)^{1/2} - \eta^{1/2}}$$

$$= \frac{(1+1/\eta)^{1/2}+1}{(1+1/\eta)^{1/2}-1},$$

which concludes the proof because of $1+1/\eta = \kappa$. □

For large κ we have again

$$\frac{\kappa^{1/2}-1}{\kappa^{1/2}+1} \approx 1 - \frac{2}{\kappa^{1/2}}.$$

Compared with (5.58), κ has been improved to $\kappa^{1/2}$.

From (5.4) and (5.34) the complexity of the five-point stencil discretization of the Poisson equation on the square results in

$$\ln\left(\frac{1}{\varepsilon}\right) O\big(\kappa^{1/2}\big)\, O(m) = O(n)\, O(m) = O\big(m^{3/2}\big).$$

This is the same behaviour as that of the SOR method with optimal relaxation parameter. The advantage of the above method lies in the fact that the determination of parameters is not necessary for applying the CG method. For quasi-uniform triangulations, Theorem 3.45 implies an analogous general statement.

A relation to the semi-iterative methods follows from (5.71): The estimate (5.66) can also be expressed as

$$\big\|e^{(k)}\big\|_A \le \big\|p(I-A)e^{(0)}\big\|_A \tag{5.74}$$

for any $p \in \mathcal{P}_k$ with $p(1) = 1$.

This is the same estimate as (5.42) for the Richardson iteration (5.28) as basis method, with the Euclidean norm $|\cdot|_2$ replaced by the energy norm $\|\cdot\|_A$. While the semi-iterative methods are defined by minimization of upper bounds in (5.42), the CG method is optimal in the sense of (5.74), without knowledge of the spectrum $\sigma(I-A)$. In this manner the CG method can be seen as an (optimal) acceleration method for the Richardson iteration.

Exercises

5.4 Let $A \in \mathbb{R}^{m,m}$ be a symmetric positive definite matrix.

(a) Show that for x, y with $x^T y = 0$ we have

$$\frac{\langle x, y\rangle_A}{\|x\|_A \|y\|_A} \le \frac{\kappa-1}{\kappa+1},$$

where κ denotes the spectral condition number of A.
Hint: Represent x, y in terms of an orthonormal basis consisting of eigenvectors of A.

(b) Show using the example $m = 2$ that this estimate is sharp. To this end, look for a positive definite symmetric matrix $A \in \mathbb{R}^{2,2}$ as well as vectors $x, y \in \mathbb{R}^2$ with $x^T y = 0$ and

$$\frac{\langle x, y \rangle_A}{\|x\|_A \|y\|_A} = \frac{\kappa - 1}{\kappa + 1} .$$

5.5 Prove that the computation of the conjugate direction in the CG method in the general step $k \geq 2$ is equivalent to the three-term recursion formula

$$d^{(k+1)} = [\alpha_k A + (\beta_k + 1) I]\, d^{(k)} - \beta_{k-1} d^{(k-1)} .$$

5.6 Let $A \in \mathbb{R}^{m,m}$ be a symmetric positive definite matrix with spectral condition number κ. Suppose that the spectrum $\sigma(A)$ of the matrix A satisfies $a_0 \in \sigma(A)$ as well as $\sigma(A) \setminus \{a_0\} \subset [a, b]$ with $0 < a_0 < a \leq b$.

Show that this yields the following convergence estimate for the CG method:

$$\|x^{(k)} - x\|_A \leq 2\, \frac{b - a_0}{a_0} \left(\frac{\sqrt{\hat{\kappa}} - 1}{\sqrt{\hat{\kappa}} + 1} \right)^{k-1} \|x^{(0)} - x\|_A ,$$

where $\hat{\kappa} := b/a$ ($< \kappa$).

5.3 Preconditioned Conjugate Gradient Method

Due to Theorem 5.16, $\kappa(A)$ should be small or only weakly growing in m, which is not true for a finite element stiffness matrix.

The technique of preconditioning is used — as already discussed in Section 5.1 — to transform the system of equations in such a way that the condition number of the system matrix is reduced without increasing the effort in the evaluation of the matrix vector product too much.

In a *preconditioning from the left* the system of equations is transformed to

$$C^{-1} A x = C^{-1} b$$

with a *preconditioner* C; in a *preconditioning from the right* it is transformed to

$$A C^{-1} y = b ,$$

such that $x = C^{-1} y$ is the solution of (5.1). Since the matrices are generally sparse, this always has to be interpreted as a solution of the system of equations $Cx = y$.

If A is symmetric and positive definite, then this property is generally violated by the transformed matrix for both variants, even for a symmetric

positive definite C. We assume for a moment to have a decomposition of C with a nonsingular matrix W as

$$C = WW^T .$$

Then $Ax = b$ can be transformed to $W^{-1}AW^{-T}W^T x = W^{-1}b$, i.e., to

$$By = c \quad \text{with} \quad B = W^{-1}AW^{-T},\ c = W^{-1}b . \tag{5.75}$$

The matrix B is symmetric and positive definite. The solution x is then given by $x = W^{-T}y$. This procedure is called *split preconditioning*.

Due to $W^{-T}BW^T = C^{-1}A$ and $WBW^{-1} = AC^{-1}$, B, $C^{-1}A$ and AC^{-1} have the same eigenvalues, and therefore also the same spectral condition number κ. Therefore, C should be "close" to A in order to reduce the condition number. The CG method, applied to (5.75) and then back transformed, leads to the *preconditioned conjugate gradient method (PCG)*: The terms of the CG method applied to (5.75) will all be marked by $\tilde{}$, with the exception of α_k and β_k.

Due to the back transformation

$$x = W^{-T}\tilde{x}$$

the algorithm has the search direction

$$d^{(k)} := W^{-T}\tilde{d}^{(k)}$$

for the transformed iterate

$$x^{(k)} := W^{-T}\tilde{x}^{(k)} . \tag{5.76}$$

The gradient $g^{(k)}$ of (5.44) in $x^{(k)}$ is given by

$$g^{(k)} := Ax^{(k)} - b = W\big(B\tilde{x}^{(k)} - c\big) = W\tilde{g}^{(k)} ,$$

and hence

$$g^{(k+1)} = g^{(k)} + \alpha_k WB\tilde{d}^{(k)} = g^{(k)} + \alpha_k Ad^{(k)} ,$$

so that this formula remains unchanged compared with the CG method with a new interpretation of the search direction. The search directions are updated by

$$d^{(k+1)} = -W^{-T}W^{-1}g^{(k+1)} + \beta_k d^{(k)} = -C^{-1}g^{(k+1)} + \beta_k d^{(k)} ,$$

so that in each iteration step additionally the system of equations $Ch^{(k+1)} = g^{(k+1)}$ has to be solved.

Finally, we have

$$\tilde{g}^{(k)^T}\tilde{g}^{(k)} = g^{(k)^T}C^{-1}g^{(k)} = g^{(k)^T}h^{(k)}$$

and

$$\tilde{d}^{(k)^T}B\tilde{d}^{(k)} = d^{(k)^T}Ad^{(k)} ,$$

so that the algorithm takes the form of Table 5.3.

Choose any $x^{(0)} \in \mathbb{R}^m$ and calculate

$$g^{(0)} = Ax^{(0)} - b\,, \quad d^{(0)} := -h^{(0)} := -C^{-1}g^{(0)}\,.$$

For $k = 0, 1, \ldots$ put

$$\begin{aligned}
\alpha_k &= \frac{{g^{(k)}}^T h^{(k)}}{{d^{(k)}}^T A d^{(k)}}\,,\\
x^{(k+1)} &= x^{(k)} + \alpha_k d^{(k)}\,,\\
g^{(k+1)} &= g^{(k)} + \alpha_k A d^{(k)}\,,\\
h^{(k+1)} &= C^{-1} g^{(k+1)}\,,\\
\beta_k &= \frac{{g^{(k+1)}}^T h^{(k+1)}}{{g^{(k)}}^T h^{(k)}}\,,\\
d^{(k+1)} &= -h^{(k+1)} + \beta_k d^{(k)}\,,
\end{aligned}$$

up to the termination criterion ("$|g^{(k+1)}|_2 = 0$") .

Table 5.3. PCG method.

The solution of the additional systems of equations for sparse matrices should have the complexity $O(m)$, in order not to worsen the complexity for an iteration. It is not necessary to know a decomposition $C = WW^T$.

Alternatively, the PCG method can be established by noting that $C^{-1}A$ is self-adjoint and positive definite with respect to the energy scalar product $\langle \cdot, \cdot \rangle_C$ defined by C:

$$\langle C^{-1}Ax, y\rangle_C = \left(C^{-1}Ax\right)^T Cy = x^T Ay = x^T C(C^{-1}Ay) = \langle x, C^{-1}Ay\rangle_C$$

and hence also $\langle C^{-1}Ax, x\rangle_C > 0$ for $x \neq 0$.

Choosing the CG method for (5.75) with respect to $\langle \cdot, \cdot \rangle_C$, we obtain precisely the above method.

In case the termination criterion "$\left|g^{(k+1)}\right|_2 = 0$" is used for the iteration, the scalar product must be additionally calculated. Alternatively, we may use "$\left|{g^{(k+1)}}^T h^{(k+1)}\right| = 0$". Then the residual is measured in the norm $\|\cdot\|_{C^{-1}}$.

Following the reasoning at the end of Section 5.2, the PCG method can be interpreted as an acceleration of a linear stationary method with iteration matrix

$$M = I - C^{-1}A\,.$$

For a consistent method, we have $N = C^{-1}$ or, in the formulation (5.10), $W = C$. This observation can be extended in such a way that the CG method can be used for the acceleration of iteration methods, for example also for the multigrid method, which will be introduced in Section 5.5. Due

to the deduction of the preconditioned CG method and the identity

$$\left\|x^{(k)} - x\right\|_A = \left\|\tilde{x}^{(k)} - \tilde{x}\right\|_B ,$$

which results from the transformation (5.76), the approximation properties for the CG method also hold for the PCG method if the spectral condition number $\kappa(A)$ is replaced by $\kappa(B) = \kappa(C^{-1}A)$. Therefore,

$$\left\|x^{(k)} - x\right\|_A \leq 2 \left(\frac{\kappa^{1/2} - 1}{\kappa^{1/2} + 1}\right)^k \left\|x^{(0)} - x\right\|_A$$

with $\kappa = \kappa(C^{-1}A)$.

There is a close relation between those preconditioning matrices C, which keep $\kappa(C^{-1}A)$ small, and well-convergent linear stationary iteration methods with $N = C^{-1}$ (and $M = I - C^{-1}A$) if N is symmetric and positive definite. Indeed,

$$\kappa(C^{-1}A) \leq (1 + \varrho(M))/(1 - \varrho(M))$$

if the method defined by M and N is convergent and N is symmetric for symmetric A (see Exercise 5.7).

From the considered linear stationary methods because of the required symmetry we may take

- Jacobi's method:

This corresponds exactly to the *diagonal scaling*, which means the division of each equation by its diagonal element. Indeed, from the decomposition (5.18) we have $C = N^{-1} = D$, and the PCG method is equivalent to the preconditioning from the left by the matrix C^{-1} in combination with the usage of the energy scalar product $\langle \cdot, \cdot \rangle_C$.

- The SSOR method:

According to (5.38) we have

$$C = \omega^{-1}(2 - \omega)^{-1}(D + \omega L)D^{-1}(D + \omega L^T) .$$

Hence C is symmetric and positive definite. The solution of the auxiliary systems of equations needs only forward and backward substitutions with the same structure of the matrix as for the system matrix, so that the requirement of lower complexity is also fulfilled. An exact estimation of $\kappa(C^{-1}A)$ shows (see [3, pp. 328 ff.]) that under certain requirements for A, which reflect properties of the boundary value problem and the discretization, we find a considerable improvement of the conditioning by using an estimate of the type

$$\kappa(C^{-1}A) \leq \text{const}(\kappa(A)^{1/2} + 1) .$$

The choice of the relaxation parameter ω is not critical. Instead of trying to choose an optimal one for the contraction number of the SSOR

method, we can minimize an estimation for $\kappa(C^{-1}A)$ (see [3, p. 337]), which recommends a choice of ω in $[1.2, 1.6]$.

For the five-point stencil discretization of the Poisson equation on the square we have, according to (5.34), $\kappa(A) = O(n^2)$, and the above conditions are fulfilled (see [3, pp. 330 f.]). By SSOR preconditioning this is improved to $\kappa(C^{-1}A) = O(n)$, and therefore the complexity of the method is

$$\ln\left(\frac{1}{\varepsilon}\right) O\big(\kappa^{1/2}\big) O(m) = \ln\left(\frac{1}{\varepsilon}\right) O\big(n^{1/2}\big) O(m) = O\big(m^{5/4}\big)\,. \tag{5.77}$$

As discussed in Section 2.5, direct elimination methods are not suitable in conjunction with the discretization of boundary value problems with large node numbers, because in general fill-in occurs. As discussed in Section 2.5, $L = (l_{ij})$ describes a lower triangular matrix with $l_{ii} = 1$ for all $i = 1, \ldots, m$ (the dimension is described there with the number of degrees of freedom M) and $U = (u_{ij})$ an upper triangular matrix. The idea of the *incomplete LU factorization*, or *ILU factorization*, is to allow only certain *patterns* $\mathcal{E} \in \{1, \ldots, m\}^2$ for the entries of L and U, and instead of $A = LU$, in general we can require only

$$A = LU - R.$$

Here the remainder $R = (r_{ij}) \in \mathbb{R}^{m,m}$ has to satisfy

$$r_{ij} = 0 \quad \text{for } (i,j) \in \mathcal{E}\,. \tag{5.78}$$

The requirements

$$a_{ij} = \sum_{k=1}^{m} l_{ik} u_{kj} \quad \text{for } (i,j) \in \mathcal{E} \tag{5.79}$$

mean $|\mathcal{E}|$ equations for the $|\mathcal{E}|$ entries of the matrices L and U. (Notice that $l_{ii} = 1$ for all i.) The existence of such factorizations will be discussed later.

Analogously to the close connection between the LU factorization and an $\mathrm{LDL}^{\mathrm{T}}$ or LL^{T} factorization for symmetric or symmetric positive definite matrices, as defined in Section 2.5, we can use the *IC factorization* (**i***ncomplete* **C***holesky factorization*) for such matrices. The IC factorization needs a representation in the following form:

$$A = LL^T - R\,.$$

Based on an ILU factorization a linear stationary method is defined by $N = (LU)^{-1}$ (and $M = I - NA$), the *ILU iteration*. We thus have an expansion of the old method of *iterative refinement*.

Using $C = N^{-1} = LU$ for the preconditioning, the complexity of the auxiliary systems depends on the choice of the matrix pattern $\mathcal{E}$. In general, the following is required:

$$\mathcal{E}' := \big\{(i,j) \;\big|\; a_{ij} \neq 0\,,\ i,j = 1,\ldots,m\big\} \subset \mathcal{E}\,,\ \ \big\{(i,i) \;\big|\; i = 1,\ldots,m\big\} \subset \mathcal{E}\,. \tag{5.80}$$

The requirement of equality $\mathcal{E}' = \mathcal{E}$ is most often used. Then, and also in the case of fixed expansions of $\mathcal{E}'$, it is ensured that for a sequence of systems of equations with a matrix A that is sparse in the strict sense, this will also hold for L and U. All in all, only $O(m)$ operations are necessary, including the calculation of L and U, as in the case of the SSOR preconditioning for the auxiliary system of equations. On the other hand, the remainder R should be rather small in order to ensure a good convergence of the ILU iteration and also to ensure a small spectral condition number $\kappa(C^{-1}A)$. Possible matrix patterns $\mathcal{E}$ are shown, for example, in [28, pp. 275 ff.], where a more specific structure of L and U is discussed if the matrix A is created by a discretization on a structured grid, for example by a finite difference method.

The question of the existence (and stability) of an ILU factorization remains to be discussed. It is known from (2.56) that also for the existence of an LU factorization certain conditions are necessary, as for example the M-matrix property. This is even sufficient for an ILU factorization.

Theorem 5.17 *Let $A \in \mathbb{R}^{m,m}$ be an M-matrix. Then for a given pattern $\mathcal{E}$ that satisfies* (5.80)*, an ILU factorization exists. The hereby defined decomposition of A as $A = LU - R$ is* regular *in the following sense:*

$$\left((LU)^{-1}\right)_{ij} \geq 0\,, \quad (R)_{ij} \geq 0 \quad \textit{for all } i,j = 1,\ldots,m\,.$$

Proof: See [16, p. 235]. □

An ILU (or IC) factorization can be defined by solving the equations (5.78) for l_{ij} and u_{ij} in an appropriate order. Alternatively, the elimination or Cholesky method can be used in its original form on the pattern $\mathcal{E}$.

An improvement of the eigenvalue distribution of $C^{-1}A$ is sometimes possible by using an MIC factorization (**m***odified* **i***ncomplete* **C***holesky factorization*) instead of an IC factorization. In contrast to (5.79) the updates in the elimination method for positions outside the pattern are not ignored here but have to be performed for the corresponding diagonal element.

Concerning the reduction of the condition number by the ILU or IC preconditioning for the model problem, we have the same situation as for the SSOR preconditioning. In particular (5.77) holds, too.

The auxiliary system of equations with $C = N^{-1}$, which means that

$$h^{(k+1)} = Ng^{(k+1)}\,,$$

can also be interpreted as an iteration step of the iteration method defined by N with initial value $z^{(0)} = 0$ and right-hand side $g^{(k+1)}$. An expansion of the discussed possibilities for preconditioning is therefore obtained by using a fixed number of iteration steps instead of only one.

Exercises

5.7 Let $A_1, A_2, \ldots, A_k, C_1, C_2, \ldots, C_k \in \mathbb{R}^{m,m}$ be symmetric positive semidefinite matrices with the property

$$ax^T C_i x \le x^T A_i x \le bx^T C_i x \quad \text{for } x \in \mathbb{R}^m\,,\ i = 1, \ldots, k \text{ and } 0 < a \le b\,.$$

Prove: If $A := \sum_{i=1}^k A_i$ and $C := \sum_{i=1}^k C_i$ are positive definite, then the spectral condition number κ of $C^{-1}A$ satisfies

$$\kappa(C^{-1}A) \le \frac{b}{a}\,.$$

5.8 Show that the matrix

$$A := \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$$

is positive definite and its spectral condition number is 4.

Hint: Consider the associated quadratic form.

5.9 Investigate the convergence of the (P)CG method on the basis of Theorem 3.45 and distinguish between $d = 2$ and $d = 3$.

5.4 Krylov Subspace Methods for Nonsymmetric Systems of Equations

With the different variants of the PCG method we have methods that are quite appropriate — regarding their complexity — for those systems of equations that arise from the discretization of boundary value problems. However, this holds only under the assumption that the system matrix is symmetric and positive definite, reducing the possibilities of application, for example to finite element discretizations of purely diffusive processes without convective transport mechanism (see (3.23)). Exceptions for time-dependent problems are only the (semi-)explicit time discretization (compare (7.72)) and the Lagrange–Galerkin method (see Section 9.4). For all other cases the systems of equations that arise are always nonsymmetric and *positive real*, which means that the system matrix A satisfies

$$A + A^T \quad \text{is positive definite.}$$

It is desirable to generalize the (P)CG methods for such matrices. The CG method is characterized by two properties:

- The iterate $x^{(k)}$ minimizes $f(\cdot) = \|\cdot - x\|_A$ on $x^{(0)} + K_k\big(A; g^{(0)}\big)$, where $x = A^{-1}b$.

- The basis vectors $d^{(i)}$, $i = 0, \ldots, k-1$, of $K_k\big(A; g^{(0)}\big)$ do not have to be calculated in advance (and stored in the computer), but will be calculated by a *three-term recursion* (5.61) during the iteration. An analogous relation holds by definition for $x^{(k)}$ (see (5.48)).

The first property can be preserved in the following, whereby the norm of the error or residual minimization varies in each method. The second property is partially lost, because generally all basis vectors $d^{(0)}, \ldots, d^{(k-1)}$ are necessary for the calculation of $x^{(k)}$. This will result in memory space problems for large k. As for the CG methods, preconditioning will be necessary for an acceptable convergence of the methods. The conditions for the preconditioning matrices are the same as for the CG method with the exception of symmetry and positive definiteness. All three methods of preconditioning are in principle possible. Therefore, preconditioning will not be discussed in the following; we refer to Section 5.3.

The simplest approach is the application of the CG method to a system of equations with symmetric positive definite matrix equivalent to (5.1). This is the case for the *normal equations*

$$A^T A x = A^T b\,. \tag{5.81}$$

The approach is called CGNR (**C***onjugate* **G***radient* **N***ormal* **R***esidual*), because here the iterate $x^{(k)}$ minimizes the Euclidean norm of the residual on $x^{(0)} + K_k\big(A^T A; g^{(0)}\big)$ with $g^{(0)} = A^T\big(Ax^{(0)} - b\big)$. This follows from the equation

$$\|y - x\|^2_{A^T A} = (Ay - b)^T (Ay - b) = |Ay - b|_2^2 \tag{5.82}$$

for any $y \in \mathbb{R}^m$ and the solution $x = A^{-1}b$.

All advantages of the CG method are preserved, although in (5.53) and (5.65) $Ad^{(k)}$ is to be replaced by $A^T A d^{(k)}$. Additionally to the doubling of the number of operations this may be a disadvantage if $\kappa_2(A)$ is large, since $\kappa_2(A^T A) = \kappa_2(A)^2$ can lead to problems of stability and convergence. Due to (5.34) this is to be expected for a large number of degrees of freedom.

Furthermore, in the case of list-based storage one of the operations Ay and $A^T y$ is always very expensive due to searching. It is even possible that we do not explicitly know the matrix A but that only the mapping $y \mapsto Ay$ can be evaluated, which then disqualifies this method completely (see Exercise 8.6).

The same drawback occurs if

$$AA^T \tilde{x} = b \tag{5.83}$$

with the solution $\tilde{x} = A^{-T} x$ taken instead of (5.81). If $\tilde{x}^{(k)}$ is the kth iterate of the CG method applied to (5.83), then the $x^{(k)} := A^T \tilde{x}^{(k)}$ minimizes the residual in the Euclidean norm on $x_0 + A^T\big[K_k\big(AA^T; g^{(0)}\big)\big]$: Note that

$$\|\tilde{y} - \tilde{x}\|^2_{AA^T} = \big(A^T\tilde{y} - x\big)^T \big(A^T\tilde{y} - x\big) = \big|A^T\tilde{y} - x\big|_2^2$$

Let $g^{(0)} \in \mathbb{R}^m$, $g^{(0)} \neq 0$ be given, Set

$$v_1 := g^{(0)} / |g^{(0)}|_2 .$$

For $j = 1, \ldots, k$ calculate

$$\begin{aligned} h_{ij} &:= v_i^T A v_j \quad \text{for } i = 1, \ldots, j , \\ w_j &:= A v_j - \sum_{i=1}^{j} h_{ij} v_i , \\ h_{j+1,j} &:= |w_j|_2 . \end{aligned}$$

If $h_{j+1,j} = 0$, termination; otherwise, set

$$v_{j+1} := w_j / h_{j+1,j} .$$

Table 5.4. Arnoldi algorithm.

holds for any $\tilde{y} \in \mathbb{R}^m$ and $g^{(0)} = Ax^{(0)} - b$. This explains the terminology CGNE (with E for **E***rror*).

Whether a method minimizes the error of the residual obviously depends on the norm used. For a symmetric positive definite $B \in \mathbb{R}^{m,m}$, any $y \in \mathbb{R}^m$, and $x = A^{-1}b$, we have

$$\|Ay - b\|_B = \|y - x\|_{A^T B A} .$$

For $B = A^{-T}$ and a symmetric positive definite A we get the situation of the CG method:

$$\|Ay - b\|_{A^{-T}} = \|y - x\|_A .$$

For $B = I$ we get again (5.82):

$$|Ay - b|_2 = \|y - x\|_{A^T A} .$$

The minimization of this functional on $x^{(0)} + K_k\big(A; g^{(0)}\big)$ (not $K_k\big(A^T A; g^{(0)}\big)$) leads to the GMRES method (***G**eneralized **M**inimum **RES**idual*).

This (and other) methods are founded algorithmically on the recursive construction of orthonormal bases of $K_k\big(A; g^{(0)}\big)$ by *Arnoldi's method*. This method combines the generation of a basis according to (5.61) and Schmidt's orthonormalization (see Table 5.4).

If Arnoldi's method can be performed up to the index k, then

$$\begin{aligned} h_{ij} &:= 0 \quad \text{for } j = 1, \ldots, k,\ i = j+2, \ldots, k+1 , \\ H_k &:= (h_{ij})_{ij} \in \mathbb{R}^{k,k} , \\ \bar{H}_k &:= (h_{ij})_{ij} \in \mathbb{R}^{k+1,k} , \\ V_{k+1} &:= (v_1, \ldots, v_{k+1}) \in \mathbb{R}^{m,k+1} . \end{aligned}$$

The matrix H_k is an upper Hessenberg matrix (see Appendix A.3). The basis for the GMRES method is the following theorem:

Theorem 5.18 *If Arnoldi's method can be performed up to the index k, then*

(1) *$v_1, \dots, v_{k+1}$ form an orthonormal basis of $K_{k+1}(A; g^{(0)})$.*

(2)
$$AV_k = V_k H_k + w_k e_k^T = V_{k+1} \bar{H}_k \,, \tag{5.84}$$
with $e_k = (0, \dots, 0, 1)^T \in \mathbb{R}^k$,
$$V_k^T A V_k = H_k \,. \tag{5.85}$$

(3) *The problem*
$$\textit{Minimize} \quad |Ay - b|_2 \quad \textit{for} \quad y \in x^{(0)} + K_k(A; g^{(0)})$$
with minimum $x^{(k)}$ is equivalent to
$$\textit{Minimize} \quad \left|\bar{H}_k \xi - \beta e_1\right|_2 \quad \textit{for} \quad \xi \in \mathbb{R}^k \tag{5.86}$$
with $\beta := -\left|g^{(0)}\right|_2$ and minimum $\xi^{(k)}$, and we have
$$x^{(k)} = x^{(0)} + V_k \xi^{(k)} \,.$$

If Arnoldi's method terminates at the index k, then
$$x^{(k)} = x = A^{-1} b \,.$$

Proof: (1): The vectors $v_1, \dots, v_{k+1}$ are orthonormal by construction; hence we have only to prove $v_i \in K_{k+1}\big(A; g^{(0)}\big)$ for $i = 1, \dots, k+1$. This follows from the representation
$$v_i = q_{i-1}(A) v_1 \quad \text{with polynomials} \quad q_{i-1} \in \mathcal{P}_{i-1} \,.$$
In this form we can prove the statement by induction with respect to k. For $k = 0$ the assertion is trivial. Let the statement hold for $k-1$. The validity for k then follows from
$$h_{k+1,k} v_{k+1} = A v_k - \sum_{i=1}^{k} h_{ik} v_i = \left(A q_{k-1}(A) - \sum_{i=1}^{k} h_{ik} q_{i-1}(A) \right) v_1 \,.$$
(2): Relation (5.85) follows from (5.84) by multiplication by V_k^T, since $V_k^T V_k = I$ and $V_k^T w_k = h_{k+1,k} V_k^T v_{k+1} = 0$ due to the orthonormality of the v_i.

The relation in (5.84) is the matrix representation of
$$A v_j = \sum_{i=1}^{j} h_{ij} v_i + w_j = \sum_{i=1}^{j+1} h_{ij} v_i \quad \text{for} \quad j = 1, \dots, k \,.$$

(3): Due to (1), the space $x^{(0)} + K_k\big(A; g^{(0)}\big)$ has the parametrisation
$$y = x^{(0)} + V_k \xi \quad \text{with} \quad \xi \in \mathbb{R}^k \,. \tag{5.87}$$

The assertion is a consequence of the identity

$$\begin{aligned} Ay - b &= A\big(x^{(0)} + V_k\xi\big) - b = AV_k\xi + g^{(0)} \\ &= V_{k+1}\bar{H}_k\xi - \beta v_1 = V_{k+1}\left(\bar{H}_k\xi - \beta e_1\right), \end{aligned}$$

which follows from (2), since it implies

$$|Ay - b|_2 = \left|V_{k+1}(\bar{H}_k\xi - \beta e_1)\right|_2 = \left|\bar{H}_k\xi - \beta e_1\right|_2$$

due to the orthogonality of V_{k+1}. The last assertion finally can be seen in this way: If Arnoldi's method breaks down at the index k, then relation (2) becomes

$$AV_k = V_k H_k\,,$$

and

$$AV_k = V_{k+1}\bar{H}_k$$

will further hold with v_{k+1} chosen arbitrarily (due to $h_{k+1,k} = 0$). Since A is nonsingular, this also holds for H_k. Hence the choice

$$\xi := H_k^{-1}(\beta e_1)\,,$$

which satisfies

$$\left|\bar{H}_k\xi - \beta e_1\right|_2 = |H_k\xi - \beta e_1|_2 = 0\,,$$

is possible. Hence the corresponding $y \in \mathbb{R}^m$ defined by (5.87) fulfills $y = x^{(k)} = x$. □

One problem of Arnoldi's method is that the orthogonality of the v_i is easily lost due to rounding errors. If one substitutes the assignment

$$w_j := Av_j - \sum_{i=1}^{j} h_{ij}v_i$$

in Table 5.4 by the operations

$$\begin{aligned} & & w_j &:= Av_j\,, \\ &\text{for } i = 1,\ldots,j \text{ calculate} & & \\ & & h_{ij} &:= w_j^T v_i\,, \\ & & w_j &:= w_j - h_{ij}v_i\,, \end{aligned}$$

which define the same vector, one obtains the *modified Arnoldi's method*. From this relation and from (5.86) the GMRES method is constructed in its basic form. Alternatively, Schmidt's orthonormalization can be replaced by the Householder method (see [28, pp. 159 ff.]). With exact arithmetic the GMRES algorithm terminates only after reaching the exact solution (with $h_{k+1,k} = 0$). This is not always the case for alternative methods of the same class. For an increasing iteration index k and large problem dimensions m there may be lack of enough memory for the storage of

the basis vectors $v_1, \ldots, v_k$. A remedy is offered by working with a fixed number n of iterations and then to restart the algorithm with $x^{(0)} := x^{(n)}$ and $g^{(0)} := Ax^{(0)} - b$, until finally the convergence criterion is fulfilled (*GMRES method with restart*). There is also a *truncated* version of the GMRES method, in which only the last n basis vectors are used. The minimization of the error in the energy norm (on the vector space K) as with the CG method makes sense only for symmetric positive definite matrices A. But the variational equation

$$(Ay - b)^T z = 0 \quad \text{for all } z \in K$$

that characterizes this minimum in general can be taken as defining condition for y. Further variants of Krylov subspace methods rely on this. Another large class of such methods is founded on the *Lanczos biorthogonalization*, in which apart from a basis $v_1, \ldots, v_k$ of $K_k(A; v_1)$ another basis $w_1, \ldots, w_k$ of $K_k(A^T; w_1)$ is constructed, such that

$$v_j^T w_i = \delta_{ij} \quad \text{for} \quad i, j = 1, \ldots, k\,.$$

The best-known representative of this method is the *BICGSTAB method*. For further discussion of this topic see, for example, [28].

Exercises

5.10 Consider the linear system $Ax = b$, where $A = \alpha Q$ for some $\alpha \in \mathbb{R} \setminus \{0\}$ and some orthogonal matrix Q. Show that, for an arbitrary initial iterate $x^{(0)}$, the CGNE method terminates after one step with the exact solution.

5.11 Provided that Arnoldi's method can be performed up to the index k, show that it is possible to incorporate a convergence test of the GMRES method without computing the approximate solution explicitely, i.e., prove the following formulas:

$$g^{(k)} := Ax^{(k)} - b = h_{k+1,k} e_k^T \xi^{(k)} v_{k+1}\,,$$
$$|g^{(k)}|_2 = h_{k+1,k} |e_k^T \xi^{(k)}|\,.$$

5.5 The Multigrid Method

5.5.1 The Idea of the Multigrid Method

We discuss again the model problem of the five-point stencil discretization for the Poisson equation on the square and use the relaxed Jacobi's method. Then due to (5.31) the iteration matrix is

$$M = \omega M_J + (1 - \omega) I = I - \frac{\omega}{4} A\,,$$

with A being the stiffness matrix according to (1.14). For $\tilde{\omega} = \omega/4$ this coincides with the relaxed Richardson method, which according to (5.35) has the poor convergence behaviour of Jacobi's method, even for optimal choice of the parameter. Nevertheless, for a suitable ω the method has positive properties. Due to (5.25) the eigenvalues of M are

$$\lambda_{k,l} = 1 - \omega + \frac{\omega}{2}\left(\cos\frac{k\pi}{n} + \cos\frac{l\pi}{n}\right), \quad 1 \le k, l \le n-1\,.$$

This shows that there is a relation between the size of the eigenvalues and the position of the frequency of the assigned eigenfunction depending on the choice of ω: For $\omega = 1$, which is Jacobi's method, $\varrho(M) = \lambda_{1,1} = -\lambda_{n-1,n-1}$. Thus the eigenvalues are large if k and l are close to 1 or n. Hence there are large eigenvalues for eigenfunctions with low frequency as well as for eigenfunctions with high frequency. For $\omega = \frac{1}{2}$, however, we have $\varrho(M) = \lambda_{1,1}$, and the eigenvalues are large only in the case that k and l are near to 1, which means that the eigenfunctions have low frequency.

In general, if the error of the iterate $\boldsymbol{e}^{(k)}$ had a representation in terms of orthonormal eigenvectors z_ν with small eigenvalues, as for example $|\lambda_\nu| \le \frac{1}{2}$,

$$\boldsymbol{e}^{(k)} = \sum_{\nu:|\lambda_\nu|\le\frac{1}{2}} c_\nu z_\nu\,,$$

then according to (5.11) it would follow for the error measured in the Euclidean vector norm $|\cdot|_2$ that

$$\begin{aligned}\left|\boldsymbol{e}^{(k+1)}\right|_2 = \left|\sum_{\nu:|\lambda_\nu|\le\frac{1}{2}} \lambda_\nu c_\nu z_\nu\right|_2 &= \left(\sum_{\nu:|\lambda_\nu|\le\frac{1}{2}} \lambda_\nu^2 c_\nu^2\right)^{1/2} \\ &\le \frac{1}{2}\left(\sum_{\nu:|\lambda_\nu|\le\frac{1}{2}} c_\nu^2\right)^{1/2} = \frac{1}{2}\left|\boldsymbol{e}^{(k)}\right|_2\end{aligned}$$

if the eigenvectors are chosen orthonormal with respect to the Euclidean scalar product (compare (5.67)). For such an initial error and with exact arithmetic the method would thus have a "small" contraction number independent of the discretization.

For Jacobi's method damped by $\omega = \frac{1}{2}$ this means that if the initial error consists of functions of high frequency only (in the sense of an eigenvector expansion only of eigenvectors with k or l distant to 1), then the above considerations hold. But already due to rounding errors we will always find functions of low frequency in the error such that the above statement of convergence indeed does not hold, but instead the *smoothing property* for the damped Jacobi's method is valid: A few steps only lead to a low reduction of the error but smooth the error in the sense that the parts of high frequency are reduced considerably.

The very idea of the multigrid method lies in the approximative calculation of this remaining error on a coarse grid. The smooth error can still be represented on the coarser grid and should be approximated there. Generally, the dimension of the problem is greatly reduced in this way. Since the finite element discretizations are a central topic of this book, we develop the idea of multigrid methods for such an example. But it will turn out that the multigrid method can be used as well for both the finite difference and the finite volume methods. Multigrid methods have even been successfully used in areas other than the discretization of differential equations. *Algebraic multigrid methods* are generally applicable to systems of linear equations (5.1) and generate by themselves an abstract analogy of a "grid hierarchy" (see, for example, [65]).

5.5.2 Multigrid Method for Finite Element Discretizations

Let $\mathcal{T}_l = \mathcal{T}_h$ be a triangulation that originates from a coarse triangulation $\mathcal{T}_0$ by l applications of a refinement strategy, for example the strategy of Section 2.4.1. As we will see, it is not necessary that, for example, in two space dimensions going from $\mathcal{T}_k$ to $\mathcal{T}_{k+1}$ each triangle will be partitioned into four triangles. Only the relation

$$V_k \subset V_{k+1}\,, \quad k = 0, \ldots, l-1\,,$$

has to hold for finite-dimensional approximation spaces $V_0, V_1, \ldots, V_l = V_h$ generated by a fixed ansatz; i.e., the approximation spaces have to be *nested*. This holds for all approximation spaces discussed in Section 3.3 if $\mathcal{T}_{k+1}$ is still a conforming triangulation and results from $\mathcal{T}_k$ by partitioning of $K \in \mathcal{T}_k$ into a possibly varying number of elements of equal kind.

The nodes of $\mathcal{T}_k$, which are the degrees of freedom of the discretization (possibly multiple in a Hermite ansatz), are denoted by

$$a_i^k\,, \quad i = 1, \ldots, M_k\,,$$

and the corresponding basis functions of V_k are denoted by

$$\varphi_i^k\,, \quad i = 1, \ldots, M_k\,,$$

with the index $k = 0, \ldots, l$. For a quadratic ansatz on a triangle and Dirichlet boundary conditions the a_i^k are just the vertices and midpoints of the edges in the interior of the domain. Let the underlying variational equation (2.21) be defined by the bilinear form a and the linear form b on the function space V. The system of equations to be solved is

$$A_l \boldsymbol{x}_l = \boldsymbol{b}_l\,. \tag{5.88}$$

In addition, we have to consider auxiliary problems

$$A_k \overline{\boldsymbol{x}}_k = \overline{\boldsymbol{b}}_k$$

for $k = 0, \ldots, l-1$. For the discretization matrix on each refinement level we have, according to (2.34),

$$(A_k)_{ij} = a\big(\varphi_j^k, \varphi_i^k\big)\,, \quad i,j = 1, \ldots, M_k\,, \quad k = 0, \ldots, l\,,$$

and for the right side of the problem to be solved

$$(\boldsymbol{b}_l)_i = b\big(\varphi_i^l\big)\,, \quad i = 1, \ldots, M_l\,.$$

In Section 2.2, $\boldsymbol{x}_l$ is denoted by $\boldsymbol{\xi}$, and $\boldsymbol{b}_l$ is denoted by $\boldsymbol{q}_h$.

First we discuss the finite element discretization of a variational equation with symmetric bilinear form, so that in reference to Lemma 2.14 the Galerkin method to be solved is equivalent to the Ritz method, i.e., to the minimization of

$$F_l(\boldsymbol{x}_l) := \frac{1}{2}\boldsymbol{x}_l^T A_l \boldsymbol{x}_l - \boldsymbol{b}_l^T \boldsymbol{x}_l\,.$$

Note that l indicates the discretization level and is *not* an index of a component or an iteration step.

We distinguish between the function $u_l \in V_l$ and the representation vector $\boldsymbol{x}_l \in \mathbb{R}^{M_l}$, so that

$$u_l = \sum_{i=1}^{M_l} x_{l,i}\,\varphi_i^l\,. \tag{5.89}$$

For a Lagrange ansatz we have

$$x_{l,i} = u_l\big(a_i^l\big)\,, \quad i = 1, \ldots, M_l\,,$$

as illustrated by Figure 5.5.

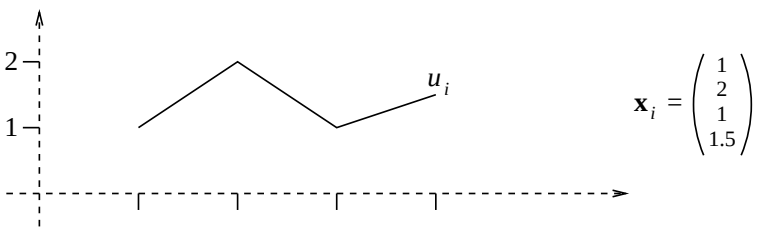


Figure 5.5. u_i and $\boldsymbol{x}_i$.

Relation (5.89) defines a linear bijective mapping

$$P_l \,:\, \mathbb{R}^{M_l} \to V_l\,. \tag{5.90}$$

Thus for $\boldsymbol{z}_l \in \mathbb{R}^{M_l}$ (compare (2.35)),

$$F_l(\boldsymbol{z}_l) = \frac{1}{2}\boldsymbol{z}_l^T A_l \boldsymbol{z}_l - \boldsymbol{b}_l^T \boldsymbol{z}_l = \frac{1}{2}a\left(P_l\boldsymbol{z}_l, P_l\boldsymbol{z}_l\right) - b(P_l\boldsymbol{z}_l) = F\left(P_l\boldsymbol{z}_l\right)\,,$$

where

$$F(u) := \frac{1}{2}a(u,u) - b(u) \quad \text{for } u \in V$$

Let $\boldsymbol{x}_l^{(k)}$ be the kth iterate to the solution of (5.88).

(1) **Smoothing step:** For fixed $\nu \in \{1, 2, \ldots\}$ calculate

$$\boldsymbol{x}_l^{(k+1/2)} = S_l^\nu \boldsymbol{x}_l^{(k)} .$$

Let the corresponding function be:

$$u_l^{(k+1/2)} = P_l \boldsymbol{x}_l^{(k+1/2)} \in V_l .$$

(2) **Coarse grid correction:** Solve (exactly)

$$F\left(u_l^{(k+1/2)} + v\right) \to \min \tag{5.93}$$

varying $v \in V_{l-1}$, with solution $\bar{v}_{l-1}$. Then set

$$\boldsymbol{x}_l^{(k+1)} = P_l^{-1}\left(u_l^{(k+1/2)} + \bar{v}_{l-1}\right) = \boldsymbol{x}_l^{(k+1/2)} + P_l^{-1}\bar{v}_{l-1} .$$

Table 5.5. $(k+1)$th step of the two-grid iteration.

is the energy functional for the variational equation.

If $\overline{\boldsymbol{x}}_l$ is an approximation of $\boldsymbol{x}_l$, then the error $\boldsymbol{y}_l := \boldsymbol{x}_l - \overline{\boldsymbol{x}}_l$ satisfies the *error equation*

$$A_l \boldsymbol{y}_l = \boldsymbol{b}_l - A_l \overline{\boldsymbol{x}}_l . \tag{5.91}$$

This equation is equivalent to the minimization problem

$$F_l\left(\overline{\boldsymbol{x}}_l + \boldsymbol{y}_l\right) = \min_{\boldsymbol{y} \in \mathbb{R}^{M_l}} F_l\left(\overline{\boldsymbol{x}}_l + \boldsymbol{y}\right)$$

and therefore to

$$F\left(P_l \overline{\boldsymbol{x}}_l + v_l\right) = \min_{v \in V_l} F\left(P_l \overline{\boldsymbol{x}}_l + v\right) , \tag{5.92}$$

with $v_l = P_l \boldsymbol{y}_l$.

If the error $\boldsymbol{y}_l$ is "smooth" in the sense that it can be well approximated also in the lower-dimensional space V_{l-1}, one can solve the error equation (5.91) approximately as part of an iteration step by solving the minimization problem (5.92) only on V_{l-1}. The starting condition of a "smooth" error will be ensured by the application of a fixed number of steps of a smoothing iteration method. Let S_l denote the application of such a smoothing operation, for example the damped Jacobi's method

$$S_l \boldsymbol{x} = \boldsymbol{x} - \omega D_l^{-1}\left(A_l \boldsymbol{x} - \boldsymbol{b}_l\right)$$

with the diagonal matrix D_l corresponding to A_l according to (5.18).

Thus we get the algorithm of the *two-grid iteration*, whose $(k+1)$th step is described in Table 5.5. Problem (5.93) from Table 5.5 is equivalent to (compare with Lemma 2.3)

$$a\left(u_l^{(k+1/2)} + v, w\right) = b(w) \quad \text{for all } w \in V_{l-1} \tag{5.94}$$

(1) **A priori smoothing:** Perform ν_1 smoothing steps:

$$\boldsymbol{x}_l^{(k+1/3)} = S_l^{\nu_1} \boldsymbol{x}_l^{(k)} ,$$

where $\nu_1 \in \{1, 2, \ldots\}$ is fixed. Let the corresponding function be

$$u_l^{(k+1/3)} := P_l \boldsymbol{x}_l^{(k+1/3)} .$$

(2) **Coarse grid correction:** Solve on V_{l-1} the Galerkin discretization

$$a(\bar{v}_{l-1}, w) = \tilde{b}(w) \quad \text{for all } w \in V_{l-1} \tag{5.95}$$

with the bilinear form a and the linear form

$$\tilde{b}(w) := b(w) - a\left(u_l^{(k+1/3)}, w\right)$$

(a) for $l = 1$ exactly,
(b) for $l > 1$ by μ steps of a multigrid iteration on level $l-1$ for a and $\tilde{b}$ and for the start approximation $\mathbf{0}$.

Set $$\boldsymbol{x}_l^{(k+2/3)} = \boldsymbol{x}_l^{(k+1/3)} + P_l^{-1} \bar{v}_{l-1}.$$

(3) **A posteriori smoothing:** Perform ν_2 smoothing steps

$$\boldsymbol{x}_l^{(k+1)} = S_l^{\nu_2} \boldsymbol{x}_l^{(k+2/3)} ,$$

with $\nu_2 \in \{1, 2, \ldots\}$ fixed.

Table 5.6. $(k+1)$th step of the multigrid iteration on level l for bilinear form a and linear form b.

and thus again to the Galerkin discretization of a variational equation with V_{l-1} instead of V, with the same bilinear form and with a linear form defined by

$$w \mapsto b(w) - a\left(u_l^{(k+1/2)}, w\right) \quad \text{for } w \in V_{l-1} .$$

Hence we can ignore the assumption of symmetry for the bilinear form a and find the approximative solution of the error equation (5.91) on grid level $l-1$ by solving the variational equation (5.94). The equivalent system of equations will be derived in the following. On the one hand, this problem has a lower dimension than the original problem, but it also must be solved for each iteration. This suggests the following recursive procedure: If we have more than two grid levels, we again approximate this variational equation by μ multigrid iterations; in the same way we treat the hereby created Galerkin discretization on level $l-2$ until level 0 is reached, where we solve exactly. Furthermore, to conclude each iteration step smoothing steps should be performed. This leads to the algorithm of the multigrid iteration. The $(k+1)$th step of the *multigrid iteration on level l for the bilinear form a, linear form b*, and starting iteration $\boldsymbol{x}_l^{(k)}$ is described in Table 5.6.

In general, $\nu_1 = \nu_2$ is used. In a convergence analysis it turns out that only the sum of smoothing steps is important. Despite the recursive definition of a multigrid iteration we have here a finite method, because the level 0 is reached after at most l recursions, where the auxiliary problem will be solved exactly. For μ usually only the values $\mu = 1$ or $\mu = 2$ are used. The terms *V-cycle* for $\mu = 1$ and *W-cycle* for $\mu = 2$ are commonly used, because for an iteration, the sequence of levels on which operations are executed have the shape of these letters (see Figure 5.6).

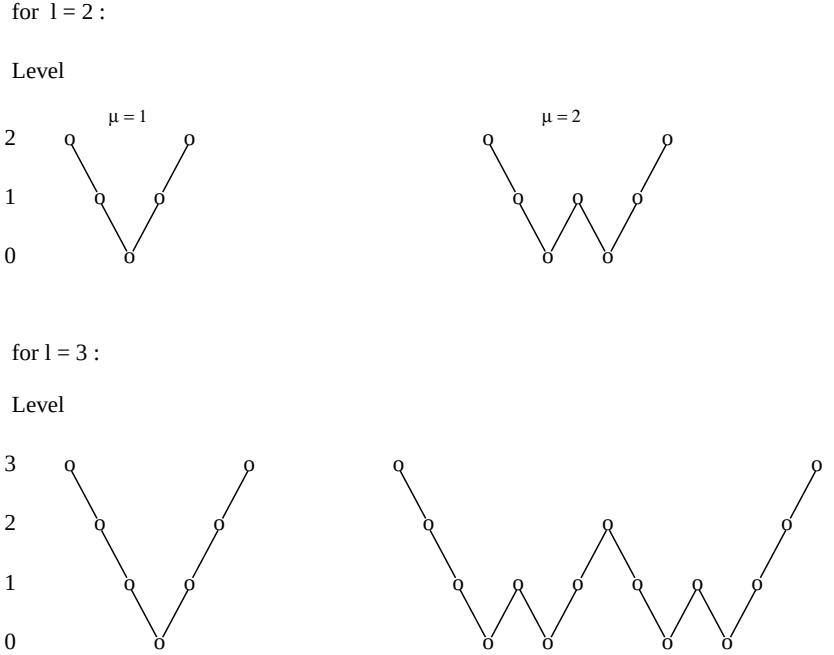


Figure 5.6. Grid levels for the V-cycle ($\mu = 1$) and the W-cycle ($\mu = 2$).

The problems in (5.94) and (5.95) (see Table 5.6) have the form

$$a\,(u + v, w) = b(w) \quad \text{for all } w \in V_{l-1}\,, \tag{5.96}$$

where $v \in V_{l-1}$ is unknown and $u \in V_l$ is known. An equivalent system of equations arises by inserting the basis functions φ_j^{l-1}, $j = 1, \ldots, M_{l-1}$, for w and an appropriate representation for v. If we again take the representation with respect to φ_j^{l-1}, we get as in (2.34)

$$A_{l-1} P_{l-1}^{-1} v = \boldsymbol{d}_{l-1}\,. \tag{5.97}$$

Here the *residual* $\boldsymbol{d}_k \in \mathbb{R}^{M_k}$ of u on the different levels $k = 0, \ldots, l$ is defined by

$$d_{k,i} := b\big(\varphi_i^k\big) - a\big(u, \varphi_i^k\big)\,, \quad i = 1, \ldots, M_k\,.$$

We now develop an alternative representation for (5.97) and the coarse grid correction for possible generalizations beyond the Galerkin approximations. Therefore, let $R \in \mathbb{R}^{M_{l-1},M_l}$ be the matrix that arises through the unique representation of the basis functions φ_j^{l-1} with respect to the basis φ_i^l, which means the elements r_{ji} of R are determined by the equations

$$\varphi_j^{l-1} = \sum_{i=1}^{M_l} r_{ji}\, \varphi_i^l\,, \quad j = 1, \ldots, M_{l-1}\,.$$

Then (5.96) is equivalent to

$$\begin{aligned}
& a(v,w) = b(w) - a(u,w) \quad \text{for all } w \in V_{l-1} \\
\Leftrightarrow\ & a\left(\sum_{s=1}^{M_{l-1}} \left(P_{l-1}^{-1} v\right)_s \varphi_s^{l-1}, \varphi_j^{l-1}\right) = b\big(\varphi_j^{l-1}\big) - a\big(u, \varphi_j^{l-1}\big)\,,\ j = 1, \ldots, M_{l-1} \\
\Leftrightarrow\ & \sum_{s=1}^{M_{l-1}} \left(P_{l-1}^{-1} v\right)_s a\left(\sum_{t=1}^{M_l} r_{st}\varphi_t^l\,, \sum_{i=1}^{M_l} r_{ji}\varphi_i^l\right) = \sum_{i=1}^{M_l} r_{ji}\left(b\big(\varphi_i^l\big) - a\big(u, \varphi_i^l\big)\right) \\
\Leftrightarrow\ & \sum_{s=1}^{M_{l-1}} \sum_{i,t=1}^{M_l} r_{ji} a\big(\varphi_t^l, \varphi_i^l\big) r_{st} \left(P_{l-1}^{-1} v\right)_s = (R\boldsymbol{d}_l)_j\,, \quad j = 1, \ldots, M_{l-1}\,.
\end{aligned}$$

Hence the system of equations has the form

$$R A_l R^T \left(P_{l-1}^{-1} v\right) = R\boldsymbol{d}_l. \tag{5.98}$$

The matrix R is easy to calculate for a node-based basis φ_i^l satisfying $\varphi_i^l\left(a_j^l\right) = \delta_{ij}$, since in this case we have for $v \in V_l$,

$$v = \sum_{i=1}^{M_l} v\big(a_i^l\big)\, \varphi_i^l\,,$$

and therefore in particular,

$$\varphi_j^{l-1} = \sum_{i=1}^{M_l} \varphi_j^{l-1}\big(a_i^l\big)\, \varphi_i^l$$

and thus

$$r_{ji} = \varphi_j^{l-1}\big(a_i^l\big)\,.$$

For the linear ansatz in one space dimension with Dirichlet boundary conditions (i.e., with $V = H_0^1(a,b)$ as basic space) this means that

$$R = \begin{pmatrix}
\frac{1}{2} & 1 & \frac{1}{2} & & & & \\
& \frac{1}{2} & 1 & \frac{1}{2} & & & \\
& & & \ddots & & & \\
& & & & \frac{1}{2} & 1 & \frac{1}{2}
\end{pmatrix}. \tag{5.99}$$

The representation (5.98) can also be interpreted in this way:
Due to $V_{l-1} \subset V_l$ the identify defines a natural *prolongation* from V_{l-1} to V_l, which means that

$$\tilde{p} : V_{l-1} \to V_l \,, \; v \mapsto v \,,$$

as illustrated by Figure 5.7.

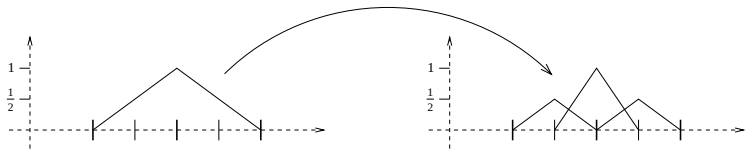


Figure 5.7. Prolongation.

This prolongation corresponds to a prolongation p from $\mathbb{R}^{M_{l-1}}$ to $\mathbb{R}^{M_l}$, the *canonical prolongation*, through the transition to the representation vectors (5.90). It is given by

$$p := P_l^{-1} P_{l-1} \,, \tag{5.100}$$

since for $\boldsymbol{x}_{l-1} \in \mathbb{R}^{M_{l-1}}$, p can be composed as follows:

$$\boldsymbol{x}_{l-1} \mapsto P_{l-1}\boldsymbol{x}_{l-1} \overset{\tilde{p}}{\mapsto} P_{l-1}\boldsymbol{x}_{l-1} \mapsto P_l^{-1} P_{l-1}\boldsymbol{x}_{l-1} \,.$$

Obviously, p is linear and can be identified with its matrix representation in $\mathbb{R}^{M_l, M_{l-1}}$. Then

$$p = R^T \tag{5.101}$$

holds, because

$$P_{l-1}\boldsymbol{y} = \sum_{j=1}^{M_{l-1}} y_j \varphi_j^{l-1} = \sum_{i=1}^{M_l} \sum_{j=1}^{M_{l-1}} y_j r_{ji} \varphi_i^l \,,$$

i.e., $R^T \boldsymbol{y} = P_l^{-1} \left(P_{l-1} \boldsymbol{y} \right)$ for any $\boldsymbol{y} \in \mathbb{R}^{M_{l-1}}$.

In the following $\mathbb{R}^{M_l}$ will be endowed with a scalar product $\langle \cdot, \cdot \rangle^{(l)}$, which is an Euclidean scalar product scaled by a factor S_l,

$$\langle \boldsymbol{x}_l, \boldsymbol{y}_l \rangle^{(l)} := S_l \sum_{i=1}^{M_l} x_{l,i} \, y_{l,i} \,. \tag{5.102}$$

The scaling factor is to be chosen such that for the induced norm $\| \cdot \|_l$ and the $L^2(\Omega)$-norm on V_l,

$$C_1 \left\| P_l \boldsymbol{x}_l \right\|_0 \leq \left\| \boldsymbol{x}_l \right\|_l \leq C_2 \left\| P_l \boldsymbol{x}_l \right\|_0 \tag{5.103}$$

for $\boldsymbol{x} \in \mathbb{R}^{M_l}$, $l = 0, 1, \ldots$, with constants C_1, C_2 independent of l: If the triangulations are members of a regular and quasi-uniform family $\mathcal{T}_h$ (see

Definition 3.28), then in d space dimensions one can choose $S_l = h_l^d$, with h_l being the maximal diameter of $K \in \mathcal{T}_l$ (see Theorem 3.43).

Let $r : \mathbb{R}^{M_l} \to \mathbb{R}^{M_{l-1}}$ be defined by

$$r = p^* \,, \tag{5.104}$$

with the *adjoint* p^* defined with respect to the scalar products $\langle \cdot, \cdot \rangle^{(l-1)}$ and $\langle \cdot, \cdot \rangle^{(l)}$; that is,

$$\left\langle r\, \boldsymbol{x}_l \,,\, \boldsymbol{y}_{l-1} \right\rangle^{(l-1)} = \left\langle p^* \, \boldsymbol{x}_l \,,\, \boldsymbol{y}_{l-1} \right\rangle^{(l-1)} = \left\langle \boldsymbol{x}_l \,,\, p\, \boldsymbol{y}_{l-1} \right\rangle^{(l)} \,.$$

If p is the canonical prolongation, then r is called the *canonical restriction*. For the representation matrices,

$$\frac{S_{l-1}}{S_l} r = p^T = R \,. \tag{5.105}$$

In example (5.102) for $d = 2$ with $h_l = h_{l-1}/2$ we have $S_{l-1}/S_l = 1/4$. Due to $P_l p = P_{l-1}$, the canonical restriction of $\mathbb{R}^{M_l}$ on $\mathbb{R}^{M_{l-1}}$ satisfies

$$r R_l = R_{l-1} \,,$$

where $R_l : V_l \to \mathbb{R}^{M_l}$ is defined as the adjoint of P_l,

$$\langle P_l \boldsymbol{x}_l, v_l \rangle_0 = \langle \boldsymbol{x}_l, R_l v_l \rangle^{(l)} \quad \text{for all } \boldsymbol{x}_l \in \mathbb{R}^{M_l} \,,\ v_l \in V_l \,,$$

because for any $\boldsymbol{y}_{l-1} \in \mathbb{R}^{M_{l-1}}$ and for $v_{l-1} \in V_{l-1} \subset V_l$,

$$\begin{aligned} \left\langle r R_l v_{l-1} \,,\, \boldsymbol{y}_{l-1} \right\rangle^{(l-1)} &= \left\langle R_l v_{l-1} \,,\, p \boldsymbol{y}_{l-1} \right\rangle^{(l)} = \left\langle v_{l-1} \,,\, P_l p \boldsymbol{y}_{l-1} \right\rangle_0 \\ &= \left\langle v_{l-1}, P_{l-1} \boldsymbol{y}_{l-1} \right\rangle_0 \ = \left\langle R_{l-1} v_{l-1}, \boldsymbol{y}_{l-1} \right\rangle^{(l-1)} . \end{aligned}$$

Using (5.105) we see the equivalence of equation (5.98) to

$$(r A_l p) \boldsymbol{y}_{l-1} = r \boldsymbol{d}_l \,. \tag{5.106}$$

Setting $v := P_{l-1} \tilde{\boldsymbol{y}}_{l-1}$ for a perhaps only approximative solution $\tilde{\boldsymbol{y}}_{l-1}$ of (5.106), the coarse grid correction will be finished by addition of $P_l^{-1} v$. Due to

$$P_l^{-1} v = P_l^{-1} P_{l-1} \left(P_{l-1}^{-1} v \right) = p \left(P_{l-1}^{-1} v \right) \,,$$

the coarse grid correction is

$$\boldsymbol{x}_l^{(k+2/3)} = \boldsymbol{x}_l^{(k+1/3)} + p(\tilde{\boldsymbol{y}}_{l-1}) \,.$$

The above-mentioned facts suggest the following structure of a general multigrid method: For discretizations defining a hierarchy of discrete problems,

$$A_l \boldsymbol{x}_l = \boldsymbol{b}_l \,,$$

one needs *prolongations*

$$p : \mathbb{R}^{M_{k-1}} \to \mathbb{R}^{M_k}$$

and *restrictions*

$$r : \mathbb{R}^{M_k} \to \mathbb{R}^{M_{k-1}}$$

for $k = 1, \ldots, l$ and the matrices $\tilde{A}_{k-1}$ for the error equations. The coarse grid correction steps (5.93) and (5.95) hence take the following form:

Solve (with μ steps of the multigrid method)

$$\tilde{A}_{l-1} \boldsymbol{y}_{l-1} = r\Big(\boldsymbol{b}_l - A_l \boldsymbol{x}_l^{(k+1/3)}\Big)$$

and set

$$\boldsymbol{x}_l^{(k+2/3)} = \boldsymbol{x}_l^{(k+1/3)} + p\boldsymbol{y}_{l-1}\,.$$

The above choice

$$\tilde{A}_{l-1} = rA_l p$$

is called the *Galerkin product*. For Galerkin approximations this coincides with the discretization matrix of the same type on the grid of level $l-1$ due to (5.97). This is also a common choice for other discretizations and then an alternative to the Galerkin product. In view of the choice of p and r we should observe the validity of (5.104). An interpolational definition of the prolongation on the basis of (finite element) basis functions as for example (5.101) (see also example (5.99)) is also common in other discretizations. In more difficult problems, as for example those with (dominant) convection in addition to diffusive transport processes, nonsymmetric problems arise with a small constant of V-ellipticity. Here the use of *matrix-dependent*, that means A_l-dependent, prolongations and restrictions is recommended.

5.5.3 Effort and Convergence Behaviour

In order to judge the efficiency of a multigrid method the number of operations per iteration and the number of iterations (required to reach an error level ε, see (5.4)) has to be estimated. Due to the recursive structure, the first number is not immediately clear. The aim is to have only the optimal amount of $O(M_l)$ operations for sparse matrices. For this the dimensions of the auxiliary problems have to decrease sufficiently. This is expressed by the following:

There exists a constant $C > 1$ such that

$$M_{l-1} \leq M_l / C \quad \text{for } l \in \mathbb{N}\,. \tag{5.107}$$

Hence we assume an infinite hierarchy of problems and/or grids, which also corresponds to the asymptotic point of view of a discretization from Section 3.4. Relation (5.107) is thus a condition for a refinement strategy. For the model problem of the Friedrichs–Keller triangulation of a rectangle (see Figure 2.9) in the case of a regular "red" refinement we have $h_l = h_{l-1}/2$. Thus $C = 4$, and for analogous constructions in d space dimensions

$C = 2^d$. The matrices that appear should be sparse, so that for level l the following holds:

$$\begin{array}{lcl} \text{smoothing step} & = & C_S M_l \text{ operations,} \\ \text{error calculation and restrictions} & = & C_D M_l \text{ operations,} \\ \text{prolongation and correction} & = & C_C M_l \text{ operations.} \end{array}$$

Then we can prove the following (see [16, p. 326]):

If the number μ of multigrid steps in the recursion satisfies

$$\mu < C\,, \tag{5.108}$$

then the number of operations for an iteration step for a problem on level l can be estimated by

$$C(\nu) M_l\,. \tag{5.109}$$

Here ν is the number of a priori and a posteriori smoothing steps and

$$C(\nu) = \frac{\nu C_S + C_D + C_S}{1 - \mu/C} + O\big((\mu/C)^l\big)\,.$$

The requirement (5.108) will be satisfied in general through the restriction to $\mu = 1$, $\mu = 2$. Analogously, the memory requirement is $O(M_l)$, since

$$\sum_{k=0}^{l} M_k \leq \frac{C}{C-1} M_l\,.$$

Whether this larger effort (of equal complexity) in comparison to other methods discussed is justified will be decided by the rate of convergence. The multigrid method is a linear stationary method. The iteration matrix M_l^{TGM} of the two-grid method results from

$$\begin{array}{rcl} \boldsymbol{x}_l^{(k+1/2)} & = & S_l^\nu \boldsymbol{x}_l^{(k)}\,, \\ \boldsymbol{x}_l^{(k+1)} & = & \boldsymbol{x}_l^{(k+1/2)} + p\left(A_{l-1}^{-1}\left(r\left(\boldsymbol{b}_l - A_l \boldsymbol{x}_l^{(k+1/2)}\right)\right)\right) \end{array}$$

to

$$M_l^{TGM} = (I - p A_{l-1}^{-1} r A_l) S_l^\nu\,. \tag{5.110}$$

Also, the consistency of the method follows immediately if the smoothing iteration is consistent.

The analysis of the convergence of the multigrid method can be reduced to the analysis of the two-grid method, since the iteration matrix is a modification of M_l^{TGM} (see [16, p. 328]). For a large class of a priori and a posteriori smoothing operators as well as of restrictions and prolongations it can be shown (see [16, p. 347]) that there exists a constant $\overline{\varrho} \in (0,1)$ independent of the discretization parameter h_l such that $\varrho(M_{TGM}) \leq \overline{\varrho}$. Combined with (5.109) this shows that multigrid methods are optimal in their complexity. This also shows their potential superiority compared with all other methods described.

In the following we will only indicate the schematic procedure to prove this assertion. It is sufficient to prove the following two properties, where the spectral norm is used as the matrix norm, that is, the matrix norm that is induced by the Euclidean vector norm.

(1) *Smoothing property:*

$$\text{There exists } C_S > 0: \quad \left\|A_l S_l^\nu\right\| \le \frac{C_S}{\nu} \left\|A_l\right\|.$$

(2) *Approximation property:*

$$\text{There exists } C_A > 0: \quad \left\|A_l^{-1} - pA_{l-1}^{-1} r\right\| \le C_A \|A_l\|^{-1}. \tag{5.111}$$

Due to

$$M_{TGM} = \left(A_l^{-1} - pA_{l-1}^{-1} r\right) A_l S_l^\nu,$$

we can conclude that

$$\left\|M_{TGM}\right\| \le \left\|A_l^{-1} - pA_{l-1}^{-1} r\right\| \left\|A_l S_l^\nu\right\| \le \frac{C_S C_A}{\nu},$$

which means that for sufficiently large ν,

$$\|M_{TGM}\| \le \overline{\varrho} < 1$$

with $\overline{\varrho}$ independent of l.

The smoothing property is of an algebraic nature, but for the proof of the approximation property we will use — at least indirectly — the original variational formulation of the boundary value problem and the corresponding error estimate. Therefore, we discuss only the smoothing property for, as an example, the relaxed Richardson method for a symmetric positive definite matrix A_l, i.e.,

$$S_l = I_l - \omega A_l \quad \text{with} \quad \omega \in \left(0, \frac{1}{\lambda_{\max}(A_l)}\right].$$

Let $\{\boldsymbol{z}_i\}_{i=1}^{M_l}$ be an orthonormal basis of eigenvectors of A_l. For any initial vector $\boldsymbol{x}^{(0)}$ represented in this basis as $\boldsymbol{x}^{(0)} = \sum_{i=1}^{M_l} c_i \boldsymbol{z}_i$ it follows that (compare (5.68))

$$\begin{aligned}\left\|A_l S_l^\nu \boldsymbol{x}^{(0)}\right\|^2 = \sum_{i=1}^{M_l} \lambda_i^2 (1-\lambda_i\omega)^{2\nu} c_i^2 &= \omega^{-2} \sum_{i=1}^{M_l} (\lambda_i\omega)^2 (1-\lambda_i\omega)^{2\nu} c_i^2 \\ &\le \omega^{-2} \left[\max_{\xi\in[0,1]} \xi(1-\xi)^\nu\right]^2 \sum_{i=1}^{M_l} c_i^2.\end{aligned}$$

The function $\xi \mapsto \xi(1-\xi)^\nu$ has its maximum at $\xi_{\max} = (\nu+1)^{-1}$; thus

$$\xi_{\max}(1-\xi_{\max})^\nu = \frac{1}{\nu+1}\left(1 - \frac{1}{\nu+1}\right)^\nu = \frac{1}{\nu}\left(\frac{\nu}{\nu+1}\right)^{\nu+1} \le \frac{1}{\mathrm{e}\nu}.$$

Hence

$$\left\|A_l S_l^\nu \boldsymbol{x}^{(0)}\right\| \leq \frac{1}{\omega \mathrm{e} \nu} \left\|\boldsymbol{x}^{(0)}\right\|,$$

which implies

$$\left\|A_l S_l^\nu\right\| \leq \frac{1}{\omega \mathrm{e} \nu}.$$

Since the inclusion $\omega \in (0, 1/\lambda_{\max}(A_l)]$ can be written in the form $\omega = \sigma/\|A_l\|$ with $\sigma \in (0, 1]$, we have $C_S = 1/(\sigma \mathrm{e})$.

The approximation property can be motivated in the following way. The fine grid solution $\boldsymbol{x}_l$ of $A_l \boldsymbol{x}_l = \boldsymbol{d}_l$ is replaced in the coarse grid correction by $p\boldsymbol{x}_{l-1}$ from $A_{l-1}\boldsymbol{x}_{l-1} = \boldsymbol{d}_{l-1} := r\boldsymbol{d}_l$. Therefore, $p\boldsymbol{x}_{l-1} \approx A_l^{-1}\boldsymbol{d}_l$ should hold. The formulation (5.111) thus is just a quantitative version of this requirement. Since in the symmetric case $\|A_l\|^{-1}$ is simply the reciprocal value of the largest eigenvalue, (3.140) in Theorem 3.45 establishes the relation to the statements of convergence in Section 3.4. For a more exact analysis of convergence and a more extensive description of this topic we refer to the cited literature (see also [17]).

Exercises

5.12 Determine the prolongation and restriction according to (5.101) and (5.104) for the case of a linear ansatz on a Friedrichs–Keller triangulation.

5.13 Prove the consistency of the two-grid method (5.110) in the case of the consistent smoothing property.

5.6 Nested Iterations

As in Section 5.5 we assume that besides the system of equations

$$A_l \boldsymbol{x}_l = \boldsymbol{b}_l$$

with M_l unknowns, there are given analogous low-dimensional systems of equations

$$A_k \boldsymbol{x}_k = \boldsymbol{b}_k, \quad k = 0, \ldots, l-1, \tag{5.112}$$

with M_k unknowns, where $M_0 < M_1 < \cdots < M_l$. Let all systems of equations be an approximation of the same continuous problem such that an error estimate of the type

$$\|u - P_l \boldsymbol{x}_l\| \leq C_A h_l^\alpha$$

holds, with P_l according to (5.90) and $\alpha > 0$. Here $\|\cdot\|$ is a norm on the basic space V, and the constant C_A generally depends on the solution u

of the continuous problem. The discretization parameter h_l determines the dimension M_l: In the simplest case of a uniform refinement, $h_l^d \sim 1/M_l$ holds in d space dimensions. One may also expect that for the discrete solution,

$$\|p\boldsymbol{x}_{k-1} - \boldsymbol{x}_k\|_k \le C_1 C_A h_k^\alpha\,, \quad k = 1, \ldots, l\,,$$

holds with a constant $C_1 > 0$. Here $\|\cdot\|_k$ is a norm on $\mathbb{R}^{M_k}$, and the mapping $p = p_{k-1,k} : \mathbb{R}^{M_{k-1}} \to \mathbb{R}^{M_k}$ is a prolongation, for example the canonical prolongation introduced in Section 5.5. In this case the estimate can be rigorously proven with the definition of the canonical prolongation $p = P_k^{-1} P_{k-1}$:

$$\begin{aligned} \|p\boldsymbol{x}_{k-1} - \boldsymbol{x}_k\|_k &= \left\|P_k^{-1}\left(P_{k-1}\boldsymbol{x}_{k-1} - P_k\boldsymbol{x}_k\right)\right\|_k \\ &\le \left\|P_k^{-1}\right\|_{L[V_k,\mathbb{R}^{M_k}]} \left\|P_{k-1}\boldsymbol{x}_{k-1} - P_k\boldsymbol{x}_k\right\| \\ &\le \left\|P_k^{-1}\right\|_{L[V_k,\mathbb{R}^{M_k}]} \left(C_A h_k^\alpha + C_A h_{k-1}^\alpha\right) \;\le\; C_1 C_A h_k^\alpha \end{aligned}$$

with

$$C_1 = \max_{j=1,\ldots,l} \left\{ \left\|P_j^{-1}\right\|_{L[V_j,\mathbb{R}^{M_j}]} \left(1 + \left(\frac{h_{j-1}}{h_j}\right)^\alpha\right)\right\}.$$

Let the system of equations be solved with an iterative method given by the fixed-point mapping Φ_k, $k = 0, \ldots, l$, which means that $\boldsymbol{x}_k$ according to (5.112) satisfies $\boldsymbol{x}_k = \Phi_k(\boldsymbol{x}_k, \boldsymbol{b}_k)$. Then it is sufficient to determine an iterate $\tilde{\boldsymbol{x}}_l$ with an accuracy

$$\|\tilde{\boldsymbol{x}}_l - \boldsymbol{x}_l\|_l \le \tilde{C}_A h_l^\alpha \tag{5.113}$$

with $\tilde{C}_A := C_A / \|P_l\|_{L[\mathbb{R}^{M_l},V]}$, because then we also have

$$\|P_l\tilde{\boldsymbol{x}}_l - P_l\boldsymbol{x}_l\| \le C_A h_l^\alpha.$$

If one does not have a good initial iterate from the concrete context, the algorithm of *nested iterations* explained in Table 5.7 can be used. It is indeed a finite process.

The question is how to choose the iteration numbers m_k such that (5.113) finally holds, and whether the arising overall effort is acceptable. An answer to this question is provided by the following theorem:

Theorem 5.19 *Let the iterative method Φ_k have the contraction number ϱ_k with respect to $\|\cdot\|_k$. Assume that there exist constants $C_2, C_3 > 0$ such that*

$$\begin{aligned} \|p\|_{L[\mathbb{R}^{M_{k-1}},\mathbb{R}^{M_k}]} &\le C_2\,, \\ h_{k-1} &\le C_3 h_k\,, \end{aligned}$$

for all $k = 1, \ldots, l$. If the iteration numbers m_k for the nested iterations are chosen in such a way that

$$\varrho_k^{m_k} \le 1/(C_2 C_3^\alpha + C_1\|P_l\|)\,, \tag{5.114}$$

Choose m_k, $k = 1, \ldots, l$.
Let $\tilde{\boldsymbol{x}}_0$ be an approximation of $\boldsymbol{x}_0$,

$$\text{for example} \quad \tilde{\boldsymbol{x}}_0 \quad = \quad \boldsymbol{x}_0 = A_0^{-1}\boldsymbol{b}_0 \,.$$

For $k = 1, \ldots, l$:

$$\tilde{\boldsymbol{x}}_k^{(0)} \quad := \quad p\,\tilde{\boldsymbol{x}}_{k-1} \,.$$

Perform m_k iterations:

$$\tilde{\boldsymbol{x}}_k^{(i)} \quad := \quad \Phi_k\big(\tilde{\boldsymbol{x}}_k^{(i-1)}, \boldsymbol{b}_k\big)\,, i = 1, \ldots, m_k \,.$$

Set $\tilde{\boldsymbol{x}}_k := \tilde{\boldsymbol{x}}_k^{(m_k)}$.

Table 5.7. Nested Iteration.

then

$$\|\tilde{\boldsymbol{x}}_k - \boldsymbol{x}_k\|_k \le \tilde{C}_A h_k^\alpha \,,$$

for all $k = 1, \ldots, l$, provided that this estimate holds for $k = 0$.

Proof: The proof is given by induction on k. Assume that the assertion is true for $k-1$. This induces

$$\begin{aligned}
\|\tilde{\boldsymbol{x}}_k - \boldsymbol{x}_k\|_k &\le \varrho_k^{m_k} \|p\tilde{\boldsymbol{x}}_{k-1} - \boldsymbol{x}_k\|_k \\
&\le \varrho_k^{m_k} \left(\|p(\tilde{\boldsymbol{x}}_{k-1} - \boldsymbol{x}_{k-1})\|_k + \|p\boldsymbol{x}_{k-1} - \boldsymbol{x}_k\|_k\right) \\
&\le \varrho_k^{m_k} \left(C_2 \tilde{C}_A h_{k-1}^\alpha + C_1 C_A h_k^\alpha\right) \\
&\le \varrho_k^{m_k} \left(C_2 C_3^\alpha + C_1 \|P_l\|\right) \tilde{C}_A h_k^\alpha \,.
\end{aligned}$$

□

Theorem 5.19 allows the calculation of the necessary number of iterations for the inner iteration from the norms $\|p\|_{L[\mathbb{R}^{M_{k-1}}, \mathbb{R}^{M_k}]}$, $\|P_k^{-1}\|_{L[V_k, \mathbb{R}^{M_k}]}$ and the constants $\frac{h_{k-1}}{h_k}$ for $k = 1, \ldots, l$, as well as the order of convergence α of the discretization.

In order to estimate the necessary effort according to (5.114) more exactly, the dependence of ϱ_k of k must be known. In the following we consider only the situation, known as the multigrid method, of a method of optimal complexity

$$\varrho_k \le \overline{\varrho} < 1 \,.$$

Here, in contrast to other methods, the number of iterations can be chosen constant ($m_k = m$ for all $k = 1, \ldots, l$). If, furthermore, the estimate (5.107) holds with the constant C, then analogously to the consideration in Section 5.5 the total number of operations for the nested iteration can

be estimated by

$$m \frac{C}{C-1} \overline{C} M_l \,.$$

Here $\overline{C} M_k$ is the number of operations for an iteration with the iteration method Φ_k.

In the model problem of the Friedrichs–Keller triangulation with uniform refinement we have $C/(C-1) = 4/3$ and $C_3 = 2$. For $\|\cdot\| = \|\cdot\|_0$ as basic norm, $\alpha = 2$ is a typical case according to Theorem 3.37. The existence of the constant C_2 will hereby finally be ensured consistently by the condition (5.103), observing (5.100). Assuming also that the constants $C_1, C_2, \|P_l\|$ are "small" and the iteration method has a "small" contraction number ϱ, only a small number of iterations m is necessary, in the ideal case $m = 1$. At least in this situation we can count on only a small increase of the necessary effort through the process of nested iterations, which provides an "appropriate" approximation $\tilde{\boldsymbol{x}}_k$ on all levels k of discretization.

Finally, it is to be observed that the sequence of the discrete problems has to be defined only during the process of the nested iteration. This offers the possibility to combine it with a posteriori error estimators as discussed in Section 4.2, in order to develop a grid $\mathcal{T}_{k+1}$ on which the discrete problem of level $k+1$ is determined, on the basis of $\tilde{\boldsymbol{x}}_k$ as a refinement of $\mathcal{T}_k$.

6
The Finite Volume Method

Finite volume methods are widely applied when differential equations in divergence form (cf. Section 0.5) or differential equations involving such differential expressions (for example, parabolic differential equations) are to be solved numerically. In the class of second-order linear elliptic differential equations, expressions of the form

$$Lu := -\nabla \cdot (K \nabla u - c\, u) + r\, u = f \tag{6.1}$$

are typical (cf. (0.33)), where

$$K : \Omega \to \mathbb{R}^{d,d}, \quad c : \Omega \to \mathbb{R}^d, \quad r, f : \Omega \to \mathbb{R}.$$

The corresponding "parabolic version" is

$$\frac{\partial u}{\partial t} + Lu = f$$

and will be treated in Chapter 7.

First-order partial differential equations such as the classical conservation laws

$$\nabla \cdot q(u) = 0,$$

where $q : \mathbb{R} \to \mathbb{R}^d$ is a nonlinear vector field depending on u, or higher-order partial differential equations (such as the biharmonic equation (3.36)), or even systems of partial differential equations can be successfully discretized by the finite volume method.

In correspondence to the comparatively large class of problems that can be treated by the finite volume method, there are rather different sources

1960	Forsythe and Wasow	computation of neutron diffusion
1961	Marčuk	computation of nuclear reactors
1971	McDonald	fluid mechanics
1972	MacCormack and Paullay	fluid mechanics
1973	Rizzi and Inouye	fluid mechanics in 3D
1977	Samarski	integro-interpolation method, balance method
	⋮	
1979	Jameson	finite volume method
1984	Heinrich	integro-balance method, generalized finite difference method
	⋮	
1987	Bank and Rose	box method
	⋮	

Table 6.1. Some sources of the finite volume method.

originating mainly from practical applications. Some of these sources are listed in Table 6.1. In contrast to finite difference or finite element methods, the theoretical understanding of the finite volume method remained at an early stage for a long time; only in recent years has essential progress been noted.

The finite volume method can be viewed as a discretization method of its own right. It includes ideas from both finite difference and finite element methods. So in the literature approaches can be found that interpret it as a "generalized finite difference method" or rather as a variant of the finite element method. In this chapter, we will consider only equations of the type (6.1).

6.1 The Basic Idea of the Finite Volume Method

Now we will describe the fundamental steps in the derivation of the finite volume method. For simplicity, we restrict ourselves to the case $d = 2$ and $r = 0$. Furthermore, we set $q(u) := -K\,\nabla u + c\,u$. Then equation (6.1) becomes

$$\nabla \cdot q(u) = f\,. \tag{6.2}$$

In order to obtain a finite volume discretization, the domain Ω will be subdivided into M subdomains Ω_i such that the collection of all those subdomains forms a *partition* of Ω, that is:

(1) each Ω_i is an open, simply connected, and polygonally bounded set without slits,

(2) $\Omega_i \cap \Omega_j = \emptyset \quad (i \neq j)$,

(3) $\cup_{i=1}^{M} \overline{\Omega_i} = \overline{\Omega}$.

These subdomains Ω_i are called *control volumes* or *control domains.*

Without going into more detail we mention that there also exist finite volume methods with a well-defined overlapping of the control volumes (that is, condition 2 is violated).

The next step, which is in common with all finite volume methods, consists in integrating equation (6.2) over each control volume Ω_i. After that, Gauss's divergence theorem is applied:

$$\int_{\partial\Omega_i} \nu \cdot q(u)\, d\sigma = \int_{\Omega_i} f\, dx\,, \quad i \in \{1, \ldots, M\}\,,$$

where ν denotes the outer unit normal to $\partial\Omega_i$. By the first condition of the partition, the boundary $\partial\Omega_i$ is formed by straight-line segments Γ_{ij} $(j = 1, \ldots, n_i)$, along which the normal $\nu|_{\Gamma_{ij}} =: \nu_{ij}$ is constant (see Figure 6.1). So the line integral can be decomposed into a sum of line integrals from which the following equation results:

$$\sum_{j=1}^{n_i} \int_{\Gamma_{ij}} \nu_{ij} \cdot q(u)\, d\sigma = \int_{\Omega_i} f\, dx\,. \tag{6.3}$$

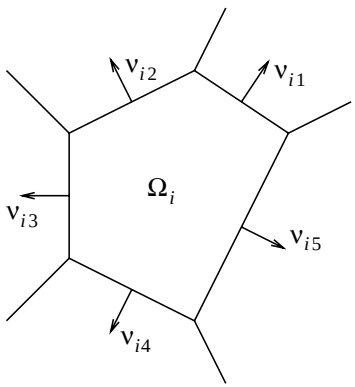


Figure 6.1. A control volume.

Now the integrals occurring in (6.3) have to be approximated. This can be done in very different ways, and so different final discretizations are obtained.

In general, finite volume methods can be distinguished by the following criteria:

(1) the geometric shape of the control volumes Ω_i,

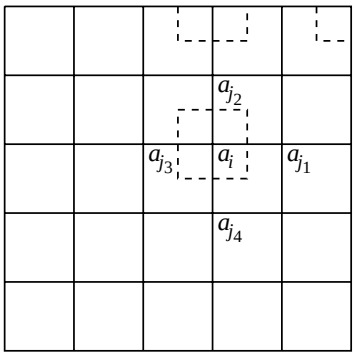
(2) the position of the unknowns ("problem variables") with respect to the control volumes,

(3) the approximation of the boundary (line ($d = 2$) or surface ($d = 3$)) integrals.

Especially the second criterion divides the finite volume methods into two large classes: the *cell-centred* and the *cell-vertex* finite volume methods. In the cell-centred methods, the unknowns are associated with the control volumes (for example, any control volume corresponds to a function value at some interior point (e.g., at the barycentre)). In the cell-vertex methods, the unknowns are located at the vertices of the control volumes. Sometimes, instead of the first-mentioned class a subdivision into two classes, the so-called *cell-centred* and *node-centred* methods, is considered. The difference is whether the problem variables are assigned to the control volumes or, given the problem variables, associated control volumes are defined.

Example 6.1 Consider the homogeneous Dirichlet problem for the Poisson equation on the unit square:

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega = (0,1)^2\,, \\ u &= 0 && \text{on } \partial\Omega\,. \end{aligned}$$



Problem variables:

Function values at the nodes a_i of a square grid with mesh width $h > 0$

Control volumes:

$\Omega_i := \{x \in \Omega : |x - a_i|_\infty < \frac{h}{2}\}$

Figure 6.2. Problem variables and control volumes in a cell-centred finite volume method.

For an inner control volume Ω_i (i.e., $a_i \in \Omega$), equation (6.3) takes the form

$$-\sum_{k=1}^{4} \int_{\Gamma_{ij_k}} \nu_{ij_k} \cdot \nabla u \, d\sigma = \int_{\Omega_i} f \, dx\,,$$

where $\Gamma_{ij_k} := \partial\Omega_i \cap \partial\Omega_{j_k}$. A closer look at the directional derivatives shows that

$$\begin{aligned} \nu_{ij_1} \cdot \nabla u &= \partial_1 u\,, & \nu_{ij_2} \cdot \nabla u &= \partial_2 u\,, \\ \nu_{ij_3} \cdot \nabla u &= -\partial_1 u\,, & \nu_{ij_4} \cdot \nabla u &= -\partial_2 u\,. \end{aligned}$$

i.e. they are just partial derivatives with respect to the first or the second variable on the corresponding parts of the boundary.

Approximating the integrals on Γ_{ij_k} by means of the midpoint rule and replacing the derivatives by difference quotients, we have

$$
\begin{aligned}
&-\sum_{k=1}^{4} \int_{\Gamma_{ij_k}} \nu_{ij_k} \cdot \nabla u \, d\sigma \approx -\sum_{k=1}^{4} \nu_{ij_k} \cdot \nabla u \left(\frac{a_i + a_{j_k}}{2} \right) h \\
&\approx -\left[\frac{u(a_{j_1}) - u(a_i)}{h} + \frac{u(a_{j_2}) - u(a_i)}{h} - \frac{u(a_i) - u(a_{j_3})}{h} - \frac{u(a_i) - u(a_{j_4})}{h} \right] h \\
&= 4\, u(a_i) - \sum_{k=1}^{4} u(a_{j_k}) \,.
\end{aligned}
$$

Thus, we obtain exactly the expression that results from the application of a finite element method with continuous, piecewise linear ansatz and test functions on a Friedrichs–Keller triangulation (cf. Figure 2.9).

Furthermore, if we approximate the integral $\int_{\Omega_i} f \, dx$ by $f(a_i)h^2$, we see that this term coincides with the trapeziodal rule applied to the right-hand side of the mentioned finite element formulation (cf. Lemma 2.13).

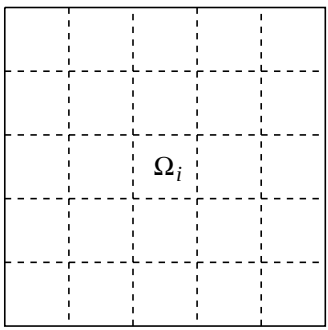
Actually, it is no accident that both discretization methods lead to the same algebraic system. Later on we will prove a more general result to confirm the above observation.

The boundary control volumes are treated as follows:

If $a_i \in \partial\Omega$, then parts of the boundary $\partial\Omega_i$ lie on $\partial\Omega$. At these nodes, the Dirichlet boundary conditions already prescribe values of the unknown function, and so there is no need to include the boundary control volumes into the balance equations (6.3).

A detailed description for the case of flux boundary conditions will be given later, in Section 6.2.4; see (6.23).

Example 6.2 We consider the same boundary value problem as in Example 6.1.



Problem variables:

Function values at the nodes a_i of a square grid with mesh width $h > 0$

Control volumes:

Subsquares of the grid

Figure 6.3. Problem variables and control volumes in a cell-vertex finite volume method.

In the interior of Ω, the resulting discretization yields a 12-point stencil (in the terminology of finite difference methods).

Remark 6.3 In the finite volume discretization of systems of partial differential equations (resulting from fluid mechanics, for example), both methods are used simultaneously for different variables; see Figure 6.4.

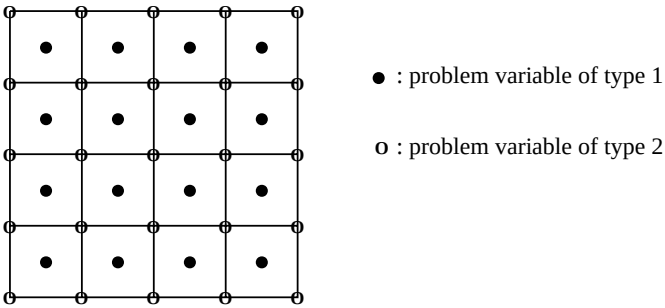


Figure 6.4. Finite volume discretization of systems of partial differential equations.

Assets and Drawbacks of the Finite Volume Method
Assets:

- Flexibility with respect to the geometry of the domain Ω (as in finite element methods).
- Admissibility of unstructured grids (as in finite element methods, important for adaptive methods).
- Simple assembling.
- Conservation of certain laws valid for the continuous problem (for example, conservation laws or maximum principles). This property is important in the numerical solution of differential equations with discontinuous coefficients or of convection-dominated diffusion-convection equations (see Section 6.2.4).
- Easy linearization of nonlinear problems (simpler than in finite element methods (Newton's method)).
- Simple discretization of boundary conditions (as in finite element methods, especially a "natural" treatment of Neumann or mixed boundary conditions).
- In principle, no restriction of the spatial dimension d of the domain Ω.

Drawbacks:

- Smaller field of applications in comparison with finite element or finite difference methods.
- Difficulties in the design of higher order methods (no so-called p-version available as in the finite element method).
- In higher spatial dimensions ($d \geq 3$), the construction of some classes or types of control volumes may be a complex task and thus may lead to a time-consuming assembling.
- Difficult mathematical analysis (stability, convergence, ...).

Exercises

6.1 Given the boundary value problem

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

with piecewise constant coefficients

$$a(x) := \begin{cases} \kappa\alpha\,, & x \in (0,\xi)\,, \\ \alpha\,, & x \in (\xi,1)\,, \end{cases}$$

where α, κ are positive constants and $\xi \in (0,1) \setminus \mathbb{Q}$:

(a) What is the weak solution $u \in H^1(0,1)$ of this problem?

(b) For general "smooth" coefficients a, the differential equation is obviously equivalent to

$$-au'' - a'u' = 0\,.$$

Therefore, the following discretization is suggested:

$$-a_i \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} - \frac{a_{i+1} - a_{i-1}}{2h}\,\frac{u_{i+1} - u_{i-1}}{2h} = 0\,,$$

where an equidistant grid with the nodes $x_i = ih$ $(i = 0, \ldots, N+1)$ and $a_i := a(x_i)$, $u_i :\approx u(x_i)$ is used.
This discretization is also formally correct in the given situation of discontinuous coefficients. Find the discrete solution $(u_i)_{i=1}^N$ in this case.

(c) Under what conditions do the values u_i converge to $u(x_i)$ for $h \to 0$?

6.2 The Finite Volume Method for Linear Elliptic Differential Equations of Second Order on Triangular Grids

In this section we will explain the development and the analysis of a finite volume method of "cell-centred" type for a model problem. Here, $\Omega \subset \mathbb{R}^2$ is a bounded, simply connected domain with a polygonal boundary, but without slits.

6.2.1 Admissible Control Volumes

The Voronoi Diagram
By $\{a_i\}_{i\in\overline{\Lambda}} \subset \overline{\Omega}$ we denote a consecutively numbered point set that includes all vertices of Ω, where $\overline{\Lambda}$ is the corresponding set of indices. Typically, the points a_i are placed at those positions where the values $u(a_i)$ of the exact solution u are to be approximated. The convex set

$$\tilde{\Omega}_i := \left\{ x \in \mathbb{R}^2 \;\middle|\; |x - a_i| < |x - a_j| \quad \text{for all } j \neq i \right\}$$

is called the *Voronoi polygon* (or *Dirichlet domain, Wigner–Seitz cell, Thiessen polygon, ...*). The family $\left\{\tilde{\Omega}_i\right\}_{i\in\overline{\Lambda}}$ is called the *Voronoi diagram* of the point set $\{a_i\}_{i\in\overline{\Lambda}}$.

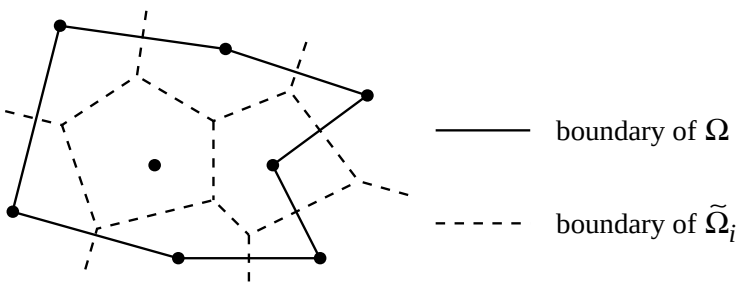


Figure 6.5. Voronoi diagram.

The Voronoi polygons are convex, but not necessarily bounded, sets (consider the situation near the boundary in Figure 6.5). Their boundaries are polygons. The vertices of these polygons are called *Voronoi vertices*.

It can be shown that at any Voronoi vertex at least three Voronoi polygons meet. According to this property, Voronoi vertices are classified into regular and degenerate Voronoi vertices: In a *regular* Voronoi vertex, the boundaries of exactly three Voronoi polygons meet, whereas a *degenerate* Voronoi vertex is shared by at least four Voronoi polygons. In the latter case, all the corresponding nodes a_i are located at some circle (they are "cocyclic", cf. Figure 6.6).

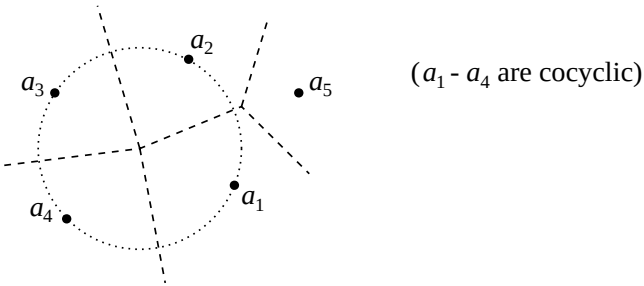


Figure 6.6. Degenerate and regular Voronoi vertex.

Now the elements Ω_i (control volumes) of the partition of Ω required for the definition of the finite volume method can be introduced as follows:

$$\Omega_i := \tilde{\Omega}_i \cap \Omega\,, \quad i \in \overline{\Lambda}\,.$$

As a consequence, the domains Ω_i need not necessarily be convex if Ω is nonconvex (cf. Figure 6.5).

Furthermore, the following notation will be used:

$$\begin{aligned}
\Lambda_i &:= \left\{ j \in \overline{\Lambda} \setminus \{i\} \,:\, \partial\Omega_i \cap \partial\Omega_j \neq \emptyset \right\}, \quad i \in \overline{\Lambda}\,, \\
&\quad \text{for the set of indices of neighbouring nodes,} \\
\Gamma_{ij} &:= \partial\Omega_i \cap \partial\Omega_j\,, \quad j \in \Lambda_i\,, \text{ for a joint piece of the} \\
&\quad \text{boundaries of neighbouring control volumes,} \\
m_{ij} &\quad \text{for the length of } \Gamma_{ij}\,.
\end{aligned}$$

The *dual graph* of the Voronoi diagram is defined as follows:

Any pair of points a_i, a_j such that $m_{ij} > 0$ is connected by a straight-line segment. In this way, a further partition of Ω with an interesting property results.

Theorem 6.4 *If all Voronoi vertices are regular, then the dual graph coincides with the set of edges of a triangulation of the convex hull of the given point set.*

This triangulation is called a *Delaunay triangulation.*

If among the Voronoi vertices there are degenerate ones, then a triangulation can be obtained from the dual graph by a subsequent local triangulation of the remaining m-polygons ($m \geq 4$). A Delaunay triangulation has the interesting property that two interior angles subtended by any given edge sum to no more than π. In this respect Delaunay triangulations satisfy the first part of the angle condition formulated in Section 3.9 for the maximum principle in finite element methods.

Therefore, if Ω is convex, then we automatically get a triangulation together with the Voronoi diagram. In the case of a nonconvex domain Ω, certain modifications could be required to achieve a correct triangulation.

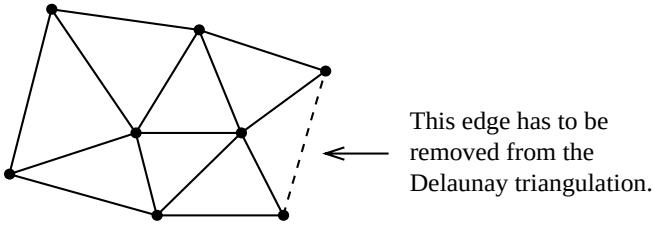


Figure 6.7. Delaunay triangulation to the Voronoi diagram from Figure 6.5.

The implication

$$\text{Voronoi diagram} \quad \Rightarrow \quad \text{Delaunay triangulation}\,,$$

which we have just discussed, suggests that we ask about the converse statement. We do not want to answer it completely at this point, but we give the following sufficient condition.

Theorem 6.5 *If a conforming triangulation of* Ω *(in the sense of finite element methods) consists of nonobtuse triangles exclusively, then it is a Delaunay triangulation, and the corresponding Voronoi diagram can be constructed by means of the perpendicular bisectors of the triangles' edges.*

We mention that the centre of the circumcircle of a nonobtuse triangle is located within the closure of that triangle.

In the analysis of the finite volume method, the following relation is important.

Lemma 6.6 *Given a nonobtuse triangle* K *with vertices* a_{i_k}, $k \in \{1,2,3\}$, *then for the corresponding parts* $\Omega_{i_k,K} := \Omega_{i_k} \cap K$ *of the control volumes* Ω_{i_k}, *we have*

$$\frac{1}{4}|K| \le |\Omega_{i_k,K}| \le \frac{1}{2}|K|\,, \quad k \in \{1,2,3\}\,.$$

The Donald diagram

In contrast to the Voronoi diagram, where the construction starts from a given point set, the starting point here is a triangulation $\mathcal{T}_h$ of Ω, which is allowed to contain obtuse triangles.

Again, let K be a triangle with vertices a_{i_k}, $k \in \{1,2,3\}$. We define

$$\Omega_{i_k,K} := \left\{ x \in K \;\middle|\; \lambda_j(x) < \lambda_k(x),\ j \ne k \right\}\,,$$

where λ_k denote the barycentric coordinates with respect to a_{i_k} (cf. (3.51)).

Obviously, the barycentre satisfies $a_S = \frac{1}{3}(a_{i_1} + a_{i_2} + a_{i_3})$, and (see, for comparison, Lemma 6.6)

$$3\,|\Omega_{i_k,K}| = |K|\,, \quad k \in \{1,2,3\}\,. \tag{6.4}$$

This relation is a simple consequence of the geometric interpretation of the barycentric coordinates as area coordinates given in Section 3.3. The

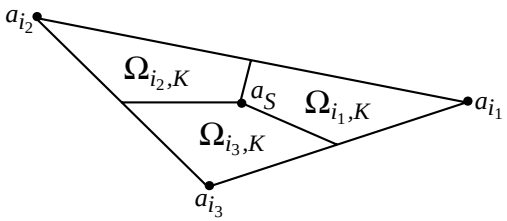


Figure 6.8. The subdomains $\Omega_{i_k,K}$.

required control volumes are defined as follows (see Figure 6.8):

$$\Omega_i := \operatorname{int}\left(\bigcup_{K:\partial K \ni a_i} \overline{\Omega}_{i,K}\right), \quad i \in \overline{\Lambda}.$$

The family $\{\Omega_i\}_{i\in\overline{\Lambda}}$ is called a *Donald diagram.*

The quantities Γ_{ij}, m_{ij}, and Λ_i are defined similarly as in the case of the Voronoi diagram. We mention that the boundary pieces Γ_{ij} are not necessarily straight, but polygonal in general.

6.2.2 Finite Volume Discretization

The model under consideration is a special case of equation (6.1). Instead of the matrix-valued diffusion coefficient K we will take a scalar coefficient $k : \Omega \to \mathbb{R}$, that is, $K = kI$. Moreover, homogeneous Dirichlet boundary conditions are to be satisfied. So the boundary value problem reads as follows:

$$\begin{aligned} -\nabla \cdot (k\,\nabla u - c\,u) + r\,u &= f \quad \text{in } \Omega\,, \\ u &= 0 \quad \text{on } \partial\Omega\,, \end{aligned} \tag{6.5}$$

with $k, r, f : \Omega \to \mathbb{R}$, $c : \Omega \to \mathbb{R}^2$.

The Case of the Voronoi Diagram
Let the domain Ω be partioned by a Voronoi diagram and the corresponding Delaunay triangulation. Due to the homogeneous Dirichlet boundary conditions, it is sufficient to consider only those control volumes Ω_i that are associated with inner nodes $a_i \in \Omega$. Therefore, we denote the set of indices of all inner nodes by

$$\Lambda := \left\{ i \in \overline{\Lambda} \;\middle|\; a_i \in \Omega \right\}.$$

In the first step, the differential equation (6.5) is integrated over the single control volumes Ω_i :

$$-\int_{\Omega_i} \nabla \cdot (k\,\nabla u - c\,u)\; dx + \int_{\Omega_i} r\,u\,dx = \int_{\Omega_i} f\,dx\,, \quad i \in \Lambda\,. \tag{6.6}$$

The application of Gauss's divergence theorem to the first integral of the left-hand side of (6.6) yields

$$\int_{\Omega_i} \nabla \cdot (k\,\nabla u - c\,u)\; dx = \int_{\partial\Omega_i} \nu \cdot (k\,\nabla u - c\,u)\; d\sigma\,.$$

Due to $\partial\Omega_i = \cup_{j\in\Lambda_i}\Gamma_{ij}$ (cf. Figure 6.9), it follows that

$$\int_{\Omega_i} \nabla \cdot (k\,\nabla u - c\,u)\; dx = \sum_{j\in\Lambda_i} \int_{\Gamma_{ij}} \nu_{ij} \cdot (k\,\nabla u - c\,u)\; d\sigma\,,$$

where ν_{ij} is the (constant) outer unit normal to Γ_{ij} (with respect to Ω_i). In the next step we approximate the line integrals over Γ_{ij}.

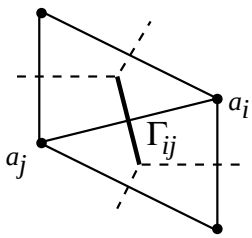


Figure 6.9. The edge Γ_{ij}.

First, the coefficients k and $\nu_{ij} \cdot c$ are approximated on Γ_{ij} by constants $\mu_{ij} > 0$, respectively γ_{ij}:

$$k|_{\Gamma_{ij}} \approx \mu_{ij} = \text{const} > 0\,, \quad \nu_{ij} \cdot c|_{\Gamma ij} \approx \gamma_{ij} = \text{const}\,.$$

In the simplest case, the approximation can be realized by the corresponding value at the midpoint $a_{\Gamma_{ij}}$ of the straight-line segment Γ_{ij}. A better choice is

$$\gamma_{ij} := \begin{cases} \dfrac{1}{m_{ij}} \displaystyle\int_{\Gamma_{ij}} \nu_{ij} \cdot c\, d\sigma\,, & m_{ij} > 0\,, \\ \nu_{ij} \cdot c(a_{\Gamma_{ij}})\,, & m_{ij} = 0\,. \end{cases} \tag{6.7}$$

We thus obtain

$$\int_{\Omega_i} \nabla \cdot (k\,\nabla u - c\,u)\; dx \approx \sum_{j\in\Lambda_i} \int_{\Gamma_{ij}} [\mu_{ij}\,(\nu_{ij} \cdot \nabla u) - \gamma_{ij}\,u]\, d\sigma\,.$$

The normal derivatives are approximated by difference quotients; that is,

$$\nu_{ij} \cdot \nabla u \approx \frac{u(a_j) - u(a_i)}{d_{ij}} \quad \text{with } d_{ij} := |a_i - a_j|\;.$$

This formula is exact for such functions that are linear along the straight-line segment between the points a_i, a_j. So it remains to approximate the integral of u over Γ_{ij}. For this, a convex combination of the values of u at

the nodes a_i and a_j is taken:

$$u|_{\Gamma_{ij}} \approx r_{ij}\, u(a_i) + (1 - r_{ij})\; u(a_j)\,,$$

where $r_{ij} \in [0,1]$ is a parameter to be defined subsequently. In general, r_{ij} depends on μ_{ij}, γ_{ij}, and d_{ij}.

Collecting all the above approximations, we arrive at the following relation:

$$\int_{\Omega_i} \nabla \cdot (k\,\nabla u - c\,u)\; dx \approx \sum_{j\in\Lambda_i} \left\{ \mu_{ij}\, \frac{u(a_j) - u(a_i)}{d_{ij}} - \gamma_{ij}\; [r_{ij}\, u(a_i) + (1 - r_{ij})\; u(a_j)] \right\} m_{ij}\,.$$

To approximate the remaining integrals from (6.6), the following formulas are used:

$$\begin{aligned} \int_{\Omega_i} r\, u\, dx &\approx r(a_i)\, u(a_i)\, m_i =: r_i\, u(a_i)\, m_i\,, \quad \text{with } m_i := |\Omega_i|\,, \\ \int_{\Omega_i} f\, dx &\approx f(a_i)\, m_i =: f_i\, m_i\,. \end{aligned}$$

Instead of $r_i := r(a_i)$ or $f_i := f(a_i)$, the approximations

$$r_i := \frac{1}{m_i} \int_{\Omega_i} r\, dx \quad \text{respectively} \quad f_i := \frac{1}{m_i} \int_{\Omega_i} f\, dx \tag{6.8}$$

can also be used. Denoting the unknown approximate values for $u(a_i)$ by u_i, we obtain the following linear system of equations:

$$\begin{aligned} \sum_{j\in\Lambda_i} \left\{ \mu_{ij} \frac{u_i - u_j}{d_{ij}} + \gamma_{ij}\, [r_{ij} u_i + (1 - r_{ij})\, u_j] \right\} m_{ij} + r_i u_i m_i \\ = f_i m_i\,, \quad i \in \Lambda\,. \end{aligned} \tag{6.9}$$

This representation clearly indicates the affinity of the finite volume method to the finite difference method. However, for the subsequent analysis it is more convenient to rewrite this system of equations in terms of a discrete variational equality.

Multiplying the ith equation in (6.9) by arbitrary numbers $v_i \in \mathbb{R}$ and summing the results up over $i \in \Lambda$, we get

$$\sum_{i\in\Lambda} v_i \left\{ \sum_{j\in\Lambda_i} \left\{ \mu_{ij} \frac{u_i - u_j}{d_{ij}} + \gamma_{ij}\, [r_{ij}\, u_i + (1 - r_{ij})\; u_j] \right\} m_{ij} + r_i\, u_i\, m_i \right\} = \sum_{i\in\Lambda} f_i\, v_i\, m_i\,.$$

Further, let V_h denote the space of continuous functions that are piecewise linear over the (Delaunay) triangulation of Ω and that vanish on $\partial\Omega$. Then the values u_i and v_i can be interpolated in V_h; that is, there are unique

$u_h, v_h \in V_h$ such that $u_h(a_i) = u_i$, $v_h(a_i) = v_i$ for all $i \in \Lambda$. The following discrete bilinear forms on $V_h \times V_h$ can then be defined:

$$\begin{aligned} a_h^0(u_h, v_h) &:= \sum_{i\in\Lambda} v_i \sum_{j\in\Lambda_i} \mu_{ij}(u_i - u_j)\frac{m_{ij}}{d_{ij}}, \\ b_h(u_h, v_h) &:= \sum_{i\in\Lambda} v_i \sum_{j\in\Lambda_i} [r_{ij}\, u_i + (1 - r_{ij})\, u_j]\,\gamma_{ij}\, m_{ij}, \\ d_h(u_h, v_h) &:= \sum_{i\in\Lambda} r_i\, u_i\, v_i\, m_i, \\ a_h(u_h, v_h) &:= a_h^0(u_h, v_h) + b_h(u_h, v_h) + d_h(u_h, v_h). \end{aligned}$$

Finally, for two continuous functions $v, w \in C(\overline{\Omega})$, we set

$$\langle w, v\rangle_{0,h} := \sum_{i\in\Lambda} w_i\, v_i\, m_i,$$

where $v_i := v(a_i)$, $w_i := w(a_i)$.

Remark 6.7 $\langle\cdot,\cdot\rangle_{0,h}$ is a scalar product on V_h. In particular, the following norm can be introduced:

$$\|v_h\|_{0,h} := \sqrt{\langle v_h, v_h\rangle_{0,h}}, \quad v_h \in V_h. \tag{6.10}$$

In (3.136) a discrete (L^2-) norm for a general finite element space v_h has been defined using the same notation. This multiple use seems to be acceptable, since for regular triangulations both norms are equivalent uniformly in h (see Remark 6.16 below).

Now the discrete variational formulation of the finite volume method is this:

Find $u_h \in V_h$ such that

$$a_h(u_h, v_h) = \langle f, v_h\rangle_{0,h} \quad \text{for all } v_h \in V_h. \tag{6.11}$$

Up to now, the choice of the weighting parameters r_{ij} has remained open.

For this, two cases can be roughly distinguished:

(1) There exists a pair of indices $(i, j) \in \Lambda \times \overline{\Lambda}$ such that $\mu_{ij} \ll |\gamma_{ij}| d_{ij}$.

(2) There is no such pair (i, j) with $\mu_{ij} \ll |\gamma_{ij}| d_{ij}$.

In the second case, an appropriate choice is $r_{ij} \equiv \frac{1}{2}$. To some extent, this can be seen as a generalization of the central difference method to nonuniform grids. The first case corresponds to a locally *convection-dominated* situation and requires a careful selection of the weighting parameters r_{ij}. This will be explained in more detail in Section 9.3.

In general, the weighting parameters are of the following structure:

$$r_{ij} = R\left(\frac{\gamma_{ij}\, d_{ij}}{\mu_{ij}}\right), \tag{6.12}$$

where $R : \mathbb{R} \to [0,1]$ is some function to be specified. The argument $\frac{\gamma_{ij}\, d_{ij}}{\mu_{ij}}$ is called the *local Péclet number.* Typical examples for this function R are

$$\begin{aligned} R(z) &= \frac{1}{2}\left[\operatorname{sign}(z)+1\right], && \textit{full upwinding},\\ R(z) &= \begin{cases} (1-\tau)/2\,, & z<0\,,\\ (1+\tau)/2\,, & z\geq 0\,, \end{cases} && \tau(z) := \max\left\{0, 1-\frac{2}{|z|}\right\},\\ R(z) &= 1-\frac{1}{z}\left(1-\frac{z}{e^z-1}\right), && \textit{exponential upwinding}. \end{aligned}$$

All these functions possess many common properties. For example, for all $z \in \mathbb{R}$,

$$\begin{aligned} &\text{(P1)} && [1-R(z)-R(-z)]\, z &&= 0\,,\\ &\text{(P2)} && \left[R(z)-\tfrac{1}{2}\right] z &&\geq 0\,,\\ &\text{(P3)} && 1-[1-R(z)]\, z &&\geq 0\,. \end{aligned} \tag{6.13}$$

Note that the constant function $R = \frac{1}{2}$ satisfies the conditions (P1) and (P2) but not (P3).

The Case of the Donald Diagram
Let the domain Ω be triangulated as in the finite element method. Then, following the explanations given in the second part of Section 6.2.1, the corresponding Donald diagram can be created.

The discrete bilinear form in this case is defined by

$$a_h(u_h, v_h) := \langle k\, \nabla u_h, \nabla v_h \rangle_0 + b_h(u_h, v_h) + d_h(u_h, v_h)\,;$$

that is, the principal part of the differential expression is discretized as in the finite element method, where b_h, d_h, and V_h are defined as in the first part of this section.

6.2.3 Comparison with the Finite Element Method

As we have already seen in Example 6.1, it may happen that a finite volume discretization coincides with a finite difference or finite element discretization. We also mention that the control volumes from that example are exactly the Voronoi polygons to the grid points (i.e., to the nodes of the triangulation).

Here we will consider this observation in more detail. By $\{\varphi_i\}_{i\in\Lambda}$ we denote the nodal basis of the space V_h of continuous, piecewise linear functions on a conforming triangulation of the domain Ω.

Lemma 6.8 *Let $\mathcal{T}_h$ be a conforming triangulation of Ω (in the sense of finite element methods), all triangles of which are nonobtuse, and consider the corresponding Voronoi diagram in accordance with Theorem 6.5. Then, for an arbitrary triangle $K \in \mathcal{T}_h$ with vertices a_i, a_j $(i \neq j)$, the following*

relation holds:

$$\int_K \nabla\varphi_j \cdot \nabla\varphi_i \, dx = -\frac{m_{ij}^K}{d_{ij}} \,,$$

where m_{ij}^K *is the length of the segment of* Γ_{ij} *that intersects* K.

Proof: Here we use some of the notation and the facts prepared at the beginning of Section 3.9. In particular, α_{ij}^K denotes the interior angle of K that is located in opposite the edge with vertices a_i, a_j. Next, the following equality is an obvious fact from elementary geometry: $2 \sin\alpha_{ij}^K \, m_{ij}^K = \cos\alpha_{ij}^K \, d_{ij}$. It remains to recall the relation

$$\int_K \nabla\varphi_j \cdot \nabla\varphi_i \, dx = -\frac{1}{2} \cot\alpha_{ij}^K$$

from Lemma 3.47, and the statement immediately follows. □

Corollary 6.9 *Under the assumptions of Lemma* 6.8*, we have for* $k \equiv 1$,

$$\langle \nabla u_h, \nabla v_h \rangle_0 = a_h^0 \, (u_h, v_h) \quad \textit{for all } u_h, v_h \in V_h \,.$$

Proof: It is sufficient to verify the relation for $v_h = \varphi_i$ and arbitrary $i \in \Lambda$. First, we see that

$$\langle \nabla u_h, \nabla\varphi_i \rangle_0 = \sum_{K \subset \mathrm{supp}\varphi_i} \int_K \nabla u_h \cdot \nabla\varphi_i \, dx \,.$$

Furthermore,

$$\begin{aligned} \int_K \nabla u_h \cdot \nabla\varphi_i \, dx &= \sum_{j:\partial K \ni a_j} u_j \int_K \nabla\varphi_j \cdot \nabla\varphi_i \, dx \\ &= u_i \int_K \nabla\varphi_i \cdot \nabla\varphi_i \, dx + \sum_{j \neq i:\partial K \ni a_j} u_j \int_K \nabla\varphi_j \cdot \nabla\varphi_i \, dx \,. \end{aligned}$$

Since

$$1 = \sum_{j:\partial K \ni a_j} \varphi_j$$

over K, it follows that

$$\nabla\varphi_i = - \sum_{j \neq i:\partial K \ni a_j} \nabla\varphi_j \,; \tag{6.14}$$

that is, by means of Lemma 6.8,

$$\int_K \nabla u_h \cdot \nabla\varphi_i \, dx = \sum_{j \neq i:\partial K \ni a_j} (u_j - u_i) \int_K \nabla\varphi_j \cdot \nabla\varphi_i \, dx$$

$$= \sum_{j \neq i : \partial K \ni a_j} (u_i - u_j) \frac{m_{ij}^K}{d_{ij}} . \tag{6.15}$$

Summing over all $K \subset \operatorname{supp} \varphi_i$, we get

$$\langle \nabla u_h, \nabla \varphi_i \rangle_0 = \sum_{j \in \Lambda_i} (u_i - u_j) \frac{m_{ij}}{d_{ij}} = a_h^0 (u_h, \varphi_i) . \qquad \square$$

Remark 6.10 By a more sophisticated argumentation it can be shown that the above corollary remains valid if the diffusion coefficient k is constant on all triangles $K \in \mathcal{T}_h$ and if the approximation μ_{ij} is chosen according to

$$\mu_{ij} := \begin{cases} \dfrac{1}{m_{ij}} \displaystyle\int_{\Gamma_{ij}} k \, d\sigma = \frac{k|_K \, m_{ij}^K + k|_{K'} \, m_{ij}^{K'}}{m_{ij}} , & m_{ij} > 0 , \\ 0 , & m_{ij} = 0 , \end{cases} \tag{6.16}$$

where K, K' are both triangles sharing the vertices a_i, a_j.

Treatment of Matrix-valued Diffusion Coefficients
Corollary 6.9 and Remark 6.10 are valid only in the spatial dimension $d = 2$. However, for more general control volumes, higher spatial dimensions, or not necessarily scalar diffusion coefficients, weaker statements can be proven.

As an example, we will state the following fact. As a by-product, we also obtain an idea for how to derive discretizations in the case of matrix-valued diffusion coefficients. For a better distinction between the elements K of the triangulation and the diffusion coefficient, we keep the notation k for the diffusion coefficient, even if k is allowed to be a matrix-valued function temporarily.

Lemma 6.11 *Let $\mathcal{T}_h$ be a conforming triangulation of Ω, where in the case of the Voronoi diagram it is additionally required that all triangles be nonobtuse. Furthermore, assume that the diffusion matrix $k : \Omega \to \mathbb{R}^{2,2}$ is constant on the single elements of $\mathcal{T}_h$. Then for any $i \in \Lambda$ and $K \in \mathcal{T}_h$ we have*

$$\int_K (k \nabla u_h) \cdot \nabla \varphi_i \, dx = - \int_{\partial \Omega_i \cap K} (k \nabla u_h) \cdot \nu \, d\sigma \quad \text{for all } u_h \in V_h,$$

where $\{\Omega_i\}_{i \in \overline{\Lambda}}$ is either a Voronoi or a Donald diagram and ν denotes the outer unit normal with respect to Ω_i.

Without difficulties, the proof can be carried over from the proof of a related result in [20, Lemma 6.1].

Now we will show how to use this fact to formulate discretizations for the case of matrix-valued diffusion coefficients. Namely, using relation (6.14),

we easily see that

$$\begin{aligned}\int_{\partial\Omega_i\cap K}(k\nabla u_h)\cdot\nu\,d\sigma &= \sum_{j:\partial K\ni a_j}\int_{\partial\Omega_i\cap K}u_j(k\nabla\varphi_j)\cdot\nu\,d\sigma\\ &= \sum_{j\neq i:\partial K\ni a_j}(u_j-u_i)\int_{\partial\Omega_i\cap K}(k\nabla\varphi_j)\cdot\nu\,d\sigma\,.\end{aligned}$$

Summing over all triangles that lie in the support of φ_i, we obtain by Lemma 6.11 the relation

$$\int_\Omega(k\nabla u_h)\cdot\nabla\varphi_i\,dx=\sum_{j\in\Lambda_i}(u_i-u_j)\int_{\partial\Omega_i}(k\nabla\varphi_j)\cdot\nu\,d\sigma\,. \tag{6.17}$$

With the definition

$$\mu_{ij}:=\begin{cases}\dfrac{d_{ij}}{m_{ij}}\displaystyle\int_{\partial\Omega_i}(k\nabla\varphi_j)\cdot\nu\,d\sigma\,, & m_{ij}>0\,,\\ 0\,, & m_{ij}=0\,,\end{cases} \tag{6.18}$$

it follows that

$$\int_\Omega(k\nabla u_h)\cdot\nabla\varphi_i\,dx=\sum_{j\in\Lambda_i}\mu_{ij}\,(u_i-u_j)\,\frac{m_{ij}}{d_{ij}}\,.$$

Note that, in the case of Voronoi diagrams, (6.16) is a special case of the choice (6.18).

Consequently, in order to obtain a discretization for the case of a matrix-valued diffusion coefficient, it is sufficient to replace in the bilinear form b_h and, if the Voronoi diagram is used, also in a_h^0, the terms involving μ_{ij} according to formula (6.18).

Implementation of the Finite Volume Method

In principle, the finite volume method can be implemented in different ways. If the linear system of equations is implemented in a node-orientated manner (as in finite difference methods), the entries of the system matrix A_h and the components of the right-hand side $\boldsymbol{q}_h$ can be taken directly from (6.9).

On the other hand, an element-orientated assembling is possible, too. This approach is preferable, especially in the case where an existing finite element program will be extended by a finite volume module. The idea of how to do this is suggested by equation (6.17). Namely, for any triangle $K\in\mathcal{T}_h$, the restricted bilinear form $a_{h,K}$ with the appropriate definition of μ_{ij} according to (6.18) is defined as follows:

$$a_{h,K}(u_h,v_h):=$$

$$\sum_{i\in\Lambda}v_i\left\{\sum_{\substack{j\neq i:\\ \partial K\ni a_j}}\left\{\mu_{ij}\frac{u_i-u_j}{d_{ij}}+\gamma_{ij}\left[r_{ij}u_i+(1-r_{ij})\,u_j\right]\right\}m_{ij}^K+r_iu_im_i^K\right\},$$

where $m_i^K := |\Omega_i \cap K|$. Then the contribution of the triangle K to the matrix entry $(A_h)_{ij}$ of the matrix A_h is equal to $a_{h,K}(\varphi_j, \varphi_i)$. In the same way, the right-hand side of (6.9) can be split elementwise.

6.2.4 Properties of the Discretization

Here we will give a short overview of basic properties of finite volume methods. For the sake of simplicity, we restrict ourselves to the case of a constant scalar diffusion coefficient $k > 0$. Then, in particular, it is useful to set $\mu_{ij} := k$ for all $i \in \Lambda$, $j \in \Lambda_i$.

Lemma 6.12 *Suppose the approximations γ_{ij} of $\nu_{ij} \cdot c|_{\Gamma_{ij}}$ satisfy $\gamma_{ji} = -\gamma_{ij}$ and the r_{ij} are defined by* (6.12) *with a function R satisfying* (P1). *Then we get for all $u_h, v_h \in V_h$,*

$$b_h(u_h, v_h) = \frac{1}{2} \sum_{i \in \Lambda} \sum_{j \in \Lambda_i} u_i \, v_i \, \gamma_{ij} \, m_{ij}$$
$$+ \frac{1}{2} \sum_{i \in \Lambda} \sum_{j \in \Lambda_i} \left[\left(r_{ij} - \frac{1}{2} \right) (u_i - u_j)(v_i - v_j) + \frac{1}{2} (u_j v_i - u_i v_j) \right] \gamma_{ij} \, m_{ij} \, .$$

Proof: First, we observe that b_h can be rewritten as follows:

$$\begin{aligned} b_h(u_h, v_h) &= \sum_{i \in \Lambda} \sum_{j \in \Lambda_i} v_i \left[(1 - r_{ij}) \, u_j - \left(\frac{1}{2} - r_{ij} \right) u_i \right] \gamma_{ij} \, m_{ij} \\ &\quad + \frac{1}{2} \sum_{i \in \Lambda} \sum_{j \in \Lambda_i} u_i \, v_i \, \gamma_{ij} \, m_{ij} \, . \end{aligned} \tag{6.19}$$

In the first term, we change the order of summation and rename the indices:

$$\begin{aligned} b_h(u_h, v_h) &= \sum_{i \in \Lambda} \sum_{j \in \Lambda_i} v_j \left[(1 - r_{ji}) \, u_i - \left(\frac{1}{2} - r_{ji} \right) u_j \right] \gamma_{ji} \, m_{ji} \\ &\quad + \frac{1}{2} \sum_{i \in \Lambda} \sum_{j \in \Lambda_i} u_i \, v_i \, \gamma_{ij} \, m_{ij} \, . \end{aligned}$$

Next we make use of the following relations, which easily result from $d_{ji} = d_{ij}$ and the assumptions on γ_{ij} and r_{ij}:

$$(1 - r_{ji}) \, \gamma_{ji} = -r_{ij} \, \gamma_{ij} \, , \qquad \left(\frac{1}{2} - r_{ji} \right) \gamma_{ji} = \left(\frac{1}{2} - r_{ij} \right) \gamma_{ij} \, .$$

So we get, due to $m_{ji} = m_{ij}$,

$$\begin{aligned} b_h(u_h, v_h) &= \sum_{i \in \Lambda} \sum_{j \in \Lambda_i} v_j \left[-r_{ij} u_i - \left(\frac{1}{2} - r_{ij} \right) u_j \right] \gamma_{ij} \, m_{ij} \\ &\quad + \frac{1}{2} \sum_{i \in \Lambda} \sum_{j \in \Lambda_i} u_i \, v_i \, \gamma_{ij} \, m_{ij} \, . \end{aligned}$$

Taking the arithmetic mean of both representations of b_h, we arrive at

$$\begin{aligned} b_h(u_h, v_h) &= \frac{1}{2}\sum_{i\in\Lambda}\sum_{j\in\Lambda_i} u_i\, v_i\, \gamma_{ij}\, m_{ij} \\ &+ \frac{1}{2}\sum_{i\in\Lambda}\sum_{j\in\Lambda_i}\left[(1-r_{ij})\, u_j v_i - r_{ij} u_i v_j - \left(\frac{1}{2}-r_{ji}\right)(u_i v_i + u_j v_j)\right]\gamma_{ij} m_{ij} \\ &= \frac{1}{2}\sum_{i\in\Lambda}\sum_{j\in\Lambda_i}\left[\left(\frac{1}{2}-r_{ij}\right)(u_j v_i + u_i v_j - u_i v_i - u_j v_j)\right. \\ &\qquad \left. + \frac{1}{2}(u_j v_i - u_i v_j)\right]\gamma_{ij} m_{ij} + \frac{1}{2}\sum_{i\in\Lambda}\sum_{j\in\Lambda_i} u_i v_i \gamma_{ij} m_{ij}\,. \end{aligned}$$

□

Corollary 6.13 *Let $c_1, c_2, \nabla\cdot c \in C(\overline{\Omega})$. Under the assumptions of Lemma* 6.12 *and also assuming property* (P2) *for R, the bilinear form b_h satisfies for all $v_h \in V_h$ the estimate*

$$b_h(v_h, v_h) \geq \frac{1}{2}\sum_{i\in\Lambda} v_i^2 \left[\int_{\Omega_i} \nabla\cdot c\, dx + \sum_{j\in\Lambda_i}\int_{\Gamma_{ij}} (\gamma_{ij} - \nu_{ij}\cdot c)\, d\sigma\right]. \quad (6.20)$$

Proof: Due to $\left(r_{ij} - \frac{1}{2}\right)\gamma_{ij} \geq 0$, because of property (P2) in (6.13), it immediately follows that

$$b_h(v_h, v_h) \geq \frac{1}{2}\sum_{i\in\Lambda}\sum_{j\in\Lambda_i} v_i^2\, \gamma_{ij}\, m_{ij} = \frac{1}{2}\sum_{i\in\Lambda} v_i^2 \sum_{j\in\Lambda_i} \gamma_{ij}\, m_{ij}\,.$$

For the inner sum, we can write

$$\begin{aligned} \sum_{j\in\Lambda_i} \gamma_{ij}\, m_{ij} &= \sum_{j\in\Lambda_i}\int_{\Gamma_{ij}} \gamma_{ij}\, d\sigma \\ &= \sum_{j\in\Lambda_i}\int_{\Gamma_{ij}} \nu_{ij}\cdot c\, d\sigma + \sum_{j\in\Lambda_i}\int_{\Gamma_{ij}} (\gamma_{ij} - \nu_{ij}\cdot c)\, d\sigma\,. \end{aligned}$$

The first term can be rewritten as an integral over the boundary of Ω_i, i.e.,

$$\sum_{j\in\Lambda_i}\int_{\Gamma_{ij}} \nu_{ij}\cdot c\, d\sigma = \int_{\partial\Omega_i} \nu\cdot c\, d\sigma\,.$$

By Gauss's divergence theorem, it follows that

$$\int_{\partial\Omega_i} \nu\cdot c\, d\sigma = \int_{\Omega_i} \nabla\cdot c\, dx\,.$$

□

Remark 6.14 If the approximations γ_{ij} are chosen according to (6.7), then $\gamma_{ji} = -\gamma_{ij}$, and (6.20) simplifies to

$$b_h(v_h, v_h) \geq \frac{1}{2} \sum_{i \in \Lambda} v_i^2 \int_{\Omega_i} \nabla \cdot c \, dx \, .$$

Using a similar argument as in the treatment of the term $\sum_{j \in \Lambda_i} \gamma_{ij} \, m_{ij}$ in the proof of Corollary 6.13, the value $d_h(v_h, v_h)$ can be represented as follows:

$$\begin{aligned} d_h(v_h, v_h) &= \sum_{i \in \Lambda} r_i \, v_i^2 \, m_i = \sum_{i \in \Lambda} v_i^2 \int_{\Omega_i} r_i \, dx \\ &= \sum_{i \in \Lambda} v_i^2 \int_{\Omega_i} r \, dx + \sum_{i \in \Lambda} v_i^2 \int_{\Omega_i} (r_i - r) \, dx \, . \end{aligned} \tag{6.21}$$

The second term vanishes if the approximations r_i are defined as in (6.8).

Theorem 6.15 *Let the r_{ij} be defined by (6.12) with R satisfying* (P1) *and* (P2). *Suppose $k > 0$, $c_1, c_2, \nabla \cdot c, r \in C(\overline{\Omega})$, $r + \frac{1}{2} \nabla \cdot c \geq r_0 = \text{const} \geq 0$ on Ω and that the approximations γ_{ij}, respectively r_i, are chosen according to* (6.7), *respectively* (6.8). *Under the assumptions of Lemma* 6.8, *we have for all $v_h \in V_h$,*

$$a_h \left(v_h, v_h \right) \geq k \left\langle \nabla v_h, \nabla v_h \right\rangle_0 + r_0 \sum_{i \in \Lambda} v_i^2 \, m_i = k \left| v_h \right|_1^2 + r_0 \| v_h \|_{0,h}^2 \, ;$$

that is, the bilinear form a_h is V_h-elliptic uniformly with respect to h.

Proof: We start with the consideration of $a_h^0 \left(v_h, v_h \right)$. Due to Corollary 6.9, the relation

$$a_h^0 \left(v_h, v_h \right) = k \left\langle \nabla v_h, \nabla v_h \right\rangle_0 = k \left| v_h \right|_1^2$$

holds. Furthermore, by Remark 6.14 and equation (6.21), we have

$$b_h \left(v_h, v_h \right) + d_h \left(v_h, v_h \right) \geq \sum_{i \in \Lambda} v_i^2 \int_{\Omega_i} \left(\frac{1}{2} \nabla \cdot c + r \right) dx \geq r_0 \sum_{i \in \Lambda} v_i^2 \, m_i \, .$$

Since by definition,

$$a_h \left(v_h, v_h \right) = a_h^0 \left(v_h, v_h \right) + b_h \left(v_h, v_h \right) + d_h \left(v_h, v_h \right) \, ,$$

both relations yield the assertion. □

Remark 6.16 Let the family of triangulations $(\mathcal{T}_h)_h$ be regular. Then the norms defined in (3.136) and in (6.10) and also the norms $\| \cdot \|_{0,h}$ and $\| \cdot \|_0$ are equivalent on V_h uniformly with respect to h; i.e., there exist two constants $C_1, C_2 > 0$ independent of h such that

$$C_1 \| v \|_0 \leq \| v \|_{0,h} \leq C_2 \| v \|_0 \quad \text{for all } v \in V_h \, .$$

Proof: Due to Theorem 3.43 (i) only the uniform equivalence of the discrete L^2-norms has to be shown. Denoting such an equivalence by $\cong$, we have for $v \in V_h$ with $v_i := v(a_i)$ for $i \in \Lambda$,

$$\begin{aligned}
\left(\sum_{i\in\Lambda} |v_i|^2 m_i \right)^{1/2} &= \left(\sum_{i\in\Lambda} |v_i|^2 \sum_{\substack{K\in\mathcal{T}_h:\\ K\cap\Omega_i\neq\emptyset}} |\Omega_{i,K}| \right)^{1/2} \\
&\cong \left(\sum_{i\in\Lambda} \sum_{\substack{K\in\mathcal{T}_h:\\ K\cap\Omega_i\neq\emptyset}} |v_i|^2 |K|^{1/2} \right)^{1/2} \\
&\qquad \text{due to Lemma 6.6 or (6.4)} \\
&\cong \left(\sum_{K\in\mathcal{T}_h} |K| \sum_{\substack{i:\\ a_i\in K}} |v_i|^2 \right)^{1/2} \\
&\cong \left(\sum_{K\in\mathcal{T}_h} h_K^2 \sum_{\substack{i:\\ a_i\in K}} |v_i|^2 \right)^{1/2} ,
\end{aligned}$$

since due to the regularity of $(\mathcal{T}_h)_h$ there is a uniform lower bound for the angles of $K \in \mathcal{T}_h$ (see (3.93)) and thus a uniform upper bound on the number of $K \in \mathcal{T}_h$ such that $K \cap \Omega_i \neq \emptyset$.

□

Corollary 6.17 *Under the assumptions of Theorem* 6.15 *and for a regular family of triangulations* $(\mathcal{T}_h)_h$ *there exists a constant* $\alpha > 0$ *independent of* h *such that*

$$a_h(v_h, v_h) \geq \alpha \|v_h\|_1^2 \quad \textit{for all } v_h \in V_h .$$

Proof: By Remark 6.16 and Theorem 6.15,

$$a_h(v_h, v_h) \geq k\,|v_h|_1^2 + r_0 C_1^2 \|v_h\|_0^2 ,$$

i.e., we can take $\alpha := \min\{k; r_0 C_1^2\}$. □

Theorem 6.15 (or Corollary 6.17) asserts the stability of the method. It is the fundamental result for the proof of an error estimate.

Theorem 6.18 *Let* $\{\mathcal{T}_h\}_{h\in(0,\bar{h}]}$ *be a regular family of conforming triangulations, where in the case of the Voronoi diagram it is additionally required that all triangles be nonobtuse. Furthermore, suppose in* (6.5) *that* $k > 0$, $c_1, c_2, \nabla\cdot c, r \in C(\overline{\Omega})$, $r + \frac{1}{2}\nabla\cdot c \geq r_0 = \text{const} > 0$ *on* Ω, $f \in C^1(\overline{\Omega})$, *and that the approximations* γ_{ij}, *respectively* r_i, *are chosen according to* (6.7),

respectively (6.8). *Let the* r_{ij} *be defined by* (6.12) *with* R *satisfying* (P1) *and* (P2). *If the exact solution* u *of* (6.5) *belongs to* $H^2(\Omega)$ *and* $u_h \in V_h$ *denotes the solution of* (6.11), *then*

$$\|u - u_h\|_1 \le C\, h\, [\|u\|_2 + |f|_{1,\infty}] \ ,$$

where the constant $C > 0$ *is independent of* h.

Proof: The proof rests on a similar idea to those in the proof and the application of Strang's first lemma (Theorem 3.38) in Section 3.6.

Denoting by $I_h : C\left(\bar{\Omega}\right) \to V_h$ the interpolation operator defined in (3.71) and setting $v_h := u_h - I_h(u)$, we have

$$\begin{aligned} a_h(v_h, v_h) &= a_h(u_h, v_h) - a_h(I_h(u), v_h) \\ &= \langle f, v_h\rangle_{0,h} - a_h(I_h(u), v_h) \\ &= \langle f, v_h\rangle_{0,h} - \sum_{i\in\Lambda} v_i \int_{\Omega_i} f\, dx + \sum_{i\in\Lambda} v_i \int_{\Omega_i} f\, dx - a_h(I_h(u), v_h). \end{aligned}$$

By the definition of the discrete form $\langle f, v_h\rangle_{0,h}$ and by the differential equation (6.5), considered as an equation in $L^2(\Omega)$, we get

$$a_h(v_h, v_h) = \sum_{i\in\Lambda} v_i \int_{\Omega_i} (f_i - f)\, dx + \sum_{i\in\Lambda} v_i \int_{\Omega_i} Lu\, dx - a_h(I_h(u), v_h)\,,$$

where $Lu = -\nabla\cdot(k\,\nabla u - c\,u) + r\,u$.

For $f \in C^1(\overline{\Omega})$ and the choice $f_i := f(a_i)$, it is easy to see that

$$|f_i - f(x)| \le |f|_{1,\infty} \max_{K: a_i \in K} h_K \le C\, h |f|_{1,\infty} \quad \text{for all } x \in \Omega_i\,.$$

So it follows that

$$\begin{aligned} \left|\sum_{i\in\Lambda} v_i \int_{\Omega_i} (f_i - f)\, dx\right| &\le C\, h |f|_{1,\infty} \sum_{i\in\Lambda} |v_i| m_i \\ &\le C\, h |f|_{1,\infty} \left\{\sum_{i\in\Lambda} v_i^2 m_i\right\}^{1/2} \underbrace{\left\{\sum_{i\in\Lambda} m_i\right\}^{1/2}}_{\le \sqrt{|\Omega|}} \\ &\le C\, h |f|_{1,\infty} \|v_h\|_{0,h}. \end{aligned}$$

For the other choice of f_i (see (6.8)), the same estimate is trivially satisfied. The difficult part of the proof is to get an estimate of the consistency error

$$\left|\sum_{i\in\Lambda} v_i \int_{\Omega_i} Lu\, dx - a_h(I_h(u), v_h)\right| .$$

This is very extensive, and so we will omit the details. A complete proof of the following result is given in the paper [40]:

$$\left| \sum_{i \in \Lambda} v_i \int_{\Omega_i} Lu \, dx - a_h(I_h(u), v_h) \right| \leq C \, h \|u\|_2 \left\{ |v_h|_1^2 + \|v_h\|_{0,h}^2 \right\}^{1/2} . \quad (6.22)$$

Putting both estimates together and taking into consideration Remark 6.16, we arrive at

$$\begin{aligned} a_h(v_h, v_h) &\leq C \, h \left[\|u\|_2 + |f|_{1,\infty} \right] \left\{ |v_h|_1^2 + \|v_h\|_{0,h}^2 \right\}^{1/2} \\ &\leq C \, h \left[\|u\|_2 + |f|_{1,\infty} \right] \|v_h\|_1 . \end{aligned}$$

By Corollary 6.17, we conclude from this that

$$\|v_h\|_1 \leq C \, h \left[\|u\|_2 + |f|_{1,\infty} \right] .$$

It remains to apply the triangle inequality and the standard interpolation error estimate (cf. Theorem 3.29 with $k = 1$ or Theorem 3.35)

$$\|u - u_h\|_1 \leq \|u - I_h(u)\|_1 + \|v_h\|_1 \leq C \, h \left[\|u\|_2 + |f|_{1,\infty} \right] .$$

□

We point out that the error measured in the H^1-seminorm is of the same order as for the finite element method with linear finite elements.

Now we will turn to the investigation of some interesting properties of the method.

Global Conservativity

Here we consider the boundary value problem

$$\begin{aligned} -\nabla \cdot (k \nabla u - c \, u) &= f \quad \text{in } \Omega , \\ \nu \cdot (k \nabla u - c \, u) &= g \quad \text{on } \partial\Omega . \end{aligned}$$

Integrating the differential equation over Ω, we conclude from Gauss's divergence theorem that

$$- \int_\Omega \nabla \cdot (k \nabla u - c \, u) \, dx = - \int_{\partial\Omega} \nu \cdot (k \nabla u - c \, u) \, d\sigma = - \int_{\partial\Omega} g \, d\sigma ,$$

and hence

$$\int_{\partial\Omega} g \, d\sigma + \int_\Omega f \, dx = 0 .$$

This is a necessary compatibility condition for the data describing the balance between the total flow over the boundary and the distributed sources. We will demonstrate that the discretization requires a discretized version of this compatibility condition, which is called *discrete global conservativity.*

Therefore, we first have to define the discretization for the above type of boundary conditions. Obviously, for inner control volumes Ω_i ($i \in \Lambda$), there

is no need for any modifications. So we have to consider only the boundary control volumes Ω_i ($i \in \partial\Lambda := \overline{\Lambda} \setminus \Lambda$).

In the case of the Voronoi diagram we have

$$\begin{aligned} -\int_{\Omega_i} \nabla \cdot (k\,\nabla u - c\,u)\; dx &= -\int_{\partial\Omega_i} \nu \cdot (k\,\nabla u - c\,u)\; d\sigma \\ &= -\sum_{j\in\Lambda_i} \int_{\Gamma_{ij}} \nu \cdot (k\,\nabla u - c\,u)\; d\sigma - \int_{\partial\Omega_i\cap\partial\Omega} \nu \cdot (k\,\nabla u - c\,u)\; d\sigma \quad (6.23) \\ &= -\sum_{j\in\Lambda_i} \int_{\Gamma_{ij}} \nu \cdot (k\,\nabla u - c\,u)\; d\sigma - \int_{\partial\Omega_i\cap\partial\Omega} g\, d\sigma\,. \end{aligned}$$

Since the line integrals over Γ_{ij} can be approximated in the standard way, we get the following equation:

$$\begin{aligned} \sum_{i\in\overline{\Lambda}} v_i \sum_{j\in\Lambda_i} \left\{ \mu_{ij}\,\frac{u_i - u_j}{d_{ij}} + \gamma_{ij}\;[r_{ij}\,u_i + (1-r_{ij})\;u_j] \right\} m_{ij} \qquad (6.24) \\ -\sum_{i\in\overline{\Lambda}} v_i \int_{\partial\Omega_i\cap\partial\Omega} g\, d\sigma \;=\; \sum_{i\in\overline{\Lambda}} f_i\, v_i\, m_i\,, \end{aligned}$$

where the ansatz and test space V_h consists of all continuous functions over $\bar{\Omega}$ that are piecewise linear with respect to the underlying triangulation (that is, in the boundary nodes no function values are prescribed).

It is again assumed that the r_{ij} are defined by (6.12) with a function R satisfying (P1) and $\gamma_{ji} = -\gamma_{ij}$.

Obviously, the particular function $i_h :\equiv 1$ belongs to V_h. So we are allowed to set $v_h = i_h$ in the discretization. Then, repeating the above symmetry argument (cf. the proof of Lemma 6.12), we get

$$\sum_{i\in\overline{\Lambda}} \sum_{j\in\Lambda_i} \mu_{ij}\;(u_i - u_j)\,\frac{m_{ij}}{d_{ij}} = -\sum_{i\in\overline{\Lambda}} \sum_{j\in\Lambda_i} \mu_{ij}\;(u_i - u_j)\,\frac{m_{ij}}{d_{ij}}\,,$$

that is,

$$\sum_{i\in\overline{\Lambda}} \sum_{j\in\Lambda_i} \mu_{ij}\;(u_i - u_j)\,\frac{m_{ij}}{d_{ij}} = 0\,.$$

On the other hand, using the same argument, we have

$$\begin{aligned} &\sum_{i\in\overline{\Lambda}} \sum_{j\in\Lambda_i} [r_{ij}\,u_i + (1-r_{ij})\,u_j]\,\gamma_{ij}\,m_{ij} \\ &\quad = \sum_{i\in\overline{\Lambda}} \sum_{j\in\Lambda_i} [r_{ji}\,u_j + (1-r_{ji})\;u_i]\;\gamma_{ji}\,m_{ji} \\ &\quad = -\sum_{i\in\overline{\Lambda}} \sum_{j\in\Lambda_i} [(1-r_{ij})\,u_j + r_{ij}\,u_i]\,\gamma_{ij}\,m_{ij}\,. \qquad (6.25) \end{aligned}$$

Consequently, this term vanishes, too. Because of

$$\sum_{i\in\overline{\Lambda}} v_i \int_{\partial\Omega_i\cap\partial\Omega} g\,d\sigma = \int_{\partial\Omega} g\,d\sigma\,,$$

it follows that

$$-\int_{\partial\Omega} g\,d\sigma = \sum_{i\in\overline{\Lambda}} f_i\,v_i\,m_i = \sum_{i\in\overline{\Lambda}} f_i\,m_i \quad \left(\approx \int_\Omega f\,dx\right). \tag{6.26}$$

This is the mentioned compatibility condition. It ensures the solvability of the discrete system (6.24).

In the case of the Donald diagram, we obviously have

$$\langle k\nabla u_h, \nabla v_h\rangle_0 = 0\,.$$

Since the proof of (6.25) does not depend on the particular type of the control volumes, the property of discrete global conservativity in the sense of (6.26) is satisfied for the Donald diagram, too.

Inverse Monotonicity

The so-called *inverse monotonicity* is a further important property of the boundary value problem (6.5) that is inherited by the finite volume discretization without any additional restrictive assumptions. Namely, it is well known that under appropriate assumptions on the coefficients, the solution u is nonnegative if the (continuous) right-hand side f in (6.5) is nonnegative in Ω.

We will demonstrate that this remains true for the approximative solution u_h. Only at this place is the property (P3) of the weighting function R used; the preceding results are also valid for the simple case $R(z) \equiv \frac{1}{2}$.

There is a close relation to the maximum principles investigated in Sections 1.4 and 3.9. However, the result given here is weaker, and the proof is based on a different technique.

Theorem 6.19 *Let the assumptions of Theorem* 6.15 *be satisfied, but R in (6.12) has to satisfy* (P1)–(P3). *Further, suppose that $f \in C(\overline{\Omega})$ and $f(x) \geq 0$ for all $x \in \Omega$. Moreover, in the case of the Donald diagram, only the weighting function $R(z) = \frac{1}{2}\,[\mathrm{sign}\,(z) + 1]$ is permitted.*

Then

$$u_h(x) \geq 0 \quad \textit{for all } x \in \Omega\,.$$

Proof: We start with the case of the Voronoi diagram. Let u_h be the solution of (6.11) with $f(x) \geq 0$ for all $x \in \Omega$. Then we have the following additive decomposition of u_h:

$$u_h = u_h^+ - u_h^-\,, \qquad \text{where } u_h^+ := \max\{0, u_h\}\,.$$

In general, u_h^+, u_h^- do not belong to V_h. So we interpolate them in V_h and set in (6.11) $v_h := I_h(u_h^-)$, where $I_h : C\left(\bar{\Omega}\right) \to V_h$ is the interpolation

operator (3.71). It follows that

$$0 \le \langle f, v_h \rangle_{0,h} = a_h(u_h, v_h) = a_h\left(I_h(u_h^+), I_h(u_h^-)\right) - a_h\left(I_h(u_h^-), I_h(u_h^-)\right) .$$

By Theorem 6.15, we have

$$k\left|I_h(u_h^-)\right|_1^2 \le a_h\left(I_h(u_h^-), I_h(u_h^-)\right) \le a_h\left(I_h(u_h^+), I_h(u_h^-)\right) .$$

If we were able to show that $a_h\left(I_h(u_h^+), I_h(u_h^-)\right) \le 0$, then the theorem would be proven, because this relation implies $\left|I_h(u_h^-)\right|_1 = 0$, and from this we immediately get $u_h^- = 0$, and so $u_h = u_h^+ \ge 0$.

Since $u_i^+ u_i^- = 0$ for all $i \in \Lambda$, it follows from (6.19) in the proof of Lemma 6.12 that

$$b_h\left(I_h(u_h^+), I_h(u_h^-)\right) = \sum_{i\in\Lambda} \sum_{j\in\Lambda_i} (1 - r_{ij})\; u_j^+ u_i^- \gamma_{ij} m_{ij} . \tag{6.27}$$

Furthermore, obviously $d_h\left(I_h(u_h^+), I_h(u_h^-)\right) = 0$ holds. Thus

$$\begin{aligned} a_h\left(I_h(u_h^+), I_h(u_h^-)\right) &= \sum_{i\in\Lambda} \sum_{j\in\Lambda_i} \left[-\frac{\mu_{ij}}{d_{ij}}\, u_j^+ + \gamma_{ij}\,(1 - r_{ij})\, u_j^+\right] u_i^- m_{ij} \\ &= -\sum_{i\in\Lambda} \sum_{j\in\Lambda_i} \frac{\mu_{ij}}{d_{ij}} \left[1 - \frac{\gamma_{ij}\, d_{ij}}{\mu_{ij}}\,(1 - r_{ij})\right] u_j^+ u_i^- m_{ij} . \end{aligned}$$

Due to $1 - [1 - R(z)]\, z \ge 0$ for all $z \in \mathbb{R}$ (cf. property (P3) in (6.13)) and $u_j^+ u_i^- \ge 0$, it follows that

$$a_h\left(I_h(u_h^+), I_h(u_h^-)\right) \le 0 .$$

So it remains to investigate the case of the Donald diagram. The function $R(z) = \frac{1}{2}\left[\operatorname{sign}(z) + 1\right]$ has the property

$$[1 - R(z)]\, z = \frac{1}{2}\left[1 - \operatorname{sign}(z)\right] z \le 0 \quad \text{for all } z \in \mathbb{R} ,$$

that is (cf. (6.27)),

$$b_h\left(I_h(u_h^+), I_h(u_h^-)\right) \le 0 .$$

Taking $u_i^+ u_i^- = 0$ into consideration, we get

$$\begin{aligned} a_h\left(I_h(u_h^+), I_h(u_h^-)\right) &\le \left\langle k\, \nabla I_h(u_h^+), \nabla I_h(u_h^-)\right\rangle_0 \\ &= k \sum_{i\in\Lambda} \sum_{j\in\Lambda_i} u_j^+ u_i^- \left\langle \nabla\varphi_j, \nabla\varphi_i\right\rangle_0 . \end{aligned}$$

Now Lemma 3.47 implies that

$$a_h\left(I_h(u_h^+), I_h(u_h^-)\right) \le -\frac{k}{2} \sum_{i\in\Lambda} \sum_{j\in\Lambda_i} u_j^+ u_i^- \left(\cot\alpha_{ij}^K + \cot\alpha_{ij}^{K'}\right) ,$$

where K and K' are a pair of triangles sharing a common edge with vertices a_i, a_j.

Since all triangles are nonobtuse, we have $\cot \alpha_{ij}^K \geq 0$, $\cot \alpha_{ij}^{K'} \geq 0$, and hence

$$a_h \left(I_h(u_h^+), I_h(u_h^-) \right) \leq 0 \,.$$

□

Exercises

6.2 Suppose that the domain $\Omega \subset \mathbb{R}^2$ can be triangulated by means of equilateral triangles with edge length $h > 0$ in an admissible way.

(a) Give the shape of the control domains in the case of the Voronoi and the Donald diagrams.

(b) Using the control domains from subproblem (a), discretize the Poisson equation with homogeneous Dirichlet boundary conditions by means of the finite volume method.

6.3 Formulate an existence result for the weak solution in $H_0^1(\Omega)$ of the boundary value problem (6.5) similar to Theorem 3.12. In particular, what form will condition (3.17) take?

6.4 Verify Remark 6.7; i.e., show that $\langle \cdot, \cdot \rangle_{0,h}$ possesses the properties of a scalar product on V_h.

6.5 Prove Remark 6.16 in detail.

6.6 Verify or disprove the properties (P1)–(P3) for the three weighting functions given before (6.13) and for $R \equiv \frac{1}{2}$.

6.7 Let K be a nonobtuse triangle with the vertices a_1, a_2, a_3. The length of the segments $\Gamma_{ij}^K := \Gamma_{ij} \cap K$ is denoted by m_{ij}^K, and d_{ij} is the length of the edge connecting a_i with a_j. Finally, α_{ij}^K is the interior angle of K opposite that edge.

Demonstrate the following relation: $2m_{ij}^K = d_{ij} \cot \alpha_{ij}^K$.

6.8

(a) Formulate problem (6.11) in terms of an algebraic system of type (1.31).

(b) Show that for the resulting matrix $A_h \in \mathbb{R}^{M_1, M_1}$, where M_1 is the number of elements of the index set Λ, the following relation is valid: $A_h^T \mathbf{1} \geq \mathbf{0}$. Here, as in Section 1.4, $\mathbf{0}$, respectively $\mathbf{1}$, denotes a vector of dimension M_1 whose components are *all* equal to 0, respectively 1. (This is nothing other than the property (1.32)(3)(i) except for the transpose of A_h.)

7

Discretization Methods for Parabolic Initial Boundary Value Problems

7.1 Problem Setting and Solution Concept

In this section initial boundary value problems for the linear case of the differential equation (0.33) are considered. We choose the form (3.12) together with the boundary conditions (3.18)–(3.20), which have already been discussed in Section 0.5. In Section 3.2 conditions have been developed to ensure a unique weak solution of the stationary boundary value problem. In contrast to Chapter 3, the heterogeneities are now allowed also to depend on time t, but for the sake of simplicity we do not do so for the coefficients in the differential equations and the boundary conditions, which covers most of the applications, for example from Chapter 0. Also for the sake of simplicity, we take the coefficient in front of the time derivative to be constant and thus 1 by a proper scaling. From time to time we will restrict attention to homogeneous Dirichlet boundary conditions for further ease of exposition. Thus the problem reads as follows:

The domain Ω is assumed to be a bounded Lipschitz domain and we suppose that $\Gamma_1, \Gamma_2, \Gamma_3$ form a disjoint decomposition of the boundary $\partial\Omega$ (cf. (0.39)):

$$\partial\Omega = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3 \ ,$$

where Γ_3 is a closed subset of the boundary.

In the space-time cylinder $Q_T = \Omega \times (0,T)$, $T > 0$, and its boundary $S_T = \partial\Omega \times (0,T)$ there are given functions $f : Q_T \to \mathbb{R}$, $g : S_T \to \mathbb{R}$, $g(x,t) = g_i(x,t)$ for $x \in \Gamma_i$, $i = 1,2,3$, and $u_0 : \Omega \to \mathbb{R}$. The problem is to

find a function $u : Q_T \to \mathbb{R}$ such that

$$\begin{aligned} \frac{\partial u}{\partial t} + Lu &= f \quad \text{in} \quad Q_T\,, \\ Ru &= g \quad \text{on} \quad S_T\,, \\ u &= u_0 \quad \text{on} \quad \Omega \times \{0\}\,, \end{aligned} \tag{7.1}$$

where Lv denotes the differential expression for some function $v : \Omega \to \mathbb{R}$,

$$(Lv)(x) := -\nabla \cdot (K(x)\,\nabla v(x)) + c(x) \cdot \nabla v(x) + r(x)v(x) \tag{7.2}$$

with sufficiently smooth, time-independent coefficients

$$K : \Omega \to \mathbb{R}^{d,d}\,, \quad c : \Omega \to \mathbb{R}^d\,, \quad r : \Omega \to \mathbb{R}\,.$$

The boundary condition is expressed by the shorthand notation $Ru = g$, which means, for a function $\alpha : \Gamma_2 \to \mathbb{R}$ on $\partial\Omega$,

- Neumann boundary condition (cf. (0.41) or (0.36))

$$K\nabla u \cdot \nu = \partial_{\nu_K} u = g_1 \quad \text{on } \Gamma_1 \times (0,T)\,, \tag{7.3}$$

- mixed boundary condition (cf. (0.37))

$$K\nabla u \cdot \nu + \alpha u = \partial_{\nu_K} u + \alpha u = g_2 \quad \text{on } \Gamma_2 \times (0,T)\,, \tag{7.4}$$

- Dirichlet boundary condition (cf. (0.38))

$$u = g_3 \quad \text{on } \Gamma_3 \times (0,T)\,. \tag{7.5}$$

Thus the stationary boundary problem considered so far reads

$$\begin{aligned} Lu(x) &= f(x) \quad \text{for } x \in \Omega\,, \\ Ru(x) &= g(x) \quad \text{for } x \in \partial\Omega\,. \end{aligned} \tag{7.6}$$

It is to be expected that both for the analysis and the discretization there are strong links between (7.6) and (7.1). The formulation (7.1) in particular includes the heat equation (cf. (0.20))

$$\frac{\partial u}{\partial t} - \nabla \cdot (K\nabla u) = f \quad \text{in } Q_T, \tag{7.7}$$

or for constant scalar coefficients in the form (cf. (0.19))

$$\frac{\partial u}{\partial t} - \Delta u = f \quad \text{in } Q_T \tag{7.8}$$

with appropriate initial and boundary conditions.
Again as in Chapter 1, one of the simplest cases will be, for two space dimensions ($d = 2$), the case of a rectangle $\Omega = (0,a) \times (0,b)$ or even the case $d = 1$ (with $\Omega = (0,a)$), for which (7.8) further reduces to

$$\frac{\partial u}{\partial t} - \frac{\partial^2}{\partial x^2} u = 0 \quad \text{in } Q_T = (0,a) \times (0,T). \tag{7.9}$$

For problem (7.1), the following typical analytical questions arise:

- existence of (classical) solutions,
- properties of the (classical) solutions,
- weaker concepts of the solution.

As in the case of elliptic boundary value problems, the theory of classical solutions requires comparatively strong assumptions on the data of the initial-boundary value problem. In particular, along the edge $\partial\Omega \times \{0\}$ of the space-time cylinder initial and boundary conditions meet, so that additional compatibility conditions have to be taken into account.

Representation of Solutions in a Special Case
To enhance the familiarity with the problem and for further comparison we briefly sketch a method, named *separation of variables*, by which closed-form solutions in the form of infinite series can be obtained for special cases. Also in these cases, the representations are not meant to be a numerical method (by its evaluation), but only serve as a theoretical tool.

We start with the case of homogeneous data, i.e., $f = 0$, $g_i = 0$ $(i = 1, 2, 3)$, so that the process is determined only by the initial data u_0.

We assume a solution of (7.1) to have the form $u(x,t) = v(t)w(x)$ with $v(t) \neq 0,\ w(x) \neq 0$. This leads to

$$\frac{v'(t)}{v(t)} = \frac{-Lw(x)}{w(x)}\ , \quad x \in \Omega,\ t \in (0,T)\ . \tag{7.10}$$

Therefore, the expressions in (7.10) must be constant, for example, equal to $-\lambda$ for $\lambda \in \mathbb{R}$. Therefore,

$$v'(t) = -\lambda v(t)\ , \quad t \in (0,T), \tag{7.11}$$

which for the initial conditions $v(0) = 1$ has the solution

$$v(t) = e^{-\lambda t}\ .$$

Furthermore, w has to satisfy

$$\begin{aligned} Lw(x) &= \lambda w(x)\ , && x \in \Omega\ , \\ Rw(x) &= 0\ , && x \in \partial\Omega\ . \end{aligned} \tag{7.12}$$

Such a function $w : \bar{\Omega} \to \mathbb{R},\ w \neq 0$, is called an *eigenfunction* for the *eigenvalue* λ of the boundary value problem (7.6). If $(w_i, \lambda_i),\ i = 1, \ldots, N$, are eigenfunctions/values for (7.6), then because of the superposition principle, the function

$$u(x,t) := \sum_{i=1}^{N} c_i e^{-\lambda_i t} w_i(x) \tag{7.13}$$

is a solution of the homogeneous initial-boundary value problem for the initial value

$$u_0(x) := \sum_{i=1}^{N} c_i w_i(x) \,, \tag{7.14}$$

where the $c_i \in \mathbb{R}$ are arbitrary. If there are infinitely many eigenfunctions/values (w_i, λ_i) and if the sums in (7.13) and (7.14) converge in such a way that also the infinite series possesses the derivatives appearing in (7.6), then also

$$u(x,t) = \sum_{i=1}^{\infty} c_i e^{-\lambda_i t} w_i(x) \tag{7.15}$$

is a solution to

$$u_0(x) = \sum_{i=1}^{\infty} c_i w_i(x) \,. \tag{7.16}$$

For an inhomogeneous right-hand side of the form

$$f(x,t) = \sum_{i=1}^{N} f_i(t) w_i(x) \tag{7.17}$$

the solution representation can be extended to *(variation of constants formula)*

$$u(x,t) := \sum_{i=1}^{N} c_i e^{-\lambda_i t} w_i(x) + \sum_{i=1}^{N} \int_0^t f_i(s) e^{-\lambda_i (t-s)} ds \; w_i(x) \,, \tag{7.18}$$

and at least formally the sum can be replaced by the infinite series. To verify (7.18) it suffices to consider the case $u_0 = 0$, for which we have

$$\begin{aligned} (\partial_t u)(x,t) &= \sum_{i=1}^{N} f_i(t) w_i(x) - \sum_{i=1}^{N} \int_0^t f_i(s) e^{-\lambda_i (t-s)} ds \; \lambda_i w_i(x) \\ &= f(x,t) - L\left(\sum_{i=1}^{N} \int_0^t f_i(s) e^{-\lambda_i (t-s)} ds \; w_i \right)(x) \\ &= f(x,t) - L(u)(x,t) \,. \end{aligned} \tag{7.19}$$

From these solution representations we can conclude that initial data (and thus also perturbances contained in it) and also the influence of the right-hand side act only exponentially damped if all eigenvalues are positive.

For $d = 1$, $\Omega = (0, a)$ and Dirichlet boundary conditions we have the eigenfunctions

$$w^\nu(x) = \sin\left(\nu \, \frac{\pi}{a} \, x \right) \,, \quad \nu \in \mathbb{N} \,, \tag{7.20}$$

for the eigenvalues

$$\lambda^\nu = \left(\frac{\nu\pi}{a}\right)^2 \ . \tag{7.21}$$

If the initial data u_0 has the representation

$$u_0(x) = \sum_{\nu=1}^{\infty} c_i \sin\left(\nu\,\frac{\pi}{a}\,x\right) \ , \tag{7.22}$$

then for example for $f = 0$ the (formal) solution reads

$$u(x,t) = \sum_{\nu=1}^{\infty} c_i e^{-\lambda_\nu t} \sin\left(\nu\,\frac{\pi}{a}\,x\right) \ . \tag{7.23}$$

The eigenfunctions w^ν are orthogonal with respect to the scalar product $\langle\cdot,\cdot\rangle_0$ in $L^2(\Omega)$, since they satisfy

$$\left\langle \sin\left(\nu\frac{\pi}{a}\cdot\right), \sin\left(\mu\frac{\pi}{a}\cdot\right)\right\rangle_0 = \begin{cases} 0 & \text{for} \quad \nu \neq \mu\,, \\ \frac{a}{2} & \text{for} \quad \nu = \mu\,, \end{cases} \tag{7.24}$$

which can by checked by means of well-known identities for the trigonometric functions.
Therefore (see below (7.57)),

$$c_i = \frac{\langle u_0, w^\nu\rangle_0}{\langle w^\nu, w^\nu\rangle_0} \ , \tag{7.25}$$

which is called the *Fourier coefficient* in the *Fourier expansion* of u_0.

Of course, the (w^ν, λ^ν) depend on the boundary conditions. For Neumann boundary conditions in $x = 0$ and $x = a$ we have

$$\begin{aligned} w^\nu(x) &= \cos\left(\nu\tfrac{\pi}{a}\,x\right) \ , \qquad && \nu = 0, 1, \ldots, \\ \lambda^\nu &= \left(\nu\tfrac{\pi}{a}\right)^2 \ , && \nu = 0, 1, \ldots . \end{aligned} \tag{7.26}$$

The occurrence of $w^0 = 1,\ \lambda^0 = 0$ reflects the nontrivial solvability of the pure Neumann problem (which therefore is excluded by the conditions of Theorem 3.15).

For $Lu = -\Delta u$ and $\Omega = (0,a) \times (0,b)$, eigenfunctions and eigenvalues can be derived from the one-dimensional case because of

$$-\Delta(v^\nu(x)\tilde{v}^\mu(y)) = -v^{\nu\prime\prime}(x)\tilde{v}^\mu(y) - v^\nu(x)\tilde{v}^{\mu\prime\prime}(y) = (\lambda^\nu + \tilde{\lambda}^\mu)v^\nu(x)\tilde{v}^\mu(y).$$

Therefore, for $\Omega = (0,a)\times(0,b)$ one has to choose the eigenfunctions/values (v^ν, λ^ν) (in x, on $(0,a)$) for the required boundary conditions at $x = 0$ and $x = a$, and $(\tilde{v}^\mu, \tilde{\lambda}^\mu)$ (in y, on $(0,b)$) for the required boundary conditions at $y = 0,\ y = b$.

For Dirichlet boundary conditions everywhere this leads to

$$w^{\nu\mu}(x,y) = \sin\left(\nu\frac{\pi}{a}x\right)\sin\left(\mu\frac{\pi}{b}x\right) \tag{7.27}$$

for the eigenvalues

$$\lambda^{\nu\mu} = \left(\frac{\nu\pi}{a}\right)^2 + \left(\frac{\mu\pi}{b}\right)^2$$

(i.e., the smallest eigenvalue is $\left(\frac{\pi}{a}\right)^2 + \left(\frac{\pi}{b}\right)^2$ and $\lambda^{\nu\mu} \to \infty$ for $\nu \to \infty$ or $\mu \to \infty$).

As a further concluding example we note the case

$x = 0$ or $x = a$: $u(x, y) = 0$ for $y \in [0, b]$,
$y = 0 : \nabla u \cdot \nu(x, y) = -\partial_2 u(x, y) = 0$ for $x \in (0, a)$,
$y = b : \nabla u \cdot \nu(x, y) = \partial_2 u(x, y) = 0$ for $x \in (0, a)$.

Eigenfunctions:

$$w^{\nu\mu}(x, y) = \sin\left(\nu\frac{\pi}{a}x\right)\cos\left(\mu\frac{\pi}{b}y\right) , \qquad (7.28)$$
$$\nu = 1, 2, \ldots , \quad \mu = 0, 1, 2, \ldots .$$

Eigenvalues:

$$\lambda^{\nu\mu} = \left(\nu\frac{\pi}{a}\right)^2 + \left(\mu\frac{\pi}{b}\right)^2 .$$

A Sketch of the Theory of Weak Solutions

As in the study of the elliptic boundary value problems (3.12), (3.18)–(3.20), for equation (7.1) a weak formulation can be given that reduces the requirements with respect to the differentiability properties of the solution.

The idea is to treat time and space variables in a different way:

(1)
- For fixed $t \in (0, T)$, the function $x \mapsto u(x, t)$ is interpreted as a parameter-dependent element $u(t)$ of some space V whose elements are functions of $x \in \Omega$. An obvious choice is (see Subsection 3.2.1, (I)) the space

$$V = \{v \in H^1(\Omega) : \ v = 0 \text{ on } \Gamma_3\}.$$

- In a next step, that is, for varying t, a function $t \mapsto u(t)$ results with values in the (function) space V.

(2) In addition to V, a further space $H = L^2(\Omega)$ occurs, from which the initial value u_0 is taken and which contains V as a dense subspace. A subspace V is called *dense* in H if the closure of V with respect to the norm on H coincides with H.

(3) The time derivative is understood in a generalized sense; see (7.29).

(4) The generalized solution $t \mapsto u(t)$ is sought as an element of a function space, the elements of which are "function-valued" (cf. (1)).

Definition 7.1 Let X denote one of the spaces H or V (in particular, this means that the elements of X are functions on $\Omega \subset \mathbb{R}^d$).

(i) The space $C^l([0,T],X)$, $l \in \mathbb{N}_0$, consists of all continuous functions $v : [0,T] \to X$ that have continuous derivatives up to the order l on $[0,T]$ with the norm

$$\sum_{i=0}^{l} \sup_{t\in(0,T)} \|v^{(i)}(t)\|_X \,.$$

For the sake of simplicity, the notation $C([0,T],X) := C^0([0,T],X)$ is used.

(ii) The space $L^p((0,T),X)$ with $1 \le p \le \infty$ consists of all functions on $(0,T)\times\Omega$ with the following properties:

$v(t,\cdot) \in X$ for any $t \in (0,T)$, $F \in L^p(0,T)$ with $F(t) := \|v(t,\cdot)\|_X$.

Furthermore,

$$\|v\|_{L^p((0,T),X)} := \|F\|_{L^p(0,T)} \,.$$

Remark 7.2 $f \in L^2(Q_T) \Rightarrow f \in L^2((0,T),H)$.

Proof:
Basically, the proof is a consequence of Fubini's theorem (see [1]). □

Concerning the interpretation of the time derivative and of the weak formulation, a comprehensive treatment is possible only within the framework of the theory of distributions; thus a detailed explanation is beyond the scope of this book. A short but mathematically rigorous introduction can be found in the book [39, Chapter 23].

The basic idea consists in the following definition:

A function $u \in L^2((0,T),V)$ is said to have a *weak derivative* w if the following holds:

$$\int_0^T u(t)\,\Psi'(t)\,dt = -\int_0^T w(t)\,\Psi(t)\,dt \quad \text{for all } \Psi \in C_0^\infty(0,T)\,. \tag{7.29}$$

Usually, this derivative w is denoted by $\frac{du}{dt}$ or u'.

Remark 7.3 The integrals occurring above are to be understood as so-called *Bochner* integrals and are extensions of the Lebesgue integral to function-valued mappings. Therefore, equation (7.29) is an equality of functions.

Before we give a weak formulation of (7.1), the following notion is worth recalling:

$$\langle u,v\rangle_0 := \int_\Omega u\,v\,dx \;\text{ for }\; u,v \in H\,, \tag{7.30}$$

$$a(u,v) := \int_\Omega [K\nabla u\cdot\nabla v + (c\cdot\nabla u + ru)\,v]\,dx + \int_{\Gamma_2} \alpha uv\,d\sigma,\; u,v\in V. \tag{7.31}$$

Let $u_0 \in H$, $f \in L^2((0,T),H)$, and in case of Dirichlet conditions we restrict ourselves to the homogeneous case.

An element $u \in L^2((0,T),V)$ is called a *weak solution* of (7.1) if it has a weak derivative $\frac{du}{dt} = u' \in L^2((0,T),H)$ and the following holds

$$\begin{aligned} \left\langle \frac{d}{dt}u(t), v \right\rangle_0 + a(u(t),v) &= \langle f(t), v\rangle_0 + \int_{\Gamma_1} g_1(\cdot,t)v \, d\sigma \\ &\quad + \int_{\Gamma_2} g_2(\cdot,t)v \, d\sigma \\ &\quad \text{for all } v \in V \text{ and } t \in (0,T), \\ u(0) &= u_0 . \end{aligned} \tag{7.32}$$

Due to $u \in L^2((0,T),V)$ and $u' \in L^2((0,T),H)$, we also have $u \in C([0,T],H)$ (see [12, p. 287]), so that the initial condition is meaningful in the classical sense.

In what follows, the bilinear form a is assumed to be continuous on $V \times V$ (see (3.2)) and V-elliptic (see (3.3)). The latter means that there exists a number $\alpha > 0$ such that

$$a(v,v) \geq \alpha \|v\|_V^2 \quad \text{for all } v \in V .$$

Lemma 7.4 *Let a be a V-elliptic, continuous bilinear form, $u_0 \in H$, and $f \in C([0,T],H)$, and suppose the considered boundary conditions are homogeneous. Then, for the solution $u(t)$ of (7.32) the following estimate holds:*

$$\|u(t)\|_0 \leq \|u_0\|_0 \, e^{-\alpha t} + \int_0^t \|f(s)\|_0 \, e^{-\alpha(t-s)} \, ds \quad \textit{for all } t \in (0,T) .$$

Proof: The following equations are valid almost everywhere in $(0,T)$. Setting $v = u(t)$, (7.32) reads as

$$\langle u'(t), u(t)\rangle_0 + a(u(t),u(t)) = \langle f(t), u(t)\rangle_0 .$$

Using the relation

$$\langle u'(t), u(t)\rangle_0 = \frac{1}{2}\frac{d}{dt}\langle u(t), u(t)\rangle_0 = \frac{1}{2}\frac{d}{dt}\|u(t)\|_0^2 = \|u(t)\|_0 \frac{d}{dt}\|u(t)\|_0$$

and the V-ellipticity, it follows that

$$\|u(t)\|_0 \frac{d}{dt}\|u(t)\|_0 + \alpha \|u(t)\|_V^2 \leq \langle f(t), u(t)\rangle_0 .$$

Now the simple inequality

$$\|u(t)\|_0 \leq \|u(t)\|_V$$

and the Cauchy–Schwarz inequality

$$\langle f(t), u(t)\rangle_0 \leq \|f(t)\|_0 \|u(t)\|_0$$

yield, after division by $\|u(t)\|_0$, the estimate

$$\frac{d}{dt}\|u(t)\|_0 + \alpha\|u(t)\|_0 \leq \|f(t)\|_0 \,.$$

Multiplying this relation by $e^{\alpha t}$, the relation

$$\frac{d}{dt}(e^{\alpha t}\|u(t)\|_0) = e^{\alpha t}\frac{d}{dt}\|u(t)\|_0 + \alpha e^{\alpha t}\|u(t)\|_0$$

leads to

$$\frac{d}{dt}(e^{\alpha t}\|u(t)\|_0) \leq e^{\alpha t}\|f(t)\|_0 \,.$$

The integration over $(0,t)$ results in

$$e^{\alpha t}\|u(t)\|_0 - \|u(0)\|_0 \leq \int_0^t e^{\alpha s}\|f(s)\|_0 \, ds$$

for all $t \in (0,T)$. Multiplying this by $e^{-\alpha t}$ and taking into consideration the initial condition, we get the asserted relation

$$\|u(t)\|_0 \leq \|u_0\|_0 \, e^{-\alpha t} + \int_0^t \|f(s)\|_0 \, e^{-\alpha(t-s)} \, ds \,.$$

□

As a consequence of this lemma, the uniqueness of the solution of (7.32) is obtained.

Corollary 7.5 *Let a be a V-elliptic, continuous bilinear form. Then there exists at most one solution of* (7.32).

Proof: Suppose there are two different solutions $u_1(t), u_2(t) \in V$. Then the difference $v(t) := u_1(t) - u_2(t)$ solves a homogeneous problem of the type (7.32) (i.e., with $f = 0$, $u_0 = 0$). Lemma 7.4 immediately implies $\|v(t)\|_0 = 0$ in $[0,T)$; that is, $u_1(t) = u_2(t)$ for all $t \in [0,T)$. □

There is a close relation between Lemma 7.4 and solution representations such as (7.18) (with the sum being infinite). The eigenvalue problem (7.12) is defined as follows in its variational form (see also the end of Section 2.2):

Definition 7.6 A number $\lambda \in \mathbb{R}$ is called an *eigenvalue* for the *eigenvector* $w \in V$, $w \neq 0$, if

$$a(w,v) = \lambda\langle w, v\rangle_0 \quad \text{for all} \quad v \in V \,.$$

Assume that additionally to our assumptions the bilinear form is symmetric and the embedding of V into H is compact (see [26]), which is the case here. Then there are enough eigenvectors in the sense that a sequence (w_i, λ_i),

$0 < \lambda_1 \le \lambda_2 \le \ldots$, exists such that the w_i are orthonormal with respect to $\langle \cdot, \cdot \rangle_0$ and every $v \in V$ has a unique representation (in H) as

$$v = \sum_{i=1}^{\infty} c_i w_i \,. \tag{7.33}$$

As in (7.25) the Fourier coefficients c_i are given by

$$c_i = \langle v, w_i \rangle_0 \,. \tag{7.34}$$

In fact, (7.33) gives a rigorous framework to the specific considerations in (7.16) and subsequent formulas. From (7.33) and (7.34) we conclude *Parseval's identity*

$$\|v\|_0^2 = \sum_{i=1}^{\infty} |\langle v, w_i \rangle_0|^2. \tag{7.35}$$

Furthermore, the sequence $v_i := \lambda_i^{-1/2} w_i$ is orthogonal with respect to $a(\cdot, \cdot)$, and a representation corresponding to (7.33), (7.34) holds such that

$$a(v,v) = \sum_{i=1}^{\infty} |a(v,v_i)|^2 = \sum_{i=1}^{\infty} \lambda_i^{-1} |a(v,w_i)|^2 = \sum_{i=1}^{\infty} \lambda_i |\langle v, w_i \rangle_0|^2. \tag{7.36}$$

From (7.35) and (7.36) we see that the ellipticity constant can be interpreted as the smallest eigenvalue λ. In fact, the solution representation (7.18) (with the sum being infinite in H) also holds true under the assumptions mentioned and also leads to the estimate of Lemma 7.4. But note that the proof there does not require symmetry of the bilinear form.

Exercises

7.1 Consider the initial-boundary value problem

$$\begin{aligned} u_t - u_{xx} &= 0 && \text{in } (0,\infty) \times (0,\infty)\,, \\ u(0,t) &= h(t)\,, && t \in (0,\infty)\,, \\ u(x,0) &= 0\,, && x \in (0,\infty)\,, \end{aligned}$$

where $h : (0,\infty) \to \mathbb{R}$ is a differentiable function, the derivative of which has at most exponential growth.

(a) Show that the function

$$u(x,t) = \sqrt{\frac{2}{\pi}} \int_{x/\sqrt{2t}}^{\infty} \mathrm{e}^{-s^2/2} h\left(t - \frac{x^2}{2s^2}\right) ds$$

is a solution.

(b) Is u_t bounded in the domain of definition? If not, give conditions on h that guarantee the boundedness of u_t.

7.2 Consider the initial-boundary value problem in one space dimension

$$\begin{aligned} u_t - u_{xx} &= 0 && \text{in } (0,\pi)\times(0,\infty)\,,\\ u(0,t) = u(\pi,t) &= 0\,, && t\in(0,\infty)\,,\\ u(x,0) &= u_0(x)\,, && x\in(0,\pi)\,. \end{aligned}$$

(a) Solve it by means of the method of separation.

(b) Give a representation for $\|u_t(t)\|_0$.

(c) Consider the particular initial condition $u_0(x) = \pi - x$ and investigate, using the result from subproblem (b), the asymptotic behaviour of $\|u_t(t)\|_0$ near $t = 0$.

7.3 Let the domain $\Omega \subset \mathbb{R}^d$ be bounded by a sufficiently smooth boundary and set $V := H_0^1(\Omega)$, $H := L^2(\Omega)$. Furthermore, $a : V \times V \to \mathbb{R}$ is a continuous, V-elliptic, symmetric bilinear form and $u_0 \in H$. Prove by using the so-called *energy method* (cf. the proof of Lemma 7.4) the following a priori estimate for the solution u of the initial boundary value problem

$$\begin{aligned} \langle u_t(t), v\rangle_0 + a(u(t), v) &= 0 \quad \text{for all } v\in V,\ t\in(0,T)\,,\\ u(0) &= u_0\,. \end{aligned}$$

(a) $\alpha t\|u(t)\|_1^2 + 2\displaystyle\int_0^t s\|u_t(s)\|_0^2\,ds \le M\int_0^t \|u(s)\|_1^2\,ds\,.$

(b) $\|u_t(t)\|_0 \le \sqrt{\dfrac{M}{2\alpha}}\dfrac{1}{t}\|u_0\|_0\,.$

Here M and α denote the corresponding constants in the continuity and ellipticity conditions, respectively.

7.2 Semidiscretization by the Vertical Method of Lines

For solving parabolic equations numerically, a wide variety of methods exists. The most important classes of these methods are the following:

- *Full discretizations:*
  - Application of finite difference methods to the classical initial boundary value problem (as of the form (7.1)).
  - Application of so-called space-time finite element methods to a variational formulation that includes the time variable, too.
- *Semidiscretizations:*
  - *The vertical method of lines:* Here the discretization starts with respect to the spatial variable(s) (e.g., by means of the finite dif-

ference method, the finite element method, or the finite volume method).

– *The horizontal method of lines* (Rothe's method): Here the discretization starts with respect to the time variable.

As the name indicates, a semidiscretization has to be followed by a further discretization step to obtain a full discretization, which may be one of the above-mentioned or not. The idea behind semidiscretization methods is to have intermediate problems that are of a well-known structure. In the case of the vertical method of lines, a system of ordinary differential equations arises for the solution of which appropriate solvers are often available. Rothe's method generates a sequence of elliptic boundary value problems for which efficient solution methods are known, too.

The attributes "vertical" and "horizontal" of the semidiscretizations are motivated by the graphical representation of the domain of definition of the unknown function $u = u(x,t)$ in one space dimension (i.e., $d = 1$), namely, assigning the abscissa (horizontal axis) of the coordinate system to the variable x and the ordinate (vertical axis) to the variable t, so that the spatial discretization yields problems that are setted along vertical lines.

In what follows, the vertical method of lines will be considered in more detail.

In the following, and similarly in the following sections, we will develop the analogous (semi)discretization approaches for the finite difference method, the finite element method, and the finite volume method. This will allow us to analyze these methods in a uniform way, as far as only the emerging (matrix) structure of the discrete problems will play a role. On the other hand, different techniques of analysis as in Chapters 1, 3 and 6 will further elucidate advantages and disadvantages of the methods. Readers who are interested only in a specific approach may skip some of the following subsections.

The Vertical Method of Lines for the Finite Difference Method
As a first example we start with the heat equation (7.8) with Dirichlet boundary conditions on a rectangle $\Omega = (0,a) \times (0,b)$. As in Section 1.2 we apply the five-point stencil discretizations at the grid points $x \in \Omega_h$ (according to (1.5)) for every fixed $t \in [0,T]$. This leads to the approximation

$$\begin{aligned} &\partial_t u_{ij}(t) + \frac{1}{h^2}\Big(- u_{i,j-1}(t) - u_{i-1,j}(t) + 4u_{ij}(t) - u_{i+1,j}(t) - u_{i,j+1}(t)\Big) \\ &= f_{ij}(t)\,, \qquad i = 1,\dots,l-1,\ j = 1,\dots,m-1,\ t \in (0,T)\,, \end{aligned} \tag{7.37}$$

$$\begin{aligned} u_{ij}(t) = g_{ij}(t), \qquad & i \in \{0,l\},\ j = 0,\dots,m\,, \\ & j \in \{0,m\},\ i = 0,\dots,l\,. \end{aligned} \tag{7.38}$$

Here we use

$$\begin{aligned} f_{ij}(t) &:= f(ih, jh, t)\,, \\ g_{ij}(t) &:= g(ih, jh, t)\,, \end{aligned} \tag{7.39}$$

and the index 3 in the boundary condition is omitted. Additionally, the initial condition (at the grid points) will be prescribed, that is,

$$u_{ij}(0) = u_0(ih, jh),\ (ih, jh) \in \overline{\Omega}_h\,. \tag{7.40}$$

The system (7.37), (7.38), (7.40) is a system of (linear) ordinary differential equations (in the "index" (i, j)). If, as in Section 1.2, we fix an ordering of the grid points, the system takes the form

$$\begin{aligned} \frac{d}{dt}\boldsymbol{u}_h(t) + A_h\boldsymbol{u}_h(t) &= \boldsymbol{q}_h(t)\,, \quad t \in (0, T)\,, \\ \boldsymbol{u}_h(0) &= \boldsymbol{u}_0\,, \end{aligned} \tag{7.41}$$

with $A_h, \boldsymbol{q}_h$ as in (1.10), (1.11) (but now $\boldsymbol{q}_h = \boldsymbol{q}_h(t)$ because of the t-dependence of f and g).

The unknown is the function

$$\boldsymbol{u}_h : [0, T] \to \mathbb{R}^{M_1}, \tag{7.42}$$

which means that the Dirichlet boundary conditions are eliminated as in Section 1.2.

For a simplification of the notation we use in the following M instead of M_1, which also includes the eliminated degrees of freedom. Only in Sections 7.5 and 7.6 will we return to the original notation.

More generally, if we consider a finite difference approximation, which applied to the stationary problem (7.6) will lead to the system of equations

$$A_h\boldsymbol{u}_h = \boldsymbol{q}_h,$$

with $\boldsymbol{u}_h \in \mathbb{R}^M$, then the same method applied to (7.1) for every fixed $t \in (0, T)$ leads to (7.41). In particular, the system (7.41) has a unique solution due to the theorem of Picard–Lindelöf (cf. [26]).

The Vertical Method of Lines for the Finite Element Method
We proceed as for the finite difference method by now applying the finite element method to (7.1) in its weak formulation (7.32) for every fixed $t \in (0, T)$, using the abbrevation

$$b(t, v) := \langle f(t), v\rangle_0 + \int_{\Gamma_1} g_1(\cdot, t)v\, d\sigma + \int_{\Gamma_2} g_2(\cdot, t)v\, d\sigma\,. \tag{7.43}$$

So let $V_h \subset V$ denote a finite-dimensional subspace with $\dim V_h = M = M(h)$ and let $u_{0h} \in V_h$ be some approximation to u_0. Then the *semidiscrete problem* reads as follows:

Find $u_h \in L^2((0,T),V_h)$ with $u_h' \in L^2((0,T),H)$, $u_h(0) = u_{0h}$ and

$$\left\langle \frac{d}{dt}u_h(t), v_h \right\rangle_0 + a(u_h(t),v_h) = b(t,v_h) \text{ for all } v_h \in V_h\,,\ t \in (0,T)\,. \quad (7.44)$$

To gain a more specific form of (7.44), again we represent the unknown $u_h(t)$ by its degrees of freedom:

Let $\{\varphi_i\}_{i=1}^M$ be a basis of V_h, $u_h(t) = \sum_{i=1}^M \xi_i(t)\,\varphi_i$ and $u_{0h} = \sum_{i=1}^M \xi_{0i}\,\varphi_i$. Then for any $t \in (0,T)$, the discrete variational equality (7.44) is equivalent to

$$\sum_{j=1}^M \langle \varphi_j, \varphi_i\rangle_0 \frac{d\xi_j(t)}{dt} + \sum_{j=1}^M a(\varphi_j,\varphi_i)\,\xi_j(t) = b(t,\varphi_i) \quad \text{for all } i \in \{1,\dots,M\}\,.$$

Denoting by $\hat{A}_h := (a(\varphi_j,\varphi_i))_{ij}$ the *stiffness matrix*, by $B_h := \left(\langle \varphi_j,\varphi_i\rangle_0\right)_{ij}$ the *mass matrix*, and by

$$\boldsymbol{\beta}_h(t) := (b(t,\varphi_i))_i\,,$$

respectively $\boldsymbol{\xi}_{0h} := (\xi_{0i})_i$, the vectors of the right-hand side and of the initial value, we obtain for $\boldsymbol{\xi}_h(t) := (\xi_i(t))_i$ the following system of linear ordinary differential equations with constant coefficients:

$$\begin{aligned} B_h \frac{d}{dt}\boldsymbol{\xi}_h(t) + \hat{A}_h\,\boldsymbol{\xi}_h(t) &= \boldsymbol{\beta}_h(t)\,, \quad t \in (0,T)\,, \\ \boldsymbol{\xi}_h(0) &= \boldsymbol{\xi}_{0h}\,. \end{aligned} \quad (7.45)$$

Since the matrix B_h is symmetric and positive definite, it can be factored (e.g., by means of Cholesky's decomposition) as $B_h = E_h^T E_h$. Introducing the new variable $\boldsymbol{u}_h := E_h\boldsymbol{\xi}_h$ (to maintain the possible definiteness of A_h), the above system (7.45) can be written as follows:

$$\begin{aligned} \frac{d}{dt}\boldsymbol{u}_h(t) + A_h\,\boldsymbol{u}_h(t) &= \boldsymbol{q}_h(t)\,, \quad t \in (0,T)\,, \\ \boldsymbol{u}_h(0) &= \boldsymbol{u}_{h0}\,, \end{aligned} \quad (7.46)$$

where $A_h := E_h^{-T}\hat{A}_h E_h^{-1}$ is an $\mathbb{R}^M$-elliptic matrix and $\boldsymbol{q}_h := E_h^{-T}\boldsymbol{\beta}_h$, $\boldsymbol{u}_{h0} := E_h\boldsymbol{\xi}_{0h}$.

Thus again the discretization leads us to a system (7.41).

Remark 7.7 By means of the same arguments as in the proof of Lemma 7.4, an estimate of $\|u_h(t)\|_0$ can be derived.

The Vertical Method of Lines for the Finite Volume Method

Based on the finite volume methods introduced in Chapter 6, in this subsection a finite volume semidiscretization is given for the problem (7.1) in its weak formulation (7.32) for every fixed $t \in (0,T)$ in the special case $\Gamma_3 = \partial\Omega$ and of homogeneous Dirichlet boundary conditions. As in Chapter 6, the only essential difference to problem (7.1) is that here the

differential expression L is in divergence form, i.e.,

$$Lu := -\nabla \cdot (K \nabla u - c\, u) + r\, u = f\,,$$

where the data K, c, r, and f are as in (7.2).

Correspondingly, the bilinear form a in the weak formulation (7.32) is to be replaced by

$$a(u,v) = \int_\Omega [(K \nabla u - c\, u) \cdot \nabla v + ruv]\, dx\,. \tag{7.47}$$

In order to obtain a finite volume semidiscretization of the problem (7.1) in divergence form, and of (7.32) with the modification (7.47), we recall the way that it was done in the elliptic situation. Namely, comparing the weak formulation of the elliptic problem (see Definition 2.2) with the finite volume method in the discrete variational formulation (6.11), we see that the bilinear form a and the linear form $b(\cdot) := \langle f, \cdot \rangle_0$ have been replaced by certain discrete forms a_h and $\langle f, \cdot \rangle_{0,h}$, respectively. This formal procedure can be applied to the weak formulation (7.32) of the parabolic problem, too.

So let $V_h \subset V$ denote a finite-dimensional subspace as introduced in Section 6.2 with $\dim V_h = M = M(h)$ and let $u_{0h} \in V_h$ be some approximation to u_0. Then, the *semidiscrete finite volume method* reads as follows:

Find $u_h \in L^2((0,T), V_h)$ with $u_h' \in L^2((0,T), H)$, $u_h(0) = u_{0h}$ and

$$\left\langle \frac{d}{dt} u_h(t), v_h \right\rangle_{0,h} + a_h(u_h(t), v_h) = \langle f(t), v_h \rangle_{0,h} \quad \text{for all } v_h \in V_h\,,\ t \in (0,T)\,, \tag{7.48}$$

where both the bilinear form a_h and the form $\langle \cdot, \cdot \rangle_{0,h}$ have been formally defined in Section 6.2. However, to facilitate the comparison of the finite volume discretization with the previously described methods, here we set $\Lambda := \{1, \ldots, M\}$.

As in Section 6.2 we consider the following discrete L^2-scalar product $\langle \cdot, \cdot \rangle_{0,h}$:

$$\left\langle \frac{d}{dt} u_h(t), v_h \right\rangle_{0,h} = \sum_{j=1}^{M} \frac{d}{dt} u_h(a_j, t) v_h(a_j) m_j\,. \tag{7.49}$$

In analogy to the case of the finite element method (cf. Remark 7.7), a stability estimate for the finite volume method can be obtained. Namely, under the assumptions of Theorem 6.15, we have that

$$a_h(v_h, v_h) \geq \alpha \|v_h\|_{0,h}^2 \quad \text{for all } v_h \in V_h$$

with some constant $\alpha > 0$ independent of h. Then, taking $v_h = u_h(t)$ in (7.48), we get

$$\left\langle \frac{d}{dt} u_h(t), u_h(t) \right\rangle_{0,h} + a_h(u_h(t), u_h(t)) = \langle f(t), u_h(t) \rangle_{0,h}\,,$$

and, after some calculations,

$$\frac{d}{dt}\|u_h(t)\|_{0,h} + \alpha\|u_h(t)\|_{0,h} \le \|f(t)\|_{0,h}\,.$$

The subsequent arguments are as in the proof of Lemma 7.4; i.e., we obtain

$$\|u_h(t)\|_{0,h} \le \|u_0\|_{0,h}\, e^{-\alpha t} + \int_0^t \|f(s)\|_{0,h}\, e^{-\alpha(t-s)}\, ds\,.$$

If the right-hand side of (7.48) is a general bounded linear form, i.e., instead of $\langle f(t), v_h\rangle_{0,h}$ we have the term $b(t, v_h)$, where $b:\ (0,T)\times V_h \to \mathbb{R}$ is such that

$$|b(t,v)| \le \|b(t)\|_*\|v\|_{0,h} \quad \text{for all } v \in V_h\,,\ t \in (0,T),$$

with $\|b(t)\|_* < \infty$ for all $t \in (0,T)$, then an analogous estimate holds:

$$\|u_h(t)\|_{0,h} \le \|u_0\|_{0,h}\, e^{-\alpha t} + \int_0^t \|b(s)\|_*\, e^{-\alpha(t-s)}\, ds\,. \tag{7.50}$$

As in the previous subsection, we now want to give a more specific form of (7.48).

Given a basis $\{\varphi_i\}_{i=1}^M$ of the space V_h, such that $\varphi_i(a_j) = \delta_{ij}$ for the underlying nodes, we have the unique expansions

$$u_h(t) = \sum_{i=1}^M \xi_i(t)\,\varphi_i \quad \text{and} \quad u_{0h} = \sum_{i=1}^M \xi_{0i}\,\varphi_i\,.$$

Then for any $t \in (0,T)$, the discrete variational equality (7.48) is equivalent to

$$m_i \frac{d\xi_i(t)}{dt} + \sum_{j=1}^M a_h(\varphi_j,\varphi_i)\,\xi_j(t) = \langle f(t),\varphi_i\rangle_{0,h} \quad \text{for all } i \in \{1,\dots,M\}\,,$$

where $m_i = |\Omega_i|$. Using the notation $\hat{A}_h := (a_h(\varphi_j,\varphi_i))_{ij}$ for the finite volume *stiffness matrix*, $B_h := \operatorname{diag}(m_i)$ for the finite volume *mass matrix*, $\boldsymbol{\beta}_h(t) := \Big(\langle f(t),\varphi_i\rangle_{0,h}\Big)_i$ for the vector of the right-hand side, and $\boldsymbol{\xi}_{0h} := (\xi_{0i})_i$ for the vector of the initial value, we obtain for the unknown vector function $\boldsymbol{\xi}_h(t) := (\xi_i(t))_i$ the following system of linear ordinary differential equations with constant coefficients:

$$\begin{aligned} B_h\,\frac{d}{dt}\boldsymbol{\xi}_h(t) + \hat{A}_h\,\boldsymbol{\xi}_h(t) &= \boldsymbol{\beta}_h(t)\,, \quad t \in (0,T)\,, \\ \boldsymbol{\xi}_h(0) &= \boldsymbol{\xi}_{0h}\,. \end{aligned} \tag{7.51}$$

In contrast to the system (7.45) arising in the finite element semidiscretization, here the matrix B_h is diagonal. Therefore, it is very easy to introduce the new variable $\boldsymbol{u}_h := E_h\boldsymbol{\xi}$ with $E_h := \operatorname{diag}(\sqrt{m_i})$, and the above system

(7.51) can be written as follows:

$$\begin{aligned} \frac{d}{dt}\boldsymbol{u}_h(t) + A_h\,\boldsymbol{u}_h(t) &= \boldsymbol{q}_h(t)\,, \quad t \in (0,T)\,, \\ \boldsymbol{u}_h(0) &= \boldsymbol{u}_{h0}\,, \end{aligned} \tag{7.52}$$

where $A_h := E_h^{-1}\hat{A}_h E_h^{-1}$ is an $\mathbb{R}^M$-elliptic matrix and $\boldsymbol{q}_h := E_h^{-1}\boldsymbol{\beta}_h$, $\boldsymbol{u}_{h0} := E_h\boldsymbol{\xi}_{0h}$.

Thus again we have arrived at a system of the form (7.41).

Representation of Solutions in a Special Case

The solution of system (7.41) can be represented explicitly if there is a basis of $\mathbb{R}^M$ composed of eigenvectors of A_h. This will be developed in the following, but is not meant for numerical use, since only in special cases can eigenvectors and values be given explicitly. Rather, it will serve as a tool for comparison with the continuous and the fully discrete cases.

Let $(\boldsymbol{w}_i, \lambda_i)$, $i = 1, \ldots, M$, be the eigenvectors and real eigenvalues of A_h. Then the following representation exists uniquely:

$$\boldsymbol{u}_0 = \sum_{i=1}^{M} c_i \boldsymbol{w}_i \qquad \text{and} \qquad \boldsymbol{q}_h(t) = \sum_{i=1}^{M} q_h^i(t)\boldsymbol{w}_i\,. \tag{7.53}$$

Again by a separation of variables approach (cf. (7.18)) we see that

$$\boldsymbol{u}_h(t) = \sum_{i=1}^{M}(c_i e^{-\lambda_i t} + \int_0^t q_h^i(t) e^{-\lambda_i(t-s)}ds)\boldsymbol{w}_i\,. \tag{7.54}$$

A more compact notation is given by

$$\boldsymbol{u}_h(t) = e^{-A_h t}\boldsymbol{u}_0 + \int_0^t e^{-A_h(t-s)}\boldsymbol{q}_h(s)ds \tag{7.55}$$

if we define for a matrix $B \in \mathbb{R}^{M,M}$,

$$e^B := \sum_{\nu=0}^{\infty} \frac{B^\nu}{\nu!}\,.$$

This can be seen as follows:
Let

$$T := (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_M) \in \mathbb{R}^{M,M}\,,$$

$$\Lambda := \operatorname{diag}(\lambda_i) \in \mathbb{R}^{M,M}\,.$$

Then

$$A_h T = T\Lambda\,, \quad \text{i.e.,} \quad T^{-1}A_h T = \Lambda\,,$$

and therefore

$$T^{-1}e^{-A_h t}T = \sum_{\nu=0}^{\infty} \frac{t^\nu}{\nu!} T^{-1}(-A_h)^\nu T = \sum_{\nu=0}^{\infty} \frac{t^\nu}{\nu!} (-\Lambda)^\nu,$$

since $T^{-1}(-A_h)^\nu T = T^{-1}(-A_h)TT^{-1}(-A_h)TT^{-1}\dots T$ and thus

$$T^{-1}e^{-A_h t}T = \operatorname{diag}\left(\sum_{\nu=0}^{\infty} \frac{(-\lambda_i t)^\nu}{\nu!}\right) = \operatorname{diag}\left(e^{-\lambda_i t}\right) .$$

Then for $\boldsymbol{c} = (c_1, \dots, c_M)^T \in \mathbb{R}^M$, because of $T\boldsymbol{c} = \boldsymbol{u}_0$ we conclude for the case $\boldsymbol{q}_h = 0$ that

$$\boldsymbol{u}_h(t) = T\operatorname{diag}(e^{-\lambda_i t})\boldsymbol{c} = TT^{-1}e^{-A_h t}T\boldsymbol{c} = e^{-A_h t}\boldsymbol{u}_0 ,$$

and similarly in general.

A basis of eigenvalues exists if A_h is *self-adjoint* with respect to a scalar product $\langle\cdot,\cdot\rangle_h$ in $\mathbb{R}^M$, meaning that

$$\langle \boldsymbol{v}, A_h\boldsymbol{u}\rangle_h = \langle A_h\boldsymbol{v}, \boldsymbol{u}\rangle_h \quad \text{for all} \quad \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^M .$$

Then the eigenvectors even are *orthogonal*; that is,

$$\langle \boldsymbol{w}_i, \boldsymbol{w}_j\rangle_h = 0 \quad \text{for} \quad i \neq j \tag{7.56}$$

because of

$$\lambda_i\langle \boldsymbol{w}_i, \boldsymbol{w}_j\rangle_h = \langle A_h\boldsymbol{w}_i, \boldsymbol{w}_j\rangle = \langle \boldsymbol{w}_i, A_h\boldsymbol{w}_j\rangle = \lambda_j\langle \boldsymbol{w}_i, \boldsymbol{w}_j\rangle_h ,$$

and thus (7.56) if $\lambda_i \neq \lambda_j$. But eigenvectors belonging to one eigenvalue can always be orthonormalized. For orthogonal $\boldsymbol{w}_i$ the coefficient c_i from (7.53) has the form

$$c_i = \frac{\langle \boldsymbol{u}_0, \boldsymbol{w}_i\rangle_h}{\langle \boldsymbol{w}_i, \boldsymbol{w}_i\rangle_h} , \tag{7.57}$$

and analogously for q_h^i.

Order of Convergence Estimates for the Finite Difference Method in a Special Case

As an illustrative example, we consider a case where the eigenvectors and eigenvalues of A_h are known explicitly: the five-point stencil discretization of the Poisson equation with Dirichlet conditions in $\Omega = (0, a) \times (0, b)$. Instead of considering a fixed ordering of the grid points, we prefer to use the "natural" two-dimensional indexing; i.e., we regard the eigenvectors as grid functions. As seen in Section 1.2, A_h is symmetric and thus self-adjoint with respect to the Euclidean scalar product scaled with h^d if $\Omega \subset R^d$, i.e., $d = 2$ here:

$$\langle \boldsymbol{u}, \boldsymbol{v}\rangle_h = h^d \sum_{i=1}^{M} u_i v_i . \tag{7.58}$$

The norm induced by this scalar product is exactly the discrete L^2-norm defined in (1.18) (for $d = 2$ and for the vectors representing the grid functions):

$$|\boldsymbol{u}|_{0,h} = \langle \boldsymbol{u}, \boldsymbol{u} \rangle_h^{1/2} = h^{d/2} \left(\sum_{i=1}^{M} |u_i|^2 \right)^{1/2} . \tag{7.59}$$

If we mean the grid function U, we denote the norm by $\|U\|_{0,h}$.

The eigenvectors, which have already been noted for a special case after Theorem 5.4, are written as grid functions

$$\begin{aligned} &u^{\nu\mu}(x,y) = \sin\left(\nu\frac{\pi}{a}x\right)\sin\left(\mu\frac{\pi}{b}y\right) \quad \text{for} \quad (x,y) \in \Omega_h \,, \\ &\text{and} \;\; \nu = 1,\ldots,l-1, \;\; \mu = 1,\ldots,m-1 \end{aligned} \tag{7.60}$$

for the eigenvalues

$$\lambda_h^{\nu\mu} = \frac{2}{h^2}\left(2 - \cos\left(\nu\frac{\pi}{a}h\right) - \cos\left(\mu\frac{\pi}{b}h\right)\right).$$

Note that the eigenvectors are the eigenfunctions of the continuous problem evaluated at the grid points, but the grid points can distinguish only the maximal frequencies $\frac{l-1}{2}$ and $\frac{m-1}{2}$, so that for other indices ν, μ the given grid functions would be repeated.

Due to $2\sin^2\left(\frac{\xi}{2}\right) = 1 - \cos(\xi)$, an alternative representation is

$$\lambda_h^{\nu\mu} = \frac{4}{h^2}\left(\sin^2\left(\nu\frac{\pi}{a}\,\frac{h}{2}\right) + \sin^2\left(\mu\frac{\pi}{b}\,\frac{h}{2}\right)\right),$$

so that for $h \to 0$,

$$\begin{aligned} \lambda_h^{\nu\mu} \;&=\; \left(\frac{\nu\pi}{a}\right)^2 \left(\sin\left(\nu\frac{\pi}{a}\,\frac{h}{2}\right) \Big/ \left(\nu\frac{\pi}{a}\,\frac{h}{2}\right)\right)^2 \\ &\quad + \left(\frac{\mu\pi}{b}\right)^2 \left(\sin\left(\mu\frac{\pi}{b}\,\frac{h}{2}\right) \Big/ \left(\mu\frac{\pi}{b}\,\frac{h}{2}\right)\right)^2 \\ &\to \;\; \left(\frac{\nu\pi}{a}\right)^2 \cos^2(0) + \left(\frac{\mu\pi}{b}\right)^2 \cos^2(0) \end{aligned} \tag{7.61}$$

holds; i.e., the eigenvalues converge to the eigenvalues (7.27) of the boundary value problem, with an order of convergence estimate of $O(h^2)$.

The eigenvectors are orthogonal with respect to $\langle \cdot, \cdot \rangle_h$, since they belong to different eigenvalues (see (7.56)). To specify the Fourier coefficients according to (7.57), we need

$$\langle \boldsymbol{u}^{\nu\mu}, \boldsymbol{u}^{\nu\mu} \rangle_h = \frac{ab}{4} \tag{7.62}$$

(see Exercise 7.5).

To investigate the accuracy of the semidiscrete approximation, the solution representations can be compared. To simplify the exposition, we

consider only $f = 0$, so that because of (7.18), (7.27) we have

$$u(x,y,t) = \sum_{\substack{\nu=1\\ \mu=1}}^{\infty} c_{\nu\mu} e^{-\lambda^{\nu\mu} t} \sin\left(\nu \frac{\pi}{a} x\right) \sin\left(\mu \frac{\pi}{b} y\right) ,$$

and

$$c_{\nu\mu} = \frac{4}{ab} \int_0^b \int_0^a u_0(x,y) \sin\left(\nu \frac{\pi}{a} x\right) \sin\left(\mu \frac{\pi}{b} y\right) dx\, dy$$

because of (7.25) and (7.24) (applied in every space direction), and finally,

$$\lambda^{\nu\mu} = \left(\frac{\nu\pi}{a}\right)^2 + \left(\frac{\mu\pi}{b}\right)^2$$

for the continuous solution. For the semidiscrete approximation at a grid point $(x,y) \in \Omega_h$ we have, due to (7.54),

$$u_h(x,y,t) = \sum_{\nu=1}^{l-1} \sum_{\mu=1}^{m-1} c_{\nu\mu}^h e^{-\lambda_h^{\nu\mu} t} \sin\left(\nu \frac{\pi}{a} x\right) \sin\left(\mu \frac{\pi}{b} y\right)$$

and

$$\begin{aligned} c_{\nu\mu}^h &= \frac{4}{ab} h^2 \sum_{i=1}^{l-1} \sum_{j=1}^{m-1} u_0(ih, jh) \sin\left(\nu \frac{\pi}{a} ih\right) \sin\left(\mu \frac{\pi}{b} jh\right) , \\ \lambda_h^{\nu\mu} &= \frac{4}{h^2} \left(\sin^2\left(\nu \frac{\pi}{a} \frac{h}{2}\right) + \sin^2\left(\mu \frac{\pi}{b} \frac{h}{2}\right)\right) . \end{aligned}$$

Compared at the grid points u has additionally the terms in the infinite series for $\nu = l, \ldots,$ or $\mu = m, \ldots$.

They can be estimated by

$$\begin{aligned} &\left| \left(\sum_{\nu=l}^{\infty} \sum_{\mu=1}^{\infty} + \sum_{\nu=1}^{\infty} \sum_{\mu=m}^{\infty} \right) c_{\nu\mu} e^{-\lambda^{\nu\mu} t} \sin\left(\nu \frac{\pi}{a} x\right) \sin\left(\mu \frac{\pi}{b} y\right) \right| \\ &\quad \leq \quad C_1 \left(\sum_{\nu=l}^{\infty} \sum_{\mu=1}^{\infty} + \sum_{\nu=1}^{\infty} \sum_{\mu=m}^{\infty} \right) e^{-\lambda^{\nu\mu} t} \end{aligned}$$

with $C_1 := \max\{|c_{\nu\mu}|\, , \nu, \mu \in \mathbb{N}, \nu \notin \{1, \ldots, l-1\}$ or $\mu \notin \{1, \ldots, m-1\}\}$

$$\leq C_1 \left(C_2 \sum_{\nu=l}^{\infty} e^{-\left(\frac{\nu\pi}{a}\right)^2 t} + C_3 \sum_{\mu=m}^{\infty} e^{-\left(\frac{\mu\pi}{b}\right)^2 t} \right)$$

with $C_2 := \sum_{\mu=1}^{\infty} e^{-\left(\frac{\mu\pi}{b}\right)^2 t} \leq \frac{q_2}{1-q_2}$, where $q_2 := e^{-\left(\frac{\pi}{b}\right)^2 \overline{t}}$ because of $\sum_{\mu=1}^{\infty} q^{\mu} = \frac{q}{1-q}$ for $|q| < 1$, and C_3 is defined analogously ($\mu \longleftrightarrow \nu$, $a \longleftrightarrow b$) with an estimate by $\frac{q_1}{1-q_1}$, $q_1 := e^{-\left(\frac{\pi}{a}\right)^2 \overline{t}}$ for $t \geq \overline{t} > 0$.

Finally, we conclude the estimate because of $\sum_{\mu=l}^{\infty} q^{\mu} = \frac{q^l}{1-q}$ by

$$\leq C_1 \left(C_2 \frac{q_1^l}{1-q_1} + C_3 \frac{q_2^m}{1-q_2} \right) .$$

Therefore, this error contribution for $t \geq \overline{t}$ (for a fixed $\overline{t} > 0$) approaches 0 for $l \to \infty$ and $m \to \infty$. The larger $\overline{t}$ is, the more this error term will decrease. Because of, for example, $l = a/h$ and thus $q_1^l = \exp\left(-\frac{\pi^2}{a}\overline{t}\frac{1}{h}\right)$, the decay in h is exponential and thus much stronger than a term like $O(h^2)$. Therefore, we have to compare the terms in the sum only for $\nu = 1, \ldots, l-1, \;\; \mu = 1, \ldots, m-1$, i.e., the error in the Fourier coefficient and in the eigenvalue:

$$c_{\nu\mu} e^{-\lambda^{\nu\mu} t} - c_{\nu\mu}^h e^{-\lambda_h^{\nu\mu} t} = \left(c_{\nu\mu} - c_{\nu\mu}^h\right) e^{-\lambda^{\nu\mu} t} + c_{\nu\mu}^h \left(e^{-\lambda^{\nu\mu} t} - e^{-\lambda_h^{\nu\mu} t} \right) .$$

Note that $c_{\nu\mu}^h$ can be perceived as an approximation of $c_{\nu\mu}$ by the trapezoidal sum with step size h in each spatial direction (see, e.g., [30], p. 129), since the integrand in the definition of $c_{\nu\mu}$ vanishes for $x = 0$ or $x = a$ and $y \in [0, b]$, and $y = 0$ or $y = b$ and $x \in [0, a]$. Thus we have for $u_0 \in C^2(\bar{\Omega})$,

$$|c_{\nu\mu} - c_{\nu\mu}^h| = O(h^2).$$

Because of

$$e^{-\lambda^{\nu\mu} t} - e^{-\lambda_h^{\nu\mu} t} = e^{-\lambda^{\nu\mu} t} \left(1 - e^{-(\lambda_h^{\nu\mu} - \lambda^{\nu\mu}) t} \right),$$

and $|\lambda_h^{\nu\mu} - \lambda^{\nu\mu}| = O(h^2)$ (see (7.61)), also this term is of order $O(h^2)$ and will be damped exponentially (depending on t and the size of the smallest eigenvalue $\lambda^{\nu\mu}$).

Summarizing, we expect

$$O(h^2)$$

to be the dominating error term in the discrete maximum norm $\|\cdot\|_\infty$ at the grid points (cf. definition (1.17)), which will also be damped exponentially for increasing t. Note that we have given only heuristic arguments and that the considerations cannot be transferred to the Neumann case, where the eigenvalue $\lambda = 0$ appears.

We now turn to the finite element method.

Order of Convergence Estimates for the Finite Element Method We will investigate the finite element method on a more abstract level as in the previous subsection, but we will achieve a result (in different norms) of similar character. As worked out at the end of Section 7.1, there is a strong relation between the V-ellipticity of the bilinear form a with the parameter α and a positive lower bound of the eigenvalues. Here we rely on the results already achieved in Section 2.3 and Section 3.4 for the stationary case.

For that, we introduce the so-called *elliptic projection* of the solution $u(t)$ of (7.32) as a very important tool in the proof.

Definition 7.8 For a V-elliptic, continuous bilinear form $a : V \times V \to \mathbb{R}$, the *elliptic*, or *Ritz, projection* $R_h : V \to V_h$ is defined by

$$v \mapsto R_h v \quad \Longleftrightarrow \quad a(R_h v - v, v_h) = 0 \quad \text{for all } v_h \in V_h \,.$$

Theorem 7.9 *Under the assumptions of Definition* 7.8*:*
(i) $R_h : V \to V_h$ *is linear and continuous.*
(ii) R_h *yields* quasi-optimal *approximations; that is,*

$$\|v - R_h v\|_V \le \frac{M}{\alpha} \inf_{v_h \in V_h} \|v - v_h\|_V \,,$$

where M *and* α *are the Lipschitz and ellipticity constants according to* (2.42) *and* (2.43)*.*

Proof: The linearity of R_h is obvious. The remaining statements immediately follow from Lemma 2.16 and Theorem 2.17; see Exercise 7.6. □

Making use of the elliptic projection, we are able to prove the following result.

Theorem 7.10 *Suppose* a *is a* V*-elliptic, continuous bilinear form,* $f \in C([0,T], H)$*,* $u_0 \in V$*, and* $u_{0h} \in V_h$*. Then if* $u(t)$ *is sufficiently smooth,*

$$\begin{aligned} \|u_h(t) - u(t)\|_0 \quad &\le \quad \|u_{0h} - R_h u_0\|_0 \, e^{-\alpha t} + \|(I - R_h)u(t)\|_0 \\ &\quad + \int_0^t \|(I - R_h)u'(s)\|_0 \; e^{-\alpha(t-s)} \, ds \,. \end{aligned}$$

Proof: First, the error is decomposed as follows:

$$u_h(t) - u(t) = u_h(t) - R_h u(t) + R_h u(t) - u(t) =: \theta(t) + \varrho(t) \,.$$

We take $v = v_h \in V_h$ in (7.32) and obtain, by the definition of R_h,

$$\langle u'(t), v_h \rangle_0 + a(u(t), v_h) = \langle u'(t), v_h \rangle_0 + a(R_h u(t), v_h) = b(t, v_h) \,.$$

Here $b(t, \cdot)$ is as defined in (7.43).

Subtracting this equation from (7.44), we get

$$\langle u_h'(t), v_h \rangle_0 - \langle u'(t), v_h \rangle_0 + a(\theta(t), v_h) = 0 \,,$$

and thus

$$\langle \theta'(t), v_h \rangle_0 + a(\theta(t), v_h) = \langle u'(t), v_h \rangle_0 - \left\langle \frac{d}{dt} R_h u(t), v_h \right\rangle_0 = - \langle \varrho'(t), v_h \rangle_0 \,.$$

The application of Lemma 7.4 yields

$$\|\theta(t)\|_0 \le \|\theta(0)\|_0 \, e^{-\alpha t} + \int_0^t \|\varrho'(s)\|_0 \, e^{-\alpha(t-s)} \, ds \,.$$

Since the elliptic projection is continuous (Theorem 7.9, (i)) and $u(t)$ is sufficiently smooth, R_h and the time derivative $\frac{d}{dt}$ commute; that is, $\varrho'(t) =$

$(R_h - I)u'(t)$. It remains to apply the triangle inequality to get the stated result. □

Theorem 7.10 has the following interpretation:
The error norm $\|u_h(t) - u(t)\|_0$ is estimated by

- the initial error (exponentially decaying in t), which occurs only if u_{0h} does not coincide with the elliptic projection of u_0,
- the projection error of the exact solution $u(t)$ measured in the norm of H,
- the projection error of $u'(t)$ measured in the norm of H and integrally weighted by the factor $\mathrm{e}^{-\alpha(t-s)}$ on $(0,t)$.

Remark 7.11 If the bilinear form a defines an elliptic problem such that for the elliptic projection an error estimate of the type

$$\|(I - R_h)w\|_0 \le Ch^2\|w\|_2 \quad \text{for all } w \in V \cap H^2(\Omega)$$

is valid, if u_{0h} approximates the elliptic projection $R_h u_0$ of the initial value u_0 at least with the same asymptotic quality, and if the solution u of (7.44) is sufficiently smooth, then an optimal L^2-error estimate results:

$$\|u_h(t) - u(t)\|_0 \le C(u(t))h^2\,.$$

We see that in order to obtain semidiscrete error estimates, we need estimates of the projection error measured in the norm of $H = L^2(\Omega)$. Due to $\|\cdot\|_0 \le \|\cdot\|_V$, the quasi-optimality of R_h (Theorem 7.9, (ii)) in conjunction with the corresponding approximation error estimates (Theorem 3.29) already yield some error estimate. Unfortunately, this result is not optimal. However, if the adjoint boundary value problem is regular in the sense of Definition 3.36, the duality argument (Theorem 3.37) can be successfully used to derive an optimal result.

Theorem 7.12 *Suppose the bilinear form a is V-elliptic and continuous, and the solution of the adjoint boundary value problem is regular.*

Furthermore, let the space $V_h \subset V$ be such that for any function $w \in V \cap H^2(\Omega)$,

$$\inf_{v_h \in V_h} \|w - v_h\|_V \le C\,h\,|w|_2\,,$$

where the constant $C > 0$ does not depend on h and w. If $u_0 \in V \cap H^2(\Omega)$, then for a sufficiently smooth solution u of (7.44) we have

$$\begin{aligned}\|u_h(t) - u(t)\|_0 \;\le\; & \|u_{0h} - u_0\|_0\,\mathrm{e}^{-\alpha t} \\ & + C\,h^2\Big(\|u_0\|_2\,\mathrm{e}^{-\alpha t} + \|u(t)\|_2 + \int_0^t \|u'(s)\|_2\,\mathrm{e}^{-\alpha(t-s)}\,ds\Big)\,.\end{aligned}$$

Proof: The first term in the error bound from Theorem 7.10 is estimated by means of the triangle inequality:

$$\|u_{0h} - R_h u_0\|_0 \leq \|u_{0h} - u_0\|_0 + \|(I - R_h)u_0\|_0 \,.$$

Then the projection error estimate (Theorem 3.37, (1)) yields the given bounds of the resulting second term as well as of the remaining two terms in the error bound from Theorem 7.10. □

Order of Convergence Estimates for the Finite Volume Method
For simplicity we restrict attention to pure homogeneous Dirichlet conditions ($\Gamma_3 = \partial\Omega$). The idea is similar to the proof given in the finite element case. However, here we will meet some additional difficulties, which are caused by the use of perturbed bilinear and linear forms.

We take $v = v_h \in V_h$ in (7.32) and subtract the result from (7.48):

$$\begin{aligned}\langle u_h'(t), v_h\rangle_{0,h} - \langle u'(t), v_h\rangle_0 + a_h(u_h(t), v_h) - a(u(t), v_h)\\ = \langle f(t), v_h\rangle_{0,h} - \langle f(t), v_h\rangle_0 \,.\end{aligned}$$

In analogy to the finite element method, we introduce the following auxiliary problem: Given some $v \in V$, find an element $R_h v \in V_h$ such that

$$a_h(R_h v, v_h) = a(v, v_h) \quad \text{for all } v_h \in V_h \,. \tag{7.63}$$

With this, the above identity can be rewritten as follows:

$$\begin{aligned}\langle u_h'(t), v_h\rangle_{0,h} - \langle u'(t), v_h\rangle_0 + a_h(u_h(t) - R_h u(t), v_h)\\ = \langle f(t), v_h\rangle_{0,h} - \langle f(t), v_h\rangle_0 \,.\end{aligned}$$

Subtracting from both sides of this relation the term $\left\langle \frac{d}{dt} R_h u(t), v_h\right\rangle_{0,h}$ and assuming that $u'(t)$ is a sufficiently smooth function of x, a slight rearrangement yields

$$\begin{aligned}\langle \theta'(t), v_h\rangle_{0,h} + a_h(\theta(t), v_h) \;=\; &- \langle \varrho'(t), v_h\rangle_{0,h} + \langle u'(t), v_h\rangle_0 \\ &- \langle u'(t), v_h\rangle_{0,h} + \langle f(t), v_h\rangle_{0,h} - \langle f(t), v_h\rangle_0 \,,\end{aligned} \tag{7.64}$$

where, as in the finite element case, $\theta(t) = u_h(t) - R_h u(t)$ and $\varrho(t) = R_h u(t) - u(t)$. Furthermore, we define, for $v \in V_h$, $b_1(t, v) := \langle u'(t), v\rangle_0 - \langle u'(t), v\rangle_{0,h}$ and $b_2(t, v) := \langle f(t), v\rangle_{0,h} - \langle f(t), v\rangle_0$.

In order to be able to apply the discrete stability estimate (7.50) to this situation, we need an error estimate for $R_h u'(t)$ as in Remark 7.11 and bounds (consistency error estimates) for $|b_1(t, v)|$, $|b_2(t, v)|$.

So we turn to the first problem. In fact, the estimate is very similar to the error estimate for the finite volume method given in the proof of Theorem 6.18.

For an arbitrary function $v \in V \cap H^2(\Omega)$ and $v_h := R_h v - I_h(v)$, we have by (7.63) that

$$a_h(v_h, v_h) = a_h(R_h v, v_h) - a_h(I_h(v), v_h) = a(v, v_h) - a_h(I_h(v), v_h).$$

By partial integration in the first term of the right-hand side, it follows that

$$a_h(v_h, v_h) = \langle Lv, v_h \rangle_0 - a_h(I_h(v), v_h).$$

From [40] an estimate of the right-hand side is known (cf. also (6.22)); thus

$$a_h(v_h, v_h) \le Ch\|v\|_2 \left\{ |v_h|_1^2 + \|v_h\|_{0,h}^2 \right\}^{1/2}.$$

So Theorem 6.15 yields

$$\left\{ |v_h|_1^2 + \|v_h\|_{0,h}^2 \right\}^{1/2} \le Ch\|v\|_2 .$$

By the triangle inequality,

$$\|(R_h - I)v\|_{0,h} \le \|R_h v - I_h(v)\|_{0,h} + \|I_h(v) - v\|_{0,h}.$$

Since the second term vanishes by the definitions of $\|\cdot\|_{0,h}$ and I_h, we get in particular

$$\|(R_h - I)v\|_{0,h} \le Ch\|v\|_2 . \tag{7.65}$$

Remark 7.13 In contrast to the finite element case (Remark 7.11), this estimate is not optimal.

To estimate $|b_1(t, v)|$ and $|b_2(t, v)|$, we prove the following result.

Lemma 7.14 *Assume $w \in C^1(\overline{\Omega})$ and $v \in V_h$. Then, if the finite volume partition of Ω is a Donald diagram,*

$$|\langle w, v \rangle_{0,h} - \langle w, v \rangle_0| \le Ch|w|_{1,\infty}\|v\|_{0,h}.$$

Proof: We start with a simple rearrangement of the order of summation:

$$\langle w, v \rangle_{0,h} = \sum_{j=1}^{M} w_j v_j m_j = \sum_{K \in \mathcal{T}_h} \sum_{j:\partial K \ni a_j} w_j v_j |\Omega_{j,K}| ,$$

where $\Omega_{j,K} = \Omega_j \cap \operatorname{int} K$. First, we will consider the inner sum. For any triangle $K \in \mathcal{T}_h$ with barycentre $a_{S,K}$, we can write

$$\begin{aligned}
\sum_{j:\partial K \ni a_j} w_j v_j |\Omega_{j,K}| \quad &= \quad \sum_{j:\partial K \ni a_j} [w_j - w(a_{S,K})] v_j |\Omega_{j,K}| \\
&+ \sum_{j:\partial K \ni a_j} w(a_{S,K}) \left[v_j |\Omega_{j,K}| - \int_{\Omega_{j,K}} v \, dx \right] \\
&+ \sum_{j:\partial K \ni a_j} \int_{\Omega_{j,K}} [w(a_{S,K}) - w] v \, dx + \sum_{j:\partial K \ni a_j} \int_{\Omega_{j,K}} wv \, dx \\
&=: I_{1,K} + I_{2,K} + I_{3,K} + \int_K wv \, dx .
\end{aligned}$$

To estimate $I_{1,K}$, we apply the Cauchy–Schwarz inequality and get

$$|I_{1,K}| \le \left\{ \sum_{j:\partial K \ni a_j} |w_j - w(a_{S,K})|^2 |\Omega_{j,K}| \right\}^{1/2} \|v\|_{0,h,K},$$

where

$$\|v\|_{0,h,K} := \left\{ \sum_{j:\partial K \ni a_j} v_j^2 |\Omega_{j,K}| \right\}^{1/2}.$$

Since $|w_j - w(a_{S,K})| \le h_K |w|_{1,\infty}$, it follows that

$$|I_{1,K}| \le h_K |w|_{1,\infty} \sqrt{|K|} \|v\|_{0,h,K}.$$

Similarly, for $I_{3,K}$ we easily get

$$\begin{aligned} |I_{3,K}| &= \left| \int_{\Omega_K} [w(a_{S,K}) - w] v \, dx \right| \\ &\le \|w(a_{S,K}) - w\|_{0,K} \|v\|_{0,K} \le h_K |w|_{1,\infty} \sqrt{|K|} \|v\|_{0,K}. \end{aligned}$$

So it remains to consider $I_{2,K}$. Obviously,

$$I_{2,K} = w(a_{S,K}) \sum_{j:\partial K \ni a_j} \int_{\Omega_{j,K}} [v_j - v] \, dx \, .$$

We will show that if Ω_j belongs to a Donald diagram, then the sum vanishes. To do so, let us suppose that the triangle under consideration has the vertices a_i, a_j, and a_k. The set $\Omega_{j,K}$ can be decomposed into two subtriangles by drawing a straight line between $a_{S,K}$ and a_j . We will denote the interior of these triangles by $\Omega_{j,K,i}$ and $\Omega_{j,K,k}$; i.e.,

$$\Omega_{j,K,i} := \operatorname{int}\big(\operatorname{conv}\{a_j, a_{S,K}, a_{ij}\}\big), \quad \Omega_{j,K,k} := \operatorname{int}\big(\operatorname{conv}\{a_j, a_{S,K}, a_{kj}\}\big) \, .$$

On each subtriangle, the integral of v can be calculated exactly by means of the trapezoidal rule. Since $|\Omega_{j,K,i}| = |\Omega_{j,K,k}| = |K|/6$ in the case of the Donald diagram (cf. also (6.4)), we have

$$\begin{aligned} \int_{\Omega_{j,K,i}} v \, dx &= \frac{|K|}{18} \Big[v_j + \frac{v_j + v_i}{2} + \frac{v_j + v_i + v_k}{3} \Big] \\ &= \frac{|K|}{18} \Big[\frac{11}{6} v_j + \frac{5}{6} v_i + \frac{1}{3} v_k \Big] , \\ \int_{\Omega_{j,K,k}} v \, dx &= \frac{|K|}{18} \Big[\frac{11}{6} v_j + \frac{5}{6} v_k + \frac{1}{3} v_i \Big] . \end{aligned}$$

Consequently,

$$\int_{\Omega_{j,K}} v \, dx = \frac{|K|}{18} \Big[\frac{11}{3} v_j + \frac{7}{6} v_i + \frac{7}{6} v_k \Big],$$

and thus

$$\sum_{j:\partial K\ni a_j}\int_{\Omega_{j,K}} v\,dx = \frac{|K|}{3}\sum_{j:\partial K\ni a_j} v_j\,.$$

On the other hand, since $3|\Omega_{j,K}| = |K|$ (cf. (6.4)), we have

$$\sum_{j:\partial K\ni a_j}\int_{\Omega_{j,K}} v_j\,dx = \frac{|K|}{3}\sum_{j:\partial K\ni a_j} v_j\,,$$

and so $I_{2,K} = 0$. In summary, we have obtained the following estimate:

$$|I_{1,K} + I_{2,K} + I_{3,K}| \le h_K|w|_{1,\infty}\sqrt{|K|}\big[\|v\|_{0,h,K} + \|v\|_{0,K}\big]\,.$$

So it follows that

$$\begin{aligned}|\,\langle w,v\rangle_{0,h} - \langle w,v\rangle_0\,| &\le \sum_{K\in\mathcal{T}_h}|I_{1,K}+I_{2,K}+I_{3,K}|\\ &\le h|w|_{1,\infty}\sum_{K\in\mathcal{T}_h}\sqrt{|K|}\big[\|v\|_{0,h,K}+\|v\|_{0,K}\big]\,.\end{aligned}$$

By the Cauchy–Schwarz inequality,

$$\sum_{K\in\mathcal{T}_h}\sqrt{|K|}\|v\|_{0,h,K} \le \left\{\sum_{K\in\mathcal{T}_h}|K|\right\}^{1/2}\left\{\sum_{K\in\mathcal{T}_h}\|v\|^2_{0,h,K}\right\}^{1/2} = \sqrt{|\Omega|}\|v\|_{0,h}$$

and, similarly,

$$\sum_{K\in\mathcal{T}_h}\sqrt{|K|}\|v\|_{0,K} \le \sqrt{|\Omega|}\|v\|_0\,.$$

So we finally arrive at

$$|\,\langle w,v\rangle_{0,h} - \langle w,v\rangle_0\,| \le Ch|w|_{1,\infty}\big[\|v\|_{0,h} + \|v\|_0\big]\,.$$

Since the norms $\|\cdot\|_{0,h}$ and $\|\cdot\|_0$ are equivalent on V_h (see Remark 6.16), we get

$$|\,\langle w,v\rangle_{0,h} - \langle w,v\rangle_0\,| \le Ch|w|_{1,\infty}\|v\|_{0,h}.$$

□

Now we are prepared to apply the discrete stability estimate (7.50) to equation (7.64):

$$\begin{aligned}\|\theta(t)\|_{0,h} &\le \|\theta(0)\|_{0,h}\,e^{-\alpha t}\\ &\quad + \int_0^t\left[\|\varrho'(s)\|_{0,h} + \|b_1(s)\|_* + \|b_2(s)\|_*\right]e^{-\alpha(t-s)}ds\,,\end{aligned}$$

where $|b_j(t,v)| \le \|b_j(t)\|_*\|v\|_{0,h}$ for all $v\in V_h$, $t\in(0,T)$, and $j=1,2$. The first term in the integral can be estimated by means of (7.65), whereas the

estimates of $\|b_1(s)\|_*$, $\|b_2(s)\|_*$ result from Lemma 7.14:

$$\begin{aligned}\|\theta(t)\|_{0,h} \;\leq\; & \|\theta(0)\|_{0,h}\, e^{-\alpha t} \\ & + Ch \int_0^t \left[\|u'(s)\|_2 + |u'(s)|_{1,\infty} + \|f(s)\|_{1,\infty}\right] e^{-\alpha(t-s)} ds\,.\end{aligned}$$

If $u_0 \in V \cap H^2(\Omega)$, we can write, by (7.65),

$$\|\theta(0)\|_{0,h} \leq \|u_{h0} - u_0\|_{0,h} + \|(I - R_h)u_0\|_{0,h} \leq \|u_{h0} - u_0\|_{0,h} + Ch\|u_0\|_2.$$

So we get

$$\begin{aligned}\|\theta(t)\|_{0,h} \;\leq\; & \|u_{h0} - u_0\|_{0,h}\, e^{-\alpha t} + Ch\Big[\|u_0\|_2 e^{-\alpha t} \\ & + \int_0^t \left[\|u'(s)\|_2 + |u'(s)|_{1,\infty} + \|f(s)\|_{1,\infty}\right] e^{-\alpha(t-s)} ds\Big].\end{aligned}$$

Since

$$\|u_h(t) - u(t)\|_{0,h} \leq \|\theta(t)\|_{0,h} + \|(R_h - I)u(t)\|_{0,h},$$

the obtained estimate and (7.65) yield the following result.

Theorem 7.15 *In addition to the assumptions of Theorem* 6.15*, let* $f \in C([0,T], C^1(\overline{\Omega}))$, $u_0 \in V \cap H^2(\Omega)$, *and* $u_{0h} \in V_h$. *Then if* $u(t)$ *is sufficiently smooth, the solution* $u_h(t)$ *of the semidiscrete finite volume method* (7.48) *on Donald diagrams satisfies the following estimate:*

$$\begin{aligned}\|u_h(t) - u(t)\|_{0,h} \;\leq\; & \|u_{h0} - u_0\|_{0,h}\, e^{-\alpha t} + Ch\Big[\|u_0\|_2 e^{-\alpha t} + \|u(t)\|_2 \\ & + \int_0^t \left[\|u'(s)\|_2 + |u'(s)|_{1,\infty} + \|f(s)\|_{1,\infty}\right] e^{-\alpha(t-s)} ds\Big].\end{aligned}$$

Remark 7.16 In comparison with the finite element method, the result is not optimal in h. The reason is that, in general, the finite volume method does not yield optimal L^2-error estimates even in the elliptic case, but this type of result is necessary to obtain optimal estimates.

Exercises

7.4 Let $A \in \mathbb{R}^{M,M}$ be an $\mathbb{R}^M$-elliptic matrix and let the symmetric positive definite matrix $B \in \mathbb{R}^{M,M}$ have the Cholesky decomposition $B = E^T E$. Show that the matrix $\hat{A} := E^{-T} A E^{-1}$ is $\mathbb{R}^M$-elliptic.

7.5 Prove identity (7.62) by first proving the corresponding identity for one space dimension:

$$h \sum_{i=1}^{l-1} \sin^2\left(\nu \frac{\pi}{a} ih\right) = \frac{a}{2}\,.$$

7.6 Let V be a Banach space and $a : V \times V \to \mathbb{R}$ a V-elliptic, continuous bilinear form. Show that the Ritz projection $R_h : V \to V_h$ in a subspace $V_h \subset V$ (cf. Definition 7.8) has the following properties:

(i) $R_h : V \to V_h$ is continuous because of $\|R_h u\|_V \leq \frac{M}{\alpha}\|u\|_V$,

(ii) R_h yields quasi-optimal approximations; that is,

$$\|u - R_h u\|_V \leq \frac{M}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|_V .$$

Here M and α denote the constants in the continuity and ellipticity conditions, respectively.

7.7 Let $u \in C^1([0,T], V)$. Show that $R_h u \in C^1([0,T], V)$ and $\frac{d}{dt} R_h u(t) = R_h \frac{d}{dt} u(t)$.

7.8 Transfer the derivation of the finite volume method given in Section 6.2.2 for the case of an elliptic boundary value problem to the parabolic initial-boundary value problem (7.1) in divergence form; i.e., convince yourself that the formalism of obtaining (7.48) indeed can be interpreted as a finite volume semidiscretization of (7.1).

7.3 Fully Discrete Schemes

As we have seen, the application of the vertical method of lines results in the following situation:

- There is a linear system of ordinary differential equations of high order (dimension) to be solved.
- There is an error estimate for the solution u of the initial-boundary value problem (7.1) by means of the exact solution $\boldsymbol{u}_h$ of the system (7.41).

A difficulty in the choice and in the analysis of an appropriate discretization method for systems of ordinary differential equations is that many standard estimates involve the Lipschitz constant of the corresponding right-hand side, here $\boldsymbol{q}_h - A_h \boldsymbol{u}_h$ (cf. (7.41)). But this constant is typically large for small spatial parameters h, and so we would obtain nonrealistic error estimates (cf. Theorem 3.45).

There are two alternatives. For comparatively simple time discretizations, certain estimates can be derived in a direct way (i.e., without using standard estimates for systems of ordinary differential equations). The second way is to apply specific time discretizations in conjunction with refined methods of proof.

Here we will explain the first way for the so-called *one-step-theta method.*

One-Step Discretizations in Time, in Particular for the Finite Difference Method

We start from the problem (7.41), which resulted from spatial discretization techniques. Provided that $T < \infty$, the time interval $(0, T)$ is subdivided into $N \in \mathbb{N}$ subintervals of equal length $\tau := T/N$. Furthermore, we set $t_n := n\tau$ for $n \in \{0, \ldots, N\}$ and $\boldsymbol{u}_h^n \in \mathbb{R}^M$ for an approximation of $\boldsymbol{u}_h(t_n)$. If the time interval is unbounded, the *time step* $\tau > 0$ is given, and the number $n \in \mathbb{N}$ is allowed to increase without bounded; that is, we set formally $N = \infty$.

The values $t = t_n$, where an approximation is to be determined, are called *time levels*. The restriction to equidistant time steps is only for the sake of simplicity. We approximate $\frac{d}{dt}\boldsymbol{u}_h$ by the difference quotient

$$\frac{d}{dt}\boldsymbol{u}_h(t) \sim \frac{1}{\tau}(\boldsymbol{u}_h(t_{n+1}) - \boldsymbol{u}_h(t_n)).$$

If we interpret this approximation to be at $t = t_n$, we take the forward difference quotient; at $t = t_{n+1}$ we take the backward difference quotient; at $t = t_n + \frac{1}{2}\tau$ we take the symmetric difference quotient. Again we obtain a generalization and unification by introducing a parameter $\Theta \in [0, 1]$ and interpreting the approximation to be taken at $t = t_n + \Theta\tau$. As for $\Theta \neq 0$ or 1, we are not at a time level, and so we need the further approximation

$$A_h \boldsymbol{u}_h\left((n + \Theta)\,\tau\right) \sim \overline{\Theta} A_h \boldsymbol{u}_h(t_n) + \Theta A_h \boldsymbol{u}_h(t_{n+1}).$$

Here we use the abbreviation $\overline{\Theta} := 1 - \Theta$. The *(one-step-)theta method* defines a sequence of vectors $\boldsymbol{u}_h^0, \ldots, \boldsymbol{u}_h^N$ by, for $n = 0, 1, \ldots, N-1$,

$$\begin{aligned} \frac{1}{\tau}\left(\boldsymbol{u}_h^{n+1} - \boldsymbol{u}_h^n\right) + \overline{\Theta} A_h \boldsymbol{u}_h^n + \Theta A_h \boldsymbol{u}_h^{n+1} &= \boldsymbol{q}_h((n+\Theta)\tau)\,, \quad (7.66) \\ \boldsymbol{u}_h^0 &= \boldsymbol{u}_0\,. \end{aligned}$$

If we apply this discretization to the more general form (7.45), we get correspondingly

$$\frac{1}{\tau}\left(B_h \boldsymbol{u}_h^{n+1} - B_h \boldsymbol{u}_h^n\right) + \overline{\Theta} \hat{A}_h \boldsymbol{u}_h^n + \Theta \hat{A}_h \boldsymbol{u}_h^{n+1} = \boldsymbol{q}_h((n+\Theta)\tau)\,. \quad (7.67)$$

Analogously to (7.45), the more general form can be transformed to (7.66), assuming that B_h is regular: either by multiplying (7.67) by B_h^{-1} or in the case of a decomposition $B_h = E_h^T E_h$ (for a symmetric positive definite B_h) by multiplying by E_h^{-T} and a change of variables to $E_h \boldsymbol{u}_h^n$. We will apply two techniques in the following:

One is based on the eigenvector decomposition of A_h; thus for (7.67), this means to consider the generalized eigenvalue problem

$$\hat{A}_h \boldsymbol{v} = \lambda B_h \boldsymbol{v}\,. \quad (7.68)$$

Note that the Galerkin approach for the eigenvalue problems according to Definition 7.6 leads to such a generalized eigenvalue problem with the stiffness matrix $\hat{A}_h$ and the mass matrix B_h.

The other approach is based on the matrix properties (1.32)* or (1.32). For the most important case,

$$B_h = \operatorname{diag}(b_i)\,, \quad b_i > 0 \quad \text{for} \quad i = 1, \ldots, M\,, \tag{7.69}$$

which corresponds to the mass lumping procedure, the above-mentioned transformation reduces to a diagonal scaling, which does not influence any of their properties.

Having this in mind, in the following we will consider explicitly only the formulation (7.66).

In the case $\Theta = 0$, the *explicit Euler method*, $\boldsymbol{u}_h^n$ can be determined explicitly by

$$\boldsymbol{u}_h^{n+1} = \tau(\boldsymbol{q}_h(t_n) - A_h\boldsymbol{u}_h^n) + \boldsymbol{u}_h^n = (I - \tau A_h)\boldsymbol{u}_h^n + \tau\boldsymbol{q}(t_n)\,.$$

Thus the effort for one time step consists of a SAXPY operation, a vector addition, and a matrix-vector operation. For dimension M the first of these is of complexity $O(M)$, and also the last one if the matrix is sparse in the sense defined at the beginning of Chapter 5. On the other hand, for $\Theta \neq 0$, the method is *implicit*, since for each time step a system of linear equations has to be solved with the system matrix $I + \Theta\tau A_h$. Here the cases $\Theta = 1$, *the implicit Euler method*, and $\Theta = \frac{1}{2}$, the *Crank–Nicolson method*, will be of interest. Due to our restriction to time-independent coefficients, the matrix is the same for every time step (for constant τ). If direct methods (see Section 2.5) are used, then the LU factorization has to be computed only once, and only forward and backward substitutions with changing right-hand sides are necessary, where computation for $\Theta \neq 1$ also requires a matrix-vector operation. For band matrices, for example, operations of the complexity bandwidth $\times$ dimension are necessary, which means for the basic example of the heat equation on a rectangle $O(M^{3/2})$ operations instead of $O(M)$ for the explicit method. Iterative methods for the resolution of (7.66) cannot make use of the constant matrix, but with $\boldsymbol{u}_h^n$ there is a good initial iterate if τ is not too large.

Although the explicit Euler method $\Theta = 0$ seems to be attractive, we will see later that with respect to accuracy or stability one may prefer $\Theta = \frac{1}{2}$ or $\Theta = 1$.

To investigate further the theta method, we resolve recursively the relations (7.66) to gain the representation

$$\begin{aligned} \boldsymbol{u}_h^n \;=\; & \left(\left(I + \Theta\tau A_h\right)^{-1}\left(I - \overline{\Theta}\tau A_h\right)\right)^n \boldsymbol{u}_0 \\ & + \tau\sum_{k=1}^{n}\left(\left(I + \Theta\tau A_h\right)^{-1}\left(I - \overline{\Theta}\tau A_h\right)\right)^{n-k}\left(I + \Theta\tau A_h\right)^{-1}\boldsymbol{q}_h\left(t_k - \overline{\Theta}\tau\right). \end{aligned} \tag{7.70}$$

Here we use the abbreviation $A^{-n} = (A^{-1})^n$ for a matrix A. Comparing this with the solution (7.55) of the semidiscrete problem, we see the approximations

$$e^{-A_h t_n} \sim E_{h,\tau}^n \,,$$

where

$$E_{h,\tau} := (I + \Theta\tau A_h)^{-1} \left(I - \overline{\Theta}\tau A_h\right)$$

and

$$\begin{aligned} \int_0^{t_n} e^{-A_h(t_n-s)} \boldsymbol{q}_h(s)ds &= \int_0^{t_n} \left(e^{-A_h\tau}\right)^{(t_n-s)/\tau} \boldsymbol{q}_h(s)ds \\ &\sim \tau \sum_{\substack{k=1 \\ s=k\tau}}^{n} E_{h,\tau}^{(t_n-s)/\tau} \left(I + \Theta\tau A_h\right)^{-1} \boldsymbol{q}_h\left(s - \overline{\Theta}\tau\right) . \end{aligned} \tag{7.71}$$

The matrix $E_{h,\tau}$ thus is the solution operator of (7.66) for one time step and homogeneous boundary conditions and right-hand side. It is to be expected that it has to capture the qualitative behaviour of $e^{-A_h\tau}$ that it is approximating. This will be investigated in the next section.

One-Step Discretizations for the Finite Element Method

The fully discrete scheme can be achieved in two ways: Besides applying (7.66) to (7.41) in the transformed variable or in the form (7.67), the discretization approach can applied directly to (7.44):

With $\partial U^{n+1} := (U^{n+1} - U^n)/\tau$, $f^{n+s} := sf(t_{n+1}) + (1-s)f(t_n)$, $b^{n+s}(v) := sb(t_{n+1}, v) + (1-s)b(t_n, v)$, b according to (7.43), $s \in [0,1]$, and with a fixed number $\Theta \in [0,1]$, the fully discrete method for (7.44) then reads as follows:

Find a sequence $U^0, \ldots, U^N \in V_h$ such that for $n \in \{0, \ldots, N-1\}$,

$$\begin{aligned} \left\langle \partial U^{n+1}, v_h \right\rangle_0 + a(\Theta U^{n+1} + \overline{\Theta} U^n, v_h) &= b^{n+\Theta}(v_h) \\ &\quad \text{for all } v_h \in V_h \,, \\ U^0 &= u_{0h} \,. \end{aligned} \tag{7.72}$$

An alternative choice for the right-hand side, closer to the finite difference method, is the direct evaluation at $t_n + \Theta\tau$, e.g., $f(t_n + \Theta\tau)$. The version here is chosen to simplify the order of convergence estimate in Section 7.6.

By representing the U^n by means of a basis of V_h as after (7.44), again we get the form (7.67) (or (7.66) in the transformed variable). Note that also for $\Theta = 0$ the problem here is implicit if B_h is not diagonal. Therefore, *mass lumping* is often applied, and the scalar product $\langle \cdot, \cdot \rangle_0$ in (7.72) is

replaced by an approximation due to numerical quadrature, i.e.,

$$\begin{aligned} \langle \partial U^{n+1}, v_h \rangle_{0,h} + a(\Theta U^{n+1} + \overline{\Theta} U^n, v_h) &= b^{n+\Theta}(v_h) \\ &\quad \text{for all } v_h \in V_h\,, \qquad (7.73) \\ U^0 &= u_{0h}\,. \end{aligned}$$

As explained in Section 3.5.2, $\langle u_h, v_h \rangle_{0,h}$ is the sum over all contributions from elements $K \in \mathcal{T}_h$, which takes the form (3.112) for the reference element. In the case of Lagrange elements and a nodal quadrature formula we have for the nodal basis functions φ_i:

$$\langle \varphi_j, \varphi_i \rangle_{0,h} = \langle \varphi_i, \varphi_i \rangle_{0,h}\, \delta_{ij} =: b_i \delta_{ij} \quad \text{for} \quad i,j = 1, \ldots, M, \qquad (7.74)$$

since for $i \neq j$ the integrand $\varphi_i \varphi_j$ vanishes at all quadrature points. In this case we arrive at the form (7.67) with a matrix B_h satisfying (7.69).

One-Step Discretizations for the Finite Volume Method

As in the previous subsection on the finite element approach, the semidiscrete formulation (7.48) can be discretized in time directly:

Find a sequence $U^0, \ldots, U^N \in V_h$ such that for $n \in \{0, \ldots, N-1\}$,

$$\begin{aligned} \left\langle \partial U^{n+1}, v_h \right\rangle_{0,h} + a_h(\Theta U^{n+1} + \overline{\Theta} U^n, v_h) &= \left\langle f^{n+\Theta}, v_h \right\rangle_{0,h} \\ &\quad \text{for all } v_h \in V_h, \qquad (7.75) \\ U^0 &= u_{0h}\,, \end{aligned}$$

where ∂U^{n+1}, Θ, $f^{n+\Theta}$ are defined as before (7.72).

Remember that here we consider only homogeneous boundary conditions.

If the elements U^n, U^{n+1} are represented by means of a basis of V_h, we recover the form (7.67).

Since the mass matrix B_h is diagonal, the problem can be regarded as being explicit for $\Theta = 0$.

Exercise

7.9 Consider linear simplicial elements defined on a general conforming triangulation of a polygonally bounded domain $\Omega \subset \mathbb{R}^2$.

(a) Determine the entries of the mass matrix B_h.

(b) Using the trapezoidal rule, determine the entries of the lumped mass matrix $\mathrm{diag}(b_i)$.

7.4 Stability

In Section 7.3 we have seen that at least if a basis of eigenvectors of the discretization matrix A_h allows for the solution representation (7.55) for

the semidiscrete method, the qualitative behaviour of $e^{-A_h\tau}\boldsymbol{u}_0$ should be preserved by $E_{h,\tau}\boldsymbol{u}_0$, being one time step τ for homogeneous boundary conditions and right-hand side ($\boldsymbol{q}_h = 0$) in the semi- and fully discrete cases. It is sufficient to consider the eigenvector $\boldsymbol{w}_i$ instead of a general $\boldsymbol{u}_0$. Thus, we have to compare

$$\left(e^{-A_h\tau}\right)\boldsymbol{w}_i = (e^{-\lambda_i\tau})\boldsymbol{w}_i \tag{7.76}$$

with

$$\left((I+\Theta\tau A_h)^{-1}\left(I-\overline{\Theta}\tau A_h\right)\right)\boldsymbol{w}_i = \left(\frac{1-\overline{\Theta}\tau\lambda_i}{1+\Theta\tau\lambda_i}\right)\boldsymbol{w}_i\,. \tag{7.77}$$

We see that the exponential function is approximated by

$$R(z) = \frac{1+(1-\Theta)z}{1-\Theta z}\,, \tag{7.78}$$

the *stability function*, at the points $z = -\lambda_i\tau \in \mathbb{C}$, given by the eigenvalues λ_i, and the time step size τ.

For n time steps and $\boldsymbol{q}_h = 0$ we have

$$\left(e^{-A_h\tau}\right)^n\boldsymbol{w}_i = e^{-\lambda_i t_n}\boldsymbol{w}_i \sim R(-\lambda_i\tau)^n\boldsymbol{w}_i\,. \tag{7.79}$$

Thus, the restriction to eigenvectors $\boldsymbol{w}_i$ with eigenvalues λ_i has diagonalized the system of ordinary differential equations (7.41) for $\boldsymbol{q}_h = 0$ to the scalar problems

$$\begin{aligned} \xi' + \lambda_i\xi &= 0\,, \quad t\in(0,T)\,, \\ \xi(0) &= \xi_0 \end{aligned} \tag{7.80}$$

(for $\xi_0 = 1$) with its solution $\xi(t) = e^{-\lambda_i t\xi_0}$, for which the one-step-theta method gives the approximation

$$\xi_{n+1} = R(-\lambda_i\tau)\xi_n = (R(-\lambda_i\tau))^{n+1}\xi_0 \tag{7.81}$$

at $t = t_{n+1}$. A basic requirement for a discretion method is the following:

Definition 7.17 A one-step method is called *nonexpansive* if for two numerical approximations $\boldsymbol{u}_h^n$ and $\tilde{\boldsymbol{u}}_h^n$, generated under the same conditions except for two discrete initial values $\boldsymbol{u}_0$ and $\tilde{\boldsymbol{u}}_0$, respectively, the following estimate is valid:

$$|\boldsymbol{u}_h^{n+1} - \tilde{\boldsymbol{u}}_h^{n+1}| \le |\boldsymbol{u}^n - \tilde{\boldsymbol{u}}^n|\,, \quad n\in\{0,\dots,N-1\}\,.$$

A recursive application of this estimate immediately results in

$$|\boldsymbol{u}^n - \tilde{\boldsymbol{u}}^n| \le |\boldsymbol{u}_0 - \tilde{\boldsymbol{u}}_0|\,, \quad n\in\{1,\dots,N\}\,.$$

Here a general one-step method has the form

$$\boldsymbol{u}_h^{n+1} = \boldsymbol{u}_h^n + \tau\boldsymbol{\Phi}(\tau,t_n,\boldsymbol{u}_h^n)\,, \quad n\in\{0,\dots,N-1\}\,,$$

with $\boldsymbol{u}_h^0 = \boldsymbol{u}_0$ and a so-called *generating function* $\boldsymbol{\Phi}:\mathbb{R}_+\times[0,T)\times\mathbb{R}^M\to\mathbb{R}^M$ that characterizes the particular method. The generating function of

the one-step-theta method applied to the system (7.41) is

$$\mathbf{\Phi}(\tau, t, \boldsymbol{\xi}) = -(I + \tau\Theta A_h)^{-1} \left[A_h\boldsymbol{\xi} - \boldsymbol{q}_h(t + \Theta\tau)\right] .$$

Thus nonexpansiveness models the fact that perturbances, i.e., in particular errors, are not amplified in time by the numerical method. This is considerably weaker than the exponential decay in the continuous solution (see (7.18)), which would be too strong a request.

Having in mind (7.79)–(7.81), and expecting the (real parts of the) eigenvalues to be positive, the following restriction is sufficient:

Definition 7.18 A one-step method is called *A-stable* if its application to the scalar model problem (7.80)

$$\begin{aligned} \xi' + \lambda\xi &= 0\,, \quad t \in (0, T)\,, \\ \xi(0) &= \xi_0\,, \end{aligned}$$

yields a nonexpansive method for all complex parameters λ with $\Re\lambda > 0$ and arbitrary step sizes $\tau > 0$.

Because of (7.81) we have

$$\xi_{n+1} - \tilde{\xi}_{n+1} = R(-\lambda\tau)[\xi_n - \tilde{\xi}_n]$$

for two approximations of the one-step-theta method applied to (7.80). This shows that the condition

$$|R(z)| \leq 1 \quad \text{for all } z \text{ with } \Re z < 0$$

is sufficient for the A-stability of the method. More generally, any one-step method that can be written for (7.80) in the form

$$\xi_{n+1} = R(-\lambda_i\tau)\xi_n \tag{7.82}$$

is nonexpansive iff

$$|R(-\lambda_i\tau)| \leq 1\,. \tag{7.83}$$

The one-step-theta method is nonexpansive for (7.41) in the case of an eigenvector basis if (7.83) holds for all eigenvalues λ_i and step size τ. A convenient formulation can be achieved by the notion of the domain of stability.

Definition 7.19 Given a stability function $R : \mathbb{C} \to \mathbb{C}$, the set

$$S_R := \{z \in \mathbb{C} : |R(z)| < 1\}$$

is called a *domain of (absolute) stability* of the one-step method $\xi_{n+1} = R(-\lambda\tau)\xi_n$.

Example 7.20
For the one-step-theta method we have:

(1) For $\Theta = 0$, S_R is the (open) unit disk with centre $z = -1$.

(2) For $\Theta = \frac{1}{2}$, S_R coincides with the left complex half-plane (except for the imaginary axis).

(3) For $\Theta = 1$, S_R is the whole complex plane except for the closed unit disk with centre $z = 1$.

The notion of A-stability reflects the fact that the property $|\mathrm{e}^{-\lambda\tau}| \leq 1$ for $\Re\lambda > 0$ is satisfied by the function $R(-\lambda\tau)$, too:

Corollary 7.21 *For a continuous stability function R the one-step method $\xi^{n+1} = R(-\lambda\tau)\xi^n$ is A-stable if the closure $\overline{S}_R$ of its domain of stability contains the left complex half-plane.*

Thus the Crank–Nicolson and the implicit Euler methods are A-stable, but not the explicit Euler method. To have nonexpansiveness, we need the requirement

$$|1 - \lambda_i\tau| = |R(-\lambda_i\tau)| \leq 1\,, \tag{7.84}$$

which is a *step size restriction:* For positive λ_i it reads

$$\tau \leq 2/\max\{\lambda_i \,|\, i = 1, \ldots, M\}\,. \tag{7.85}$$

For the example of the five-point stencil discretization of the heat equation on a rectangle with Dirichlet boundary conditions according to (7.37)–(7.39), equation (7.84) reads

$$\left|1 - \frac{\tau}{h^2} 2\left(2 - \cos\left(\nu\frac{\pi}{a}h\right) - \cos\left(\mu\frac{\pi}{b}h\right)\right)\right| \leq 1 \tag{7.86}$$

for all $\nu = 1, \ldots, l-1$, $\mu = 1, \ldots, m-1$.

The following condition is sufficient (and for $l, m \to \infty$ also necessary):

$$\frac{\tau}{h^2} \;\leq\; \frac{1}{4}\,. \tag{7.87}$$

For the finite element method a similar estimate holds in a more general context. Under the assumptions of Theorem 3.45 we conclude from its proof (see (3.141)) that the following holds:

$$\max\{\lambda_i \,|\, i = 1, \ldots, M\} \leq C \left(\min_{K \in \mathcal{T}_h} h_K\right)^{-2}$$

for the eigenvalues of $B_h^{-1}\hat{A}_h$, where $B_h = E_h^T E_h$ is the mass matrix and $\hat{A}_h$ the stiffness matrix, and thus also for $A_h = E_h B_h^{-1}\hat{A}_h E_h^{-1}$. Here C is a constant independent of h.

Therefore, we have

$$\tau \Big/ \left(\min_{K \in \mathcal{T}_h} h_K\right)^2 \;\leq\; 2/C \tag{7.88}$$

as a sufficient condition for the nonexpansiveness of the method with a specific constant depending on the stability constant of the bilinear form and the constant from Theorem 3.43, (2).

These step size restrictions impede the attractivity of the explicit Euler method, and so implicit versions are often used. But also in the A-stable case there are distinctions in the behaviour (of the stability functions). Comparing them, we see that

$$\begin{aligned} &\text{for} \quad \Theta = \tfrac{1}{2}: \qquad R(-x) \to -1 \qquad \text{for} \qquad x \to \infty\,; \\ &\text{for} \quad \Theta = 1: \qquad R(-x) \to 0 \qquad\ \ \text{for} \qquad x \to \infty\,. \end{aligned} \tag{7.89}$$

This means that for the implicit Euler method the influence of large eigenvalues will be more greatly damped, the larger they are, corresponding to the exponential function to be approximated, but the Crank–Nicolson method preserves these components nearly undamped in an oscillatory manner. This may lead to a problem for "rough" initial data or discontinuities between initial data and Dirichlet boundary conditions. On the other hand, the implicit Euler method also may damp solution components too strongly, making the solution "too" smooth.

The corresponding notion is the following:

Definition 7.22 One-step methods whose stability function satisfies

$$R(z) \to 0 \quad \text{for } \Re z \to -\infty,$$

are called *L-stable.*

An intermediate position is filled by the *strongly A-stable methods.* They are characterized by the properties

- $|R(z)| < 1$ for all z with $\Re z < 0$,
- $\lim\limits_{\Re z \to -\infty} |R(z)| < 1$.

Example 7.23

(1) Among the one-step-theta methods, only the implicit Euler method ($\Theta = 1$) is L-stable.
(2) The Crank–Nicolson method ($\Theta = \frac{1}{2}$) is not strongly A-stable, because of (7.89).

The nonexpansiveness of a one-step method can also be characterized by a norm condition for the solution operator $E_{h,\tau}$.

Theorem 7.24 *Let the spatial discretization matrix A_h have a basis of eigenvectors $\boldsymbol{w}_i$ orthogonal with respect to the scalar product $\langle \cdot, \cdot \rangle_h$, with eigenvalues λ_i, $i = 1, \ldots, M$. Consider the problem* (7.41) *and its discretization in time by a one-step method with a linear solution representation*

$$\boldsymbol{u}_h^n = E_{h,\tau}^n \boldsymbol{u}_0 \tag{7.90}$$

for $\boldsymbol{q}_h = 0$, where $E_{h,\tau} \in \mathbb{R}^{M,M}$, and a stability function R such that (7.82) *and*

$$E_{h,\tau} \boldsymbol{w}_i = R(-\lambda_i \tau) \boldsymbol{w}_i \tag{7.91}$$

for $i = 1, \ldots, M$. Then the following statements are equivalent:

(1) *The one-step method is nonexpansive for the model problem* (7.80) *and all eigenvalues λ_i of A_h.*

(2) *The one-step method is nonexpansive for the problem* (7.41)*, with respect to the norm $\|\cdot\|_h$ induced by $\langle\cdot,\cdot\rangle_h$.*

(3) *$\|E_{h,\tau}\|_h \leq 1$ in the matrix norm $\|\cdot\|_h$ induced by the vector norm $\|\cdot\|_h$.*

Proof: We prove (1) $\Rightarrow$ (3) $\Rightarrow$ (2) $\Rightarrow$ (1):
(1) $\Rightarrow$ (3): According to (7.83) (1) is characterized by

$$|R(-\lambda_i\tau)| \leq 1\,, \tag{7.92}$$

for the eigenvalues λ_i.

For the eigenvector, $\boldsymbol{w}_i$ with eigenvalue λ_i we have (7.91), and thus, for an arbitrary $\boldsymbol{u}_0 = \sum_{i=1}^{M} c_i \boldsymbol{w}_i$,

$$\begin{aligned} \|E_{h,\tau}\boldsymbol{u}_0\|_h^2 &= \|\sum_{i=1}^{M} c_i E_{h,\tau}\boldsymbol{w}_i\|_h^2 \\ &= \|\sum_{i=1}^{M} c_i R(-\lambda_i\tau)\boldsymbol{w}_i\|_h^2 = \sum_{i=1}^{M} c_i^2 |R(-\lambda_i\tau)|^2 \, \|\boldsymbol{w}_i\|_h^2\,, \end{aligned}$$

because of the orthogonality of the $\boldsymbol{w}_i$, and analogously,

$$\|\boldsymbol{u}_0\|_h^2 = \sum_{i=1}^{M} c_i^2 \|\boldsymbol{w}_i\|_h^2\,,$$

and finally, because of (7.92),

$$\|E_{h,\tau}\boldsymbol{u}_0\|_h^2 \leq \sum_{i=1}^{M} c_i^2 \|\boldsymbol{w}_i\|_h^2 = \|\boldsymbol{u}_0\|_h^2\,,$$

which is assertion (3).
(3) $\Rightarrow$ (2): is obvious.
(2) $\Rightarrow$ (3):

$$|R(-\lambda_i\tau)| \, \|\boldsymbol{w}_i\|_h = \|R(-\lambda_i\tau)\boldsymbol{w}_i\|_h = \|E_{h,\tau}\boldsymbol{w}_i\|_h \leq \|\boldsymbol{w}_i\|_h\,.$$

□

Thus, nonexpansiveness is often identical to what is (vaguely) called *stability*:

Definition 7.25 A one-step method with a solution representation $E_{h,\tau}$ for $q_h = 0$ is called *stable* with respect to the vector norm $\|\cdot\|_h$ if

$$\|E_{h,\tau}\|_h \leq 1$$

in the induced matrix norm $\|\cdot\|_h$.

Till now we have considered only homogeneous boundary data and right-hand sides. At least for the one-step-theta method this is not a restriction:

Theorem 7.26 *Consider the one-step-theta method under the assumption of Theorem 7.24, with $\lambda_i \geq 0$, $i = 1, \ldots, M$, and with τ such that the method is stable. Then the solution is stable in initial condition $\boldsymbol{u}_0$ and right-hand side $\boldsymbol{q}_h$ in the following sense:*

$$\|\boldsymbol{u}_h^n\|_h \leq \|\boldsymbol{u}_0\|_h + \tau \sum_{k=1}^{n} \left\|\boldsymbol{q}_h\left(t_k - \overline{\Theta}\tau\right)\right\|_h . \tag{7.93}$$

Proof: From the solution representation (7.70) we conclude that

$$\|\boldsymbol{u}_h^n\|_h \leq \|E_{h,\tau}\|_h^n \|\boldsymbol{u}_0\|_h + \tau \sum_{k=1}^{n} \|E_{h,\tau}\|_h^{n-k} \|(I + \tau\Theta A_h)^{-1}\|_h \|\boldsymbol{q}_h(t_k - \overline{\Theta}\tau)\|_h \tag{7.94}$$

using the submultiplicativity of the matrix norm.

We have the estimate

$$\|(I + \Theta\tau A_h)^{-1}\boldsymbol{w}_i\|_h = \left|\frac{1}{1 + \Theta\,\tau\lambda_i}\right| \|\boldsymbol{w}_i\|_h \;\leq\; \|\boldsymbol{w}_i\|_h ,$$

and thus as in the proof of Theorem 7.24, (1) $\Rightarrow$ (3),

$$\|(I + \Theta\,\tau A_h)^{-1}\|_h \;\leq\; 1$$

concludes the proof. □

The stability condition requires step size restrictions for $\Theta < \frac{1}{2}$, which have been discussed above for $\Theta = 0$.

The requirement of stability can be weakened to

$$\|E_{h,\tau}\|_h \leq 1 + K\tau \tag{7.95}$$

for some constant $K > 0$, which in the situation of Theorem 7.24 is equivalent to

$$|R(-\lambda\tau)| \leq 1 + K\tau,$$

for all eigenvalues λ of A_h. Because of

$$(1 + K\tau)^n \leq \exp(Kn\tau),$$

in (7.93) the additional factor $\exp(KT)$ appears and correspondingly $\exp(K(n-k)\tau)$ in the sum. If the process is to be considered only in a small time interval, this becomes part of the constant, but for large time horizons the estimate becomes inconclusive.

On the other hand, for the one-step-theta method for $\frac{1}{2} < \Theta \leq 1$ the estimate $\|E_{h,\tau}\|_h \leq 1$ and thus the constants in (7.93) can be sharpened to $\|E_{h,\tau}\|_h \;\leq\; R(-\lambda_{\min}\tau)$, where $\lambda_{\min}$ is the smallest eigenvalue of A_h,

reflecting the exponential decay. For example, for $\Theta = 1$, the (error in the) initial data is damped with the factor

$$\|E_{h,\tau}\|_h^n = R(-\lambda_{\min}\tau)^n = \frac{1}{(1+\lambda_{\min}\tau)^n}\,,$$

which for $\tau \le \tau_0$ for some fixed $\tau_0 > 0$ can be estimated by

$$\exp(-\lambda n\tau) \quad \text{for some} \quad \lambda > 0.$$

We conclude this section with an example.

Example 7.27 (Prothero-Robinson model) Let $g \in C^1[0,T]$ be given. We consider the initial value problem

$$\begin{aligned} \xi' + \lambda(\xi - g) &= g'\,, \quad t \in (0,T)\,, \\ \xi(0) &= \xi_0\,. \end{aligned}$$

Obviously, g is a particular solution of the differential equation, so the general solution is

$$\xi(t) = \mathrm{e}^{-\lambda t}[\xi_0 - g(0)] + g(t)\,.$$

In the special case $g(t) = \arctan t$, $\lambda = 500$, and for the indicated values of ξ_0, Figure 7.1 shows the qualitative behaviour of the solution.

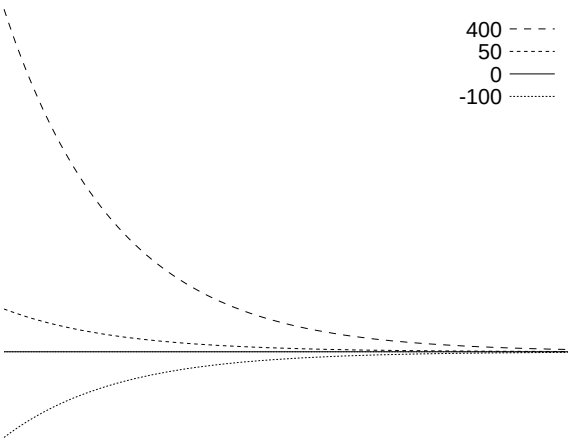


Figure 7.1. Prothero–Robinson model.

It is worth mentioning that the figure is extremely scaled: The continuous line (to $\xi_0 = 0$) seems to be straight, but it is the graph of g.

The explicit Euler method for this model is

$$\xi^{n+1} = (1-\lambda\tau)\xi^n + \tau\left[g'(t_n) + \lambda g(t_n)\right]\,.$$

According to the above considerations, it is nonexpansive only if $\lambda\tau \leq 1$ holds. For large numbers λ, this is a very restrictive step size condition; see also the discussion of (7.85) to (7.87).

Due to their better stability properties, implicit methods such as the Crank–Nicolson and the implicit Euler methods do not have such step size restrictions. Nevertheless, the application of implicit methods is not free from surprises. For example, in the case of large numbers λ, an order reduction can occur.

Exercises

7.10 Determine the corresponding domain of stability S_R of the one-step-theta method for the following values of the parameter $\Theta : 0, \frac{1}{2}, 1$.

7.11 Show the L-stability of the implicit Euler method.

7.12 (a) Show that the discretization

$$\xi^n = \xi^{n-2} + 2\tau f(t_{n-1}, \xi^{n-1}), \quad n = 2, \dots N$$

(*midpoint rule*), applied to the model equation $\xi' = f(t, \xi)$ with $f(t, \xi) = -\lambda\xi$ and $\lambda > 0$ leads, for a sufficiently small step size $\tau > 0$, to a general solution that can be additively decomposed into a decaying and an increasing (by absolute value) oscillating component.

(b) Show that the oscillating component can be damped if additionally the quantity ξ_*^N is computed (*modified midpoint rule*):

$$\xi_*^N = \frac{1}{2}\left[\xi^N + \xi^{N-1} + \tau f(t_N, \xi^N)\right] .$$

7.13 Let $m \in \mathbb{N}$ be given. Find a polynomial $R_m(z) = 1 + z + \sum_{j=2}^{m} \gamma_j z^j$ ($\gamma_j \in \mathbb{R}$) such that the corresponding domain of absolute stability for $R(z) := R_m(z)$ contains an interval of the negative real axis that is as large as possible.

7.5 The Maximum Principle for the One-Step-Theta Method

In Section 1.4 we have seen that for a discrete problem of the form (1.31) there is a hierarchy of properties ranging from a comparison principle to a strong maximum principle, which is in turn applied by a hierarchy of conditions, partly summarized as (1.32) or (1.32)*. To remind the reader, we regroup these conditions accordingly:

The collection of conditions (1.32), (1), (2), (3) i), $(4)^*$ is called *(IM)*.
(IM) implies the inverse monotonicity of A_h (Theorem 1.12, (1.39)).
The collection of conditions *(IM)*, (5) is called *(CP)*.
(CP) implies a comparison principle in the sense of Corollary 1.13.
The collection of conditions *(CP)*, $(6)^*$ is called $(MP)^*$.
$(MP)^*$ implies a maximum principle in the form of Theorem 1.10 (1.38).
Alternatively, the collection of conditions *(CP)* $(6)^\#$ (see Exercise 1.13) is called *(MP)*.
(MP) implies a maximum principle in the form of Theorem 1.9 (1.34).
Finally, the collection of conditions *(CP)*, (6), (4) (instead of $(4)^*$), (7) is called *(SMP)*.
(SMP) implies a strong maximum principle in the sense of Theorem 1.9.

An L^∞-stability estimate in the sense of Theorem 1.14 is closely related. This will be taken up in the next section.

In the following we will discuss the above-mentioned properties for the one-step-theta method, cast into the form (1.31), on the basis of corresponding properties of the underlying elliptic problem and its discretization. It will turn out that under a reasonable condition (see (7.100)), condition $(4)^*$ (and thus (3) ii)) will not be necessary for the elliptic problem. This reflects the fact that contrary to the elliptic problem, for the parabolic problem also the case of a pure Neumann boundary condition (where no degrees of freedom are given and thus eliminated) is allowed, since the initial condition acts as a Dirichlet boundary condition.

In assuming that the discretization of the underlying elliptic problem is of the form (1.31), we return to the notation $M = M_1 + M_2$, where M_2 is the number of degrees of freedom eliminated, and thus $A_h, B_h \in \mathbb{R}^{M_1, M_1}$.

We write the discrete problem according to (7.66) as one large system of equations for the unknown

$$\boldsymbol{u}_h = \begin{pmatrix} \boldsymbol{u}_h^1 \\ \boldsymbol{u}_h^2 \\ \vdots \\ \boldsymbol{u}_h^N \end{pmatrix}, \tag{7.96}$$

in which the vector of grid values $\boldsymbol{u}_h^i \in \mathbb{R}^{M_1}$ are collected to one large vector of dimension $\overline{M}_1 := N \cdot M_1$. Thus the grid points in $\Omega \times (0, T)$ are the points (x_j, t_n), $n = 1, \ldots, N$, $x_j \in \Omega_h$, e.g., for the finite difference method. The defining system of equations has the form

$$C_h \boldsymbol{u}_h = \boldsymbol{p}_h \,, \tag{7.97}$$

where

$$C_h = \begin{pmatrix} I + \tau\Theta A_h & & & \\ -I + \tau\overline{\Theta}A_h & \ddots & & \mathbf{0} \\ & \ddots & \ddots & \\ \mathbf{0} & & \ddots & \ddots \\ & & -I + \tau\overline{\Theta}A_h & I + \tau\Theta A_h \end{pmatrix},$$

again with $\overline{\Theta} := 1 - \Theta$,

$$\boldsymbol{p}_h = \begin{pmatrix} \tau\boldsymbol{q}_h(\Theta\tau) + (I - \tau\overline{\Theta}A_h)\boldsymbol{u}_0 \\ \tau\boldsymbol{q}_h((1+\Theta)\tau) \\ \vdots \\ \vdots \\ \tau\boldsymbol{q}_h(N - 1 + \Theta)\tau) \end{pmatrix}.$$

Since the spatial discretization is performed as in the stationary case, and in the nth step the discretization relates to $t = t_{n-1} + \Theta\tau$ and also the approximation

$$A_h u(t_{n-1} + \Theta\tau) \sim \overline{\Theta}A_h u(t_{n-1}) + \Theta A_h u(t_n)$$

enters the formulation (7.66), we can assume to have the following structure of the right-hand side of (7.66):

$$\boldsymbol{q}_h((n-1+\Theta)\tau) = -\hat{A}_h(\overline{\Theta}\hat{\boldsymbol{u}}_h^{n-1} + \Theta\hat{\boldsymbol{u}}_h^n) + \boldsymbol{f}((n-1+\Theta)\tau) \text{ for } n = 1, \ldots, N. \tag{7.98}$$

Here the $\hat{\boldsymbol{u}}_h^n \in \mathbb{R}^{M_2}$ are the known spatial boundary values on time level t_n, which have been eliminated from the equation as explained, e.g., in Chapter 1 for the finite difference method. But as noted, we allow also for the case where such values do not appear (i.e., $M_2 = 0$) then (7.98) reduces to

$$\boldsymbol{q}_h((n-1+\Theta)\tau) = \boldsymbol{f}((n-1+\Theta)\tau) \quad \text{for} \quad n = 1, \ldots, N\,.$$

For the continuous problem, data are prescribed at the *parabolic boundary* $\Omega \times \{0\} \cup \partial\Omega \times [0, T]$; correspondingly, the known values $\hat{\boldsymbol{u}}_h^i$ are collected with the initial data $\boldsymbol{u}_0 \in \mathbb{R}^{M_1}$ to a large vector

$$\hat{\boldsymbol{u}}_h = \begin{pmatrix} \boldsymbol{u}_0 \\ \hat{\boldsymbol{u}}_h^0 \\ \hat{\boldsymbol{u}}_h^1 \\ \vdots \\ \hat{\boldsymbol{u}}_h^N \end{pmatrix},$$

i.e., a vector of dimension $\overline{M}_2 := M_1 + (N+1)M_2$, which may reduce to $\hat{\boldsymbol{u}}_h = \boldsymbol{u}_0 \in \mathbb{R}^{M_1}$.

With this notation we have

$$\boldsymbol{p}_h = -\hat{C}_h \hat{\boldsymbol{u}}_h + \boldsymbol{e} \tag{7.99}$$

if we define

$$\hat{C}_h = \begin{pmatrix} -I + \tau\overline{\Theta}A_h & \tau\overline{\Theta}\hat{A}_h & \tau\Theta\hat{A}_h & & O \\ O & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ O & & & \tau\overline{\Theta}\hat{A}_h & \tau\Theta\hat{A}_h \end{pmatrix},$$

$$\boldsymbol{e} = \begin{pmatrix} \tau\boldsymbol{f}(\Theta\tau) \\ \tau\boldsymbol{f}((1+\Theta)\tau) \\ \vdots \\ \vdots \\ \tau\boldsymbol{f}((N-1+\Theta)\tau) \end{pmatrix}.$$

In the following the validity of (1.32)* or (1.32) for

$$\tilde{C}_h = (C_h, \hat{C}_h)$$

will be investigated on the basis of corresponding properties of

$$\tilde{A}_h = (A_h, \hat{A}_h)\ .$$

Note that even if A_h is irreducible, the matrix C_h is always reducible, since $\boldsymbol{u}_h^n$ depends only on $\boldsymbol{u}_h^1, \ldots, \boldsymbol{u}_h^{n-1}$, but not on the future time levels. (Therefore, (7.97) serves only for the theoretical analysis, but not for the actual computation.)

In the following we assume that

$$\tau\overline{\Theta}(A_h)_{jj} < 1 \quad \text{for } j = 1, \ldots, M_1\ , \tag{7.100}$$

which is always satisfied for the implicit Euler method ($\Theta = 1$). Then:

(1) $(C_h)_{rr} > 0$ *for* $r = 1, \ldots, \overline{M}_1$
holds if (1) is valid for A_h. Actually, also $(A_h)_{jj} > -1/(\tau\Theta)$ would be sufficient.

(2) $(C_h)_{rs} \le 0$ *for* $r, s = 1, \ldots, \overline{M}_1$, $r \ne s$:
If (2) is valid for A_h, then only the nonpositivity of the diagonal elements of the off-diagonal block of C_h, $-I + \tau\overline{\Theta}A_h$, is in question. This is ensured by (7.100) (weakened to "$\le$").

(3) (i) $C_r := \Big(\sum_{s=1}^{\overline{M}_1} C_h\Big)_{rs} \ge 0$ *for* $r = 1, \ldots, \overline{M}_1$:

(ii) $C_r > 0$ *for at least one* $r \in \{1, \ldots, \overline{M}_1\}$:
We set

$$A_j := \sum_{k=1}^{M_1} (A_h)_{jk} \,,$$

so that condition (3) (i) for A_h means that $A_j \geq 0$ for $j = 1, \ldots, M_1$. Therefore, we have

$$C_r = 1 + \tau\Theta A_j > 0 \tag{7.101}$$

for the indices r of the first time level, where the "global" index r corresponds to the "local" spatial index j. For the following time levels, the relation

$$C_r = 1 - 1 + \tau(\Theta + \overline{\Theta})A_j = \tau A_j \geq 0 \tag{7.102}$$

holds, i.e., (3) (i) and (ii).

$(4)^*$ *For every* $r_1 \in \{1, \ldots, \overline{M}_1\}$ *satisfying*

$$\sum_{r=1}^{\overline{M}_1} (C_h)_{rs} = 0 \tag{7.103}$$

there exist indices $r_2, \ldots, r_{l+1}$ *such that*

$$(C_h)_{r_i r_{i+1}} \neq 0 \quad \text{for} \quad i = 1, \ldots, l$$

and

$$\sum_{s=1}^{\overline{M}_1} (C_h)_{r_{l+1} s} > 0 \,. \tag{7.104}$$

To avoid too many technicalities, we adopt the background of a finite difference method. Actually, only matrix properties enter the reasoning. We call (space-time) grid points satisfying (7.103) *far from the boundary*, and those satisfying (7.104) *close to the boundary*. Due to (7.101), all points of the first time level are close to the boundary (consistent with the fact that the grid points for $t_0 = 0$ belong to the parabolic boundary). For the subsequent time level n, due to (7.102), a point (x_i, t_n) is close to the boundary if x_i is close to the boundary with respect to $\tilde{A}_h$. Therefore, the requirement of $(4)^*$, that a point far from the boundary can be connected via a chain of neighbours to a point close to the boundary, can be realized in two ways: Firstly, within the time level n, i.e., the diagonal block of C_h if A_h satisfies condition $(4)^*$. Secondly, without this assumption a chain of neighbours exist by (x, t_n), (x, t_{n-1}) up to (x, t_1), i.e., a point close to the boundary, since the diagonal element of $-I + \tau\overline{\Theta}A_h$ does not vanish due to (7.100). This reasoning additionally has established the following:

(4)$^{\#}$ *If A_h is irreducible, then a grid point (x, t_n), $x \in \Omega_h$ can be connected via a chain of neighbours to every grid point (y, t_k), $y \in \Omega_h$ and $0 \le k \le n$.*

(5) $(\hat{C}_h)_{rs} \le 0$ *for* $r = 1, \dots, \overline{M}_1$, $s = \overline{M}_1 + 1, \dots, \overline{M}_2$:
Analogously to (2), this follows from (5) for $\hat{A}_h$ and (7.100).

(6)* $\tilde{C}_r := \sum_{s=1}^{M} (\tilde{C}_h)_{rs} = 0$ *for* $r = 1, \dots, M$:
Analogously to (7.102), we have

$$\tilde{C}_r = \tau \tilde{A}_j := \tau \sum_{k=1}^{M} (\tilde{A}_h)_{jk} \,,$$

so that the property is equivalent to the corresponding one of $\tilde{A}_h$.

(6) $\tilde{C}_r \ge 0$ *for* $r = 1, \dots, \overline{M}$
is equivalent to (6) for $\tilde{A}_h$ by the above argument.

(7) *For every $s \in \overline{M}_1 + 1, \dots, \overline{M}$ there exists an $r \in \{1, \dots, \overline{M}_1\}$ such that $(\hat{C}_h)_{rs} \ne 0$:*
Every listed boundary value should influence the solution: For the values from $\hat{\boldsymbol{u}}_h^0, \dots, \hat{\boldsymbol{u}}_h^N$ this is the case iff $\hat{A}_h$ satisfies this property. Furthermore, the "local" indices of the equation, where the boundary values appear, are the same for each time level. For the values from $\boldsymbol{u}_0 \in \mathbb{R}^{M_1}$ the assertion follows from (7.100).

From the considerations we have the following theorem:

Theorem 7.28 *Consider the one-step-theta method in the form* (7.66). *Let* (7.100) *hold. If the spatial discretization $\hat{A}_h$ satisfies* (1.32) (1), (2), (3) (i), *and* (5), *then a* comparison principle *holds:*

(1) *If for two sets of data $\boldsymbol{f}_i$, $\boldsymbol{u}_{0i}$ and $\hat{\boldsymbol{u}}_{hi}^n$, $n = 0, \dots, N$ and $i = 1, 2$, we have*

$$\boldsymbol{f}_1((n - 1 + \Theta)\tau) \le \boldsymbol{f}_2((n - 1 + \Theta)\tau) \quad \text{for } n = 1, \dots, N \,,$$

and

$$\boldsymbol{u}_{01} \le \boldsymbol{u}_{02};, \quad \hat{\boldsymbol{u}}_{h1}^n \le \hat{\boldsymbol{u}}_{h1}^n \quad \text{for } n = 0, \dots, N,$$

then

$$\hat{\boldsymbol{u}}_{h1}^n \le \hat{\boldsymbol{u}}_{h2}^n \quad \text{for } n = 1, \dots, N$$

for the corresponding solutions.
If $\hat{\boldsymbol{u}}_{h1}^n = \hat{\boldsymbol{u}}_{h2}^n$ for $n = 1, \dots, N$, then condition (1.32) (5) *can be omitted.*

(2) *If* $\tilde{A}_h$ *additionally satisfies* (1.32) $(6)^*$, *then the following* weak maximum principle *holds:*

$$\max_{\substack{r\in\{1,\dots,M\}\\ n=0,\dots,N}} (\tilde{\boldsymbol{u}}_h^n)_r \;\le\; \max\Big(\max_{r\in\{1,\dots,M_1\}} (\boldsymbol{u}_0)_r,\; \max_{\substack{r\in\{M_1+1,\dots,M\}\\ n=0,\dots,N}} (\hat{\boldsymbol{u}}_h^n)_r \Big),$$

where

$$\tilde{\boldsymbol{u}}_h^n := \begin{pmatrix} \boldsymbol{u}_h^n \\ \hat{\boldsymbol{u}}_h \end{pmatrix} .$$

(3) *If* $\tilde{A}_h$ *satisfies* (1.32) (1), (2), (3) (i), (4), (5), (6), (7), *then a* strong maximum principle *in the following sense holds:*
If the components of $\tilde{\boldsymbol{u}}_h^n$, $n = 0,\dots,N$, *attain a nonnegative maximum for some spatial index* $r \in \{1,\dots,M_1\}$ *and at some time level* $k \in \{1,\dots,N\}$, *then all components for the time levels* $n = 0,\dots,k$ *are equal.*

Proof: Only part (3) needs further consideration. Theorem 1.9 cannot be applied directly to (7.97), since C_h is reducible. Therefore, the proof of Theorem 1.9 has to be repeated: We conclude that the solution is constant at all points that are connected via a chain of neighbours to the point where the maximum is attained. According to $(4)^\#$ these include all grid points (x, t_l) with $x \in \Omega_h$ and $l \in \{0,\dots,k\}$. From (7.100) and the discussion of (7) we see that the connection can also be continued to boundary values up to level k. □

The additional condition (7.100), which may be weakened to nonstrict inequality, as seen above, actually is a time step restriction: Consider again the example of the five-point stencil discretization of the heat equation on a rectangle, for which we have $(A_h)_{jj} = 4/h^2$. Then the condition takes the form

$$\frac{\tau}{h^2} < \frac{1}{4(1-\Theta)} \tag{7.105}$$

for $\Theta < 1$. This is very similar to the condition (7.87), (7.88) for the explicit Euler method, but the background is different.

As already noted, the results above also apply to the more general form (7.67) under the assumption (7.69). The condition (7.100) then takes the form

$$\tau\overline{\Theta}(A_h)_{jj} \;\le\; b_j \quad \text{for} \quad j = 1,\dots,M_1 .$$

Exercises

7.14 Formulate the results of this section, in particular condition (7.100), for the problem in the form (7.67) with B_h according to (7.69) (i.e.

appropriate for finite element discretizations with mass lumping, see (7.74)).

7.15 Show the validity of $(6)^{\#}$ from Exercise 1.13 for C_h if it holds here for A_h and conclude as in Exercise 1.13 a weak maximum principle for the one-step-theta method.

7.16 Consider the initial-boundary value problem in one space dimension

$$\left\{\begin{array}{rcll} u_t - \varepsilon u_{xx} + c u_x & = & f & \text{in } (0,1)\times(0,T)\,, \\ u(0,t) \;=\; g_-(t),\; u(1,t) & = & g_+(t)\,, & t\in(0,T)\,, \\ u(x,0) & = & u_0(x)\,, & x\in(0,1)\,, \end{array}\right.$$

where $T > 0$ and $\varepsilon > 0$ are constants, and $c, f : (0,1)\times(0,T) \to \mathbb{R}$, $u_0 : (0,1) \to \mathbb{R}$, and $g_-, g_+ : (0,T) \to \mathbb{R}$ are sufficiently smooth functions such that the problem has a classical solution.

Define $h := 1/m$ and $\tau = T/N$ for some numbers $m, N \in \mathbb{N}$. Then the so-called *full-upwind finite difference method* for this problem reads as follows: Find a sequence of vectors $\boldsymbol{u}_h^0, \ldots, \boldsymbol{u}_h^N$ by

$$\frac{u_i^{n+1} - u_i^n}{\tau} - \varepsilon \frac{u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1}}{h^2} - c^- \frac{u_{i+1}^{n+1} - u_i^{n+1}}{h} + c^+ \frac{u_i^{n+1} - u_{i-1}^{n+1}}{h} = f_i^{n+1}, \qquad i = 1, \ldots, m-1,\ n = 0, \ldots, N-1,$$

where $c = c^+ - c^-$ with $c^+ = \max\{c, 0\}$, $f_i^n = f(ih, n\tau)$, $u_i^0 = u_0(ih)$, $u_0^n = g_-(n\tau)$ and $u_m^n = g_+(n\tau)$.

Prove that a weak maximum principle holds for this method.

7.6 Order of Convergence Estimates

Based on stability results already derived, we will investigate the (order of) convergence properties of the one-step-theta method for different discretization approaches. Although the results will be comparable, they will be in different norms, appropriate for the specific discretization method, as already seen in Chapters 1, 3, and 6.

Order of Convergence Estimates for the Finite Difference Method From Section 1.4 we know that the investigation of the (order of) convergence of a finite difference method consists of two ingredients:

- (order of) convergence of the consistency error
- stability estimates.

The last tool has already been provided by Theorem 7.26 and by Theorem 1.14, which together with the considerations of Section 7.5 allow us to concentrate on the consistency error. Certain smoothness properties will be

required for the classical solution u of the initial boundary value problem (7.1), which in particular makes its evaluation possible at the grid points $x_i \in \overline{\Omega}_h$ at each instance of time $t \in [0,T]$ and also of various derivatives. The vector representing the corresponding grid function (for a fixed ordering of the grid points) will be denoted by $\boldsymbol{U}(t)$, or for short by $\boldsymbol{U}^n := \boldsymbol{U}(t_n)$ for $t = t_n$. The corresponding grid points depend on the boundary condition. For a pure Dirichlet problem, the grid points will be from Ω_h, but if Neumann or mixed boundary conditions appear, they are from the enlarged set

$$\tilde{\Omega}_h := \overline{\Omega}_h \cap (\Omega \cup \Gamma_1 \cup \Gamma_2)\,. \tag{7.106}$$

Then the error at the grid points and each time level is given by

$$\boldsymbol{e}_h^n := \boldsymbol{U}^n - \boldsymbol{u}_h^n \quad \text{for} \quad n = 0, \ldots, N\,, \tag{7.107}$$

where $\boldsymbol{u}_h^n$ is the solution of the one-step-theta method according to (7.66). The consistency error $\hat{q}_h$ as a grid function on $\Omega_h \times \{t_1, \ldots, t_N\}$ or correspondingly a sequence of vectors $\hat{\boldsymbol{q}}_h^n$ in $\mathbb{R}^{M_1}$ for $n = 1, \ldots, N$ is then defined by

$$\begin{aligned} \hat{\boldsymbol{q}}_h^{n+1} \;:=\; & \frac{1}{\tau}\left(\boldsymbol{U}^{n+1} - \boldsymbol{U}^n\right) + \Theta A_h \boldsymbol{U}^{n+1} \\ & + \overline{\Theta} A_h \boldsymbol{U}^n - \boldsymbol{q}_h((n+\Theta)\tau) \end{aligned} \tag{7.108}$$

for $n = 0, \ldots, N-1$. Then the error grid function obviously satisfies

$$\begin{aligned} \frac{1}{\tau}\left(\boldsymbol{e}_h^{n+1} - \boldsymbol{e}_h^n\right) + \Theta A_h \boldsymbol{e}_h^{n+1} + \overline{\Theta} A_h \boldsymbol{e}_h^n &= \hat{\boldsymbol{q}}_h^{n+1} \quad \text{for} \;\; n = 0, \ldots, N-1\,, \\ \boldsymbol{e}_h^0 &= \boldsymbol{0} \end{aligned} \tag{7.109}$$

(or nonvanishing initial data if the initial condition is not evaluated exactly at the grid points). In the following we estimate the grid function $\hat{q}_h$ in the discrete maximum norm

$$\begin{aligned} \|\hat{q}_h\|_\infty \;&:=\; \max\{|(\hat{\boldsymbol{q}}_h^n)_r| \mid r \in \{1, \ldots, M_1\}, n \in \{1, \ldots, N\}\} \\ &=\; \max\{|\hat{\boldsymbol{q}}_h^n|_\infty \mid n \in \{1, \ldots, N\}\}, \end{aligned} \tag{7.110}$$

i.e., pointwise in space and time. An alternative norm would be the discrete L^2-norm, i.e.,

$$\|\hat{q}_h\|_{0,h} := \left(\tau \sum_{n=1}^{N} h^d \sum_{r=1}^{M_1} |(\hat{\boldsymbol{q}}_h^n)_r|^2\right)^{1/2} = \left(\tau \sum_{n=1}^{N} |\hat{\boldsymbol{q}}_h^n|_{0,h}^2\right)^{1/2}, \tag{7.111}$$

using the spatial discrete L^2-norm from (7.59), where the same notation is employed. If for the sequence of underlying grid points considered there is a constant $C > 0$ independent of the discretization parameter h such that

$$M_1 = M_1(h) \le C h^{-d}\,, \tag{7.112}$$

then obviously,

$$\|\hat{q}_h\|_{0,h} \le (CT)^{1/2}\|\hat{q}_h\|_\infty\,,$$

so that the L^2-norm is weaker than the maximum norm. Condition (7.112) is satisfied for such uniform grids, as considered in Section 1.2. A norm in between is defined by

$$\|\hat{q}_h\|_{\infty,0,h} := \max\{|\hat{\boldsymbol{q}}_h^n|_{0,h} \mid n = 1,\dots,N\}\,, \tag{7.113}$$

which is stronger than (7.111) and in the case of (7.112) weaker than the maximum norm.

Analogously to Section 1.4, we denote $\boldsymbol{U}^n$ amended by the eliminated boundary values $\hat{\boldsymbol{U}}_h^n \in \mathbb{R}^{M_2}$ by the vector $\tilde{\boldsymbol{U}}^n \in \mathbb{R}^M$.

For simplicity we restrict attention, at the beginning, to the case of pure Dirichlet data. Taking into account (7.98) and assuming that $\boldsymbol{f}((n-1+\Theta)\tau)$ is derived from the continuous right-hand side by evaluation at the grid points, we get

$$\begin{aligned}\hat{\boldsymbol{q}}_h^{n+1} &= \frac{1}{\tau}(\boldsymbol{U}^{n+1}-\boldsymbol{U}^n) - \left(\frac{d}{dt}\boldsymbol{U}\right)(t_n+\Theta\tau)\\ &\quad + \Theta\tilde{A}_h\tilde{\boldsymbol{U}}^{n+1} + \overline{\Theta}\tilde{A}_h\tilde{\boldsymbol{U}}^n - (L\boldsymbol{U})(t_n+\Theta\tau)\\ &=: \boldsymbol{S}_1 + \boldsymbol{S}_2\,,\end{aligned} \tag{7.114}$$

so that $\boldsymbol{S}_1$, consisting of the first two terms, is the consistency error for the time discretization.

Here $\frac{d}{dt}\boldsymbol{U}$ and $L\boldsymbol{U}$ are the vectors representing the grid functions corresponding to $\frac{d}{dt}u$ and Lu, which requires the continuity of these functions as in the notion of a classical solution. *We make the following assumption:*

The spatial discretization has the order of consistency α measured in $\|\cdot\|_\infty$ (according to (1.17)) if the solution of the stationary problem (7.6) is in $C^p(\overline{\Omega})$ for some $\alpha > 0$ and $p \in \mathbb{N}$.

For example, for the Dirichlet problem of the Poisson equation and the five-point stencil discretization on a rectangle, we have seen in Chapter 1 that $\alpha = 2$ is valid for $p = 4$. If we assume for $u(\cdot,t)$, u being the solution of (7.1), that

$$\begin{aligned}&\text{the spatial derivatives up to order } p \text{ exist continuously}\\ &\text{and are bounded uniformly in } t \in [0,T]\,,\end{aligned} \tag{7.115}$$

then there exists a constant $C > 0$ such that

$$|(\tilde{A}_h\tilde{\boldsymbol{U}}(t))_i - (Lu(\cdot,t))(x_i)| \le Ch^\alpha \tag{7.116}$$

for every grid point $x_i \in \Omega_h$ and $t \in [0,T]$.

In the case of Neumann or mixed boundary conditions, then some of the equations will correspond to discretizations of these boundary conditions.

This discretization may be directly a discretization of (7.3) or (7.4) (typically, if one-sided difference quotients are used) or a linear combination

of the discretizations of the differential operator at $x_i \in \tilde{\Omega}_h$ and of the boundary differential operator of (7.3) or (7.4) (to eliminate "artificial" grid points) (see Section 1.3).

Thus we have to take $x_i \in \tilde{\Omega}_h$ and interpret Lu in (7.116) as this modified differential operator for $x_i \in \Gamma_1 \cup \Gamma_2$ just described to extend all the above reasoning to the general case.

The estimation of the contribution $\boldsymbol{S}_2$ on the basis of (7.116) is directly possible for $\Theta = 0$ or $\Theta = 1$, but requires further smoothness for $\Theta \in (0,1)$.

We have

$$\boldsymbol{S}_2 = \boldsymbol{S}_3 + \boldsymbol{S}_4\,,$$

where

$$\begin{aligned} \boldsymbol{S}_3 &:= \Theta(\tilde{A}_h \tilde{\boldsymbol{U}}^{n+1} - (L\boldsymbol{U})(t_{n+1})) + \overline{\Theta}(\tilde{A}_h \tilde{\boldsymbol{U}}^{n} - (L\boldsymbol{U})(t_n))\,, \\ \boldsymbol{S}_4 &:= \Theta(L\boldsymbol{U})(t_{n+1}) + \overline{\Theta}(L\boldsymbol{U})(t_n) - (L\boldsymbol{U})(t_n + \Theta\tau)\,. \end{aligned}$$

By Taylor expansion we conclude for a function $v \in C^2[0,T]$ that

$$\Theta v(t_{n+1}) + \overline{\Theta} v(t_n) = v(t_n + \Theta\tau) + \tau^2 \left(\overline{\Theta} \frac{\Theta^2}{2}\, v''(t_n^1) + \Theta \frac{\overline{\Theta}^2}{2}\, v''(t_n^2) \right)$$

for some $t_n^1 \in (t_n, t_n + \Theta\tau)$, $t_n^2 \in (t_n + \Theta\tau,\ t_{n+1})$, so that

$$|\boldsymbol{S}_4|_\infty \le C\tau^2 \tag{7.117}$$

for some constant $C > 0$ independent of τ and h if for $\Theta \in (0,1)$ the solution u of (7.1) satisfies

$$\frac{\partial}{\partial t}\, Lu\,, \quad \frac{\partial^2}{\partial t^2}\, Lu \in C(\overline{Q}_T)\,. \tag{7.118}$$

This is a quite severe regularity assumption, which often does not hold.

For $\boldsymbol{S}_3$ we conclude directly from (7.116) that

$$|\boldsymbol{S}_3|_\infty \le Ch^\alpha\,. \tag{7.119}$$

To estimate $\boldsymbol{S}_1$ we have to distinguish between $\Theta = \frac{1}{2}$ and $\Theta \ne \frac{1}{2}$: If

$$\frac{\partial}{\partial t} u,\ \frac{\partial^2}{\partial t^2} u \in C(\overline{Q}_T) \ \text{ and for } \ \Theta = \frac{1}{2} \ \text{ also } \ \frac{\partial^3}{\partial t^3} u \in C(\overline{Q}_T), \tag{7.120}$$

then Lemma 1.2 implies (for $\Theta = 0, 1, \frac{1}{2}$, for $\Theta \in (0,1)$ again with a Taylor expansion)

$$|\boldsymbol{S}_1|_\infty \le C\tau^\beta \tag{7.121}$$

for some constant C, independent of τ and h, with $\beta = 1$ for $\Theta \ne \frac{1}{2}$ and $\beta = 2$ for $\Theta = \frac{1}{2}$.

Thus, under the additional regularity assumptions (7.115), (7.118), (7.120), and if the spatial discretization has order of consistency α in

the maximum norm, i.e., (7.116), then the one-step-theta method has the following order of consistency:

$$\|\hat{q}_h\|_\infty \le C(h^\alpha + \tau^\beta) \tag{7.122}$$

for some constant C, independent of τ and h, with β as in (7.121).

By using a weaker norm one might hope to achieve a higher order of convergence. If this is, for example, the case for the spatial discretization, e.g., by considering the discrete L^2-norm $\|\cdot\|_{0,h}$ instead of $\|\cdot\|_\infty$, then instead of (7.116) we have

$$\|\tilde{A}_h \tilde{\boldsymbol{U}}(t) - Lu(\cdot,t)\|_{0,h} \le\ Ch^\alpha\,, \tag{7.123}$$

where the terms in the norm denote the corresponding grid functions.

Then again under (weaker forms of) the additional regularity assumptions (7.115), (7.118), (7.120) and assuming (7.112), we have

$$\|\hat{q}_h\|_{0,h} \le C(h^\alpha + \tau^\beta)\,. \tag{7.124}$$

By means of Theorem 7.26 we can conclude the first order of convergence result:

Theorem 7.29 *Consider the one-step-theta method and assume that the spatial discretization matrix A_h has a basis of eigenvectors $\boldsymbol{w}_i$ with eigenvalues $\lambda_i \ge 0$, $i = 1, \dots, M_1$, orthogonal with respect to the scalar product $\langle\cdot,\cdot\rangle_h$, defined in* (7.58)*. The spatial discretization has order of consistency α in $\|\cdot\|_{0,h}$ for solutions in $C^p(\overline{\Omega})$. If τ is such that the method is stable according to* (7.95)*, then for a sufficiently smooth solution u of* (7.1) *(e.g.,* (7.115)*,* (7.118)*,* (7.120)*), and for a sequence of grid points satisfying* (7.112)*, the method converges in the norm $\|\cdot\|_{\infty,0,h}$ with the order*

$$O(h^\alpha + \tau^\beta)\,,$$

where $\beta = 2$ for $\Theta = \frac{1}{2}$ and $\beta = 1$ otherwise.

Proof: Due to Theorem 7.26 and (7.109) we have to estimate the consistency error in a norm defined by $\tau \sum_{n=1}^{N} |\hat{\boldsymbol{q}}_h^n|_{0,h}$ (i.e., a discrete L^1-L^2-norm), which is weaker than $\|\hat{q}_h\|_{0,h}$, in which the estimate has been verified in (7.124). □

Again we see here a smoothing effect in time: The consistency error has to be controlled only in a discrete L^1-sense to gain a convergence result in a discrete L^∞-sense.

If a consistency estimate is provided in $\|\cdot\|_\infty$ as in (7.122), a convergence estimate still needs the corresponding stability. Instead of constructing a vector as in Theorem 1.14 for the formulation (7.97), we will argue directly with the help of the comparison principle (Theorem 7.28, 1)), which would have been possible also in Section 1.4 (see Exercise 1.14).

Theorem 7.30 *Consider the one-step-theta method and assume that the spatial discretization matrix A_h satisfies* (1.32) (1), (2), (3) (i) *and assume its L^∞-stability by the existence of vectors $\boldsymbol{w}_h \in \mathbb{R}^{M_1}$ and a constant $C > 0$ independent of h such that*

$$A_h \boldsymbol{w}_h \geq \mathbf{1} \quad \textit{and} \quad |\boldsymbol{w}_h|_\infty \leq C\,. \tag{7.125}$$

The spatial discretization has order of consistency α in $\|\cdot\|_\infty$ for solutions in $C^p(\overline{\Omega})$. If (7.100) *is satisfied, then for a sufficiently smooth solution u of* (7.1) *(e.g.,* (7.115), (7.118), (7.120)*) the method converges in the norm $\|\cdot\|_\infty$ with the order*

$$O(h^\alpha + \tau^\beta)\,,$$

where $\beta = 2$ for $\Theta = \frac{1}{2}$ and $\beta = 1$ otherwise.

Proof: From (7.122) we conclude that

$$-\hat{C}(h^\alpha + \tau^\beta)\mathbf{1} \leq \hat{\boldsymbol{q}}_h^n \leq \hat{C}(h^\alpha + \tau^\beta)\mathbf{1} \quad \text{for} \quad n = 1, \ldots, N$$

for some constant $\hat{C}$ independent of h and τ.

Thus (7.109) implies

$$\begin{aligned} \frac{1}{\tau}\left(\boldsymbol{e}_h^{n+1} - \boldsymbol{e}_h^n\right) + \Theta A_h \boldsymbol{e}_h^{n+1} + \overline{\Theta} A_h \boldsymbol{e}_h^n &\leq \hat{C}(h^\alpha + \tau^\beta)\mathbf{1}\,, \\ \boldsymbol{e}_h^0 &= 0\,. \end{aligned}$$

Setting $\boldsymbol{w}_h^n := \hat{C}(h^\alpha + \tau^\beta)\boldsymbol{w}_h$ with $\boldsymbol{w}_h$ from (7.125), this constant sequence of vectors satisfies

$$\frac{1}{\tau}\left(\boldsymbol{w}_h^{n+1} - \boldsymbol{w}_h^n\right) + \Theta A_h \boldsymbol{w}_h^{n+1} + \overline{\Theta} A_h \boldsymbol{w}_h^n \geq \hat{C}(h^\alpha + \tau^\beta)\mathbf{1}\,.$$

Therefore, the comparison principle (Theorem 7.28, (1)) implies

$$\boldsymbol{e}_h^n \leq \boldsymbol{w}_h^n = \hat{C}(h^\alpha + \tau^\beta)\boldsymbol{w}_h$$

for $n = 0, \ldots, N$, and analogously, we see that

$$-\hat{C}(h^\alpha + \tau^\beta)\boldsymbol{w}_h \leq \boldsymbol{e}_h^n\,,$$

so that

$$\left|(\boldsymbol{e}_h^n)_j\right| \leq \hat{C}(h^\alpha + \tau^\beta)(\boldsymbol{w}_h)_j \tag{7.126}$$

for all $n = 0, \ldots, N$ and $j = 1, \ldots, M_1$, and finally,

$$|\boldsymbol{e}_h^n|_\infty \leq \hat{C}(h^\alpha + \tau^\beta)|\boldsymbol{w}_h|_\infty \leq \hat{C}(h^\alpha + \tau^\beta)C$$

with the constant C from (7.125). □

Note that the pointwise estimate (7.126) is more precise, since it also takes into account the shape of $\boldsymbol{w}_h$. In the example of the five-point stencil with Dirichlet conditions on the rectangle (see the discussion around (1.43))

the error bound is smaller in the vicinity of the boundary (which is to be expected due to the exactly fulfilled boundary conditions).

Order of Convergence Estimates for the Finite Element Method We now return to the one-step-theta method for the finite element method as introduced in (7.72). In particular, instead of considering grid functions as for the finite difference method, the finite element method allows us to consider directly a function U^n from the finite-dimensional approximation space V_h and thus from the underlying function space V.

In the following, an error analysis for the case $\Theta \in [\frac{1}{2}, 1]$ under the assumption $u \in C^2([0,T], V)$ will be given. In analogy with the decomposition of the error in the semidiscrete situation, we write

$$u(t_n) - U^n = u(t_n) - R_h u(t_n) + R_h u(t_n) - U^n =: \varrho(t_n) + \theta^n \,.$$

The first term of the right-hand side is the error of the elliptic projection at the time t_n, and for this term an estimate is already known. The following identity is used to estimate the second member of the right-hand side, which immediately results from the definition of the elliptic projection:

$$\begin{aligned}
&\left\langle \frac{1}{\tau}(\theta^{n+1} - \theta^n), v_h \right\rangle_0 + a(\Theta\theta^{n+1} + \overline{\Theta}\theta^n, v_h) \\
&= \left\langle \frac{1}{\tau}((R_h u(t_{n+1}) - R_h u(t_n)), v_h \right\rangle_0 + a(\Theta R_h u(t_{n+1}) + \overline{\Theta} R_h u(t_n), v_h) \\
&\quad - \left\langle \frac{1}{\tau}(U^{n+1} - U^n), v_h \right\rangle_0 - a(\Theta U^{n+1} + \overline{\Theta} U^n, v_h) \\
&= \left\langle \frac{1}{\tau}(R_h u(t_{n+1}) - R_h u(t_n)), v_h \right\rangle_0 + a(\Theta u(t_{n+1}) + \overline{\Theta} u(t_n), v_h) \\
&\quad - b^{n+\Theta}(v_h) \\
&= \left\langle \frac{1}{\tau}(R_h u(t_{n+1}) - R_h u(t_n)), v_h \right\rangle_0 - \langle \Theta u'(t_{n+1}) + \overline{\Theta} u'(t_n), v_h \rangle_0 \\
&= \langle w^n, v_h \rangle_0 \,,
\end{aligned}$$

where

$$w^n := \frac{1}{\tau}(R_h u(t_{n+1}) - R_h u(t_n)) - \Theta u'(t_{n+1}) - \overline{\Theta} u'(t_n) \,.$$

Taking into consideration the inequality $a(v_h, v_h) \geq 0$, the particular choice of the test function as $v_h = \Theta\theta^{n+1} + \overline{\Theta}\theta^n$ yields

$$\Theta \|\theta^{n+1}\|_0^2 + (1 - 2\Theta)\langle \theta^n, \theta^{n+1} \rangle_0 - \overline{\Theta} \|\theta^n\|_0^2 \leq \tau \langle w^n, \Theta\theta^{n+1} + \overline{\Theta}\theta^n \rangle_0 .$$

For $\Theta \in [\frac{1}{2}, 1]$ we have $(1 - 2\Theta) \leq 0$, and hence

$$\begin{aligned}
\left[\|\theta^{n+1}\|_0 - \|\theta^n\|_0\right] \left[\Theta\|\theta^{n+1}\|_0 + \overline{\Theta}\|\theta^n\|_0\right] & \\
&= \Theta\|\theta^{n+1}\|_0^2 + (1 - 2\Theta)\|\theta^n\|_0 \|\theta^{n+1}\|_0 - \overline{\Theta}\|\theta^n\|_0^2 \\
&\leq \Theta\|\theta^{n+1}\|_0^2 + (1 - 2\Theta)\langle \theta^n, \theta^{n+1} \rangle_0 - \overline{\Theta}\|\theta^n\|_0^2 \\
&\leq \tau \|w^n\|_0 \left[\Theta\|\theta^{n+1}\|_0 + \overline{\Theta}\|\theta^n\|_0\right] .
\end{aligned}$$

Dividing each side by the expression in the square brackets, we get

$$\|\theta^{n+1}\|_0 \le \|\theta^n\|_0 + \tau\|w^n\|_0 .$$

The recursive application of this inequality leads to

$$\|\theta^{n+1}\|_0 \le \|\theta^0\|_0 + \tau \sum_{j=0}^{n} \|w^j\|_0 . \tag{7.127}$$

That is, it remains to estimate the terms $\|w^j\|_0$. A simple algebraic manipulation yields

$$\begin{aligned} w^n &:= \frac{1}{\tau}((R_h - I)u(t_{n+1}) - (R_h - I)u(t_n)) + \frac{1}{\tau}(u(t_{n+1}) - u(t_n)) \\ &\quad - \Theta u'(t_{n+1}) - \overline{\Theta} u'(t_n) . \end{aligned} \tag{7.128}$$

Taylor expansion with integral remainder implies

$$u(t_{n+1}) = u(t_n) + u'(t_n)\tau + \int_{t_n}^{t_{n+1}} (t_{n+1} - s)u''(s)\, ds$$

and

$$u(t_n) = u(t_{n+1}) - u'(t_{n+1})\tau + \int_{t_{n+1}}^{t_n} (t_n - s)u''(s)\, ds .$$

Using the above relations we get the following useful representations of the difference quotient of u in t_n :

$$\begin{aligned} \frac{1}{\tau}(u(t_{n+1}) - u(t_n)) &= u'(t_n) + \frac{1}{\tau}\int_{t_n}^{t_{n+1}} (t_{n+1} - s)u''(s)\, ds , \\ \frac{1}{\tau}(u(t_{n+1}) - u(t_n)) &= u'(t_{n+1}) + \frac{1}{\tau}\int_{t_n}^{t_{n+1}} (t_n - s)u''(s)\, ds . \end{aligned}$$

Multiplying the first equation by $\overline{\Theta}$ and the second one by Θ, the summation of the results yields

$$\begin{aligned} \frac{1}{\tau}(u(t_{n+1}) - u(t_n)) &= \Theta u'(t_{n+1}) + \overline{\Theta} u'(t_n) \\ &\quad + \frac{1}{\tau}\int_{t_n}^{t_{n+1}} [\Theta t_n + \overline{\Theta} t_{n+1} - s]u''(s)\, ds . \end{aligned}$$

Since $|\Theta t_n + \overline{\Theta} t_{n+1} - s| \le \tau$, the second term in the decomposition (7.128) of w^n can be estimated as

$$\left\| \frac{1}{\tau}(u(t_{n+1}) - u(t_n)) - \Theta u'(t_{n+1}) - \overline{\Theta} u'(t_n) \right\|_0 \le \int_{t_n}^{t_{n+1}} \|u''(s)\|_0\, ds .$$

To estimate the first term in (7.128), Taylor expansion with integral remainder is applied to the function $v(t) := (R_h - I)u(t)$. Then we have

$$\frac{1}{\tau}((R_h - I)u(t_{n+1}) - (R_h - I)u(t_n)) = \frac{1}{\tau}\int_{t_n}^{t_{n+1}} [(R_h - I)u(s)]'\, ds .$$

With the assumption on u using the fact that the derivative and the elliptic projection commute, we get

$$\left\| \frac{1}{\tau}((R_h - I)u(t_{n+1}) - (R_h - I)u(t_n)) \right\|_0 \leq \frac{1}{\tau} \int_{t_n}^{t_{n+1}} \|(R_h - I)u'(s)\|_0 \, ds \,.$$

With (7.127) and summing the estimates for $\|w^n\|_0$ we obtain the following result:

Theorem 7.31 *Let a be a V-elliptic, continuous bilinear form, $u_{0h} \in V_h$, $u_0 \in V$, $\Theta \in [\frac{1}{2}, 1]$. If $u \in C^2([0,T], V)$, then*

$$\begin{aligned} \|u(t_n) - U^n\|_0 \quad &\leq \quad \|u_{0h} - R_h u_0\|_0 + \|(I - R_h)u(t_n)\|_0 \\ &\quad + \int_0^{t_n} \|(I - R_h)u'(s)\|_0 \, ds + \tau \int_0^{t_n} \|u''(s)\|_0 \, ds \,. \end{aligned}$$

Remark 7.32 (i) Under stronger smoothness assumptions on u and by detailed considerations it can also be shown that the Crank–Nicolson method ($\Theta = \frac{1}{2}$) is of order 2 in τ.

(ii) Contrary to the semidiscrete situation (Theorem 7.12), the fully discrete estimate does not reflect any exponential decay in time.

Utilizing the error estimate for the elliptic projection as in Section 7.2 (cf. Theorem 7.12) and assuming $u_0 \in V \cap H^2(\Omega)$, we have

$$\begin{aligned} \|u(t_n) - U^n\|_0 \leq \|u_{0h} - u_0\|_0 + Ch^2 \left[\|u_0\|_2 + \|u(t_n)\|_2 + \int_0^{t_n} \|u'(s)\|_2 \, ds \right] \\ + \tau \int_0^{t_n} \|u''(s)\|_0 \, ds \,. \end{aligned}$$

If, in addition, $\|u_{0h} - u_0\|_0 \leq Ch^2 \|u_0\|_2$, we obtain

$$\|u(t_n) - U^n\|_0 \leq C(u)(h^2 + \tau) \,,$$

with $C(u) > 0$ depending on the solution u (and thus on u_0) but not depending on h and τ.

To conclude this section we give without proof a summary of error estimates for all possible values of Θ:

$$\|u(t_n) - U^n\|_0 \leq \begin{cases} C(u)(h^2 + \tau) \,, & \text{if } \Theta \in [\frac{1}{2}, 1] \,, \\ C(u)(h^2 + \tau^2) \,, & \text{if } \Theta = \frac{1}{2} \,, \\ C(u)h^2 \,, & \text{if } \Theta \in [0, 1] \text{ and } \tau \leq \vartheta h^2 \,, \end{cases} \tag{7.129}$$

where $\vartheta > 0$ is a constant upper bound of the step size relation τ / h^2.

The occurrence of such a restriction is not surprising, since similar requirements have already appeared for the finite difference method.

We also mention that the above restriction to a constant step size τ is only for simplicity of the notation. If a variable step size τ_{n+1} is used (which is typically determined by a step size control strategy), then the number τ in Theorem 7.31 is to be replaced by $\max_{n=0,\ldots,N-1} \tau_n$.

Order of Convergence Estimates for the Finite Volume Method
We now consider the one-step-theta method for the finite volume method as introduced in (7.75).

The error analysis will run in a similar way as for the finite element method.

We write

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

where R_h is the auxiliary operator defined in (7.63). So for the first term of the right-hand side, an estimate is already known.

From the definition (7.63) and (7.32), we immediately derive the following identity:

$$\begin{aligned}
&\left\langle \frac{1}{\tau}(\theta^{n+1} - \theta^n), v_h \right\rangle_{0,h} + a_h(\Theta\theta^{n+1} + \overline{\Theta}\theta^n, v_h) \\
&= \left\langle \frac{1}{\tau}(R_h u(t_{n+1}) - R_h u(t_n)), v_h \right\rangle_{0,h} + a_h(\Theta R_h u(t_{n+1}) + \overline{\Theta} R_h u(t_n), v_h) \\
&\quad - \left\langle \frac{1}{\tau}(U^{n+1} - U^n), v_h \right\rangle_{0,h} - a_h(\Theta U^{n+1} + \overline{\Theta} U^n, v_h) \\
&= \left\langle \frac{1}{\tau}(R_h u(t_{n+1}) - R_h u(t_n)), v_h \right\rangle_{0,h} + a(\Theta u(t_{n+1}) + \overline{\Theta} u(t_n), v_h) \\
&\quad - \langle f^{n+\Theta}, v_h \rangle_{0,h} \\
&= \left\langle \frac{1}{\tau}(R_h u(t_{n+1}) - R_h u(t_n)), v_h \right\rangle_{0,h} - \langle \Theta u'(t_{n+1}) + \overline{\Theta} u'(t_n), v_h \rangle_0 \\
&\quad + \langle f^{n+\Theta}, v_h \rangle_0 - \langle f^{n+\Theta}, v_h \rangle_{0,h} \\
&= \left\langle \frac{1}{\tau}(R_h u(t_{n+1}) - R_h u(t_n)), v_h \right\rangle_{0,h} - \langle \Theta u'(t_{n+1}) + \overline{\Theta} u'(t_n), v_h \rangle_{0,h} \\
&\quad + \langle \Theta u'(t_{n+1}) + \overline{\Theta} u'(t_n), v_h \rangle_{0,h} - \langle \Theta u'(t_{n+1}) + \overline{\Theta} u'(t_n), v_h \rangle_0 \\
&\quad + \langle f^{n+\Theta}, v_h \rangle_0 - \langle f^{n+\Theta}, v_h \rangle_{0,h} \\
&= \langle w^n, v_h \rangle_{0,h} + r^n(v_h) \,,
\end{aligned}$$

where

$$w^n := \frac{1}{\tau}(R_h u(t_{n+1}) - R_h u(t_n)) - \Theta u'(t_{n+1}) - \overline{\Theta} u'(t_n)$$

and

$$\begin{aligned}
r^n(v_h) \quad := \quad & \langle \Theta u'(t_{n+1}) + \overline{\Theta} u'(t_n), v_h \rangle_{0,h} - \langle \Theta u'(t_{n+1}) + \overline{\Theta} u'(t_n), v_h \rangle_0 \\
& + \langle f^{n+\Theta}, v_h \rangle_0 - \langle f^{n+\Theta}, v_h \rangle_{0,h} \,.
\end{aligned}$$

Under the assumptions of Theorem 6.15, we know that $a_h(v_h, v_h) \geq 0$ for all $v_h \in V_h$. The particular choice of the test function as $v_h = v_h^\Theta := \Theta\theta^{n+1} + \overline{\Theta}\theta^n$ yields, similarly to the finite element case, for $\Theta \in [\frac{1}{2}, 1]$ the

estimate

$$\begin{aligned}
&\left[\|\theta^{n+1}\|_{0,h} - \|\theta^n\|_{0,h}\right]\left[\Theta\|\theta^{n+1}\|_{0,h} + \overline{\Theta}\|\theta^n\|_{0,h}\right] \\
&\quad \le \tau\left(\langle w^n, v_h^\Theta\rangle_{0,h} + r^n(v_h^\Theta)\right) \\
&\quad \le \tau\left(\|w^n\|_{0,h} + \sup_{v_h\in V_h}\frac{r^n(v_h)}{\|v_h\|_{0,h}}\right)\|v_h^\Theta\|_{0,h} \\
&\quad \le \tau\left(\|w^n\|_{0,h} + \sup_{v_h\in V_h}\frac{r^n(v_h)}{\|v_h\|_{0,h}}\right)\left[\Theta\|\theta^{n+1}\|_{0,h} + \overline{\Theta}\|\theta^n\|_{0,h}\right].
\end{aligned}$$

Dividing each side by the expression in the square brackets, we get

$$\|\theta^{n+1}\|_{0,h} \le \|\theta^n\|_{0,h} + \tau\left(\|w^n\|_{0,h} + \sup_{v_h\in V_h}\frac{r^n(v_h)}{\|v_h\|_{0,h}}\right).$$

The recursive application of this inequality leads to

$$\|\theta^{n+1}\|_{0,h} \le \|\theta^0\|_{0,h} + \tau\sum_{j=0}^{n}\|w^j\|_{0,h} + \tau\sum_{j=0}^{n}\sup_{v_h\in V_h}\frac{r^j(v_h)}{\|v_h\|_{0,h}}. \tag{7.130}$$

The representation of w^j obtained in the subsection on the finite element method yields the following estimate:

$$\|w^j\|_{0,h} \le \frac{1}{\tau}\int_{t_j}^{t_{j+1}}\|(R_h - I)u'(s)\|_{0,h}\,ds + \int_{t_j}^{t_{j+1}}\|u''(s)\|_{0,h}\,ds.$$

Furthermore, by Lemma 7.14, we have

$$|r^j(v_h)| \le Ch\left[\Theta|u'(t_{j+1})|_{1,\infty} + \overline{\Theta}|u'(t_j)|_{1,\infty} + |f^{j+\Theta}|_{1,\infty}\right]\|v_h\|_{0,h}.$$

Using both estimates in (7.130), we obtain

$$\begin{aligned}
&\|\theta^{n+1}\|_{0,h} \\
&\le \|\theta^0\|_{0,h} + C\left[\int_0^{t_{n+1}}\|(R_h - I)u'(s)\|_{0,h}\,ds + \tau\int_0^{t_{n+1}}\|u''(s)\|_{0,h}\,ds\right] \\
&\quad + Ch\tau\Big[\overline{\Theta}|u'(0)|_{1,\infty} + \sum_{j=1}^{n}|u'(t_j)|_{1,\infty} + \Theta|u'(t_{n+1})|_{1,\infty} \\
&\qquad\qquad + \sum_{j=0}^{n}|f^{j+\Theta}|_{1,\infty}\Big] \\
&\le \|\theta^0\|_{0,h} + C\left[\int_0^{t_{n+1}}\|(R_h - I)u'(s)\|_{0,h}\,ds + \tau\int_0^{t_{n+1}}\|u''(s)\|_{0,h}\,ds\right] \\
&\quad + Ch\Big[\sup_{s\in(0,t_{n+1})}|u'(s)|_{1,\infty} + \sup_{s\in(0,t_{n+1})}|f(s)|_{1,\infty}\Big].
\end{aligned}$$

This is the basic estimate. The final estimate is easily obtained by the same approach as in the finite element method. In summary, we have the following result.

Theorem 7.33 *In addition to the assumptions of Theorem* 6.15*, consider the finite volume method on Donald diagrams. Furthermore, let* $u_{0h} \in V_h$,

$u_0 \in V \cap H^2(\Omega)$, $f \in C([0,T], C^1(\overline{\Omega}))$, $\Theta \in [\frac{1}{2}, 1]$. *Then if* $u(t)$ *is sufficiently smooth, the following estimate is valid:*

$$\begin{aligned}\|u(t_n) - U^n\|_{0,h} \quad \leq \quad & \|u_{0h} - u_0\|_{0,h} + Ch\Big[\|u_0\|_2 + \|u(t_n)\|_2 \\ & + \int_0^{t_n} \|u'(s)\|_2 \, ds + \sup_{s\in(0,t_n)} |u'(s)|_{1,\infty} \\ & + \sup_{s\in(0,t_n)} |f(s)|_{1,\infty}\Big] + C\tau \int_0^{t_n} \|u''(s)\|_{0,h} \, ds \,.\end{aligned}$$

Exercise

7.17 Verify Remark 7.32.

8

Iterative Methods for Nonlinear Equations

In the same way as linear (initial-) boundary value problems by the discretization techniques discussed in this book lead to (sequences of) linear equations, we get nonlinear equations of similar type from nonlinear problems. Two of them will be treated in this chapter. As in the Sections 1.2, 3.4, 7.3, and 6.2.4, we have to answer the question of the quality of the approximation, and as in Section 2.5 and Chapter 5, the question of the approximative resolution of the systems of equations. We will focus on the latter in this chapter.

In general, the problem may be formulated in different equivalent settings, namely:

$$\text{Find} \quad x \in U \quad \text{with} \quad f(x) = b\,. \tag{8.1}$$

$$\text{Find} \quad x \in U \quad \text{with} \quad f(x) = 0\,. \tag{8.2}$$

Then x is called a *root* of (8.2) and a *zero* of f.

$$\text{Find} \quad x \in U \quad \text{with} \quad f(x) = x\,. \tag{8.3}$$

Then x is called a *fixed point*.

Here $U \subset \mathbb{R}^m$, $f : U \to \mathbb{R}^m$ is a mapping, and $b \in \mathbb{R}^m$. The transition from one formulation to another follows by redefining f in evident ways.

In most cases, a root or a fixed point cannot be calculated (with exact arithmetic) in a finite number of operations, but only by an *iterative method*, i.e., by a mapping

$$\Phi : U \to U\,,$$

so that (as in (5.7)) for the sequence

$$x^{(k+1)} := \Phi\big(x^{(k)}\big) \tag{8.4}$$

with given $x^{(0)}$ we get

$$x^{(k)} \to x \quad \text{for} \quad k \to \infty\,. \tag{8.5}$$

Here x is the solution of (8.1), (8.2), or (8.3).

As we already stated in Section 5.1, in the case of a continuous Φ it follows from (8.4), (8.5) that the limit x satisfies

$$x = \Phi(x)\,. \tag{8.6}$$

This means that (8.6) should imply that x is a solution of (8.1), (8.2), or (8.3). The extension of the definition of consistency in Section 5.1 requires the inverse implication.

Concerning the error level that we should achieve in relation to the approximation error of the discretization, the statements in the introduction of Chapter 5 still hold. In addition to the criteria of comparison for linear stationary methods we now have to take into account the following: Methods may, if they do at all, converge only locally, which leads to the following definition:

Definition 8.1 If in the above situation (8.5) holds for all $x^{(0)} \in U$ (i.e., for arbitrary starting values), then $(x^{(k)})_k$ is called *globally convergent*. If an open $\tilde{U} \subset U$ exists such that (8.5) holds for $x^{(0)} \in \tilde{U}$, then $(x^{(k)})_k$ is called *locally convergent*. In the latter case $\tilde{U}$ is called the *range* of the iteration.

On the other hand, we may observe a faster convergence than the linear convergence introduced in (5.3):

Definition 8.2 Let $(x^{(k)})_k$ be a sequence in $\mathbb{R}^m$, $x \in \mathbb{R}^m$, and $\|\cdot\|$ a norm on $\mathbb{R}^m$. The sequence $(x^{(k)})_k$ *converges linearly* to x with respect to $\|\cdot\|$ if there exists a C with $0 < C < 1$ such that

$$\big\|x^{(k+1)} - x\big\| \le C\big\|x^{(k)} - x\big\| \quad \text{for all } k \in \mathbb{N}\,.$$

The sequence $(x^{(k)})_k$ *converges with order of convergence* $p > 1$ to x if $x^{(k)} \to x$ for $k \to \infty$ and if there exists a $C > 0$ such that

$$\big\|x^{(k+1)} - x\big\| \le C\big\|x^{(k)} - x\big\|^p \quad \text{for all } k \in \mathbb{N}\,.$$

The sequence $(x^{(k)})_k$ *converges superlinearly* to x if

$$\lim_{k\to\infty} \frac{\|x^{(k+1)} - x\|}{\|x^{(k)} - x\|} = 0\,.$$

The case $p = 2$ is also called *quadratic convergence*. Thus, while a linearly converging method guarantees a reduction of the error by a constant factor C, this reduction is improved step by step in the case of superlinear or

higher-order convergence. When we encounter quadratic convergence, for example, the number of significant digits is doubled in every step (minus a fixed number), so that usually only a small number of iterations will be necessary. For this reason variants of the quadratically converging *Newton method* (Section 8.2) are attractive. But the restriction of local convergence may require modifications to enlarge the range of convergence.

To evaluate the complexity of a numerical method the number of elementary operations for an iteration has to be considered. By an elementary operation we want also to understand the evaluation of functions like the sine, although this is much more costly than an ordinary floating-point operation. A typical subproblem during an iteration cycle is the solution of a system of linear equations, analogously to the simpler systems in the form (5.10) occurring in linear stationary problems. Besides the effort to *assemble* this system of equations, we have to account for the work to solve it, which can be done with one of the methods described in Section 2.5 and Chapter 5, i.e., in particular, again with an iterative method. We call this a *secondary* or *inner iteration*, which is attractive because of the sparse structure of the matrices originating from the discretization, as already discussed in Chapter 5. Here an *inexact* variant may be useful, with which the inner iteration is performed only up to a precision that conserves the convergence properties of the *outer iteration*. The numerical cost for the assembling may, in fact, be more expensive than the cost for the inner iteration. Hence methods with low cost for the assembling (but worse convergence) should also be considered. Keeping this in mind, we devote an introductory chapter to the *fixed-point iterations*, which are, roughly speaking, methods in which the iteration Φ coincides with the mapping f.

8.1 Fixed-Point Iterations

For the fixed-point formulation (8.3) the choice $\Phi := f$ is evident according to (8.6); in other words, the *fixed-point iteration* reads

$$x^{(k+1)} := f\big(x^{(k)}\big)\,. \tag{8.7}$$

To diminish the distance of two succeeding members of the sequence, i.e.,

$$\big\|\Phi(x^{(k+1)}) - \Phi(x^{(k)})\big\| = \big\|x^{(k+2)} - x^{(k+1)}\big\| < \big\|x^{(k+1)} - x^{(k)}\big\|\,,$$

it is sufficient that the iteration function (here $\Phi = f$) be contractive (see Appendix A.4).

Sufficient conditions for a contraction are given by the following lemma:

Lemma 8.3 *Let $U \subset \mathbb{R}^m$ be open and convex, and $g : U \to \mathbb{R}^m$ continuously differentiable. If*

$$\sup_{x\in U} \|Dg(x)\| =: L < 1$$

holds, where $\|\cdot\|$ *in* $\mathbb{R}^{m,m}$ *is compatible with* $\|\cdot\|$ *in* $\mathbb{R}^m$*, then* g *is contracting in* U*.*

Proof: Exercise 8.1. □

Therefore, if $U \subset \mathbb{R}^m$ is open, $f : U \subset \mathbb{R}^m \to \mathbb{R}^m$ is continuously differentiable, and if there exists some $\tilde{x} \in U$ with $\|Df(\tilde{x})\| < 1$, then there exists a closed convex neighbourhood $\tilde{U}$ of $\tilde{x}$ with

$$\|Df(x)\| \le L < 1 \quad \text{for } x \in \tilde{U}$$

and, for example, $L = \|Df(\tilde{x})\| + \frac{1}{2}(1 - \|Df(\tilde{x})\|)$, guaranteeing the contractivity of f in U.

The unique existence of a fixed point and the convergence of (8.7) is guaranteed if the set U where f is a contraction is mapped into itself:

Theorem 8.4 (Banach's fixed-point theorem) *Let* $U \subset \mathbb{R}^m$*,* $U \neq \emptyset$*, and* U *be closed. Let* $f : U \to \mathbb{R}^m$ *be contractive with Lipschitz constant* $L < 1$ *and* $f[U] \subset U$*. Then we have:*

(1) There exists one and only one fixed point $x \in U$ *of* f*.*

(2) For arbitrary $x^{(0)} \in U$ *the fixed point iteration* (8.7) *converges to* x*, and we have*

$$\begin{aligned} \left\|x^{(k)} - x\right\| &\le \frac{L}{1-L}\left\|x^{(k)} - x^{(k-1)}\right\| \\ &\quad \text{(a posteriori error estimate)} \\ &\le \frac{L^k}{1-L}\left\|x^{(1)} - x^{(0)}\right\| \\ &\quad \text{(a priori error estimate).} \end{aligned}$$

Proof: The sequence $x^{(k+1)} := f(x^{(k)})$ is well-defined because of $f[U] \subset U$. We prove that $(x^{(k)})_k$ is a Cauchy sequence (see Appendix A.4).

$$\begin{aligned} &\left\|x^{(k+1)} - x^{(k)}\right\| = \left\|f(x^{(k)}) - f(x^{(k-1)})\right\| \le L\left\|x^{(k)} - x^{(k-1)}\right\| \\ &\quad \le L^2\left\|x^{(k-1)} - x^{(k-2)}\right\| \le \cdots \le L^k\left\|x^{(1)} - x^{(0)}\right\|, \end{aligned} \tag{8.8}$$

so that for any $k, l \in \mathbb{N}$

$$\begin{aligned} &\left\|x^{(k+l)} - x^{(k)}\right\| \\ &\le \left\|x^{(k+l)} - x^{(k+l-1)}\right\| + \left\|x^{(k+l-1)} - x^{(k+l-2)}\right\| + \cdots + \left\|x^{(k+1)} - x^{(k)}\right\| \\ &\le (L^{k+l-1} + L^{k+l-2} + \cdots + L^k)\left\|x^{(1)} - x^{(0)}\right\| \\ &= L^k(1 + L + \cdots + L^{l-1})\left\|x^{(1)} - x^{(0)}\right\| \\ &\le L^k \sum_{l=0}^{\infty} L^l \|x^{(1)} - x^{(0)}\| = L^k \frac{1}{1-L}\left\|x^{(1)} - x^{(0)}\right\|. \end{aligned}$$

Thus we have $\left\|x^{(k+l)} - x^{(k)}\right\| \to 0$ for $k \to \infty$; i.e., $(x^{(k)})_k$ is a Cauchy sequence and thus converges to some $x \in \mathbb{R}^m$ because of the completeness of $\mathbb{R}^m$. Due to the closedness of U we conclude that $x \in U$. Since we have

$$x^{(k+1)} \to x\,, \quad f\left(x^{(k)}\right) \to f(x) \quad \text{for } k \to \infty\,,$$

x is also a fixed point of f.

The fixed point is unique, because for fixed points $x, \bar{x}$,

$$\|x - \bar{x}\| = \|f(x) - f(\bar{x})\| \leq L\|x - \bar{x}\|,$$

which immediately implies $x = \bar{x}$ because of $L < 1$. Moreover, we have

$$\begin{aligned} \left\|x^{(k)} - x\right\| &= \left\|f(x^{(k-1)}) - f(x)\right\| \leq L\|x^{(k-1)} - x\| \\ &\leq L\left(\left\|x^{(k-1)} - x^{(k)}\right\| + \left\|x^{(k)} - x\right\|\right), \end{aligned}$$

and thus from (8.8),

$$\left\|x^{(k)} - x\right\| \leq \frac{L}{1-L}\left\|x^{(k)} - x^{(k-1)}\right\| \leq \frac{L}{1-L}L^{k-1}\left\|x^{(1)} - x^{(0)}\right\|.$$

□

Remark 8.5 The theorem can be generalized: Since we used only the completeness of $\mathbb{R}^m$, the proposition holds even in a Banach space $(X, \|\cdot\|)$, where $U \subset X$ is a closed subset.

This enables us to define iterative schemes directly in the function space for nonlinear boundary value problems, which means that the resulting (linear) problems in the iteration step are to be discretized. So instead of proceeding in the order *discretization–iteration*, we can apply the sequence *iteration–discretization*. This leads in general to different schemes, even if the approaches have been the same. We will always refer to the first strategy.

According to Lemma 8.3 we can often construct a closed U such that f is contractive on U. It remains to verify that $f[U] \subset U$. For this, the following lemma is helpful:

Lemma 8.6 *Let $U \subset \mathbb{R}^m$, $f : U \to \mathbb{R}^m$. If there exists a $y \in U$ and a $r > 0$ with*

$$\overline{B}_r(y) \subset U\,,$$

with f contractive on $\overline{B}_r(y)$ with Lipschitz constant $L < 1$, so that

$$\|y - f(y)\| \leq r(1 - L)\,,$$

then f has one and only one fixed point in $\overline{B}_r(y)$, and (8.7) converges.

Proof: Exercise 8.2. □

In the setting of Theorem 8.4 the fixed-point iteration is thus globally convergent in U. In the setting of Lemma 8.6 it is locally convergent in U (globally in $\overline{B}_r(y)$). We see that in the situation of Theorem 8.4 the sequence $(x^{(k)})$ has, because of

$$\|x^{(k+1)} - x\| = \|f(x^{(k)}) - f(x)\| \leq L\|x^{(k)} - x\|,$$

a linear order of convergence (and in general not better).

A sufficient condition for local convergence of the corresponding order is given by the following theorem:

Theorem 8.7 *Let $U \subset \mathbb{R}^m$ be open, $\Phi : U \to U$ continuous, the sequence $(x^{(k)})$ defined by $x^{(k+1)} := \Phi\big(x^{(k)}\big)$ for a given $x^{(0)} \in U$. If there exists some $\bar{x} \in U$, an open $V \subset U$ with $\bar{x} \in V$, and constants C, p with $p \geq 1$, $C \geq 0$, and $C < 1$ for $p = 1$, such that for all $x \in V$,*

$$\|\Phi(x) - \bar{x}\| \leq C\|x - \bar{x}\|^p$$

holds, then the iteration defined by Φ converges locally to $\bar{x}$ of order at least p, and $\bar{x}$ is a fixed point of Φ.

Proof: Choose $W = B_r(\bar{x}) \subset V$, with $r > 0$ sufficiently small, such that $W \subset V$ and

$$Cr^{p-1} =: L < 1\,.$$

If $x^{(k)} \in W$, then we conclude because of

$$\big\|x^{(k+1)} - \bar{x}\big\| = \big\|\Phi\big(x^{(k)}\big) - \bar{x}\big\| \leq C\big\|x^{(k)} - \bar{x}\big\|^p < Cr^p < r$$

that $x^{(k+1)} \in W$, too. This means that for $x^{(0)} \in W$ we have that $x^{(k)} \in W$ for all $k \in \mathbb{N}$. Furthermore, we have

$$\big\|x^{(k+1)} - \bar{x}\big\| \leq C\big\|x^{(k)} - \bar{x}\big\|^p < C\,r^{p-1}\big\|x^{(k)} - \bar{x}\big\| = L\,\big\|x^{(k)} - \bar{x}\big\|\,,$$

i.e.,

$$x^{(k)} \to \bar{x} \quad \text{for } k \to \infty,$$

and consequently,

$$\bar{x} = \lim_{k\to\infty} x^{(k+1)} = \lim_{k\to\infty} \Phi\big(x^{(k)}\big) = \Phi(\bar{x})\,.$$

□

The special case of a scalar equation shows that we can expect at most linear convergence for $\Phi = f$:

Corollary 8.8 *Let $U \subset \mathbb{R}$ be an open subset, Φ on U p-times continuously differentiable, and $\bar{x} \in U$ a fixed point of Φ.*
If $\Phi'(\bar{x}) \neq 0$, $|\Phi'(\bar{x})| < 1$ for $p = 1$ and $\Phi'(\bar{x}) = \cdots = \Phi^{(p-1)}(\bar{x}) = 0$, $\Phi^{(p)}(\bar{x}) \neq 0$ for $p > 1$, then the iteration defined by Φ is locally convergent to $\bar{x}$ with order of convergence p, but not better.

Proof: Taylor's expansion of Φ at $\bar{x}$ gives, for $x \in U$,

$$\Phi(x) = \Phi(\bar{x}) + \frac{\Phi^{(p)}(\xi)}{p!}(x - \bar{x})^p \quad \text{with } \xi \in (x, \bar{x}),$$

and in the case $p = 1$ we have $|\Phi'(\xi)| < 1$ for sufficiently small $|x - \bar{x}|$. Thus, there exists a neighbourhood V of $\bar{x}$ such that $|\Phi(x) - \bar{x}| \leq C|x - \bar{x}|^p$ for all $x \in V$ and $C < 1$ for $p = 1$. Theorem 8.7 implies order of convergence p. The example $\Phi(x) = Lx^p$ with $L < 1$ for $p = 1$ with the fixed point $x = 0$ shows that no improvement is possible. □

Exercises

8.1 Prove Lemma 8.3 with the help of the mean value theorem.

8.2 Prove Lemma 8.6.

8.2 Newton's Method and Its Variants

8.2.1 The Standard Form of Newton's Method

In the following we want to study the formulation stated in (8.2), i.e., the problem of finding the solutions of

$$f(x) = 0\,.$$

The simplest method of Chapter 5, the Richardson iteration (cf. (5.28)), suggests the direct application of the fixed-point iteration for, e.g., $\Phi(x) := -f(x) + x$. This approach succeeds only if, in the case of a differentiable f, the Jacobian $I - Df(x)$ is small in the sense of Lemma 8.3 close to the solution. Here we denote by $Df(x) = (\partial_j f_i(x))_{ij}$ the Jacobi or functional matrix of f. A relaxation method similar to (5.30) leads to the *damped* variants, which will be treated later.

The method in its *corrector formulation*, analogously to (5.10) with

$$\delta^{(k)} := x^{(k+1)} - x^{(k)}\,,$$

is

$$\delta^{(k)} = -f\big(x^{(k)}\big)\,, \tag{8.9}$$

or in its relaxation formulation with relaxation parameter $\omega > 0$,

$$\delta^{(k)} = -\omega f(x^{(k)})\,.$$

Now we want to introduce another approach to define Φ:

Let $x^{(0)}$ be an approximation of a zero. An improved approximation is probably given by the following:

- Replace f by a simple function g that approximates f near $x^{(0)}$ and whose zero is to be determined.
- Find $x^{(1)}$ as the solution of $g(x) = 0$.

Newton's method needs the differentiability of f, and one chooses the approximating affine-linear function given by $Df(x^{(0)})$, i.e.,

$$g(x) = f\big(x^{(0)}\big) + Df\big(x^{(0)}\big)\big(x - x^{(0)}\big)\,.$$

Under the assumption that $Df(x^{(0)})$ is nonsingular, the new iterate $x^{(1)}$ is determined by solving the system of *linear* equations

$$Df\big(x^{(0)}\big)\big(x^{(1)} - x^{(0)}\big) = -f\big(x^{(0)}\big), \tag{8.10}$$

or formally by

$$x^{(1)} := x^{(0)} - Df\big(x^{(0)}\big)^{-1} f\big(x^{(0)}\big)\,.$$

This suggests the following definition:

$$\Phi(f)(x) = x - Df(x)^{-1} f(x)\,. \tag{8.11}$$

Here Φ is well-defined only if $Df(x)$ is nonsingular. Then $x \in \mathbb{R}^m$ is a zero of f if and only if x is a fixed point of Φ. When executing the iteration, we do not calculate $Df\big(x^{(k)}\big)^{-1}$ but only the system of equations similar to (8.10).

Thus, the kth iteration of *Newton's method* reads as follows: Solve

$$Df\big(x^{(k)}\big)\,\delta^{(k)} = -f\big(x^{(k)}\big) \tag{8.12}$$

and set

$$x^{(k+1)} := x^{(k)} + \delta^{(k)}\,. \tag{8.13}$$

Equation (8.13) has the same form as (5.10) with $W = Df(x^{(k)})$, with the *residual* at $x^{(k)}$

$$d^{(k)} := f\big(x^{(k)}\big)\,.$$

Thus the subproblem of the kth iteration is easier in the sense that it consists of a system of linear equations (with the same structure of dependence as f; see Exercise 8.6). In the same sense the system of equations (5.10) in the case of linear stationary methods is "easier" to solve than the original problem of the same type. Furthermore, W is in general different for different k.

An application of (8.12), (8.13) to $Ax = b$, i.e., $Df(x) = A$ for all $x \in \mathbb{R}^m$ results in (5.10) with $W = A$, a method converging in one step, which just reformulates the original problem:

$$A\big(x - x^{(0)}\big) = -\big(Ax^{(0)} - b\big)\,.$$

The range of the iteration may be very small, as can be shown already by one-dimensional examples. But in this neighbourhood of the solution we have, e.g., for $m = 1$, the following:

Corollary 8.9 *Let $f \in C^3(\mathbb{R})$ and let $\bar{x}$ be a simple zero of f (i.e., $f'(\bar{x}) \neq 0$). Then Newton's method converges locally to $\bar{x}$, of order at least 2.*

Proof: There exists an open neighbourhood V of $\bar{x}$ such that $f'(x) \neq 0$ for all $x \in V$; i.e., Φ is well-defined by (8.11), continuous on V, and $\bar{x}$ is a fixed point of Φ. According to Corollary 8.8 it suffices to show that $\Phi'(\bar{x}) = 0$:

$$\Phi'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = f(x)\frac{f''(x)}{f'(x)^2} = 0 \quad \text{for } x = \bar{x},$$

and Φ'' exists continuously, because $f \in C^3(\mathbb{R})$. □

In the following we want to develop a general local theorem of convergence for Newton's method (according to L.V. Kantorovich). It necessitates only the Lipschitz continuity of Df and ensures the existence of a zero, too. Here we always suppose a fixed norm on $\mathbb{R}^m$ and consider a compatible norm on $\mathbb{R}^{m,m}$. As a prerequisite we need the following lemma:

Lemma 8.10 *Let $C_0 \subset \mathbb{R}^m$ be convex, open, $f : C_0 \to \mathbb{R}^m$ differentiable, and suppose there exists $\gamma > 0$ such that*

$$\|Df(x) - Df(y)\| \leq \gamma \|x - y\| \quad \text{for all } x, y \in C_0\,. \tag{8.14}$$

Then for all $x, y \in C_0$,

$$\|f(x) - f(y) - Df(y)(x - y)\| \leq \frac{\gamma}{2}\|x - y\|^2\,.$$

Proof: Let $\varphi : [0,1] \to \mathbb{R}^m$ be defined by $\varphi(t) := f(y + t(x - y))$, for arbitrary, fixed $x, y \in C_0$. Then φ is differentiable on $[0,1]$ and

$$\varphi'(t) = Df(y + t(x - y))(x - y)\,.$$

Thus for $t \in [0,1]$ we have

$$\begin{aligned} \|\varphi'(t) - \varphi'(0)\| &= \|(Df(y + t(x - y)) - Df(y))\,(x - y)\| \\ &\leq \|Df(y + t(x - y)) - Df(y)\|\|x - y\| \leq \gamma\, t\, \|x - y\|^2\,. \end{aligned}$$

For

$$\Delta := f(x) - f(y) - Df(y)(x - y) = \varphi(1) - \varphi(0) - \varphi'(0) = \int_0^1 (\varphi'(t) - \varphi'(0))\, dt$$

we also get

$$\|\Delta\| \leq \int_0^1 \|\varphi'(t) - \varphi'(0)\|\, dt \leq \gamma \|x - y\|^2 \int_0^1 t\, dt = \frac{1}{2}\gamma\, \|x - y\|^2\,.$$

□

Now we are able to conclude local, quadratic convergence:

Theorem 8.11 *Let $C \subset \mathbb{R}^m$ be convex, open and $f : C \to \mathbb{R}^m$ differentiable.*
For $x^{(0)} \in C$ there exist $\alpha, \beta, \gamma > 0$ such that

$$h := \alpha\,\beta\,\gamma/2 < 1\,,$$
$$r := \alpha/(1-h)\,,$$
$$\bar{B}_r\big(x^{(0)}\big) \subset C\,.$$

Furthermore, we require:

(i) Df is Lipschitz continuous on $C_0 = B_{r+\varepsilon}\big(x^{(0)}\big)$ for some $\varepsilon > 0$ with constant γ in the sense of (8.14).

(ii) For all $x \in B_r\big(x^{(0)}\big)$ there exists $Df(x)^{-1}$ and $\big\|Df(x)^{-1}\big\| \le \beta$.

(iii) $\big\|Df\big(x^{(0)}\big)^{-1} f\big(x^{(0)}\big)\big\| \le \alpha\,.$

Then:

(1) The Newton iteration

$$x^{(k+1)} := x^{(k)} - Df\big(x^{(k)}\big)^{-1} f\big(x^{(k)}\big)$$

is well-defined and

$$x^{(k)} \in B_r\big(x^{(0)}\big) \quad \text{for all } k \in \mathbb{N}\,.$$

(2) $x^{(k)} \to \bar{x}$ for $k \to \infty$ and $f(\bar{x}) = 0$.

(3) $\big\|x^{(k+1)} - \bar{x}\big\| \le \dfrac{\beta\gamma}{2}\big\|x^{(k)} - \bar{x}\big\|^2$ and $\big\|x^{(k)} - \bar{x}\big\| \le \alpha \dfrac{h^{2^k-1}}{1-h^{2k}}$ for $k \in \mathbb{N}\,.$

Proof: (1): To show that $x^{(k+1)}$ is well-defined it is sufficient to verify

$$x^{(k)} \in B_r\big(x^{(0)}\big) (\subset C) \quad \text{for all } k \in \mathbb{N}\,.$$

By induction we prove the extended proposition

$$x^{(k)} \in B_r\big(x^{(0)}\big) \text{ and } \big\|x^{(k)} - x^{(k-1)}\big\| \le \alpha\, h^{2^{k-1}-1} \text{ for all } k \in \mathbb{N}\,. \quad (8.15)$$

The proposition (8.15) holds for $k = 1$, because according to (iii),

$$\big\|x^{(1)} - x^{(0)}\big\| = \big\|Df\big(x^{(0)}\big)^{-1} f\big(x^{(0)}\big)\big\| \le \alpha < r\,.$$

Let (8.15) be valid for $l = 1, \ldots, k$. Then $x^{(k+1)}$ is well-defined, and by the application of the Newton iteration for $k-1$ we get

$$\begin{aligned}\big\|x^{(k+1)} - x^{(k)}\big\| &= \big\|Df\big(x^{(k)}\big)^{-1} f\big(x^{(k)}\big)\big\| \le \beta\,\big\|f\big(x^{(k)}\big)\big\| \\ &= \beta\,\big\|f\big(x^{(k)}\big) - f\big(x^{(k-1)}\big) - Df\big(x^{(k-1)}\big)\big(x^{(k)} - x^{(k-1)}\big)\big\| \\ &\le \frac{\beta\gamma}{2}\,\big\|x^{(k)} - x^{(k-1)}\big\|^2\end{aligned}$$

according to Lemma 8.10 with $C_0 = B_r\big(x^{(0)}\big)$, and

$$\big\|x^{(k+1)} - x^{(k)}\big\| \le \frac{\beta\gamma}{2}\big\|x^{(k)} - x^{(k-1)}\big\|^2 \le \frac{\beta\gamma}{2}\alpha^2 h^{2^k-2} = \alpha h^{2^k-1}\,.$$

Thus the second part of (8.15) holds for $k+1$, and also

$$\begin{aligned}\big\|x^{(k+1)} - x^{(0)}\big\| &\le \big\|x^{(k+1)} - x^{(k)}\big\| + \big\|x^{(k)} - x^{(k-1)}\big\| + \cdots + \big\|x^{(1)} - x^{(0)}\big\| \\ &\le \alpha\big(h^{2^k-1} + h^{2^{k-1}-1} + \cdots + h^7 + h^3 + h + 1\big) \\ &< \alpha/(1-h) = r\,.\end{aligned}$$

Hence (8.15) holds for $k+1$.

(2): Using (8.15) we are able to verify that $(x^{(k)})_k$ is a Cauchy sequence, because for $l \ge k$ we have

$$\begin{aligned}\big\|x^{(l+1)} - x^{(k)}\big\| &\le \big\|x^{(l+1)} - x^{(l)}\big\| + \big\|x^{(l)} - x^{(l-1)}\big\| + \cdots + \big\|x^{(k+1)} - x^{(k)}\big\| \\ &\le \alpha\, h^{2^k-1}\Big(1 + h^{2^k} + \big(h^{2^k}\big)^3 + \cdots\Big) \qquad (8.16) \\ &< \frac{\alpha\, h^{2^k-1}}{1 - h^{2^k}} \to 0 \quad \text{for } k \to \infty\,,\end{aligned}$$

since $h < 1$. Hence there exists $\bar{x} = \lim_{k\to\infty} x^{(k)}$ and $\bar{x} \in \overline{B}_r\big(x^{(0)}\big)$.

Furthermore, $f(\bar{x}) = 0$, because we can conclude from $x^{(k)} \in B_r\big(x^{(0)}\big)$ that

$$\big\|Df\big(x^{(k)}\big) - Df\big(x^{(0)}\big)\big\| \le \gamma\big\|x^{(k)} - x^{(0)}\big\| < \gamma r\,;$$

thus

$$\big\|Df\big(x^{(k)}\big)\big\| \le \gamma r + \big\|Df\big(x^{(0)}\big)\big\| =: K$$

and from $f\big(x^{(k)}\big) = -Df\big(x^{(k)}\big)\big(x^{(k+1)} - x^{(k)}\big)$, we obtain

$$\big\|f\big(x^{(k)}\big)\big\| \le K\big\|x^{(k+1)} - x^{(k)}\big\| \to 0$$

for $k \to \infty$. Thus we also have

$$f(\bar{x}) = \lim_{k\to\infty} f(x^{(k)}) = 0\,.$$

(3): With $l \to \infty$ in (8.16) we can prove the second part in (3); the first part follows from

$$\begin{aligned}x^{(k+1)} - \bar{x} &= x^{(k)} - Df\big(x^{(k)}\big)^{-1} f\big(x^{(k)}\big) - \bar{x} \\ &= x^{(k)} - \bar{x} - Df\big(x^{(k)}\big)^{-1}\big(f\big(x^{(k)}\big) - f(\bar{x})\big) \\ &= Df\big(x^{(k)}\big)^{-1}\Big(f(\bar{x}) - f\big(x^{(k)}\big) - Df\big(x^{(k)}\big)\big(\bar{x} - x^{(k)}\big)\Big)\,,\end{aligned}$$

which implies, according to Lemma 8.10 with $C_0 = B_{r+\varepsilon}\big(x^{(0)}\big) \subset C$,

$$\big\|x^{(k+1)} - \bar{x}\big\| \le \beta\,\frac{\gamma}{2}\,\big\|x^{(k)} - \bar{x}\big\|^2.$$

□

The termination criterion (5.15), which is oriented at the residual, may also be used for the nonlinear problem (and not just for the Newton iteration). This can be deduced in analogy to (5.16):

Theorem 8.12 *Let the following be valid:*

$$\text{There exists a zero } \bar{x} \text{ of } f \text{ such that } Df(\bar{x}) \text{ is nonsingular and } Df \text{ is Lipschitz continuous in an open neighbourhood } C \text{ of } \bar{x}. \tag{8.17}$$

Then for every $\varrho > 0$ there exists a $\delta > 0$ such that for $x, y \in B_\delta(\bar{x})$,

$$\|f(y)\| \, \|x - \bar{x}\| \leq (1 + \varrho) \, \kappa(Df(\bar{x})) \, \|f(x)\| \, \|y - \bar{x}\| \, .$$

Proof: See [22, p. 69, p. 72] and Exercise 8.4. □

Here κ is the condition number in a matrix norm that is consistent with the chosen vector norm. For $x = x^{(k)}$ and $y = x^{(0)}$ we get (locally) the generalization of (5.16).

8.2.2 Modifications of Newton's Method

Modifications of Newton's method aim in two directions:

- Reduction of the cost of the assembling and the solution of the system of equations (8.12) (without a significant deterioration of the properties of convergence).
- Enlargement of the range of convergence.

We can account for the first aspect by simplifying the matrix in (8.12) (*modified* or *simplified Newton's method*). The extreme case is the replacement of $Df\big(x^{(k)}\big)$ by the identity matrix; this leads us to the fixed-point iteration (8.9). If the mapping f consists of a nonlinear and a linear part,

$$f(x) := Ax + g(x) = 0 \, , \tag{8.18}$$

then the system of equations (8.12) of the Newton iteration reads as

$$\big(A + Dg\big(x^{(k)}\big)\big) \, \delta^{(k)} = -f\big(x^{(k)}\big) \, .$$

A straightforward simplification in this case is the fixed-point iteration

$$A \, \delta^{(k)} = -f\big(x^{(k)}\big) \, . \tag{8.19}$$

It may be interpreted as the fixed-point iteration (8.9) of the system that is preconditioned with A, i.e., of

$$A^{-1} f(x) = 0.$$

In (8.19) the matrix is identical in every iteration step; therefore, it has to be assembled only once, and if we use a direct method (cf. Section 2.5), the LU factorization has to be carried out only once. Thus with forward

and backward substitution we have only to perform methods with lower computational cost. For iterative methods we cannot rely on this advantage, but we can expect that $x^{(k+1)}$ is close to $x^{(k)}$, and consequently $\delta^{(k,0)} = 0$ constitutes a good initial guess. Accordingly, the assembling of the matrix gets more important with respect to the overall computational cost, and savings during the assembling become relevant.

We get a system of equations similar to (8.19) by applying the *chord method* (see Exercise 8.3), where the linear approximation of the initial iterate is maintained, i.e.,

$$Df\big(x^{(0)}\big)\,\delta^{(k)} = -f\big(x^{(k)}\big)\,. \tag{8.20}$$

If the matrix $B\big(x^{(k)}\big)$, which approximates $Df\big(x^{(k)}\big)$, is changing in each iteration step, i.e.,

$$B\big(x^{(k)}\big)\,\delta^{(k)} = -f\big(x^{(k)}\big)\,, \tag{8.21}$$

then the only advantage can be a possibly easier assembling or solvability of the system of equations. If the partial derivatives $\partial_j f_i(x)$ are more difficult to evaluate than the function $f_i(y)$ itself (or possibly not evaluable at all), then the approximation of $Df(x^{(k)})$ by difference quotients can be taken into consideration. This corresponds to

$$B\big(x^{(k)}\big)\,e_j = \frac{1}{h}\big(f(x + he_j) - f(x)\big) \tag{8.22}$$

for column j of $B\big(x^{(k)}\big)$ with a fixed $h > 0$. The number of computations for the assembling of the matrix remains the same: m^2 for the full matrix and analogously for the sparse matrix (see Exercise 8.6). Observe that numerical differentiation is an ill-posed problem, which means that we should ideally choose $h \sim \delta^{1/2}$, where $\delta > 0$ is the error level in the evaluation of f. Even then we can merely expect

$$\big\|Df\big(x^{(k)}\big) - B\big(x^{(k)}\big)\big\| \le C\delta^{1/2}$$

(see [22, pp. 80 f.]). Thus in the best case we can expect only half of the significant digits of the machine precision. The second aspect of facilitated solvability of (8.21) occurs if there appear "small" entries in the Jacobian, due to a problem-dependent weak coupling of the components, and these entries may be skipped. Take, for example, a $Df\big(x^{(k)}\big)$ with a block structure as in (5.39):

$$Df\big(x^{(k)}\big) = \big(A_{ij}\big)_{ij}\,, \quad A_{ij} \in \mathbb{R}^{m_i,m_j}\,,$$

such that the blocks A_{ij} may be neglected for $j > i$. Then there results a nested system of equations of the dimensions $m_1, m_2, \ldots, m_p$.

The possible advantages of such simplified Newton's methods have to be weighted against the disadvantage of a deterioration in the order of convergence: Instead of an estimation like that in Theorem 8.11, (3), we

have to expect an additional term

$$\left\|B\left(x^{(k)}\right) - Df\left(x^{(k)}\right)\right\| \left\|x^{(k)} - x\right\|.$$

This means only linear or — by successive improvement of the approximation — superlinear convergence (see [22, pp. 75 ff.]). If we have a good initial iterate, it may often be advantageous to perform a small number of steps of Newton's method. So in the following we will treat again Newton's method, although the subsequent considerations can also be transferred to its modifications.

If the linear problems (8.12) are solved with an iterative scheme, we have the possibility to adjust the accuracy of the algorithm in order to reduce the number of inner iterations, without a (severe) deterioration of the convergence of the outer iteration of the Newton iteration. So dealing with such *inexact Newton's methods*, we determine instead of $\delta^{(k)}$ from (8.12) only $\tilde{\delta}^{(k)}$, which fulfils (8.12) only up to an *inner residual* $r^{(k)}$, i.e.,

$$Df\left(x^{(k)}\right)\tilde{\delta}^{(k)} = -f\left(x^{(k)}\right) + r^{(k)}\,.$$

The new iterate is given by

$$x^{(k+1)} := x^{(k)} + \tilde{\delta}^{(k)}\,.$$

The accuracy of $\tilde{\delta}^{(k)}$ is estimated by the requirement

$$\left\|r^{(k)}\right\| \le \eta_k \left\|f\left(x^{(k)}\right)\right\| \tag{8.23}$$

with certain properties for the sequence $(\eta_k)_k$ that still have to be determined. Since the natural choice of the initial iterate for solving (8.12) is $\delta^{(k,0)} = 0$, (8.23) corresponds to the termination criterion (5.15). Conditions for η_k can be deduced from the following theorem:

Theorem 8.13 *Let* (8.17) *hold and consider compatible matrix and vector norms. Then there exists for every* $\varrho > 0$ *a* $\delta > 0$ *such that for* $x^{(k)} \in B_\delta(\bar{x})$,

$$\begin{aligned}\left\|x^{(k+1)} - \bar{x}\right\| \;\le\; & \left\|x^{(k)} - Df\left(x^{(k)}\right)^{-1} f\left(x^{(k)}\right) - \bar{x}\right\| \\ & + (1+\varrho)\,\kappa\left(Df(\bar{x})\right)\eta_k \left\|x^{(k)} - \bar{x}\right\|.\end{aligned} \tag{8.24}$$

Proof: By the choice of δ we can ensure the nonsingularity of $Df(x^{(k)})$. From

$$\tilde{\delta}^{(k)} = -Df\left(x^{(k)}\right)^{-1} f\left(x^{(k)}\right) + Df\left(x^{(k)}\right)^{-1} r^{(k)}$$

it follows that

$$\begin{aligned}\left\|x^{(k+1)} - \bar{x}\right\| &= \left\|x^{(k)} - \bar{x} + \tilde{\delta}^{(k)}\right\| \\ &\le \left\|x^{(k)} - \bar{x} - Df\left(x^{(k)}\right)^{-1} f\left(x^{(k)}\right)\right\| + \left\|Df\left(x^{(k)}\right)^{-1} r^{(k)}\right\|.\end{aligned}$$

The assertion can be deduced from the estimation

$$\left\|Df\left(x^{(k)}\right)^{-1} r^{(k)}\right\| \le (1+\varrho)^{1/2} \left\|Df(\bar{x})^{-1}\right\| \left\|r^{(k)}\right\|$$

$$\leq \quad (1+\varrho)^{1/2} \left\| Df(\bar{x})^{-1} \right\| \eta_k \, (1+\varrho)^{1/2} \left\| Df(\bar{x}) \right\| \left\| x^{(k)} - \bar{x} \right\| .$$

Here we used Exercise 8.4 (2), (3) and (8.23). □

The first part of the approximation corresponds to the error of the exact Newton step, which can be estimated using the same argument as in Theorem 8.11, (3) (with Exercise 8.4, (2)) by

$$\left\| x^{(k)} - Df\big(x^{(k)}\big)^{-1} f\big(x^{(k)}\big) - \bar{x} \right\| \leq (1+\varrho)^{1/2} \left\| Df(\bar{x})^{-1} \right\| \frac{\gamma}{2} \left\| x^{(k)} - \bar{x} \right\|^2 .$$

This implies the following result:

Corollary 8.14 *Let the assumptions of Theorem* 8.13 *be satisfied. Then there exist* $\delta > 0$ *and* $\bar{\eta} > 0$ *such that for* $x^{(0)} \in B_\delta(\bar{x})$ *and* $\eta_k \leq \bar{\eta}$ *for all* $k \in \mathbb{N}$ *for the inexact Newton's method the following hold:*

(1) The sequence $\big(x^{(k)}\big)_k$ *converges linearly to* $\bar{x}$.

(2) If $\eta_k \to 0$ *for* $k \to \infty$, *then* $\big(x^{(k)}\big)_k$ *converges superlinearly.*

(3) If $\eta_k \leq K \left\| f\big(x^{(k)}\big) \right\|$ *for a* $K > 0$, *then* $\big(x^{(k)}\big)_k$ *converges quadratically.*

Proof: Exercise 8.5. □

The estimation (8.24) suggests that we carefully choose a very fine level of accuracy $\bar{\eta}$ of the inner iteration to guarantee the above statements of convergence. This is particularly true for ill-conditioned $Df(\bar{x})$ (which is common for discretization matrices: See (5.34)). In fact, the analysis in the weighted norm $\| \cdot \| = \| Df(\bar{x}) \cdot \|$ shows that only $\eta_k \leq \bar{\eta} < 1$ has to be ensured (cf. [22, pp. 97 ff.]). With this and on the basis of

$$\tilde{\eta}_k = \alpha \left\| f\big(x^{(k)}\big) \right\|^2 / \left\| f\big(x^{(k-1)}\big) \right\|^2$$

for some $\alpha \leq 1$ we can construct η_k in an adaptive way (see [22, p. 105]). Most of the iterative methods introduced in Chapter 5 do not require the explicit knowledge of the matrix $Df\big(x^{(k)}\big)$. It suffices that the operation $Df\big(x^{(k)}\big)y$ be feasible for vectors y, in general for fewer than m of them; i.e., the directional derivative of f in $x^{(k)}$ in direction y is needed. Thus in case a difference scheme for the derivatives of f should be necessary or reasonable, it is more convenient to choose directly a difference scheme for the directional derivative.

Since we cannot expect convergence of Newton's method in general, we require indicators for the convergence behaviour of the iteration. The solution $\bar{x}$ is in particular also the solution of

$$\text{Minimize} \quad \|f(x)\|^2 \quad \text{for } x \in \mathbb{R}^m .$$

Let $x^{(0)}$, $\tau > 0$, η_0, $\bar{\Theta} \in (0,1)$, $k = 0$, $i = 0$ be given.

(1) $$\tilde{\delta}^{(k,0)} := 0\,,\ i := 1\,.$$

(2) Determine the ith iterate $\tilde{\delta}^{(k,i)}$ for $Df(x^{(k)})\tilde{\delta}^{(k)} = -f(x^{(k)})$ and calculate
$$r^{(i)} := Df(x^{(k)})\tilde{\delta}^{(k,i)} + f(x^{(k)})\,.$$

(3) If $\|r^{(i)}\| \le \eta_k \|f(x^{(k)})\|$, then go to (4), else set $i := i + 1$ and go to (2).

(4) $$\tilde{\delta}^{(k)} := \tilde{\delta}^{(k,i)}\,.$$

(5) $$x^{(k+1)} := x^{(k)} + \tilde{\delta}^{(k)}\,.$$

(6) If $\|f(x^{(k+1)})\| > \overline{\Theta}\|f(x^{(k)})\|$, interrupt.

(7) If $\|f(x^{(k+1)})\| \le \tau\|f(x^{(0)})\|$, end. Else calculate η_{k+1}, set $k := k + 1$, and go to (1).

Table 8.1. Inexact Newton's method with monotonicity test.

Thus we could expect a descent of the sequence of iterates $(x^{(k)})$ in this functional, i.e.,

$$\left\|f(x^{(k+1)})\right\| \le \bar{\Theta}\left\|f(x^{(k)})\right\| \quad \text{for a } \bar{\Theta} < 1.$$

If this *monotonicity test* is not fulfilled, the iteration is terminated. Such an example of an inexact Newton's method is given in Table 8.1.

In order to avoid the termination of the method due to divergence, the *continuation methods* have been developed. They attribute the problem $f(x) = 0$ to a family of problems to provide successively good initial iterates. The approach presented at the end of Section 8.3 for time-dependent problems is similar to the continuation methods. Another approach (which can be combined with the latter) modifies the (inexact) Newton's method, so that the range of convergence is enlarged: Applying the *damped (inexact) Newton's method* means reducing the step length of $x^{(k)}$ to $x^{(k+1)}$ as long as we observe a decrease conformable to the monotonicity test. One strategy of damping, termed *Armijo's rule*, is described in Table 8.2 and replaces the steps (1), (5), and (6) in Table 8.1.

Thus damping Newton's method means also a relaxation similar to (5.30), where $\omega = \lambda_k$ is being adjusted to the iteration step as in (5.41).

In the formulation of Table 8.2 the iteration may eventually not terminate if in (5) λ_k is successively reduced. This must be avoided in a practical implementation of the method. But except for situations where divergence is obvious, this situation will not appear, because we have the following theorem:

Let additionally $\alpha, \beta \in (0,1)$ be given.

(1) $\tilde{\delta}^{(k,0)} := 0\,,\ i := 1\,,\ \lambda_k := 1.$

(5) If $\|f(x^{(k)} + \lambda_k \tilde{\delta}^{(k)})\| \geq (1 - \alpha\lambda_k)\|f(x^{(k)})\|$, set $\lambda_k := \beta\lambda_k$ and go to (5).

(6) $x^{(k+1)} := x^{(k)} + \lambda_k \tilde{\delta}^{(k)}\,.$

Table 8.2. Damped inexact Newton step according to Armijo's rule.

Theorem 8.15 *Let $\alpha, \beta, \gamma > 0$ exist such that conditions (i), (ii) of Theorem* 8.11 *on $\bigcup_{k\in\mathbb{N}} B_r(x^{(k)})$ hold for the sequence $(x^{(k)})_k$ defined according to Table* 8.2. *Let $\eta_k \leq \bar{\eta}$ for an $\bar{\eta} < 1 - \alpha$.*
Then if $f\big(x^{(0)}\big) \neq 0$, there exists a $\bar{\lambda} > 0$ such that $\lambda_k \geq \bar{\lambda}$ for all $k \in \mathbb{N}$. If furthermore $\big(x^{(k)}\big)_k$ is bounded, then there exists a zero $\bar{x}$, satisfying (8.17) *and*

$$x^{(k)} \to \bar{x} \quad \text{for } k \to \infty\,.$$

There exists a $k_0 \in \mathbb{N}$ such that for $k \geq k_0$ the relation

$$\lambda_k = 1$$

holds.

Proof: See [22, pp. 139 ff.]. □

We see that in the final stage of the iteration we again deal with the (inexact) Newton's method with the previously described behaviour of convergence.

Finally, the following should be mentioned: The problem $f(x) = 0$ and Newton's method are *affine-invariant* in the sense that a transition to $Af(x) = 0$ with a nonsingular $A \in \mathbb{R}^{m,m}$ changes neither the problem nor the iteration method, since

$$D(Af)(x)^{-1} Af(x) = Df(x)^{-1} f(x)\,.$$

Among the assumptions of Theorem 8.11, (8.14) is not affine-invariant. A possible alternative would be

$$\|Df(y)^{-1}(Df(x) - Df(y))\| \leq \gamma\|x - y\|\,,$$

which fulfils the requirement. With the proof of Lemma 8.10 it follows that

$$\|Df(y)^{-1}(f(x) - f(y) - Df(y)(x - y))\| \leq \frac{\gamma}{2}\|x - y\|^2\,.$$

With this argument a similar variant of Theorem 8.11 can be proven.

The test of monotonicity is not affine-invariant, so probably the *natural test of monotonicity*

$$\left\|Df\left(x^{(k)}\right)^{-1} f\left(x^{(k+1)}\right)\right\| \leq \bar{\Theta}\left\|Df\left(x^{(k)}\right)^{-1} f\left(x^{(k)}\right)\right\|$$

has to be preferred. The vector on the right-hand side has already been calculated, being, except for the sign, the Newton correction $\delta^{(k)}$. But for the vector in the left-hand side, $-\bar{\delta}^{(k+1)}$, the system of equations

$$Df\left(x^{(k)}\right)\bar{\delta}^{(k+1)} = -f\left(x^{(k+1)}\right)$$

additionally has to be resolved.

Exercises

8.3 Consider the chord method as described in (8.20). Prove the convergence of this method to the solution $\bar{x}$ under the following assumptions:

(1) Let (8.17) with $\overline{B}_r(\bar{x}) \subset C$ hold,

(2) $\left\|\left[Df(x^{(0)})\right]^{-1}\right\| \leq \beta$,

(3) $2\beta\gamma r < 1$,

(4) $x^{(0)} \in \overline{B}_r(\bar{x})$.

8.4 Let assumption (8.17) hold. Prove for compatible matrix and vector norms that for every $\varrho > 0$ there exists a $\delta > 0$ such that for every $x \in B_\delta(\bar{x})$,

(1) $\|Df(x)\| \leq (1+\varrho)^{1/2}\|Df(\bar{x})\|$,

(2) $\|Df(x)^{-1}\| \leq (1+\varrho)^{1/2}\|Df(\bar{x})^{-1}\|$
(employ $\|(I-M)^{-1}\| \leq 1/(1-\|M\|)$ for $\|M\| < 1$),

(3) $(1+\varrho)^{-1/2}\|Df(\bar{x})^{-1}\|^{-1}\|x-\bar{x}\| \leq \|f(x)\|$
$\leq (1+\varrho)^{1/2}\|Df(\bar{x})\|\|x-\bar{x}\|$,

(4) Theorem 8.12.

8.5 Prove Corollary 8.14.

8.6 Let $U \subset \mathbb{R}^m$ be open and convex. Consider problem (8.2) with continuously differentiable $f : U \to \mathbb{R}^m$. For $i = 1, \ldots, m$ let $J_i \subset \{1, \ldots, m\}$ be defined by

$$\partial_j f_i(x) = 0 \quad \text{for } j \notin J_i \text{ and every } x \in U\,.$$

Then the operator f is *sparsely occupied* if $l_i := |J_i| < m$, or *sparsely occupied* in the strict sense if $l_i \le l$ for all $i = 1, \dots, m$ and $l < m$ is independent of m for a sequence of problems (8.2) of dimension m.

Then the evaluation of $Df(x)$ and its approximation according to (8.22) both need $\sum_{k=1}^{m} l_k$ evaluations of $\partial_j f_i$ or of f_l, respectively. What is the computational effort for a difference approximation

$$\frac{f(x + h\delta/\|\delta\|) - f(x)}{h} \|\delta\|$$

of the directional derivative $Df(x)\delta$?

8.3 Semilinear Boundary Value Problems for Elliptic and Parabolic Equations

In this section we treat *semilinear* problems as the simplest nonlinear case, where nonlinearities do not occur in parts containing derivatives. Hence we want to examine differential equations of the form (0.33) that satisfy (0.42) and (0.43).

Stationary Problems

As a stationary problem we consider the differential equation

$$Lu(x) + \psi(u(x)) = 0 \quad \text{for } x \in \Omega \tag{8.25}$$

with the linear elliptic differential operator L according to (3.12) and linear boundary conditions on $\partial\Omega$ according to (3.18)–(3.20). Here $\psi : \mathbb{R} \to \mathbb{R}$ denotes a mapping that is supposed to be continuously differentiable.

A Galerkin discretization in $V_h \subset V$ with $H_0^1(\Omega) \subset V \subset H^1(\Omega)$ according to the type of boundary condition and $V_h = \text{span}\,\{\varphi_1, \dots, \varphi_M\}$ with the approximative solution $u_h \in V_h$ in the representation $u_h = \sum_{i=1}^{M} \xi_i \varphi_i$ gives

$$S\boldsymbol{\xi} + G(\boldsymbol{\xi}) = \boldsymbol{b} \tag{8.26}$$

with the stiffness matrix $S = \left(a\left(\varphi_j, \varphi_i\right)\right)_{i,j}$ and a vector $\boldsymbol{b}$ that contains the contributions of the inhomogeneous boundary conditions. Here the nonlinear mapping $G : \mathbb{R}^M \to \mathbb{R}^M$ is defined by

$$G(\boldsymbol{\xi}) := \left(G_j(\boldsymbol{\xi})\right)_j \quad \text{with} \quad G_j(\boldsymbol{\xi}) := \int_\Omega \psi\left(\sum_{i=1}^{M} \xi_i \varphi_i\right) \varphi_j \, dx\,.$$

Note that this notation differs from that in Section 2.2 and the subsequent chapters: There we denoted S by A_h and $\boldsymbol{b} - G(\boldsymbol{\xi})$ by $\boldsymbol{q}_h$. For reasons of brevity we omit the index h.

For the moment we want to suppose that the mapping G can be evaluated exactly. The system of equations (8.26) with

$$A := S \quad \text{and} \quad g(\boldsymbol{\xi}) := G(\boldsymbol{\xi}) - \boldsymbol{b}$$

is of the type introduced in (8.18) in the variable $\boldsymbol{\xi}$. Thus we may apply, besides the Newton iteration, the fixed-point iteration, introduced in (8.19), and the variants of Newton's method, namely the modified and inexact versions with their already discussed advantages and drawbacks. We have to examine the question of how the properties of the matrix will change by the transition from A to $A + DG(\bar{\boldsymbol{\xi}})$, where $\bar{\boldsymbol{\xi}}$ stands for the current iterate. We have

$$\big(DG(\bar{\boldsymbol{\xi}})\big)_{ij} = \int_\Omega \psi'(\bar{u})\varphi_i\varphi_j \, dx \,, \tag{8.27}$$

where $\bar{\boldsymbol{u}} = P\bar{\boldsymbol{\xi}} = \sum_{i=1}^M \bar{\boldsymbol{\xi}}_i\varphi_i \in V_h$ denotes the function belonging to the representing vector $\bar{\boldsymbol{\xi}}$. This means that $DG(\bar{\boldsymbol{\xi}})$ is symmetric and positive semidefinite, respectively definite, if the following condition for $\alpha = 0$, respectively $\alpha > 0$, holds:

$$\text{There exists some } \alpha \geq 0 \text{ such that } \psi'(u) \geq \alpha \text{ for all } u \in \mathbb{R} \,. \tag{8.28}$$

More precisely, we have for $\boldsymbol{\eta} \in \mathbb{R}^M$, if (8.28) is valid,

$$\boldsymbol{\eta}^T DG(\bar{\boldsymbol{\xi}})\boldsymbol{\eta} = \int_\Omega \psi'(\bar{u}) \, |P\boldsymbol{\eta}|^2 \, dx \geq \alpha \, \|P\boldsymbol{\eta}\|_0^2 \,.$$

For such a monotone nonlinearity the properties of definiteness of the stiffness matrix S may be "enforced". If, on the other hand, we want to make use of the properties of an M-matrix that can be ensured by the conditions (1.32) or $(1.32)^*$, then it is not clear whether these properties are conserved after addition of $DG(\bar{\boldsymbol{\xi}})$. This is due to the fact that $DG(\bar{\boldsymbol{\xi}})$ is a sparse matrix of the same structure as S, but it also entails a spatial coupling that is not contained in the continuous formulation (8.25).

Numerical Quadrature

Owing to the above reason, the use of a node-oriented quadrature rule for the approximation of $G(\boldsymbol{\xi})$ is suggested, i.e., a quadrature formula of the type

$$Q(f) := \sum_{i=1}^M \omega_i f(a_i) \quad \text{for } f \in C(\bar{\Omega}) \tag{8.29}$$

with weights $\omega_i \in \mathbb{R}$. Such a quadrature formula results from

$$Q(f) := \int_\Omega I(f) \, dx \quad \text{for } f \in C(\bar{\Omega}) \,, \tag{8.30}$$

where

$$I : C(\bar{\Omega}) \to V_h \,, \quad I(f) := \sum_{i=1}^M f(a_i)\varphi_i \,,$$

is the interpolation operator of the degrees of freedom. For this consideration we thus assume that only Lagrangian elements enter the definition of

V_h. In the case of (8.30) the weights in (8.29) are hence given by

$$\omega_i = \int_\Omega \varphi_i \, dx \, .$$

This corresponds to the local description (3.116). More specifically, we get, for example, for the linear approach on simplices as a generalization of the *composite trapezoidal rule*,

$$\omega_i = \frac{1}{d+1} \sum_{\substack{K \in \mathcal{T}_h \\ \text{with } a_i \in K}} |K| \, , \tag{8.31}$$

with d denoting the spatial dimension and $\mathcal{T}_h$ the underlying triangulation. Approximation of the mapping G by a quadrature rule of the type (8.29) gives

$$\tilde{G}(\boldsymbol{\xi}) = \left(\tilde{G}_j(\boldsymbol{\xi}) \right)_j \quad \text{with} \quad \tilde{G}_j(\boldsymbol{\xi}) = \omega_j \psi(\xi_j) \, ,$$

because of $\varphi_j(a_i) = \delta_{ij}$. We see that the approximation $\tilde{G}$ has the property that $\tilde{G}_j$ depends only on ξ_j. We call such a $\tilde{G}$ a *diagonal field*. Qualitatively, this corresponds better to the continuous formulation (8.25) and leads to the fact that $D\tilde{G}(\bar{\boldsymbol{\xi}})$ is diagonal:

$$D\tilde{G}(\bar{\boldsymbol{\xi}})_{ij} = \omega_j \psi'(\bar{\xi}_j) \delta_{ij} \, . \tag{8.32}$$

If we impose that all quadrature weights ω_i are positive, which is the case in (8.31) and also in other examples in Section 3.5.2, all of the above considerations about the properties of $D\tilde{G}(\bar{\boldsymbol{\xi}})$ and $S + D\tilde{G}(\bar{\boldsymbol{\xi}})$ remain valid; additionally, if S is an M-matrix, because the conditions (1.32) or $(1.32)^*$ are fulfilled, then $S + D\tilde{G}(\bar{\boldsymbol{\xi}})$ remains an M-matrix, too. This is justified by the following fact (compare [34] and [5]; cf. (1.33) for the notation):

> If A is an M-matrix and $B \geq A$ with $b_{ij} \leq 0$ for $i \neq j$, then B is an M-matrix as well. (8.33)

Conditions of Convergence

Comparing the requirements for the fixed-point iteration and Newton's method stated in the (convergence) Theorems 8.4 and 8.11, we observe that the conditions in Theorem 8.4 can be fulfilled only in special cases, where $S^{-1} D\tilde{G}(\bar{\boldsymbol{\xi}})$ is small according to a suitable matrix norm (see Lemma 8.3). But it is also difficult to draw general conclusions about requirement (iii) in Theorem 8.11, which together with $h < 1$ quantifies the closeness of the initial iterate to the solution. The postulation (i), on the other hand, is met for (8.27) and (8.32) if ψ' is Lipschitz continuous (see Exercise 8.7). Concerning the postulation (ii) we have the following: Let ψ be monotone nondecreasing (i.e., (8.28) holds with $\alpha \geq 0$) and let S be symmetric and positive definite, which is true for a problem without convection terms

(compare (3.27)). Then we have in the spectral norm

$$\left\|S^{-1}\right\|_2 = 1/\lambda_{\min}(S)\,.$$

Here $\lambda_{\min}(S) > 0$ denotes the smallest eigenvalue of S. Hence

$$\left\| \left(S + DG(\boldsymbol{\xi})\right)^{-1} \right\|_2 = 1/\lambda_{\min}\left(S + DG(\bar{\boldsymbol{\xi}})\right) \le 1/\lambda_{\min}(S) = \left\|S^{-1}\right\|_2\,,$$

and consequently, requirement (ii) is valid with $\beta = \left\|S^{-1}\right\|_2$.

Concerning the choice of the initial iterate, there is no generally successful strategy. We may choose the solution of the linear subproblem, i.e.,

$$S\boldsymbol{\xi}^{(0)} = \boldsymbol{b}\,. \tag{8.34}$$

Should it fail to converge even with damping, then we may apply, as a generalization of (8.34), the continuation method to the family of problems

$$f(\lambda, \boldsymbol{\xi}) := S + \lambda G(\boldsymbol{\xi}) - \boldsymbol{b} = 0$$

with continuation parameter $\lambda \in [0, 1]$. If all these problems have solutions $\boldsymbol{\xi} = \boldsymbol{\xi}_\lambda$ so that $Df(\boldsymbol{\xi}; \lambda)$ exists and is nonsingular in a neighbourhood of $\boldsymbol{\xi}_\lambda$, and if there exists a continuous solution trajectory without bifurcation, then $[0, 1]$ can be discretized by $0 = \lambda_0 < \lambda_1 < \cdots < \lambda_N = 1$, and solutions $\boldsymbol{\xi}_{\lambda_i}$ of $f(\boldsymbol{\xi}; \lambda_i) = 0$ can be obtained by performing a Newton iteration with the (approximative) solution for $\lambda = \lambda_{i-1}$ as starting iterate. Since the $\boldsymbol{\xi}_{\lambda_i}$ for $i < N$ are just auxiliary means, they should be obtained rather coarsely, i.e., with one or two Newton steps. The stated conditions are fulfilled under the supposition (8.28). If this condition of monotonicity does not hold, we may encounter a *bifurcation* of the continuous solution (see, for example, [29, pp. 28 ff.]).

Instationary Problems

The elliptic boundary value problem (8.25) corresponds to the parabolic initial value problem

$$\partial_t u(x,t) + Lu(x,t) + \psi(u(x,t)) = 0 \quad \text{for } (x,t) \in Q_T \tag{8.35}$$

with linear boundary conditions according to (3.18)–(3.20) and the initial condition

$$u(x,0) = u_0(x) \quad \text{for } x \in \Omega\,. \tag{8.36}$$

We have already met an example for (8.35), (8.36) in (0.32). Analogously to (8.26) and (7.45), the Galerkin discretization in V_h (i.e., the semidiscretization) leads to the nonlinear system of ordinary differential equations

$$B\frac{d}{dt}\boldsymbol{\xi}(t) + S\boldsymbol{\xi}(t) + G(\boldsymbol{\xi}(t)) = \boldsymbol{\beta}(t) \quad \text{for } t \in (0, T]\,, \quad \boldsymbol{\xi}(0) = \boldsymbol{\xi}_0$$

for the representing vector $\boldsymbol{\xi}(t)$ of the approximation $u_h(\cdot, t) = \sum_{i=1}^M \xi_i(t)\varphi_i$, where $u_{0h} = \sum_{i=1}^M \xi_{0i}\varphi_i$ is an approximation of the initial value u_0 (see

Section 7.2). The matrix B is the mass matrix

$$B = \left(\langle \varphi_j, \varphi_i \rangle_0\right)_{ij} ,$$

and $\boldsymbol{\beta}(t)$ contains the contributions of the inhomogeneous boundary conditions analogously to $\boldsymbol{b}$ in (8.26).

To obtain the fully discrete scheme we use the one-step-theta method as in Section 7.3. Here we allow the time step size τ_n to vary in each step, in particular determined by a time step control before the execution of the nth time step. So, if the approximation U^n is known for $t = t_n$, then the approximation U^{n+1} for $t = t_{n+1} := t_n + \tau_n$ is given in generalization of (7.72) as the solution of

$$\begin{aligned} \left\langle \frac{1}{\tau_n}\left(U^{n+1} - U^n\right), v_h \right\rangle_0 + a\left(\Theta U^{n+1} + (1-\Theta)U^n, v_h\right) \\ + \left\langle \psi^{n+\Theta}, v_h \right\rangle = \Theta\boldsymbol{\beta}(t_{n+1}) + (1-\Theta)\boldsymbol{\beta}(t_n). \end{aligned} \tag{8.37}$$

Here $\Theta \in [0,1]$ is the fixed parameter of implicity. For the choice of $\psi^{n+\Theta}$ we have two possibilities:

$$\psi^{n+\Theta} = \Theta\psi(U^{n+1}) + (1-\Theta)\psi(U^n) \tag{8.38}$$

or

$$\psi^{n+\Theta} = \psi\left(\Theta U^{n+1} + (1-\Theta)U^n\right) . \tag{8.39}$$

In the explicit case, i.e., $\Theta = 0$, (8.37) represents a linear system of equations for U^{n+1} (with the system matrix B) and does not have to be treated further here. In the implicit case $\Theta \in (0,1]$ we obtain again a nonlinear system of the type (8.18), i.e.,

$$A\boldsymbol{\xi} + g(\boldsymbol{\xi}) = 0\,,$$

in the variable $\boldsymbol{\xi} = \boldsymbol{\xi}^{n+1}$, where $\boldsymbol{\xi}^{n+1}$ is the representation vector of U^{n+1}: $U^{n+1} = \sum_{i=1}^{M} \xi_i^{n+1}\varphi_i$. Now we have for the variant (8.38),

$$A := B + \Theta\tau_n S\,, \tag{8.40}$$

$$g(\boldsymbol{\xi}) := \Theta\tau_n G(\boldsymbol{\xi}) - \boldsymbol{b}\,, \tag{8.41}$$

with

$$\begin{aligned} \boldsymbol{b} := \left(B - (1-\Theta)\tau_n S\right)\boldsymbol{\xi}^n - (1-\Theta)\tau_n G(\boldsymbol{\xi}^n) \\ + \Theta\boldsymbol{\beta}(t_{n+1}) + (1-\Theta)\boldsymbol{\beta}(t_n)\,. \end{aligned} \tag{8.42}$$

For the variant (8.39) g changes to

$$g(\boldsymbol{\xi}) := \tau_n G\left(\Theta\boldsymbol{\xi} + (1-\Theta)\boldsymbol{\xi}^n\right) - \boldsymbol{b}\,,$$

and in the definition of $\boldsymbol{b}$ the second summation term drops out. The vector $\boldsymbol{\xi}^n$ is the representation vector of the already known approximation U^n.

Numerical Quadrature

As in the stationary case we can approximate g by a quadrature rule of the form (8.29), which leads to

$$\tilde{g}(\boldsymbol{\xi}) = \Theta\tau_n \tilde{G}(\boldsymbol{\xi}) - \boldsymbol{b}$$

in (8.38) and to

$$\tilde{g}(\boldsymbol{\xi}) = \tau_n \tilde{G}\left(\Theta\boldsymbol{\xi} + (1-\Theta)\boldsymbol{\xi}^n\right) - \boldsymbol{b}$$

in (8.39). The functional matrices of g and $\tilde{g}$ are thus equal for (8.38) and (8.39), except to the point where ψ' is being evaluated. Consequently, it suffices in the following to refer to (8.38). Based on the same motivation, a quadrature rule of the form (8.29) can be applied to the mass matrix B. Such a *mass lumping* results in a diagonal approximation of the mass matrix

$$\tilde{B} = \operatorname{diag}(\omega_i)\,.$$

In contrast to the stationary case we get the factor $\Theta\tau_n$ in front of the nonlinearity, where the time step size τ_n may be chosen arbitrarily small. Of course, we have to take into account that the number of time steps necessary to achieve a fixed time T is respectively raised. All of the above considerations about the matrix properties of $A + Dg(\bar{\boldsymbol{\xi}})$ are conserved, where A is no longer the stiffness matrix, but represents the linear combination (8.40) with the mass matrix. This reduces the requirements concerning the V-ellipticity of a (see (3.27)) and thus the positive definiteness of A.

Admittedly, A is not necessarily an M-matrix if S is one, because the conditions (1.32) or $(1.32)^*$ are not valid. Here the approximation $\tilde{B}$ is advantageous, because using nonnegative weights will conserve this property due to (8.33).

Conditions of Convergence

Clear differences arise in answering the question of how to ensure the convergence of the iteration schemes. Even for the fixed-point iteration it is true that the method converges globally if only the time step size τ_n is chosen small enough. We want to demonstrate this in the following by an example of a quadrature with nonnegative weights in the mass matrix and the nonlinearity. Therefore, the Lipschitz constant of $A^{-1}g$ is estimated according to Lemma 8.3. Let the norm be a matrix norm induced by a p-norm $|\cdot|_p$ and let A be nonsingular. We get

$$\begin{aligned}
\left\|A^{-1}\right\| \sup_{\boldsymbol{\xi}\in\mathbb{R}^M} \|D\tilde{g}(\boldsymbol{\xi})\| &\leq \left\|\left(I + \Theta\tau_n \tilde{B}^{-1}S\right)^{-1}\tilde{B}^{-1}\right\| \Theta\tau_n \sup_{s\in\mathbb{R}} |\psi'(s)| \left\|\tilde{B}\right\| \\
&\leq \Theta\tau_n \sup_{s\in\mathbb{R}} |\psi'(s)|\, \kappa(\tilde{B}) \left\|\left(I + \Theta\tau_n \tilde{B}^{-1}S\right)^{-1}\right\| \\
&=: C\tau_n \left\|\left(I + \Theta\tau_n \tilde{B}^{-1}S\right)^{-1}\right\| .
\end{aligned}$$

Thus we assume the boundedness of ψ' on $\mathbb{R}$ (which may even be weakened). For a given $\vartheta \in (0,1)$ choose τ_n sufficiently small such that

$$\Theta\tau_n \|\tilde{B}^{-1}S\| \le \vartheta$$

holds. With Lemma (A3.11) it follows that

$$\left\| \left(I + \Theta\tau_n \tilde{B}S\right)^{-1} \right\| \le \frac{1}{1-\vartheta},$$

and thus we obtain

$$\gamma = \frac{C\tau_n}{1-\vartheta}$$

as a Lipschitz constant for $A^{-1}g$. We see that by choosing τ_n sufficiently small, the contraction property of $A^{-1}g$ can be guaranteed. From this fact a (heuristic) step size control can be deduced that reduces the step size when a lack of convergence is detected and repeats the step, and in case of satisfactory convergence increases the time step size.

Nevertheless, in general, Newton's method is to be preferred: Here we can expect that the quality of the initial iterate $\boldsymbol{\xi}^{(0)} = \boldsymbol{\xi}^n$ for time step $(n+1)$ improves the smaller we choose τ_n. The step size control mentioned above may thus be chosen here, too (in conjunction with the enlargement of the range of convergence via damping). Nonetheless, a problem only to be solved in numerical practice consists in coordinating the control parameters of the time step control, the damping strategy, and eventually the termination of the inner iteration in such a way that overall, an efficient algorithm is obtained.

Exercises

8.7 Study the Lipschitz property of DG defined by (8.27) and of $D\tilde{G}$ defined by (8.32), provided ψ' is Lipschitz.

8.8 Decide whether $A^{-1}g$ is contractive in case of (8.40)–(8.42).

8.9 The boundary value problem

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

is to be discretized by a finite element method using continuous, piecewise linear functions on equidistant grids. Quadrature is to be done with the trapezoidal rule.

(a) Compute the matrix $A_h \in \mathbb{R}^{m,m}$ and the nonlinear vectorvalued function $F_h : \mathbb{R}^m \to \mathbb{R}^m$, in a matrix-vector notation

$$A_h U_h + F_h(U_h) = 0$$

of the discretization. Here $U_h \in \mathbb{R}^m$ denotes the vector of unknown nodal values of the approximative solution and, for uniqueness of the representation, the elements of A_h are independent of the discretization parameter h.

(b) Study the convergence of the iterative procedure

$$(\alpha) \qquad (2+h^2)U_h^{(k+1)} = \big((2+h^2)I - A_h\big)\, U_h^{(k)} - F_h\big(U_h^{(k)}\big)\,,$$

$$(\beta) \qquad 2U_h^{(k+1)} + F_h\big(U_h^{(k+1)}\big) = (2I - A_h)\, U_h^{(k)}\,.$$

9

Discretization Methods for Convection-Dominated Problems

9.1 Standard Methods and Convection-Dominated Problems

As we have seen in the introductory Chapter 0, the modelling of transport and reaction processes in porous media results in differential equations of the form

$$\partial_t u - \nabla \cdot (K \nabla u - cu) = f \,,$$

which is a special case of the form (0.33). Similar equations occur in the modelling of the heat transport in flowing water, the carrier transport in semiconductors, and the propagation of epidemics. These application-specific equations often share the property that their so-called *global Péclet number*

$$\mathrm{Pe} := \frac{\|c\|_\infty \mathrm{diam}(\Omega)}{\|K\|_\infty} \tag{9.1}$$

is significantly larger than one. For example, representative values range from 25 (transport of a dissolved substance in ground water) up to about 10^7 (modelling of semiconductors). In such cases, the equations are called *convection-dominated.*

Therefore, in what follows, the Dirichlet boundary value problem introduced in Section 3.2 will be looked at from the point of view of large global Péclet numbers, whereas in Section 9.4, the initial boundary value problem from Chapter 7 will be considered from this aspect.

Let $\Omega \subset \mathbb{R}^d$ denote a bounded domain with a Lipschitz continuous boundary. Given a function $f : \Omega \to \mathbb{R}$, a function $u : \Omega \to \mathbb{R}$ is to be determined such that

$$\begin{aligned} Lu &= f \quad \text{in} \quad \Omega\,, \\ u &= 0 \quad \text{on} \quad \Gamma\,, \end{aligned} \tag{9.2}$$

where again

$$Lu := -\nabla \cdot (K\nabla u) + c \cdot \nabla u + ru\,,$$

with sufficiently smooth coefficients

$$K : \Omega \to \mathbb{R}^{d,d}\,, \quad c : \Omega \to \mathbb{R}^d\,, \quad r : \Omega \to \mathbb{R}\,.$$

Unfortunately, standard discretization methods (finite difference, finite element, and finite volume methods) fail when applied to convection-dominated equations. At first glance, this seems to be a contradiction to the theory of these methods presented in the preceding chapters, because there we did not have any restriction on the global Péclet number. This apparent contradiction may be explained as follows: On the one hand, the theoretical results are still true for the convection-dominated case, but on the other hand, some assumptions of the statements therein (such as "for sufficiently small h") lack sharpness. This, in turn, may lead to practically unrealistic conditions (cf. the later discussion of the estimate (9.13)). For example, it may happen that the theoretically required step sizes are so small that the resulting discrete problems are too expensive or even untreatable.

So one can ask whether the theory is insufficient or not. The following example will show that this is not necessarily the case.

Example 9.1 Given a constant diffusion coefficient $k > 0$, consider the boundary value problem

$$\begin{aligned} (-ku' + u)' &= 0 \quad \text{in } \Omega := (0,1)\,, \\ u(0) = u(1) - 1 &= 0\,. \end{aligned}$$

Its solution is

$$u(x) = \frac{1 - \exp(x/k)}{1 - \exp(1/k)}\,.$$

A rough sketch of the graph (Figure 9.1) shows that this function has a significant boundary layer at the right boundary of the interval even for the comparatively small global Péclet number $\mathrm{Pe} = 100$. In the larger subinterval (about $(0, 0.95)$) it is very smooth (nearly constant), whereas in the remaining small subinterval (about $(0.95, 1)$) the absolute value of its first derivative is large.

Given an equidistant grid of width $h = 1/(M+1)$, $M \in \mathbb{N}$, a discretization by means of symmetric difference quotients yields the difference

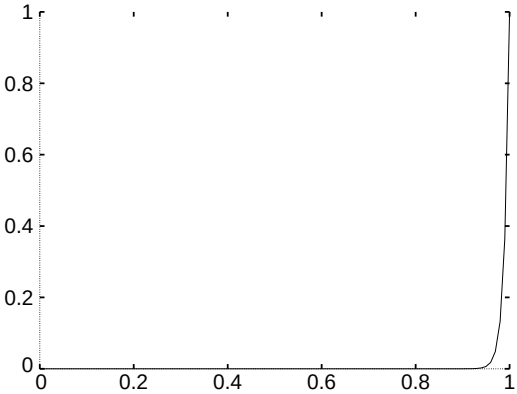


Figure 9.1. Solution for $k = 0.01$.

equations

$$-k\frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} + \frac{u_{i+1} - u_{i-1}}{2h} = 0\,, \quad i \in \{1, \ldots, M\} =: \Lambda\,,$$
$$u_0 = u_{M+1} - 1 = 0\,.$$

Collecting the coefficients and multiplying the result by $2h$, we arrive at

$$\left(-\frac{2k}{h} - 1\right) u_{i-1} + \frac{4k}{h} u_i + \left(-\frac{2k}{h} + 1\right) u_{i+1} = 0\,, \quad i \in \Lambda\,.$$

If we make the ansatz $u_i = \lambda^i$, the difference equations can be solved exactly:

$$u_i = \frac{1 - \left(\frac{2k+h}{2k-h}\right)^i}{1 - \left(\frac{2k+h}{2k-h}\right)^{M+1}}\,.$$

In the case $2k < h$, which is by no means unrealistic (e.g., for the typical value $k = 10^{-7}$), the numerical solution considerably oscillates, in contrast to the behaviour of the exact solution u. These oscillations do not disappear until $h < 2k$ is reached, but this condition is very restrictive for small values of k.

But even if the condition $h < 2k$ is satisfied, undesirable effects can be observed. For example, in the special case $h = k$ we have at the node $a_M = Mh$ that

$$\begin{aligned} u(a_M) &= \frac{1 - \exp(Mh/k)}{1 - \exp(1/k)} = \frac{1 - \exp(M)}{1 - \exp(M+1)} = \frac{\exp(-M) - 1}{\exp(-M) - \exp(1)} \\ &\to \exp(-1) \quad \text{for } h \to 0\,, \end{aligned}$$

whereas the numerical solution at this point asymptotically behaves like (note that $\lambda = (2k+h)/(2k-h) = 3$)

$$u_M = \frac{1-\lambda^M}{1-\lambda^{M+1}} = \frac{\lambda^{-M}-1}{\lambda^{-M}-\lambda} \to \frac{1}{\lambda} = \frac{1}{3} \quad \text{for } h \to 0\,.$$

So the numerical solution does not converge to the exact solution at the node a_M.

Again this is no contradiction to possible convergence results for the finite difference method in the discrete maximum norm, since now the diffusion coefficient is not fixed, but rather the discretization is to be viewed as belonging to the limit case $k = 0$, with an artificial diffusion part in the discretization (see (9.8) below).

Finite Difference Methods with Symmetric and One-Sided Difference Quotients

The oscillations in Example 9.1 show that in this case no comparison principle as in Corollary 1.13 is valid. Such a comparison principle, or more strongly a maximum principle, will lead to nonnegative solutions in the case of nonnegative right-hand side and Dirichlet data. This avoids for a homogeneous right-hand side an *undershooting,* as observed in Example 9.1, i.e., negative solution values in this case, and also an *overshooting*, i.e., solution values larger than the maximum of the Dirichlet data, provided that condition (1.32) (6)* holds.

In the following we will examine how the convective part influences the matrix properties (1.32) and thus the validity of a maximum or comparison principle and also conclude a first simple remedy.

We consider the model problem (9.2), for simplicity on a rectangle $\Omega = (0,a) \times (0,b)$, with constant, scalar $K = kI$ and equipped with an equidistant grid Ω_h. To maintain the order of consistency 2 of a spatial discretization of $-\nabla \cdot (K\nabla u) = -k\Delta u$ by the five-point stencil, the use of the symmetric difference quotient for the discretization of

$$(c \cdot \nabla u)(x) = c_1(x)\partial_1 u(x) + c_2(x)\partial_2 u(x)$$

suggests itself, i.e., for a grid point $x \in \Omega_h$,

$$c_1(x)\partial_1 u(x) \sim c_1(x)\frac{1}{2h}(u_{i+1,j} - u_{i-1,j}), \tag{9.3}$$

and similarly for $c_2(x)\partial_2 u(x)$ (cf. (1.7) for the notation). This leads to the following entries of the system matrix $\tilde{A}_h$, for example in a rowwise

numbering (compare (1.13)):

$$\begin{aligned}
&\text{left secondary diagonal:} && -\frac{c_1(x)}{2h} - \frac{k}{h^2}\,;\\
&\text{right secondary diagonal:} && +\frac{c_1(x)}{2h} - \frac{k}{h^2}\,;\\
&l+1 \text{ positions to the left:} && -\frac{c_2(x)}{2h} - \frac{k}{h^2}\,;\\
&l+1 \text{ positions to the right:} && +\frac{c_2(x)}{2h} - \frac{k}{h^2}\,;\\
&\text{diagonal:} && \frac{4k}{h^2}\,.
\end{aligned}$$

Condition (1.32) (1) and (1.32) (6)* obviously hold.

We check the conditions sufficient for a comparison principle (Corollary 1.13). To satisfy condition (1.32) (2) we require

$$\begin{aligned}
-\frac{k}{h^2} + \frac{|c_1(x)|}{2h} &< 0\,,\\
-\frac{k}{h^2} + \frac{|c_2(x)|}{2h} &< 0\,.
\end{aligned}$$

Denoting the *grid Péclet number* by

$$\mathrm{Pe}_h := \frac{\|c\|_\infty h}{2k}\,, \tag{9.4}$$

the above conditions are satisfied if

$$\mathrm{Pe}_h < 1 \tag{9.5}$$

is satisfied. Under this assumption also the conditions (1.32) (5) and (7) are satisfied, and thus also (3), i.e., (9.5), is sufficient for the validity of a comparison principle. In Example 9.1 this is just the condition $h < 2k$.

The grid Péclet number is obviously related to the global Péclet number from (9.1) by

$$\mathrm{Pe}_h = \mathrm{Pe}\,\frac{h}{2\,\mathrm{diam}(\Omega)}\,.$$

The requirement (9.4) can always be met by choosing h sufficiently small, but for large Pe this may be a severe requirement, necessary for the sake of stability of the method, whereas for the accuracy desired a larger step size may be sufficient. A simple remedy to ensure condition (1.32) (2) is to use a one-sided (*upwind*) discretization of $c_1\partial_1 u$ and $c_2\partial_2 u$, which is selected against the stream direction defined by c_1 and c_2, respectively:

$$\begin{aligned}
\text{For } c_1(x) \geq 0 &: \qquad c_1(x)\partial_1 u(x) \sim c_1(x)\frac{1}{h}(u_{i,j} - u_{i-1,j})\,, \qquad (9.6)\\
\text{for } c_1(x) < 0 &: \qquad c_1(x)\partial_1 u(x) \sim c_1(x)\frac{1}{h}(u_{i+1,j} - u_{i,j})\,,
\end{aligned}$$

and analogously for $c_2\partial_2 u$.

Due to this choice there are only additional nonnegative addends to the diagonal position and nonpositive ones to the off-diagonal positions compared to the five-point stencil or another discretization of a diffusive part. Thus all properties (1.32) (1)–(7), (4)*, and (6)* remains unaffected; i.e., the upwind discretization satisfies all qualitative properties of Section 1.4 from the inverse monotonicity to the strong maximum principle, without any restrictions to the local Péclet number.

The drawback lies in the reduced accuracy, since the one-sided difference quotient has only order of consistency 1. In Section 9.3 we will develop more refined upwind discretizations.

Due to

$$c_1 \frac{u(x,y) - u(x-h,y)}{h} = \tag{9.7}$$
$$\frac{c_1 h}{2} \frac{-u(x-h,y) + 2u(x,y) - u(x+h,y)}{h^2} + c_1 \frac{u(x+h,y) - u(x-h,y)}{2h},$$

and analogously for the forward difference quotient, the upwind discretization can be perceived as a discretization with symmetric difference quotients if a step-size-dependent diffusive part, also discretized with $\partial^-\partial^+$, is added with the diffusion coefficient

$$K_h(x) := \frac{h}{2} \begin{pmatrix} |c_1(x)| & 0 \\ 0 & |c_2(x)| \end{pmatrix}. \tag{9.8}$$

Therefore, one also speaks of adding *artificial diffusion* (or *viscosity*). The disadvantage of this *full* upwind method is that it recognizes the flow direction only if the flow is aligned to one of the coordinate axes. This will be improved in Section 9.2.

Error Estimates for the Standard Finite Element Method

In order to demonstrate the theoretical deficiencies, we will again reproduce the way for obtaining standard error estimates for a model problem. So let $K(x) \equiv \varepsilon I$ with a constant coefficient $\varepsilon > 0$, $c \in C^1(\overline{\Omega}, \mathbb{R}^d)$, $r \in C(\overline{\Omega})$, $f \in L^2(\Omega)$. Furthermore, assume that the following inequality is valid in Ω, where $r_0 > 0$ is a constant: $r - \frac{1}{2}\nabla \cdot c \geq r_0$.

Then the bilinear form $a : V \times V \to \mathbb{R}$, $V := H_0^1(\Omega)$, corresponding to the boundary value problem (9.2), reads as (cf. (3.23))

$$a(u,v) := \int_\Omega [\varepsilon \nabla u \cdot \nabla v + c \cdot \nabla u\, v + r\, uv]\, dx, \quad u, v \in V. \tag{9.9}$$

To get an ellipticity estimate of a, we set $u = v \in V$ in (9.9) and take the relation $2v(c \cdot \nabla v) = c \cdot \nabla v^2$ into account. Then, by partial integration of the middle term, we obtain

$$a(v,v) \quad = \quad \varepsilon |v|_1^2 + \langle c \cdot \nabla v, v \rangle_0 + \langle rv, v \rangle_0$$

$$= \varepsilon|v|_1^2 - \left\langle \frac{1}{2}\nabla\cdot c, v^2 \right\rangle_0 + \langle rv, v\rangle_0 = \varepsilon|v|_1^2 + \left\langle r - \frac{1}{2}\nabla\cdot c, v^2 \right\rangle_0 .$$

Introducing the so-called *ε-weighted H^1-norm* by

$$\|v\|_\varepsilon := \left\{\varepsilon|v|_1^2 + \|v\|_0^2\right\}^{1/2}, \tag{9.10}$$

we immediately arrive at the estimate

$$a(v,v) \geq \varepsilon|v|_1^2 + r_0\|v\|_0^2 \geq \tilde{\alpha}\|v\|_\varepsilon^2, \tag{9.11}$$

where $\tilde{\alpha} := \min\{1, r_0\}$ does not depend on ε.

Due to $c\cdot\nabla u = \nabla\cdot(cu) - (\nabla\cdot c)u$, partial integration yields for arbitrary $u, v \in V$ the identity

$$\langle c\cdot\nabla u, v\rangle_0 = -\langle u, c\cdot\nabla v\rangle_0 - \langle(\nabla\cdot c)u, v\rangle_0 .$$

So we get the continuity estimate

$$\begin{aligned} |a(u,v)| &\leq \varepsilon|u|_1|v|_1 + \|c\|_{0,\infty}\|u\|_0|v|_1 + (|c|_{1,\infty} + \|r\|_{0,\infty})\,\|u\|_0\|v\|_0 \\ &\leq (\sqrt{\varepsilon}\,|u|_1 + \|u\|_0)\,\{(\sqrt{\varepsilon} + \|c\|_{0,\infty})\,|v|_1 + (|c|_{1,\infty} + \|r\|_{0,\infty})\,\|v\|_0\} \\ &\leq \tilde{M}\|u\|_\varepsilon\|v\|_1 , \end{aligned} \tag{9.12}$$

where $\tilde{M} := 2\max\{\sqrt{\varepsilon} + \|c\|_{0,\infty}, |c|_{1,\infty} + \|r\|_{0,\infty}\}$.

Since we are interested in the case of small diffusion $\varepsilon > 0$ and present convection (i.e., $\|c\|_{0,\infty} > 0$), the continuity constant $\tilde{M}$ can be bounded independent of ε. It is not very surprising that the obtained continuity estimate is nonsymmetric, since also the differential expression L behaves like that. Passing over to a symmetric estimate results in the following relation:

$$|a(u,v)| \leq \frac{\tilde{M}}{\sqrt{\varepsilon}}\|u\|_\varepsilon\|v\|_\varepsilon .$$

Now, if $V_h \subset V$ denotes a finite element space, we can argue as in the proof of Céa's lemma (Theorem 2.17) and get an error estimate for the corresponding finite element solution $u_h \in V_h$. To do this, the nonsymmetric continuity estimate (9.12) is sufficient. Indeed, for arbitrary $v_h \in V_h$, we have

$$\tilde{\alpha}\|u-u_h\|_\varepsilon^2 \leq a(u-u_h, u-u_h) = a(u-u_h, u-v_h) \leq \tilde{M}\|u-u_h\|_\varepsilon\|u-v_h\|_1 .$$

Thus

$$\|u-u_h\|_\varepsilon \leq \frac{\tilde{M}}{\tilde{\alpha}} \inf_{v_h\in V_h} \|u-v_h\|_1 .$$

Here the constant $\tilde{M}/\tilde{\alpha}$ does not depend on ε, h, and u. This estimate is weaker than the standard estimate, because the ε-weighted H^1-norm is weaker than the H^1-norm. Moreover, the error of the best approximation is

not independent of ε, in general. For example, if we apply continuous, piecewise linear elements, then, under the additional assumption $u \in H^2(\Omega)$, Theorem 3.29 yields the estimate

$$\inf_{v_h \in V_h} \|u - v_h\|_1 \leq \|u - I_h(u)\|_1 \leq Ch|u|_2 \,,$$

where the constant $C > 0$ does not depend on ε, h, and u. So, we finally arrive at the relation

$$\|u - u_h\|_\varepsilon \leq Ch|u|_2 \,. \tag{9.13}$$

However, the H^2-seminorm of the solution u depends on ε in a disadvantageous manner; for example, it may be (cf. also [27, Lemma III.1.18]) that

$$|u|_2 = O(\varepsilon^{-3/2}) \qquad (\varepsilon \to 0) \,.$$

This result is sharp, since for examples of boundary value problems for ordinary linear differential equations the error of the best approximation already exhibits this asymptotic behaviour.

So the practical as well as the theoretical problems mentioned above indicate the necessity to use special numerical methods for solving convection-dominated equations. In the next sections, a small collection of these methods will be depicted.

9.2 The Streamline-Diffusion Method

The streamline-diffusion method is the prevalent method in the numerical treatment of stationary convection-dominated problems. The basic idea is due to Brooks and Hughes [49], who called the method the *streamline upwind Petrov–Galerkin method (SUPG method).*

We describe the idea of the method for a special case of boundary value problem (9.2) under consideration. Let the domain $\Omega \subset \mathbb{R}^d$ be a bounded polyhedron. We consider the same model as in the preceding section, that is, $K(x) \equiv \varepsilon I$ with a constant coefficient $\varepsilon > 0$, $c \in C^1(\overline{\Omega}, \mathbb{R}^d)$, $r \in C(\overline{\Omega})$, $f \in L^2(\Omega)$. We also assume that the inequality $r - \frac{1}{2}\nabla \cdot c \geq r_0$ is valid in Ω, where $r_0 > 0$ is a constant. Then the variational formulation of (9.2) reads as follows:

Find $u \in V$ such that

$$a(u, v) = \langle f, v \rangle_0 \quad \text{for all } v \in V \,, \tag{9.14}$$

where a is the bilinear form (9.9).

Given a regular family of triangulations $\{\mathcal{T}_h\}$, let $V_h \subset V$ denote the set of continuous functions that are piecewise polynomial of degree $k \in \mathbb{N}$ and satisfy the boundary conditions, i.e.,

$$V_h := \left\{ v_h \in V \;\middle|\; v_h|_K \in \mathcal{P}_k(K) \text{ for all } K \in \mathcal{T}_h \right\} \,. \tag{9.15}$$

If in addition the solution $u \in V$ of (9.14) belongs to the space $H^{k+1}(\Omega)$, we have, by (3.87), the following error estimate for the interpolant $I_h(u)$:

$$\|u - I_h(u)\|_{l,K} \leq c_{\text{int}} h_K^{k+1-l} |u|_{k+1,K} \tag{9.16}$$

for $0 \leq l \leq k+1$ and all $K \in \mathcal{T}_h$. Since the spaces V_h are of finite dimension, a so-called *inverse inequality* can be proven (cf. Theorem 3.43, (2) and Exercise 9.3):

$$\|\Delta v_h\|_{0,K} \leq \frac{c_{\text{inv}}}{h_K} |v_h|_{1,K} \tag{9.17}$$

for all $v_h \in V_h$ and all $K \in \mathcal{T}_h$. Here it is important that the constants $c_{\text{int}}, c_{\text{inv}} > 0$ from (9.16) and (9.17), respectively, do not depend on u or v_h and on the particular elements $K \in \mathcal{T}_h$.

The basic idea of the streamline-diffusion method consists in the addition of suitably weighted residuals to the variational formulation (9.14). Because of the assumption $u \in H^{k+1}(\Omega)$, $k \in \mathbb{N}$, the differential equation can be interpreted as an equation in $L^2(\Omega)$. In particular, it is valid on any element $K \in \mathcal{T}_h$ in the sense of $L^2(K)$, i.e.,

$$-\varepsilon \Delta u + c \cdot \nabla u + ru = f \quad \text{almost everywhere in } K \text{ and for all } K \in \mathcal{T}_h \,.$$

Next we take an elementwise defined mapping $\tau : V_h \to L^2(\Omega)$ and multiply the local differential equation in $L^2(K)$ by the restriction of $\tau(v_h)$ to K. Scaling by a parameter $\delta_K \in \mathbb{R}$ and summing the results over all elements $K \in \mathcal{T}_h$, we obtain

$$\sum_{K \in \mathcal{T}_h} \delta_K \left\langle -\varepsilon \Delta u + c \cdot \nabla u + ru, \tau(v_h) \right\rangle_{0,K} = \sum_{K \in \mathcal{T}_h} \delta_K \left\langle f, \tau(v_h) \right\rangle_{0,K} .$$

If we add this relation to equation (9.14) restricted to V_h, we see that the weak solution $u \in V \cap H^{k+1}(\Omega)$ satisfies the following variational equation:

$$a_h(u, v_h) = \langle f, v_h \rangle_h \quad \text{for all } v_h \in V_h \,,$$

where

$$\begin{aligned} a_h(u, v_h) &:= a(u, v_h) + \sum_{K \in \mathcal{T}_h} \delta_K \left\langle -\varepsilon \Delta u + c \cdot \nabla u + ru, \tau(v_h) \right\rangle_{0,K} , \\ \langle f, v_h \rangle_h &:= \langle f, v \rangle_0 + \sum_{K \in \mathcal{T}_h} \delta_K \left\langle f, \tau(v_h) \right\rangle_{0,K} . \end{aligned}$$

Then the corresponding discretization reads as follows:

Find $u_h \in V_h$ such that

$$a_h(u_h, v_h) = \langle f, v_h \rangle_h \quad \text{for all } v_h \in V_h \,. \tag{9.18}$$

Corollary 9.2 *Suppose the problems* (9.14) *and* (9.18) *have a solution* $u \in V \cap H^{k+1}(\Omega)$ *and* $u_h \in V_h$, *respectively. Then the following error equation is valid:*

$$a_h(u - u_h, v_h) = 0 \quad \textit{for all } v_h \in V_h \,. \tag{9.19}$$

In the *streamline-diffusion method (sdFEM),* the mapping τ used in (9.18) is chosen as $\tau(v_h) := c \cdot \nabla v_h$.

Without going into details, we mention that a further option is to set $\tau(v_h) := -\varepsilon\Delta v_h + c \cdot \nabla v_h + r v_h$. This results in the so-called *Galerkin/least squares–FEM (GLSFEM)* [54].

Especially with regard to the extension of the method to other finite element spaces, the discussion of how to choose τ and δ_K is not yet complete.

Interpretation of the Additional Term in the Case of Linear Elements

If the finite element spaces V_h are formed by piecewise linear functions (i.e., in the above definition (9.15) of V_h we have $k = 1$), we get $\Delta v_h|_K = 0$ for all $K \in \mathcal{T}_h$. If in addition there is no reactive term (i.e., $r = 0$), the discrete bilinear form is

$$a_h(u_h, v_h) = \int_\Omega \varepsilon \nabla u_h \cdot \nabla v_h \, dx + \langle c \cdot \nabla u_h, v_h \rangle_0 + \sum_{K \in \mathcal{T}_h} \delta_K \, \langle c \cdot \nabla u_h, c \cdot \nabla v_h \rangle_{0,K} \, .$$

Since the scalar product appearing in the sum can be rewritten as $\langle c \cdot \nabla u_h, c \cdot \nabla v_h \rangle_{0,K} = \int_K (cc^T \nabla u_h) \cdot \nabla v_h \, dx$, we obtain the following equivalent representation:

$$a_h(u_h, v_h) = \sum_{K \in \mathcal{T}_h} \int_K \left((\varepsilon I + \delta_K cc^T) \nabla u_h \right) \cdot \nabla v_h \, dx + \langle c \cdot \nabla u_h, v_h \rangle_0 \, .$$

This shows that the additional term introduces an element-dependent extra diffusion in the direction of the convective field c (cf. also Exercise 0.3), which motivates the name of the method. In this respect, the streamline-diffusion method can be understood as an improved version of the full upwind method, as seen, for example, in (9.6).

Analysis of the Streamline-Diffusion Method

To start the analysis of stability and convergence properties of the streamline-diffusion method, we consider the term $a_h(v_h, v_h)$ for arbitrary $v_h \in V_h$.

As in Section 3.2.1, the structure of the discrete bilinear form a_h allows us to derive the estimate

$$a_h(v_h, v_h) \geq \varepsilon |v_h|_1^2 + r_0 \|v_h\|_0^2 + \sum_{K \in \mathcal{T}_h} \delta_K \, \langle -\varepsilon \Delta v_h + c \cdot \nabla v_h + r v_h, c \cdot \nabla v_h \rangle_{0,K} \, .$$

Furthermore, neglecting for a moment the second term in the sum and using the elementary inequality $ab \leq a^2 + b^2/4$ for arbitrary $a, b \in \mathbb{R}$, we

get

$$\begin{aligned}
&\left| \sum_{K\in\mathcal{T}_h} \delta_K \left\langle -\varepsilon\Delta v_h + r v_h, c\cdot\nabla v_h \right\rangle_{0,K} \right| \\
&\qquad \le \sum_{K\in\mathcal{T}_h} \left\{ \left| \left\langle -\varepsilon\sqrt{|\delta_K|}\,\Delta v_h, \sqrt{|\delta_K|}\,c\cdot\nabla v_h \right\rangle_{0,K} \right| \right. \\
&\qquad\qquad \left. + \left| \left\langle \sqrt{|\delta_K|}\,r v_h, \sqrt{|\delta_K|}\,c\cdot\nabla v_h \right\rangle_{0,K} \right| \right\} \\
&\qquad \le \sum_{K\in\mathcal{T}_h} \left\{ \varepsilon^2|\delta_K|\,\|\Delta v_h\|_{0,K}^2 + |\delta_K|\,\|r\|_{0,\infty,K}^2 \|v_h\|_{0,K}^2 \right. \\
&\qquad\qquad \left. + \frac{|\delta_K|}{2}\,\|c\cdot\nabla v_h\|_{0,K}^2 \right\}.
\end{aligned}$$

By means of the inverse inequality (9.17) it follows that

$$\begin{aligned}
\left| \sum_{K\in\mathcal{T}_h} \delta_K \left\langle -\varepsilon\Delta v_h + r v_h, c\cdot\nabla v_h \right\rangle_{0,K} \right| &\le \sum_{K\in\mathcal{T}_h} \left\{ \varepsilon^2|\delta_K|\,\frac{c_{\mathrm{inv}}^2}{h_K^2}|v_h|_{1,K}^2 \right. \\
&\left. + |\delta_K|\,\|r\|_{0,\infty,K}^2\|v_h\|_{0,K}^2 + \frac{|\delta_K|}{2}\,\|c\cdot\nabla v_h\|_{0,K}^2 \right\}.
\end{aligned}$$

Putting things together, we obtain

$$\begin{aligned}
a_h(v_h,v_h) \;\ge\; &\sum_{K\in\mathcal{T}_h} \left\{ \left(\varepsilon - \varepsilon^2|\delta_K|\,\frac{c_{\mathrm{inv}}^2}{h_K^2} \right) |v_h|_{1,K}^2\|v_h\|_{0,K}^2 \right. \\
&\left. + \left(r_0 - |\delta_K|\,\|r\|_{0,\infty,K}^2\right) + \left(\delta_K - \frac{|\delta_K|}{2} \right) \|c\cdot\nabla v_h\|_{0,K}^2 \right\}.
\end{aligned}$$

The choice

$$0 < \delta_K \le \frac{1}{2}\min\left\{ \frac{h_K^2}{\varepsilon c_{\mathrm{inv}}^2}, \frac{r_0}{\|r\|_{0,\infty,K}^2} \right\} \tag{9.20}$$

leads to

$$a_h(v_h,v_h) \ge \frac{\varepsilon}{2}|v_h|_1^2 + \frac{r_0}{2}\|v_h\|_0^2 + \frac{1}{2}\sum_{K\in\mathcal{T}_h} \delta_K\|c\cdot\nabla v_h\|_{0,K}^2 .$$

Therefore, if the so-called *streamline-diffusion norm* is defined by

$$\|v\|_{\mathrm{sd}} := \left\{ \varepsilon|v|_1^2 + r_0\|v\|_0^2 + \sum_{K\in\mathcal{T}_h} \delta_K\|c\cdot\nabla v\|_{0,K}^2 \right\}^{1/2}, \quad v\in V,$$

then the choice (9.20) implies the estimate

$$\frac{1}{2}\|v_h\|_{\mathrm{sd}}^2 \le a_h(v_h,v_h) \quad \text{for all } v_h\in V_h . \tag{9.21}$$

Obviously, the streamline-diffusion norm $\|\cdot\|_{\rm sd}$ is stronger than the ε-weighted H^1-norm (9.10); i.e.,

$$\min\{1, \sqrt{r_0}\}\|v\|_\varepsilon \le \|v\|_{\rm sd} \quad \text{for all } v \in V\,.$$

Now an error estimate can be proven. Since estimate (9.21) holds only on the finite element spaces V_h, we consider first the norm of $I_h(u) - u_h \in V_h$ and make use of the error equation (9.19):

$$\frac{1}{2}\|I_h(u) - u_h\|_{\rm sd}^2 \le a_h(I_h(u) - u_h, I_h(u) - u_h) = a_h(I_h(u) - u, I_h(u) - u_h)\,.$$

In particular, under the assumption $u \in V \cap H^{k+1}(\Omega)$ the following three estimates are valid:

$$\begin{aligned}\varepsilon \int_\Omega \nabla(I_h(u) - u) \cdot \nabla(I_h(u) - u_h)\,dx &\le \sqrt{\varepsilon}\,|I_h(u) - u|_1 \|I_h(u) - u_h\|_{\rm sd} \\ &\le c_{\rm int}\sqrt{\varepsilon}\,h^k |u|_{k+1}\|I_h(u) - u_h\|_{\rm sd}\,,\end{aligned}$$

$$\begin{aligned}&\int_\Omega [c \cdot \nabla(I_h(u) - u) + r(I_h(u) - u)](I_h(u) - u_h)\,dx \\ &\quad= \int_\Omega (r - \nabla\cdot c)(I_h(u) - u)(I_h(u) - u_h)\,dx \\ &\qquad - \int_\Omega (I_h(u) - u)\, c \cdot \nabla(I_h(u) - u_h)\,dx \\ &\quad\le \|r - \nabla\cdot c\|_{0,\infty}\|I_h(u) - u\|_0 \|I_h(u) - u_h\|_0 \\ &\qquad + \|I_h(u) - u\|_0 \|c \cdot \nabla(I_h(u) - u_h)\|_0 \\ &\quad\le C\left\{\left\{\sum_{K\in\mathcal{T}_h} \|I_h(u) - u\|_{0,K}^2\right\}^{1/2}\right. \\ &\qquad \left. + \left\{\sum_{K\in\mathcal{T}_h} \delta_K^{-1}\|I_h(u) - u\|_{0,K}^2\right\}^{1/2}\right\} \|I_h(u) - u_h\|_{\rm sd} \\ &\quad\le Ch^k \left\{\sum_{K\in\mathcal{T}_h} \left(1 + \delta_K^{-1}\right) h_K^2 |u|_{k+1,K}^2\right\}^{1/2} \|I_h(u) - u_h\|_{\rm sd}\,,\end{aligned}$$

and

$$\begin{aligned}\Bigg| \sum_{K\in\mathcal{T}_h} \delta_K \langle -\varepsilon\Delta(I_h(u) - u) + c \cdot \nabla(I_h(u) - u) \\ + r(I_h(u) - u), c \cdot \nabla(I_h(u) - u_h)\rangle_{0,K} \Bigg|\end{aligned} \tag{9.22}$$

$$\leq \sum_{K\in\mathcal{T}_h} c_{\rm int}\sqrt{\delta_K}\,\left[\varepsilon h_K^{k-1}+\|c\|_{0,\infty,K}h_K^k+\|r\|_{0,\infty,K}h_K^{k+1}\right] \times |u|_{k+1,K}\sqrt{\delta_K}\,\|c\cdot\nabla(I_h(u)-u_h)\|_{0,K}$$

$$\leq C\left\{\sum_{K\in\mathcal{T}_h}\delta_K\left[\varepsilon h_K^{k-1}+h_K^k+h_K^{k+1}\right]^2|u|_{k+1,K}^2\right\}^{1/2}\|I_h(u)-u_h\|_{\rm sd}\,.$$

Condition (9.20), which was already required for estimate (9.21), implies that

$$\varepsilon\delta_K\leq\frac{h_K^2}{c_{\rm inv}^2},$$

and so the application to the first term of the last bound leads to

$$\left|\sum_{K\in\mathcal{T}_h}\delta_K\big\langle-\varepsilon\Delta(I_h(u)-u)+c\cdot\nabla(I_h(u)-u) +r(I_h(u)-u),c\cdot\nabla(I_h(u)-u_h)\big\rangle_{0,K}\right| \leq Ch^k\left\{\sum_{K\in\mathcal{T}_h}\left[\varepsilon+\delta_K\right]|u|_{k+1,K}^2\right\}^{1/2}\|I_h(u)-u_h\|_{\rm sd}\,.$$

Collecting the estimates and dividing by $\|I_h(u)-u_h\|_{\rm sd}$, we obtain the relation

$$\|I_h(u)-u_h\|_{\rm sd}\leq Ch^k\left\{\sum_{K\in\mathcal{T}_h}\left[\varepsilon+\frac{h_K^2}{\delta_K}+h_K^2+\delta_K\right]|u|_{k+1,K}^2\right\}^{1/2}.$$

Finally, the terms in the square brackets will be equilibrated with the help of condition (9.20). We rewrite the ε-dependent term in this condition as

$$\frac{h_K^2}{\varepsilon c_{\rm inv}^2}=\frac{2}{c_{\rm inv}^2\|c\|_{\infty,K}}\,{\rm Pe}_K h_K$$

with

$${\rm Pe}_K:=\frac{\|c\|_{\infty,K}h_K}{2\varepsilon}\,. \tag{9.23}$$

This *local Péclet number* is a refinement of the definition (9.4).

The following distinctions concerning ${\rm Pe}_K$ are convenient:

$${\rm Pe}_K\leq 1\quad\text{and}\quad{\rm Pe}_K>1\,.$$

In the first case, we choose

$$\delta_K=\delta_0{\rm Pe}_K h_K=\delta_1\frac{h_K^2}{\varepsilon}\,,\quad\delta_0=\frac{2}{\|c\|_{\infty,K}}\,\delta_1\,,$$

with appropriate constants $\delta_0 > 0$ and $\delta_1 > 0$, respectively, which are independent of K and ε. Then we have

$$\varepsilon + \frac{h_K^2}{\delta_K} + h_K^2 + \delta_K = \left(1 + \frac{1}{\delta_1}\right)\varepsilon + h_K^2 + \delta_1 \frac{2\mathrm{Pe}_K}{\|c\|_{0,\infty,K}} h_K \leq C(\varepsilon + h_K)\,,$$

where $C > 0$ is independent of K and ε. In the second case, it is sufficient to choose $\delta_K = \delta_2 h_K$ with an appropriate constant $\delta_2 > 0$ that is independent of K and ε. Then

$$\delta_K = \frac{\delta_2}{\mathrm{Pe}_K}\mathrm{Pe}_K h_K = \frac{\delta_2 \|c\|_{0,\infty,K}}{2\mathrm{Pe}_K}\,\frac{h_K^2}{\varepsilon}$$

and

$$\varepsilon + \frac{h_K^2}{\delta_K} + h_K^2 + \delta_K = \varepsilon + \left(\frac{1}{\delta_2} + \delta_2\right) h_K + h_K^2 \leq C(\varepsilon + h_K)\,,$$

with $C > 0$ independent of K and ε. Note that in both cases the constants can be chosen sufficiently small, independent of Pe_K, that the condition (9.20) is satisfied. Now we are prepared to prove the following error estimate.

Theorem 9.3 *Let the parameters δ_K be given by*

$$\delta_K = \begin{cases} \delta_1 \dfrac{h_K^2}{\varepsilon}\,, & Pe_K \leq 1\,, \\ \delta_2 h_K\,, & Pe_K > 1\,, \end{cases}$$

where $\delta_1, \delta_2 > 0$ do not depend on K and ε and are chosen such that condition (9.20) *is satisfied. If the weak solution u of* (9.14) *belongs to $H^{k+1}(\Omega)$, then*

$$\|u - u_h\|_{\mathrm{sd}} \leq C\left(\sqrt{\varepsilon} + \sqrt{h}\right) h^k |u|_{k+1},$$

where the constant $C > 0$ is independent of ε, h, and u.

Proof: By the triangle inequality, we get

$$\|u - u_h\|_{\mathrm{sd}} \leq \|u - I_h(u)\|_{\mathrm{sd}} + \|I_h(u) - u_h\|_{\mathrm{sd}}\,.$$

An estimate of the second addend is already known. To deal with the first term, the estimates of the interpolation error (9.16) are used directly:

$$\begin{aligned} &\|u - I_h(u)\|_{\mathrm{sd}}^2 \\ &\quad = \varepsilon |u - I_h(u)|_1^2 + r_0 \|u - I_h(u)\|_0^2 + \sum_{K \in \mathcal{T}_h} \delta_K \|c \cdot \nabla(u - I_h(u))\|_{0,K}^2 \\ &\quad \leq c_{\mathrm{int}}^2 \sum_{K \in \mathcal{T}_h} \left[\varepsilon h_K^{2k} + r_0 h_K^{2(k+1)} + \delta_K \|c\|_{0,\infty,K}^2 h_K^{2k}\right] |u|_{k+1,K}^2 \end{aligned}$$

$$\leq \quad Ch_K^{2k} \sum_{K \in \mathcal{T}_h} \left[\varepsilon + h_K^2 + \delta_K\right] |u|_{k+1,K}^2 \leq C(\varepsilon + h) h_K^{2k} |u|_{k+1}^2 \,.$$

□

Remark 9.4 (i) In the case of large local Péclet numbers, we have $\varepsilon \leq \frac{1}{2}\|c\|_{\infty,K} h_K$ and thus

$$\|u - u_h\|_0 + \left\{ \delta_2 \sum_{K \in \mathcal{T}_h} h_K \|c \cdot \nabla(u - u_h)\|_{0,K}^2 \right\}^{1/2} \leq C h^{k+1/2} |u|_{k+1} \,.$$

So the L^2-error of the solution is not optimal in comparison with the estimate of the interpolation error

$$\|u - I_h(u)\|_0 \leq C h^{k+1} |u|_{k+1} \,,$$

whereas the L^2-error of the directional derivative of u in the direction of c is optimal.

(ii) In general, the seminorm $|u|_{k+1}$ depends on negative powers of ε. Therefore, if $h \to 0$, the convergence in Theorem 9.3 is not uniform with respect to ε.

Comparing the estimate from Theorem 9.3 for the special case of continuous linear elements with the estimate (9.13) for the corresponding standard method given at the end of the introduction, i.e.,

$$\|u - u_h\|_\varepsilon \leq C h |u|_2 \,,$$

we see that the error of the streamline-diffusion method is measured in a stronger norm than the $\|\cdot\|_\varepsilon$-norm and additionally, that the error bound is asymptotically better in the interesting case $\varepsilon < h$. A further advantage of the streamline-diffusion method is to be seen in the fact that its implementation is not much more difficult than that of the standard finite element method.

However, there are also some disadvantages: Since the error bound involves the H^{k+1}-seminorm of the solution u, it may depend on negative powers of ε. Furthermore, there is no general rule to determine the parameters δ_1, δ_2. Usually, they are chosen more or less empirically. This may be a problem when the streamline-diffusion method is embedded into more complex programs (for example, for solving nonlinear problems). Finally, in contrast to the finite volume methods described in the next section, the property of inverse monotonicity (cf. Theorem 6.19) cannot be proven in general.

Exercises

9.1 (a) Given a constant diffusion coefficient $\varepsilon > 0$, rewrite the ordinary boundary value problem

$$\begin{aligned} (-\varepsilon u' + u)' &= 0 \quad \text{in } \Omega := (0,1)\,, \\ u(0) = u(1) - 1 &= 0\,, \end{aligned}$$

into an equivalent form but with nonnegative right-hand side and homogeneous Dirichlet boundary conditions.

(b) Compute the $H^2(0,1)$-seminorm of the solution of the transformed problem and investigate its dependence on ε.

9.2 Prove the error equation of the streamline-diffusion method (Corollary 9.2).

9.3 Given an arbitrary, but fixed, triangle K with diameter h_K, prove the inequality

$$\|\Delta p\|_{0,K} \le \frac{c_{\text{inv}}}{h_K} |p|_{1,K}$$

for arbitrary polynomials $p \in \mathcal{P}_k(K)$, $k \in \mathbb{N}$, where the constant $c_{\text{inv}} > 0$ is independent of K and p.

9.4 Verify that the streamline-diffusion norm $\|\cdot\|_{\text{sd}}$ is indeed a norm.

9.3 Finite Volume Methods

In the convection-dominated situation, the finite volume method introduced in Chapter 6 proves to be a very stable, but not so accurate, method. One reason for this stability lies in an appropriate asymptotic behaviour of the weighting function R for large absolute values of its argument.

Namely, if we consider the examples of nonconstant weighting functions given in Section 6.2.2, we see that

$$\text{(P4)} \qquad \lim_{z \to -\infty} R(z) = 0\,, \quad \lim_{z \to \infty} R(z) = 1\,.$$

In the general case of the model problem (6.5) with $k = \varepsilon > 0$, (P4) implies that for $\dfrac{\gamma_{ij} d_{ij}}{\varepsilon} \ll -1$ the term $r_{ij} u_i + (1 - r_{ij}) u_j$ in the bilinear form b_h effectively equals u_j, whereas in the case $\dfrac{\gamma_{ij} d_{ij}}{\varepsilon} \gg 1$ the quantity u_i remains.

In other words, in the case of dominating convection, the approximation b_h evaluates the "information" (u_j or u_i) *upwind*, i.e., just at that node (a_j or a_i) from which "the flow is coming".

This essentially contributes to the stabilization of the method and makes it possible to prove properties such as global conservativity or inverse monotonicity (cf. Section 6.2.4) *without* any restrictions on the size of the local Péclet number $\frac{\gamma_{ij} d_{ij}}{\varepsilon}$ and thus without any restrictions on the ratio of h and ε. This local Péclet number (note the missing factor 2 in comparison to (9.23)) also takes the direction of the flow compared to the edge $a_i a_j$ into account.

The Choice of Weighting Parameters

In order to motivate the choice of the weighting parameters in the case of the Voronoi diagram, we recall the essential step in the derivation of the finite volume method, namely the approximation of the integral

$$I_{ij} := \int_{\Gamma_{ij}} [\mu_{ij} \, (\nu_{ij} \cdot \nabla u) - \gamma_{ij} \, u] \, d\sigma \, .$$

It first suggests itself to apply a simple quadrature rule, for example

$$I_{ij} \approx q_{ij} m_{ij} \, ,$$

where q_{ij} denotes the value of the expression to be integrated at the point a_{ij} of the intersection of the boundary segment Γ_{ij} with the edge bounded by the vertices a_i and a_j (i.e., $2a_{ij} = a_i + a_j$). Next, if this edge is parametrised according to

$$x = x(\tau) = a_{ij} + \tau d_{ij} \nu_{ij} \, , \quad \tau \in \left[-\frac{1}{2}, \frac{1}{2} \right] \, ,$$

and if we introduce the composite function $w(\tau) := u(x(\tau))$, then we can write

$$\mu_{ij} \, (\nu_{ij} \cdot \nabla u) - \gamma_{ij} \, u = q(0) \quad \text{with} \quad q(\tau) := \frac{\mu_{ij}}{d_{ij}} \frac{dw}{d\tau}(\tau) - \gamma_{ij} w(\tau) \, .$$

The relation defining the function q can be interpreted as a linear ordinary differential equation for the unknown function $w : \left[-\frac{1}{2}, \frac{1}{2} \right] \to \mathbb{R}$. Provided that q is continuous on the interval $\left[-\frac{1}{2}, \frac{1}{2} \right]$, the equation can be solved exactly:

$$\begin{aligned} w(\tau) \quad = \quad & \left\{ \frac{d_{ij}}{\mu_{ij}} \int_{-1/2}^{\tau} q(s) \exp\left(-\frac{\gamma_{ij} d_{ij}}{\mu_{ij}} \left(s + \frac{1}{2} \right) \right) ds + w\left(-\frac{1}{2} \right) \right\} \\ & \times \exp\left(\frac{\gamma_{ij} d_{ij}}{\mu_{ij}} \left(\tau + \frac{1}{2} \right) \right) \, . \end{aligned}$$

Approximating q by a constant q_{ij}, we get in the case $\gamma_{ij} \neq 0$,

$$\begin{aligned} w(\tau) \quad \approx \quad & \left\{ \frac{q_{ij}}{\gamma_{ij}} \left[1 - \exp\left(-\frac{\gamma_{ij} d_{ij}}{\mu_{ij}} \left(\tau + \frac{1}{2} \right) \right) \right] + w\left(-\frac{1}{2} \right) \right\} \\ & \times \exp\left(\frac{\gamma_{ij} d_{ij}}{\mu_{ij}} \left(\tau + \frac{1}{2} \right) \right) \, . \end{aligned}$$

In particular,

$$w\left(\frac{1}{2}\right) \approx \left\{ \frac{q_{ij}}{\gamma_{ij}} \left[1 - \exp\left(-\frac{\gamma_{ij} d_{ij}}{\mu_{ij}} \right) \right] + w\left(-\frac{1}{2}\right) \right\} \exp\left(\frac{\gamma_{ij} d_{ij}}{\mu_{ij}} \right) ; \quad (9.24)$$

that is, the approximation q_{ij} of $q(0)$ can be expressed by means of the values $w(\pm\frac{1}{2})$:

$$q_{ij} \approx \gamma_{ij} \frac{w(\frac{1}{2}) - w(-\frac{1}{2}) \exp\left(\frac{\gamma_{ij} d_{ij}}{\mu_{ij}} \right)}{\exp\left(\frac{\gamma_{ij} d_{ij}}{\mu_{ij}} \right) - 1} . \quad (9.25)$$

In the case $\gamma_{ij} = 0$, it immediately follows from the exact solution and the approximation $q \approx q_{ij}$ that

$$q_{ij} \approx \mu_{ij} \frac{w(\frac{1}{2}) - w(-\frac{1}{2})}{d_{ij}} .$$

Since this is equal to the limit of (9.25) for $\gamma_{ij} \to 0$, we can exclusively work with the representation (9.25).

If we define the weighting function $R : \mathbb{R} \to [0, 1]$ by

$$R(z) := 1 - \frac{1}{z} \left(1 - \frac{z}{e^z - 1} \right) , \quad (9.26)$$

then with the choice $r_{ij} := R\left(\frac{\gamma_{ij} d_{ij}}{\mu_{ij}} \right)$, (9.25) can be written as

$$q_{ij} \approx \mu_{ij} \frac{u_j - u_i}{d_{ij}} - [r_{ij} u_i + (1 - r_{ij})\, u_j]\, \gamma_{ij} .$$

A simple algebraic manipulation shows that this is exactly the approximation scheme given in Section 6.2.

The use of the weighting function (9.26) yields a discretization method that can be interpreted as a generalization of the so-called Il'in–Allen–Southwell scheme. However, in order to avoid the comparatively expensive computation of the function values r_{ij} of (9.26), often simpler functions $R : \mathbb{R} \to [0, 1]$ are used (see Section 6.2.2), which are to some extent approximations of (9.26) keeping the properties (P1) to (P4).

At the end of this paragraph we will illustrate the importance of the properties (P1) to (P3), especially for convection-dominated problems. Property (P2) has been used in the proof of the basic stability estimate (6.20). On the other hand, we have seen at several places (e.g., in Section 1.4 or in Chapter 5) that the matrix A_h of the corresponding system of linear algebraic equations should have positive diagonal entries. For example, if in the differential equation from (9.2) the reaction term disappears, then properties (P1) and (P3) guarantee that the diagonal entries are at least nonnegative. This can be seen as follows:

From (6.9) we conclude the following formula:

$$(A_h)_{ii} = \left[\frac{\mu_{ij}}{d_{ij}} + \gamma_{ij} r_{ij}\right] m_{ij} = \frac{\mu_{ij}}{d_{ij}} \left[1 + \frac{\gamma_{ij} d_{ij}}{\mu_{ij}} r_{ij}\right] m_{ij}, \quad i \in \Lambda\,.$$

If we replace in property (P3) the number z by $-z$, then we get, by property (P1),

$$0 \le 1 + [1 - R(-z)]\, z = 1 + zR(z)\,.$$

Therefore, if the weighting function R satisfies (P1) and (P3), then we have that $(A_h)_{ii} \ge 0$ for all $i \in \Lambda$.

The simple choice $r_{ij} \equiv \frac{1}{2}$ does not satisfy property (P3). In this case, the condition $(A_h)_{ii} \ge 0$ leads to the requirement

$$-\frac{\gamma_{ij} d_{ij}}{2\mu_{ij}} \le 1\,,$$

which in the case $\gamma_{ij} \le 0$, i.e., for a local flow from a_j to a_i, is a restriction to the ratio of h and ε, and this is analogous to the condition (9.5) on the grid Péclet number, where only the sizes of K, c, and h enter.

Similarly, it can be shown that property (P3) implies the nonpositivity of the off-diagonal entries of A_h.

An Error Estimate

At the end of this section an error estimate will be cited, which can be derived similarly to the corresponding estimate of the standard method. The only special aspect is that the dependence of the occurring quantities on ε is carefully tracked (see [40]).

Theorem 9.5 *Let $\{\mathcal{T}_h\}_h$ be a regular family of conforming triangulations, all triangles of which are nonobtuse. Furthermore, in addition to the assumptions on the coefficients of the bilinear form* (9.9)*, let $f \in C^1(\overline{\Omega})$.*

If the exact solution u of the model problem belongs to $H^2(\Omega)$ and if $u_h \in V_h$ denotes the approximative solution of the finite volume method (6.11)*, where the approximations γ_{ij}, respectively r_i, are chosen according to* (6.7)*, respectively* (6.8)*, then for sufficiently small $\bar{h} > 0$ the estimate*

$$\|u - u_h\|_\varepsilon \le C\, \frac{h}{\sqrt{\varepsilon}} \left[\|u\|_2 + |f|_{1,\infty}\right], \quad h \in (0, \bar{h}]\,,$$

holds, where both the constant $C > 0$ and $\bar{h} > 0$ do not depend on ε.

In special, but practically not so relevant, cases (for example, if the triangulations are of Friedrichs–Keller type), it is possible to remove the factor $\frac{1}{\sqrt{\varepsilon}}$ in the bound above.

Comparing the finite volume method with the streamline-diffusion method, we see that the finite volume method is less accurate. However, it is globally conservative and inverse monotone.

Exercise

9.5 Using an equidistant grid, formulate both the streamline-diffusion method and the finite volume method for a one-dimensional model problem ($d = 1$, $\Omega = (0,1)$, $r = 0$) with constant coefficients and compare the resulting discretizations. Based on that comparison, what can be said about the choice of the parameters in the streamline-diffusion method?

9.4 The Lagrange–Galerkin Method

In the previous sections, discretization methods for stationary diffusion-convection equations were presented. In conjunction with the method of lines, these methods can also be applied to parabolic problems. However, since the method of lines decouples spatial and temporal variables, it cannot be expected that the peculiarities of nonstationary diffusion-convection equations are reflected adequately.

The so-called *Lagrange–Galerkin method* attempts to bypass this problem by means of an intermediate change from the Eulerian coordinates (considered up to now) to the so-called *Lagrangian coordinates*. The latter are chosen in such a way that the origin of the coordinate system (i.e., the position of the observer) is moved with the convective field, and in the new coordinates no convection occurs.

To illustrate the basic idea, the following initial-boundary value problem will be considered, where $\Omega \subset \mathbb{R}^d$ is a bounded domain with Lipschitz continuous boundary and $T > 0$:

For given functions $f : Q_T \to \mathbb{R}$ and $u_0 : \Omega \to \mathbb{R}$, find a function $u : Q_T \to \mathbb{R}$ such that

$$\begin{aligned} \frac{\partial u}{\partial t} + Lu &= f \quad \text{in} \quad Q_T\,, \\ u &= 0 \quad \text{on} \quad S_T\,, \\ u &= u_0 \quad \text{on} \quad \Omega \times \{0\}\,, \end{aligned} \tag{9.27}$$

where

$$(Lu)(x,t) := -\nabla\cdot(K(x)\,\nabla u(x,t)) + c(x,t)\cdot\nabla u(x,t) + r(x,t)u(x,t), \tag{9.28}$$

with sufficiently smooth coefficients

$$K : \Omega \to \mathbb{R}^{d,d}\,, \quad c : Q_T \to \mathbb{R}^d\,, \quad r : Q_T \to \mathbb{R}\,.$$

As usual, the differential operators ∇ and $\nabla\cdot$ act only with respect to the spatial variables.

The new coordinate system is obtained by solving the following parameter-dependent auxiliary problem:

Given $(x,s) \in \overline{Q}_T$, find a vector field $X : \overline{\Omega} \times [0,T]^2 \to \mathbb{R}^d$ such that

$$\begin{aligned} \frac{d}{dt}X(x,s,t) &= c(X(x,s,t),t)\,, \quad t \in (0,T)\,, \\ X(x,s,s) &= x\,. \end{aligned} \tag{9.29}$$

The trajectories $X(x,s,\cdot)$ are called *characteristics* (through (x,s)). If c is continuous on $\overline{Q}_T$ and, for fixed $t \in [0,T]$, Lipschitz continuous with respect to the first argument on $\overline{\Omega}$, then there exists a unique solution $X = X(x,s,t)$. Denoting by u the sufficiently smooth solution of (9.27) and setting

$$\hat{u}(x,t) := u(X(x,s,t),t) \quad \text{for fixed } s \in [0,T]\,,$$

then the chain rule implies that

$$\frac{\partial \hat{u}}{\partial t}(x,t) = \left(\frac{\partial u}{\partial t} + c \cdot \nabla u\right)(X(x,s,t),t)\,.$$

The particular value

$$\frac{\partial \hat{u}}{\partial t}(x,s) = \frac{\partial u}{\partial t}(x,s) + c(x,s) \cdot \nabla u(x,s)$$

is called the *material derivative* of u at (x,s). Thus the differential equation reads as

$$\frac{\partial \hat{u}}{\partial t} - \nabla \cdot (K \nabla u) + ru = f\,;$$

i.e., it is formally free of any convective terms.

Now the equation will be semidiscretized by means of the horizontal method of lines. A typical way is to approximate the time derivative by backward difference quotients. So let an equidistant partition of the time interval $(0,T)$ with step size $\tau := T/N$, $N \in \mathbb{N}$ (provided that $T < \infty$), be given.

Tracking the characteristics backwards in time, in the strip $\Omega \times [t_n, t_{n+1})$, $n \in \{0,1,\ldots,N-1\}$, with $x = X(x,t_{n+1},t_{n+1})$ the following approximation results:

$$\frac{\partial \hat{u}}{\partial t} \approx \frac{1}{\tau}[\hat{u}(x,t_{n+1}) - \hat{u}(x,t_n)] = \frac{1}{\tau}[u(x,t_{n+1}) - u(X(x,t_{n+1},t_n),t_n)]\,.$$

Further, if V_h denotes a finite-dimensional subspace of V in which we want to find the approximations to $u(\cdot,t_n)$, the method reads as follows:

Given $u_{0h} \in V_h$, find an element $U^{n+1} \in V_h$, $n \in \{0,\ldots,N-1\}$, such that

$$\begin{aligned} &\frac{1}{\tau}\left\langle U^{n+1} - U^n(X(\cdot,t_{n+1},t_n)), v_h\right\rangle_0 \\ &\quad + \left\langle K\nabla U^{n+1} \cdot \nabla v_h, 1\right\rangle_0 + \left\langle r(\cdot,t_{n+1})U^{n+1}, v_h\right\rangle_0 = \langle f(\cdot,t_{n+1}), v_h\rangle_0 \\ &\qquad\qquad \text{for all } v_h \in V_h, \\ &\qquad\qquad U^0 = u_{0h}\,. \end{aligned} \tag{9.30}$$

A possible extension of the method is to use time-dependent subspaces; that is, given a sequence of subspaces $V_h^n \subset V$, $n \in \{0, \ldots, N\}$, the approximations U^n to $u(\cdot, t_n)$ are chosen from V_h^n.

So the basic idea of the Lagrange–Galerkin method, namely, the elimination of convective terms by means of an appropriate transformation of coordinates, allows the application of standard discretization methods and makes the method attractive for situations where convection is dominating.

In fact, there exists a whole variety of papers dealing with error estimates for the method in the convection-dominated case, but often under the condition that the system (9.29) is integrated exactly.

In practice, the exact integration is impossible, and the system (9.29) has to be solved numerically (cf. [61]). This may lead to stability problems, so there is still a considerable need in the theoretical foundation of Lagrange–Galerkin methods.

Only recently it has been possible for a model situation to prove order of convergence estimates uniformly in the Péclet number for (9.30) (see [43]). The key is the consequent use of Lagrangian coordinates, revealing that (9.30) is just the application of the implicit Euler method to an equation arising from a transformation by characteristics defined piecewise backward in time. This equation is a pure diffusion problem, but with a coefficient reflecting the transformation. In conjunction with the backward Euler method this is not visible in the elliptic part to be discretized. Thus tracking the characteristics backward in time turns out to be important.

A
Appendices

A.1 Notation

$\mathbb{C}$	set of complex numbers
$\mathbb{N}$	set of natural numbers
$\mathbb{N}_0$	$:= \mathbb{N} \cup \{0\}$
$\mathbb{Q}$	set of rational numbers
$\mathbb{R}$	set of real numbers
$\mathbb{R}_+$	set of positive real numbers
$\mathbb{Z}$	set of integers
$\Re z$	real part of the complex number z
$\Im z$	imaginary part of the complex number z
x^T	transpose of the vector $x \in \mathbb{R}^d,\ d \in \mathbb{N}$
$\lvert x\rvert_p$	$:= \left(\sum_{j=1}^d \lvert x_j\rvert^p\right)^{1/p}, x = (x_1, \ldots, x_d)^T \in \mathbb{R}^d,\ d \in \mathbb{N},\ p \in [1, \infty)$
$\lvert x\rvert_\infty$	$:= \max_{j=1,\ldots,d} \lvert x_j\rvert$ maximum norm of the vector $x \in \mathbb{R}^d,\ d \in \mathbb{N}$
$\lvert x\rvert$	$:= \lvert x\rvert_2$ Euclidean norm of the vector $x \in \mathbb{R}^d,\ d \in \mathbb{N}$
$x \cdot y$	$:= x^T y = \sum_{j=1}^d x_j y_j$ scalar product of the vectors $x, y \in \mathbb{R}^d$
$\langle x, y\rangle_A$	$:= y^T A x = y \cdot Ax$ energy product of the vectors $x, y \in \mathbb{R}^d$ w.r.t. a symmetric, positive definite matrix A
$\lvert\alpha\rvert$	$:= \lvert\alpha\rvert_1$ order (or length) of the multi-index $\alpha \in \mathbb{N}_0^d,\ d \in \mathbb{N}$
I	identity matrix or identity operator
e_j	jth unit vector in $\mathbb{R}^m$, $j = 1, \ldots, m$
$\operatorname{diag}(\lambda_i)$	$= \operatorname{diag}(\lambda_1, \ldots, \lambda_m)$ diagonal matrix in $\mathbb{R}^{m,m}$ with diagonal entries $\lambda_1, \ldots, \lambda_m \in \mathbb{C}$

A^T	transpose of the matrix A
A^{-T}	transpose of the inverse matrix A^{-1}
$\det A$	determinant of the square matrix A
$\lambda_{\min}(A)$	minimum eigenvalue of a matrix A with real eigenvalues
$\lambda_{\max}(A)$	maximum eigenvalue of a matrix A with real eigenvalues
$\sigma(A)$	set of eigenvalues (spectrum) of the square matrix A
$\varrho(A)$	spectral radius of the square matrix A
$m(A)$	bandwidth of the symmetric matrix A
$\mathrm{Env}\,(A)$	hull of the square matrix A
$p(A)$	profile of the square matrix A
$B_\varrho(x_0)$	$:= \{x : \|x - x_0\| < \varrho\}$ open ball in a normed space
$\overline{B}_\varrho(x_0)$	$:= \{x : \|x - x_0\| \le \varrho\}$ closed ball in a normed space
$\mathrm{diam}\,(G)$	diameter of the set $G \subset \mathbb{R}^d$
$\|G\|_n$	n-dimensional (Lebesgue) measure of the $G \subset \mathbb{R}^n$, $n \in \{1, \ldots, d\}$
$\|G\|$	$:= \|G\|_d$ d-dimensional (Lebesgue) measure of the set $G \subset \mathbb{R}^d$
$\mathrm{vol}\,(G)$	length ($d = 1$), area ($d = 2$), volume ($d = 3$) of "geometric bodies" $G \subset \mathbb{R}^d$
$\mathrm{int}\,G$	interior of the set G
∂G	boundary of the set G
$\overline{G}$	closure of the set G
$\mathrm{span}\,G$	linear hull of the set G
$\mathrm{conv}\,G$	convex hull of the set G
$\|G\|$	cardinal number of the discrete set G
ν	outer unit normal w.r.t. the set $G \subset \mathbb{R}^d$
Ω	domain of $\mathbb{R}^d$, $d \in \mathbb{N}$
Γ	$:= \partial\Omega$ boundary of the domain $\Omega \subset \mathbb{R}^d$
$\mathrm{supp}\,\varphi$	support of the function φ
f^{-1}	inverse of the mapping f
$f[G]$	image of the set G under the mapping f
$f^{-1}[G]$	preimage of the set G under the mapping f
$f\|_K$	restriction of $f : G \to \mathbb{R}$ to a subset $K \subset G$
$\|v\|_X$	norm of the element v of the normed space X
$\dim X$	dimension of the finite-dimensional linear space X
$L[X, Y]$	set of linear, continuous operators acting from the normed space X in the normed space Y
X'	$:= L[X, \mathbb{R}]$ dual space of the real normed space X
$O(\cdot)$, $o(\cdot)$	Landau symbols of asymptotic analysis
δ_{ij}	$(i, j \in \mathbb{N}_0)$ Kronecker symbol, i.e., $\delta_{ii} = 1$ and $\delta_{ij} = 0$ if $i \neq j$

Differential expressions

∂_l	$(l \in \mathbb{N})$ symbol for the partial derivative w.r.t. the lth variable
∂_t	$(t \in \mathbb{R})$ symbol for the partial derivative w.r.t. the variable t
∂^α	$(\alpha \in \mathbb{N}_0^d$ multi-index) αth partial derivative
∇	$:= (\partial_1, \ldots, \partial_d)^T$ Nabla operator (symbolic vector)

Δ	Laplace operator
∂_μ	$:= \mu \cdot \nabla$ directional derivative w.r.t. the vector μ
$D\Phi$	$:= \frac{\partial \Phi}{\partial x} := (\partial_j \Phi_i)_{i,j=1}^m$ Jacobi matrix or functional matrix of a differentiable mapping $\Phi : \mathbb{R}^m \to \mathbb{R}^m$

Coefficients in differential expressions

K	diffusion coefficient (a square matrix function)
c	convection coefficient (a vector function)
r	reaction coefficient

Discretization methods

V_h	ansatz space
X_h	extended ansatz space without any homogeneous Dirichlet boundary conditions
a_h	approximated bilinear form
b_h	approximated linear form

Function spaces (see also Appendix A.5)

$\mathcal{P}_k(G)$	set of polynomials of maximum degree k on $G \subset \mathbb{R}^d$
$C(G) = C^0(G)$	set of continuous functions on G
$C^l(G)$	$(l \in \mathbb{N})$ set of l-times continuously differentiable functions on G
$C^\infty(G)$	set of infinitely often continuously differentiable functions on G
$C(\overline{G}) = C^0(\overline{G})$	set of bounded and uniformly continuous functions on G
$C^l(\overline{G})$	$(l \in \mathbb{N})$ set of functions with bounded and uniformly continuous derivatives up to the order l on G
$C^\infty(\overline{G})$	set of functions, all partial derivatives of which are bounded and uniformly continuous on G
$C_0(G) = C_0^0(G)$	set of continuous functions on G with compact support
$C_0^l(G)$	$(l \in \mathbb{N})$ set of l-times continuously differentiable functions on G with compact support
$C_0^\infty(G)$	set of infinitely often continuously differentiable functions on G with compact support
$L^p(G)$	$(p \in [1, \infty))$ set of Lebesgue-measurable functions whose pth power of their absolute value is Lebesgue-integrable on G
$L^\infty(G)$	set of measurable, essentially bounded functions
$\langle \cdot, \cdot \rangle_{0,G}$	scalar product in $L^2(G)$ †
$\| \cdot \|_{0,G}$	norm in $L^2(G)$ †
$\| \cdot \|_{0,p,G}$	$(p \in [1, \infty])$ norm in $L^p(G)$ †
$\| \cdot \|_{\infty,G}$	norm in $L^\infty(G)$ †
$W_p^l(G)$	$(l \in \mathbb{N}, p \in [1, \infty])$ set of l-times weakly differentiable functions from $L_p(G)$, with derivatives in $L^p(G)$
$\| \cdot \|_{l,p,G}$	$(l \in \mathbb{N}, p \in [1, \infty])$ norm in $W_p^l(G)$ †
$\| \cdot \|_{l,p,G}$	$(l \in \mathbb{N}, p \in [1, \infty])$ seminorm in $W_p^l(G)$ †

$H^l(G)$ $:= W_2^l(G)$ $(l \in \mathbb{N})$
$\langle \cdot, \cdot \rangle_{l,G}$ $(l \in \mathbb{N})$ scalar product in $H^l(G)$ †
$\| \cdot \|_{l,G}$ $(l \in \mathbb{N})$ norm in $H^l(G)$ †
$| \cdot |_{l,G}$ $(l \in \mathbb{N})$ seminorm in $H^l(G)$ †
$\langle \cdot, \cdot \rangle_{0,h}$ discrete $L^2(\Omega)$-scalar product
$\| \cdot \|_{0,h}$ discrete $L^2(\Omega)$-norm
$L^2(\partial G)$ set of square Lebesgue-integrable functions on the boundary ∂G
$H_0^1(G)$ set of functions from $H^1(G)$ with vanishing trace on ∂G
$C([0,T],X) = C^0([0,T],X)$ set of continuous functions on $[0,T]$ with values in the normed space X
$C^l([0,T],X)$ $(l \in \mathbb{N})$ set of l-times continuously differentiable functions on $[0,T]$ with values in the normed space X
$L^p((0,T),X)$ $(p \in [1,\infty])$ Lebesgue-space of functions on $[0,T]$ with values in the normed space X

† **Convention:** In the case $G = \Omega$, this specification is omitted.

A.2 Basic Concepts of Analysis

A subset $G \subset \mathbb{R}^d$ is called a *set of measure zero* if, for any number $\varepsilon > 0$, a countable family of balls B_j with d-dimensional volume $\varepsilon_j > 0$ exists such that

$$\sum_{j=1}^{\infty} \varepsilon_j < \varepsilon \qquad \text{and} \qquad G \subset \bigcup_{j=1}^{\infty} B_j \ .$$

Two functions $f, g : G \to \mathbb{R}$ are called *equal almost everywhere* (in short: *equal a.e.,* notation: $f \equiv g$) if the set $\{x \in G : f(x) \neq g(x)\}$ is of measure zero.

In particular, a function $f : G \to \mathbb{R}$ is called *vanishing almost everywhere* if it is equal to the constant function zero almost everywhere.

A function $f : G \to \mathbb{R}$ is called *measurable* if there exists a sequence $(f_i)_i$ of step functions $f_i : G \to \mathbb{R}$ such that $f_i \to f$ for $i \to \infty$ almost everywhere.

In what follows, G denotes a subset of $\mathbb{R}^d$, $d \in \mathbb{N}$.

(i) A point $x = (x_1, x_2, \ldots, x_d)^T \in \mathbb{R}^d$ is called a *boundary point* of G if every open neighbourhood (perhaps an open ball) of x contains a point of G as well as a point of the complementary set $\mathbb{R} \setminus G$.

(ii) The collection of all boundary points of G is called the *boundary* of G and is denoted by ∂G.

(iii) The set $\overline{G} := G \cup \partial G$ is called the *closure* of G.

(iv) The set G is called *closed* if $\overline{G} = G$.

(v) The set G is called *open* if $G \cap \partial G = \emptyset$.

(vi) The set $G \setminus \partial G$ is called the *interior* of G and is denoted by $\operatorname{int} G$.

A subset $G \subset \mathbb{R}^d$ is called *connected* if for arbitrary distinct points $x_1, x_2 \in G$ there exists a continuous curve in G connecting them.

The set G is called *convex* if any two points from G can be connected by a straight-line segment in G.

A nonempty, open, and connected set $G \subset \mathbb{R}^d$ is called a *domain* in $\mathbb{R}^d$.

By $\alpha = (\alpha_1, \ldots, \alpha_d)^T \in \mathbb{N}_0^d$ a so-called *multi-index* is denoted. Multi-indices are a popular tool to abbreviate some elaborate notation. For example,

$$\partial^\alpha := \prod_{i=1}^{d} \partial_i^{\alpha_i} , \quad \alpha! := \prod_{i=1}^{d} \alpha_i! , \quad |\alpha| := \sum_{i=1}^{d} \alpha_i .$$

The number $|\alpha|$ is called the *order* (or *length*) of the multi-index α.

For a continuous function $\varphi : G \to \mathbb{R}$, the set $\operatorname{supp} \varphi := \overline{\{x \in G : \varphi(x) \neq 0\}}$ denotes the *support* of φ.

A.3 Basic Concepts of Linear Algebra

A square matrix $A \in \mathbb{R}^{n,n}$ with entries a_{ij} is called *symmetric* if $a_{ij} = a_{ji}$ holds for all $i, j \in \{1, \ldots, n\}$.

A matrix $A \in \mathbb{R}^{n,n}$ is called *positive definite* if $x \cdot Ax > 0$ for all $x \in \mathbb{R}^n \setminus \{0\}$.

Given a polynomial $p \in \mathcal{P}_k$, $k \in \mathbb{N}_0$, of the form

$$p(z) = \sum_{j=0}^{k} a_j z^j \quad \text{with} \quad a_j \in \mathbb{C},\ j \in \{0, \ldots, k\}$$

and a matrix $A \in \mathbb{C}^{n,n}$, then the following *matrix polynomial* of A can be established:

$$p(A) := \sum_{j=0}^{k} a_j A^j .$$

Eigenvalues and Eigenvectors
Let $A \in \mathbb{C}^{n,n}$. A number $\lambda \in \mathbb{C}$ is called an *eigenvalue* of A if

$$\det(A - \lambda I) = 0 .$$

If λ is an eigenvalue of A, then any vector $x \in \mathbb{C}^n \setminus \{0\}$ such that

$$Ax = \lambda x \qquad (\Leftrightarrow (A - \lambda I)x = 0)$$

is called an *eigenvector* of A associated with the eigenvalue λ.

The polynomial $p_A(\lambda) := \det(A - \lambda I)$ is called the *characteristic polynomial* of A.

The set of all eigenvalues of a matrix A is called the *spectrum* of A, denoted by $\sigma(A)$.

If all eigenvalues of a matrix A are real, then the numbers $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the largest, respectively smallest, of these eigenvalues.

The number $\varrho(A) = \max_{\lambda\in\sigma(A)} |\lambda|$ is called the *spectral radius* of A.

Norms of Vectors and Matrices

The *norm of a vector* $x \in \mathbb{R}^n$, $n \in \mathbb{N}$, is a real-valued function $x \mapsto |x|$ satisfying the following three properties:

(i) $|x| \geq 0$ for all $x \in \mathbb{R}^n$, $|x| = 0 \Leftrightarrow x = 0$,
(ii) $|\alpha x| = |\alpha|\,|x|$ for all $\alpha \in \mathbb{R}$, $x \in \mathbb{R}^n$,
(iii) $|x + y| \leq |x| + |y|$ for all $x, y \in \mathbb{R}^n$.

For example, the most frequently used vector norms are
(a) the *maximum norm*:

$$|x|_\infty := \max_{j=1\ldots n} |x_j|\,. \tag{A3.1}$$

(b) the *ℓ_p-norm*, $p \in [1, \infty)$:

$$|x|_p := \left\{ \sum_{j=1}^{n} |x_j|^p \right\}^{1/p}. \tag{A3.2}$$

The important case $p = 2$ yields the so-called *Euclidean norm*:

$$|x|_2 := \left\{ \sum_{j=1}^{n} x_j^2 \right\}^{1/2}. \tag{A3.3}$$

The three most important norms (that is, $p = 1, 2, \infty$) in $\mathbb{R}^n$ are *equivalent* in the following sense: The inequalities

$$\begin{aligned}
\frac{1}{\sqrt{n}}\,|x|_2 &\leq |x|_\infty \leq |x|_2 \leq \sqrt{n}\,|x|_\infty\,,\\
\frac{1}{n}\,|x|_1 &\leq |x|_\infty \leq |x|_1 \leq n\,|x|_\infty\,,\\
\frac{1}{\sqrt{n}}\,|x|_1 &\leq |x|_2 \leq |x|_1 \leq \sqrt{n}\,|x|_2
\end{aligned}$$

are valid for all $x \in \mathbb{R}^n$.

The *norm of the matrix* $A \in \mathbb{R}^{n,n}$ is a real-valued function $A \mapsto \|A\|$ satisfying the following four properties:

(i) $\|A\| \geq 0$ for all $A \in \mathbb{R}^{n,n}$, $\|A\| = 0 \Leftrightarrow A = 0$,
(ii) $\|\alpha A\| = |\alpha|\,\|A\|$ for all $\alpha \in \mathbb{R}$, $A \in \mathbb{R}^{n,n}$,
(iii) $\|A + B\| \leq \|A\| + \|B\|$ for all $A, B \in \mathbb{R}^{n,n}$,
(iv) $\|AB\| \leq \|A\|\,\|B\|$ for all $A, B \in \mathbb{R}^{n,n}$.

In comparison with the definition of a vector norm, we include here an additional property (iv), which is called the *submultiplicative property*. It restricts the general set of matrix norms to the practically important class of *submultiplicative norms*.

The most common matrix norms are
(a) the *total norm*:

$$\|A\|_G := n \max_{1\le i,k\le n} |a_{ik}| \,, \tag{A3.4}$$

(b) the *Frobenius norm*:

$$\|A\|_F := \left\{ \sum_{i,k=1}^{n} a_{ik}^2 \right\}^{1/2} , \tag{A3.5}$$

(c) the *maximum row sum*:

$$\|A\|_\infty := \max_{1\le i\le n} \sum_{k=1}^{n} |a_{ik}| \,, \tag{A3.6}$$

(d) the *maximum column sum*:

$$\|A\|_1 := \max_{1\le k\le n} \sum_{i=1}^{n} |a_{ik}| \,. \tag{A3.7}$$

All these matrix norms are equivalent. For example, we have

$$\frac{1}{n}\|A\|_G \le \|A\|_p \le \|A\|_G \le n\|A\|_p \,, \quad p \in \{1,\infty\} \,,$$

or

$$\frac{1}{n}\|A\|_G \le \|A\|_F \le \|A\|_G \le n\|A\|_F \,.$$

Note that the spectral radius $\varrho(A)$ is not a matrix norm, as the following simple example shows:

For $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, we have that $A \neq 0$ but $\varrho(A) = 0$.

However, for any matrix norm $\|\cdot\|$ the following relation is valid:

$$\varrho(A) \le \|A\| \,. \tag{A3.8}$$

Very often, matrices and vectors simultaneously appear as a product Ax. In order to be able to handle such situations, there should be a certain correlation between matrix and vector norms.

A matrix norm $\|\cdot\|$ is called *mutually consistent* or *compatible* with the vector norm $|\cdot|$ if the inequality

$$|Ax| \le \|A\|\,|x| \tag{A3.9}$$

is valid for all $x \in \mathbb{R}^n$ and all $A \in \mathbb{R}^{n,n}$.

Examples of mutually consistent norms are

$$\|A\|_G \quad \text{or} \quad \|A\|_\infty \quad \text{with} \quad |x|_\infty \,,$$

$$\|A\|_G \quad \text{or} \quad \|A\|_1 \quad \text{with} \quad |x|_1 \,,$$

$$\|A\|_G \quad \text{or} \quad \|A\|_F \quad \text{with} \quad |x|_2 \,.$$

In many cases, the bound for $|Ax|$ given by (A3.9) is not sharp enough; i.e., for $x \neq 0$ we just have that

$$|Ax| < \|A\| \, |x| \,.$$

Therefore, the question arises of how to find, for a given vector norm, a compatible matrix norm such that in (A3.9) the equality holds for at least one element $x \neq 0$.

Given a vector norm $|x|$, the number

$$\|A\| := \sup_{x \in \mathbb{R}^n \setminus \{0\}} \frac{|Ax|}{|x|} = \sup_{x \in \mathbb{R}^n \colon |x|=1} |Ax|$$

is called the *induced* or *subordinate* matrix norm.

The induced norm is a compatible matrix norm with the given vector norm. It is the smallest norm among all matrix norms that are compatible with the given vector norm $|x|$.

To illustrate the definition of the induced matrix norm, the matrix norm induced by the Euclidean vector norm is derived:

$$\|A\|_2 := \max_{|x|_2=1} |Ax|_2 = \max_{|x|_2=1} \sqrt{x^T (A^T A) x} = \sqrt{\lambda_{\max}(A^T A)} = \sqrt{\varrho(A^T A)} \,. \tag{A3.10}$$

The matrix norm $\|A\|_2$ induced by the Euclidean vector norm is also called the *spectral norm.* This term becomes understandable in the special case of a symmetric matrix A. If $\lambda_1, \ldots, \lambda_n$ denote the real eigenvalues of A, then the matrix $A^T A = A^2$ has the eigenvalues λ_i^2 satisfying

$$\|A\|_2 = |\lambda_{\max}(A)| \,.$$

For symmetric matrices, the spectral norm coincides with the spectral radius. Because of (A3.8), it is the smallest possible matrix norm in that case.

As a further example, the maximum row sum $\|A\|_\infty$ is the matrix norm induced by the maximum norm $|x|_\infty$.

The number

$$\kappa(A) := \|A\| \, \|A^{-1}\|$$

is called the *condition number* of the matrix A with respect to the matrix norm under consideration.

The following relation holds:

$$1 \leq \|I\| = \|AA^{-1}\| \leq \|A\| \, \|A^{-1}\| \, .$$

For $|\cdot| = |\cdot|_p$, the condition number is also denoted by $\kappa_p(A)$. If all eigenvalues of A are real, the number

$$\kappa(A) := \lambda_{\max}(A)/\lambda_{\min}(A)$$

is called the *spectral condition number*. Hence, for a symmetric matrix A the equality $\kappa(A) = \kappa_2(A)$ is valid.

Occasionally, it is necessary to estimate small perturbations of nonsingular matrices. For this purpose, the following result is useful (*perturbation lemma* or *Neumann's lemma*). Let $A \in \mathbb{R}^{n,n}$ satisfy $\|A\| < 1$ with respect to an arbitrary, but fixed, matrix norm. Then the inverse of $I - A$ exists and can be represented as a convergent power series of the form

$$(I - A)^{-1} = \sum_{j=0}^{\infty} A^j,$$

with

$$\|(I - A)^{-1}\| \leq \frac{1}{1 - \|A\|} \, . \tag{A3.11}$$

Special Matrices

The matrix $A \in \mathbb{R}^{n,n}$ is called an *upper,* respectively *lower, triangular matrix* if its entries satisfy $a_{ij} = 0$ for $i > j$, respectively $a_{ij} = 0$ for $i < j$.

A matrix $H \in \mathbb{R}^{n,n}$ is called an *(upper) Hessenberg matrix* if it has the following structure:

$$H := \begin{pmatrix} h_{11} & & & & \\ h_{21} & \ddots & & * & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & 0 & & h_{nn-1} & h_{nn} \end{pmatrix}$$

(that is, $h_{ij} = 0$ for $i > j + 1$).

The matrix $A \in \mathbb{R}^{n,n}$ satisfies the *strict row sum criterion* (or is *strictly row diagonally dominant*) if it satisfies

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| < |a_{ii}| \quad \text{for all } i = 1, \ldots, n \, .$$

It satisfies the *strict column sum criterion* if the following relation holds:

$$\sum_{\substack{i=1 \\ i \neq j}}^{n} |a_{ij}| < |a_{jj}| \quad \text{for all } j = 1, \ldots, n \, .$$

The matrix $A \in \mathbb{R}^{n,n}$ satisfies the *weak row sum criterion* (or is *weakly row diagonally dominant*) if

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \leq |a_{ii}| \quad \text{holds for all } i = 1, \ldots, n$$

and the strict inequality "$<$" is valid for at least one number $i \in \{1, \ldots, n\}$.

The weak column sum criterion is defined similarly.

The matrix $A \in \mathbb{R}^{n,n}$ is called *reducible* if there exist subsets $N_1, N_2 \subset \{1, \ldots, n\}$ with $N_1 \cap N_2 = \emptyset$, $N_1 \neq \emptyset \neq N_2$, and $N_1 \cup N_2 = \{1, \ldots, n\}$ such that the following property is satisfied:

$$\text{For all } i \in N_1,\ j \in N_2 \colon\ a_{ij} = 0 \ .$$

A matrix that is not reducible is called *irreducible.*

A matrix $A \in \mathbb{R}^{n,n}$ is called an L_0*-matrix* if for $i, j \in \{1, \ldots, n\}$ the inequalities

$$a_{ii} \geq 0 \quad \text{and} \quad a_{ij} \leq 0 \ (i \neq j)$$

are valid. An L_0-matrix is called an *L-matrix* if all diagonal entries are positive.

A matrix $A \in \mathbb{R}^{n,n}$ is called *monotone* (or *of monotone type*) if the relation $Ax \leq Ay$ for two (otherwise arbitrary) elements $x, y \in \mathbb{R}^n$ implies $x \leq y$. Here the relation sign is to be understood componentwise.

A matrix of monotone type is invertible.

A matrix $A \in \mathbb{R}^{n,n}$ is a matrix of monotone type if it is invertible and all entries of the inverse are nonnegative.

An important subclass of matrices of monotone type is formed by the so-called M-matrices.

A monotone matrix A with $a_{ij} \leq 0$ for $i \neq j$ is called an *M-matrix.*

Let $A \in \mathbb{R}^{n,n}$ be a matrix with $a_{ij} \leq 0$ for $i \neq j$ and $a_{ii} \geq 0$ $(i, j \in \{1, \ldots, n\})$. In addition, let A satisfy one of the following conditions:

(i) A satisfies the strict row sum criterion.

(ii) A satisfies the weak row sum criterion and is irreducible.

Then A is an M-matrix.

A.4 Some Definitions and Arguments of Linear Functional Analysis

Working with vector spaces whose elements are (classical or generalized) functions, it is desirable to have a measure for the "length" or "magnitude" of a function, and, as a consequence, for the distance of two functions.

Let V be a real vector space (in short, an $\mathbb{R}$ vector space) and let $\|\cdot\|$ be a real-valued mapping $\|\cdot\| : V \to \mathbb{R}$.

The pair $(V, \|\cdot\|)$ is called a *normed space* ("V is endowed with the *norm* $\|\cdot\|$") if the following properties hold:

$$\|u\| \geq 0 \quad \text{for all } u \in V\,, \quad \|u\| = 0 \Leftrightarrow u = 0\,, \tag{A4.1}$$

$$\|\alpha u\| = |\alpha|\,\|u\| \quad \text{for all } \alpha \in \mathbb{R}\,,\ u \in V\,, \tag{A4.2}$$

$$\|u + v\| \leq \|u\| + \|v\| \quad \text{for all } u, v \in V\,. \tag{A4.3}$$

The property (A4.1) is called *definiteness*; (A4.3) is called the *triangle inequality.* If a mapping $\|\cdot\| : V \to \mathbb{R}$ satisfies only (A4.2) and (A4.3), it is called a *seminorm.* Due to (A4.2), we still have $\|0\| = 0$, but there may exist elements $u \neq 0$ with $\|u\| = 0$.

A particularly interesting example of a norm can be obtained if the space V is equipped with a so-called *scalar product.* This is a mapping $\langle\cdot,\cdot\rangle : V \times V \to \mathbb{R}$ with the following properties:

(1) $\langle\cdot,\cdot\rangle$ is a *bilinear form,* that is,

$$\begin{aligned} \langle u, v_1 + v_2\rangle &= \langle u, v_1\rangle + \langle u, v_2\rangle \quad &&\text{for all} \quad u, v_1, v_2 \in V\,, \\ \langle u, \alpha v\rangle &= \alpha\,\langle u, v\rangle &&\text{for all} \quad u, v \in V,\ \alpha \in \mathbb{R}\,, \end{aligned} \tag{A4.4}$$

and an analogous relation is valid for the first argument.

(2) $\langle\cdot,\cdot\rangle$ is *symmetric*, that is,

$$\langle u, v\rangle = \langle v, u\rangle \quad \text{for all} \quad u, v \in V\,. \tag{A4.5}$$

(3) $\langle\cdot,\cdot\rangle$ is *positive*, that is,

$$\langle u, u\rangle \geq 0 \quad \text{for all} \quad u \in V\,. \tag{A4.6}$$

(4) $\langle\cdot,\cdot\rangle$ is *definite*, that is,

$$\langle u, u\rangle = 0 \;\Leftrightarrow\; u = 0\,. \tag{A4.7}$$

A positive and definite bilinear form is called *positive definite.*

A scalar product $\langle\cdot,\cdot\rangle$ defines a norm on V in a natural way if we set

$$\|v\| := \langle v, v\rangle^{1/2}\,. \tag{A4.8}$$

In absence of the definiteness (A4.7), only a seminorm is induced.

A norm (or a seminorm) induced by a scalar product (respectively by a symmetric and positive bilinear form) has some interesting properties. For example, it satisfies the *Cauchy–Schwarz inequality*, that is,

$$|\langle u, v\rangle| \leq \|u\|\,\|v\| \quad \text{for all } u, v \in V\,, \tag{A4.9}$$

and the *parallelogram identity*

$$\|u + v\|^2 + \|u - v\|^2 = 2(\|u\|^2 + \|v\|^2) \quad \text{for all } u, v \in V\,. \tag{A4.10}$$

Typical examples of normed spaces are the spaces $\mathbb{R}^n$ equipped with one of the ℓ^p-norms (for some fixed $p \in [1, \infty]$). In particular, the Euclidean norm (A3.3) is induced by the *Euclidean scalar product*

$$(x, y) \mapsto x \cdot y \quad \text{for all } x, y \in \mathbb{R}^n . \tag{A4.11}$$

On the other hand, infinite-dimensional function spaces play an important role (see Appendix A.5).

If a vector space V is equipped with a scalar product $\langle \cdot, \cdot \rangle$, then, in analogy to $\mathbb{R}^n$, an element $u \in V$ is said to be *orthogonal* to $v \in V$ if

$$\langle u, v \rangle = 0 . \tag{A4.12}$$

Given a normed space $(V, \|\cdot\|)$, it is easy to define the concept of *convergence* of a sequence $(u_i)_i$ in V to $u \in V$:

$$u_i \to u \quad \text{for } i \to \infty \quad \iff \quad \|u_i - u\| \to 0 \quad \text{for } i \to \infty . \tag{A4.13}$$

Often, it is necessary to consider function spaces endowed with different norms. In such situations, different kinds of convergence may occur. However, if the corresponding norms are equivalent, then there is no change in the type of convergence. Two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ in V are called *equivalent* if there exist constants $C_1, C_2 > 0$ such that

$$C_1 \|u\|_1 \le \|u\|_2 \le C_2 \|u\|_1 \quad \text{for all } u \in V . \tag{A4.14}$$

If there is only a one-sided inequality of the form

$$\|u\|_2 \le C \|u\|_1 \quad \text{for all } u \in V \tag{A4.15}$$

with a constant $C > 0$, then the norm $\|\cdot\|_1$ is called *stronger* than the norm $\|\cdot\|_2$.

In a finite-dimensional vector space, all norms are equivalent. Examples can be found in Appendix A.3. In particular, it is important to observe that the constants may depend on the dimension n of the finite-dimensional vector space. This observation also indicates that in the case of infinite-dimensional vector spaces, the equivalence of two different norms cannot be expected, in general.

As a consequence of (A4.14), two equivalent norms $\|\cdot\|_1, \|\cdot\|_2$ in V yield the same type of convergence:

$$\begin{aligned} u_i \to u \text{ w.r.t. } \|\cdot\|_1 \quad &\Leftrightarrow \quad \|u_i - u\|_1 \to 0 \\ &\Leftrightarrow \quad \|u_i - u\|_2 \to 0 \quad \Leftrightarrow \quad u_i \to u \text{ w.r.t. } \|\cdot\|_2 . \end{aligned} \tag{A4.16}$$

In this book, the finite-dimensional vector space $\mathbb{R}^n$ is used in two aspects: For $n = d$, it is the basic space of independent variables, and for $n = M$ or $n = m$ it represents the finite-dimensional trial space. In the first case, the equivalence of all norms can be used in all estimates without any side effects, whereas in the second case the aim is to obtain uniform

estimates with respect to all M and m, and so the dependence of the equivalence constants on M and m has to be followed thoroughly.

Now we consider two normed spaces $(V, \|\cdot\|_V)$ and $(W, \|\cdot\|_W)$. A mapping $f : V \to W$ is called *continuous* in $v \in V$ if for all sequences $(v_i)_i$ in V with $v_i \to v$ for $i \to \infty$ we get

$$f(v_i) \to f(v) \quad \text{for} \quad i \to \infty\,.$$

Note that the first convergence is measured in $\|\cdot\|_V$ and the second one in $\|\cdot\|_W$. Hence a change of the norm may have an influence on the continuity. As in classical analysis, we can say that

$$\begin{array}{l} f \text{ is continuous in all } v \in V \iff \\ f^{-1}[G] \text{ is closed for each closed } G \subset W\,. \end{array} \tag{A4.17}$$

Here, a subset $G \subset W$ of a normed space W is called *closed* if for any sequence $(u_i)_i$ from G such that $u_i \to u$ for $i \to \infty$ the inclusion $u \in G$ follows. Because of (A4.17), the closedness of a set can be verified by showing that it is a continuous preimage of a closed set.

The concept of continuity is a qualitative relation between the preimage and the image. A quantitative relation is given by the stronger notion of Lipschitz continuity:

A mapping $f : V \to W$ is called *Lipschitz continuous* if there exists a constant $L > 0$, the *Lipschitz constant*, such that

$$\|f(u) - f(v)\|_W \le L\|u - v\|_V \quad \text{for all } u, v \in V\,. \tag{A4.18}$$

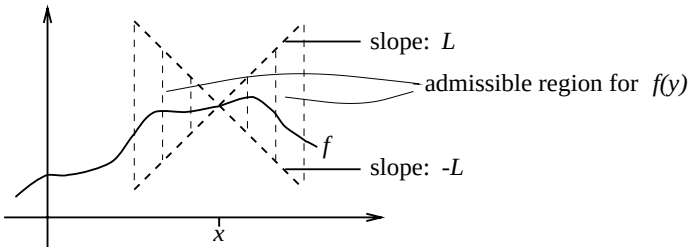


Figure A.1. Lipschitz continuity (for $V = W = \mathbb{R}$).

A Lipschitz continuous mapping with $L < 1$ is called *contractive* or a *contraction*; cf. Figure A.1.

Most of the mappings used are *linear*; that is, they satisfy

$$\left.\begin{array}{rcl} f(u+v) & = & f(u) + f(v)\,, \\ f(\lambda u) & = & \lambda f(u)\,, \end{array}\right\} \quad \text{for all } u, v \in V \text{ and } \lambda \in \mathbb{R}\,. \tag{A4.19}$$

For a linear mapping, the Lipschitz continuity is equivalent to the *boundedness*; that is, there exists a constant $C > 0$ such that

$$\|f(u)\|_W \le C\|u\|_V \quad \text{for all } u \in V\,. \tag{A4.20}$$

In fact, for a linear mapping f, the continuity at one point is equivalent to (A4.20). Linear, continuous mappings acting from V to W are also called (linear, continuous) *operators* and are denoted by capital letters, for example $S, T, \ldots$.

In the case $V = W = \mathbb{R}^n$, the linear, continuous operators in $\mathbb{R}^n$ are the mappings $x \mapsto Ax$ defined by matrices $A \in \mathbb{R}^{n,n}$. Their boundedness, for example with respect to $\|\cdot\|_V = \|\cdot\|_W = \|\cdot\|_\infty$, is an immediate consequence of the compatibility property of the $\|\cdot\|_\infty$-norm. Moreover, since all norms in $\mathbb{R}^n$ are equivalent, these mappings are bounded with respect to any norms in $\mathbb{R}^n$.

Similarly to (A4.20), a bilinear form $f : V \times V \to \mathbb{R}$ is continuous if it is *bounded,* that is, if there exists a constant $C > 0$ such that

$$|f(u,v)| \le C\,\|u\|_V \|v\|_V \quad \text{for all } u, v \in V\,. \tag{A4.21}$$

In particular, due to (A4.9) any scalar product is continuous with respect to the induced norm of V; that is,

$$u_i \to u\,,\ v_i \to v \quad \Rightarrow \quad \langle u_i, v_i\rangle \to \langle u, v\rangle\,. \tag{A4.22}$$

Now let $(V, \|\cdot\|_V)$ be a normed space and W a subspace that is (additionally to $\|\cdot\|_V$) endowed with the norm $\|\cdot\|_W$. The *embedding* from $(W, \|\cdot\|_W)$ to $(V, \|\cdot\|_V)$, i.e., the linear mapping that assigns any element of W to itself but considered as an element of V, is continuous iff the norm $\|\cdot\|_W$ is stronger than the norm $\|\cdot\|_V$ (cf. (A4.15)).

The collection of linear, continuous operators from $(V, \|\cdot\|_V)$ to $(W, \|\cdot\|_W)$ forms an $\mathbb{R}$ vector space with the following (argumentwise) operations:

$$\begin{aligned} (T+S)(u) &:= T(u) + S(u) && \text{for all } u \in V\,, \\ (\lambda T)(u) &:= \lambda T(u) && \text{for all } u \in V\,, \end{aligned}$$

for all operators T, S and $\lambda \in \mathbb{R}$. This space is denoted by

$$L[V, W]\,. \tag{A4.23}$$

In the special case $W = \mathbb{R}$, the corresponding operators are called linear, continuous *functionals,* and the notation

$$V' := L[V, \mathbb{R}] \tag{A4.24}$$

is used. The $\mathbb{R}$ vector space $L[V, W]$ can be equipped with a norm, the so-called *operator norm*, by

$$\|T\| := \sup\big\{\|T(u)\|_W \;\big|\; u \in V\,,\ \|u\|_V \le 1\big\} \quad \text{for} \quad T \in L[V, W]\,. \tag{A4.25}$$

Here $\|T\|$ is the smallest constant such that (A4.20) holds. Specifically, for a functional $f \in V'$, we have that

$$\|f\| = \sup\big\{|f(u)| \;\big|\; \|u\|_V \le 1\big\}\,.$$

For example, in the case $V = W = \mathbb{R}^n$ and $\|u\|_V = \|u\|_W$, the norm of a linear, bounded operator that is represented by a matrix $A \in \mathbb{R}^{n,n}$ coincides with the corresponding induced matrix norm (cf. Appendix A.3).

Let $(V, \|\cdot\|_V)$ be a normed space. A sequence $(u_i)_i$ in V is called a *Cauchy sequence* if for any $\varepsilon > 0$ there exists a number $n_0 \in \mathbb{N}$ such that

$$\|u_i - u_j\|_V \leq \varepsilon \quad \text{for all } i, j \in \mathbb{N} \text{ with } i, j \geq n_0 \,.$$

The space V is called *complete* or a *Banach space* if for any Cauchy sequence $(u_i)_i$ in V there exists an element $u \in V$ such that $u_i \to u$ for $i \to \infty$. If the norm $\|\cdot\|_V$ of a Banach space V is induced by a scalar product, then V is called a *Hilbert space*.

A subspace W of a Banach space is complete iff it is closed. A basic problem in the variational treatment of boundary value problems consists in the fact that the space of continuous functions (cf. the preliminary definition (2.7)), which is required to be taken as a basis, is not complete with respect to the norm ($\|\cdot\|_l$, $l = 0$ or $l = 1$). However, if in addition to the normed space $(W, \|\cdot\|)$, a larger space V is given that is complete with respect to the norm $\|\cdot\|$, then that space or the closure

$$\widetilde{W} := \overline{W} \tag{A4.26}$$

(as the smallest Banach space containing W) can be used. Such a *completion* can be introduced for any normed space in an abstract way. The problem is that the "nature" of the limiting elements remains vague.

If the relation (A4.26) is valid for some normed space W, then W is called *dense* in $\widetilde{W}$. In fact, given W, all "essential" elements of $\widetilde{W}$ are already captured. For example, if T is a linear, continuous operator T from $(\widetilde{W}, \|\cdot\|)$ to another normed space, then the identity

$$T(u) = 0 \quad \text{for all} \quad u \in W \tag{A4.27}$$

is sufficient for

$$T(u) = 0 \quad \text{for all} \quad u \in \widetilde{W} \,. \tag{A4.28}$$

The space of linear, bounded operators is complete if the image space is complete. In particular, the space V' of linear, bounded functionals on the normed space V is always complete.

A.5 Function Spaces

In this section $G \subset \mathbb{R}^d$ denotes a bounded domain.

The function space $C(G)$ contains all (real-valued) functions defined on G that are continuous in G. By $C^l(G)$, $l \in \mathbb{N}$, the set of l-times continuously differentiable functions on G is denoted. Usually, for the sake of consistency, the conventions $C^0(G) := C(G)$ and $C^\infty(G) := \bigcap_{l=0}^{\infty} C^l(G)$ are used.

Functions from $C^l(G),\ l \in \mathbb{N}_0$, and $C^\infty(G)$ need not be bounded, as for $d = 1$ the example $f(x) := x^{-1},\ x \in (0,1)$ shows.

To overcome this difficulty, further spaces of continuous functions are introduced. The space $C(\overline{G})$ contains all bounded and uniformly continuous functions on G, whereas $C^l(\overline{G}),\ l \in \mathbb{N}$, consists of functions with bounded and uniformly continuous derivatives up to order l on G. Here the conventions $C^0(\overline{G}) := C(\overline{G})$ and $C^\infty(\overline{G}) := \bigcap_{l=0}^\infty C^l(\overline{G})$ are used, too.

The space $C_0(G)$, respectively $C_0^l(G),\ l \in \mathbb{N}$, denotes the set of all those continuous, respectively l-times continuously differentiable, functions, the supports of which are contained in G. Often this set is called the set of functions with compact support in G. Since G is bounded, this means that the supports do not intersect boundary points of G. We also set $C_0^0(G) := C_0(G)$ and $C_0^\infty(G) := C_0(G) \cap C^\infty(G)$.

The linear space $L^p(G),\ p \in [1,\infty)$, contains all Lebesgue measurable functions defined on G whose pth power of their absolute value is Lebesgue integrable on G. The norm in $L^p(G)$ is defined as follows:

$$\|u\|_{0,p,G} := \left\{ \int_G |u|^p \, dx \right\}^{1/p}, \quad p \in [1,\infty) .$$

In the case $p = 2$, the specification of p is frequently omitted; that is, $\|u\|_{0,G} = \|u\|_{0,2,G}$. The $L^2(G)$-scalar product

$$\langle u, v \rangle_{0,G} := \int_G uv \, dx , \quad u, v \in L^2(G) ,$$

induces the $L^2(G)$-norm by setting $\|u\|_{0,G} := \sqrt{\langle u, u \rangle_{0,G}}$.

The space $L^\infty(G)$ contains all measurable, essentially bounded functions on G, where a function $u : G \to \mathbb{R}$ is called *essentially bounded* if the quantity

$$\|u\|_{\infty,G} := \inf_{G_0 \subset G:\, |G_0|_d = 0} \ \sup_{x \in G \setminus G_0} |u(x)|$$

is finite. For continuous functions, this norm coincides with the usual maximum norm:

$$\|u\|_{\infty,G} = \max_{x \in \overline{G}} |u(x)| , \quad u \in C(\overline{G}) .$$

For $1 \le q \le p \le \infty$, we have $L^p(G) \subset L^q(G)$, and the embedding is continuous.

The space $W_p^l(G),\ l \in \mathbb{N}$, $p \in [1,\infty]$, consists of all l-times weakly differentiable functions from $L_p(G)$ with derivatives in $L^p(G)$. In the special case $p = 2$, we also write $H^l(G) := W_2^l(G)$. In analogy to the case of continuous functions, the convention $H^0(G) := L^2(G)$ is used. The norm in $W_p^l(G)$ is

defined as follows:

$$\|u\|_{l,p,G} := \left\{ \sum_{|\alpha|\le l} \int_G |\partial^\alpha u|^p \, dx \right\}^{1/p}, \quad p \in [1,\infty) ,$$

$$\|u\|_{l,\infty,G} := \max_{|\alpha|\le l} |\partial^\alpha u|_{\infty,G} .$$

In $H^l(G)$ a scalar product can be defined by

$$\langle u, v\rangle_{l,G} := \sum_{|\alpha|\le l} \int_G \partial^\alpha u \partial^\alpha v \, dx , \quad u, v \in H^l(G) .$$

The norm induced by this scalar product is denoted by $\|\cdot\|_{l,G}$, $l \in \mathbb{N}$:

$$\|u\|_{l,G} := \sqrt{\langle u, u\rangle_{l,G}} .$$

For $l \in \mathbb{N}$, the symbol $|\cdot|_{l,G}$ stands for the corresponding $H^l(G)$-*seminorm*:

$$|u|_{l,G} := \sqrt{\sum_{|\alpha|=l} \int_G |\partial^\alpha u|^2 \, dx} .$$

The space $H_0^1(G)$ is defined as the closure (or completion) of $C_0^\infty(G)$ in the norm $\|\cdot\|_1$ of $H^1(G)$.

Convention: Usually, in the case $G = \Omega$ the specification of the domain in the above norms and scalar products is omitted.

In the study of partial differential equations, it is often desirable to speak of boundary values of functions defined on the domain G. In this respect, the Lebesgue spaces of functions that are square integrable at the boundary of G are important. To introduce these spaces, some preparations are necessary.

In what follows, a point $x \in \mathbb{R}^d$ is written in the form $x = \begin{pmatrix} x' \\ x_d \end{pmatrix}$ with $x' = (x_1, \ldots, x_{d-1})^T \in \mathbb{R}^{d-1}$.

A domain $G \subset \mathbb{R}^d$ is said to be *located at one side of* ∂G if for any $x \in \partial G$ there exist an open neighbourhood $U_x \subset \mathbb{R}^d$ and an orthogonal mapping Q_x in $\mathbb{R}^d$ such that the point x is mapped to a point $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_d)^T$, and so U_x is mapped onto a neighbourhood $U_{\hat{x}} \subset \mathbb{R}^d$ of $\hat{x}$, where in the neighbourhood $U_{\hat{x}}$ the following properties hold:

(1) The image of $U_x \cap \partial G$ is the graph of some function $\Psi_x : Y_x \subset \mathbb{R}^{d-1} \to \mathbb{R}$; that is, $\hat{x}_d = \Psi_x(\hat{x}_1, \ldots, \hat{x}_{d-1}) = \Psi_x(\hat{x}')$ for $\hat{x}' \in Y_x$.

(2) The image of $U_x \cap G$ is "above this graph" (i.e., the points in $U_x \cap G$ correspond to $\hat{x}_d > 0$).

(3) The image of $U_x \cap (\mathbb{R}^d \setminus \overline{G})$ is "below this graph" (i.e., the points in $U_x \cap (\mathbb{R}^d \setminus \overline{G})$ correspond to $\hat{x}_d < 0$).

A domain G that is located at one side of ∂G is called a C^l *domain*, $l \in \mathbb{N}$, respectively a *Lipschitz(ian) domain*, if all Ψ_x are l-times continuously differentiable, respectively Lipschitz continuous, in Y_x.

Bounded Lipschitz domains are also called *strongly Lipschitz*.

For bounded domains located at one side of ∂G, it is well known (cf., e.g. [37]) that from the whole set of neighbourhoods $\{U_x\}_{x\in\partial G}$ there can be selected a family $\{U_i\}_{i=1}^n$ of finitely many neighbourhoods covering ∂G, i.e., $n \in \mathbb{N}$ and $\partial G \subset \bigcup_{i=1}^n U_i$. Furthermore, for any such family there exists a system of functions $\{\varphi_i\}_{i=1}^n$ with the properties $\varphi_i \in C_0^\infty(U_i)$, $\varphi_i(x) \in [0,1]$ for all $x \in U_i$ and $\sum_{i=1}^n \varphi_i(x) = 1$ for all $x \in \partial G$. Such a system is called a *partition of unity*.

If the domain G is at least Lipschitzian, then Lebesgue's integral over the boundary of G is defined by means of those partitions of unity. In correspondence to the definition of a Lipschitz domain, Q_i, Ψ_i, and Y_i denote the orthogonal mapping on U_i, the function describing the corresponding local boundary, and the preimage of $Q_i(U_i \cap \partial G)$ with respect to Ψ_i.

A function $v : \partial G \to \mathbb{R}$ is called *Lebesgue integrable over* ∂G if the composite functions $\hat{x}' \mapsto v\Big(Q_i^T\Big(\begin{smallmatrix}\hat{x}'\\ \Psi_i(\hat{x}')\end{smallmatrix}\Big)\Big)$ belong to $L^1(Y_i)$. The integral is defined as follows:

$$\begin{aligned}\int_{\partial G} v(s)\,ds &:= \sum_{i=1}^n \int_{\partial G} v(s)\varphi_i(s)\,ds \\ &:= \sum_{i=1}^n \int_{Y_i} v\Big(Q_i^T\Big(\begin{smallmatrix}\hat{x}'\\ \Psi_i(\hat{x}')\end{smallmatrix}\Big)\Big)\varphi_i\Big(Q_i^T\Big(\begin{smallmatrix}\hat{x}'\\ \Psi_i(\hat{x}')\end{smallmatrix}\Big)\Big) \\ &\qquad \times\sqrt{|\det(\partial_j\Psi_i(\hat{x}')\partial_k\Psi_i(\hat{x}'))_{j,k=1}^{d-1}|}\,d\hat{x}'\,.\end{aligned}$$

A function $v : \partial G \to \mathbb{R}$ belongs to $L^2(\partial G)$ iff both v and v^2 are Lebesgue integrable over ∂G.

In the investigation of time-dependent partial differential equations, linear spaces whose elements are functions of the time variable $t \in [0,T]$, $T > 0$, with values in a normed space X are of interest.

A function $v : [0,T] \to X$ is called *continuous on* $[0,T]$ if for all $t \in [0,T]$ the convergence $\|v(t+k) - v(t)\|_X \to 0$ as $k \to 0$ holds.

The space $C([0,T],X) = C^0([0,T],X)$ consists of all continuous functions $v : [0,T] \to X$ such that

$$\sup_{t\in(0,T)} \|v(t)\|_X < \infty\,.$$

The space $C^l([0,T],X)$, $l \in \mathbb{N}$, consists of all continuous functions $v : [0,T] \to X$ that have continuous derivatives up to order l on $[0,T]$ with the

norm

$$\sum_{i=0}^{l} \sup_{t\in(0,T)} \|v^{(i)}(t)\|_X \,.$$

The space $L^p((0,T),X)$ with $1 \le p \le \infty$ consists of all functions on $(0,T)\times\Omega$ for which

$$v(t,\cdot) \in X \text{ for any } t \in (0,T)\,, \quad F \in L^p(0,T) \quad \text{with } F(t) := \|v(t,\cdot)\|_X \,.$$

Furthermore,

$$\|v\|_{L^p((0,T),X)} := \|F\|_{L^p(0,T)} \,.$$

References: Textbooks and Monographs

[1] R.A. Adams. *Sobolev Spaces*. Academic Press, New York, 1975.

[2] M. Ainsworth and J.T. Oden. *A Posteriori Error Estimation in Finite Element Analysis*. Wiley, New York, 2000.

[3] O. Axelsson and V.A. Barker. *Finite Element Solution of Boundary Value Problems. Theory and Computation*. Academic Press, Orlando, 1984.

[4] R.E. Bank. *PLTMG, a Software Package for Solving Elliptic Partial Differential Equations: Users Guide 7.0*. SIAM, Philadelphia, 1994. Frontiers in Applied Mathematics, Vol. 15.

[5] A. Berman and R.J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. Academic Press, New York, 1979.

[6] D. Braess. *Finite Elements. Theory, Fast Solvers, and Applications in Solid Mechanics*. Cambridge University Press, Cambridge, 2001 (2nd ed.).

[7] S.C. Brenner and L.R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer, New York–Berlin–Heidelberg, 2002 (2nd ed.). Texts in Applied Mathematics, Vol. 15.

[8] V.I. Burenkov. *Sobolev Spaces on Domains*. Teubner, Stuttgart, 1998.

[9] P.G. Ciarlet. Basic Error Estimates for Elliptic Problems. In: P.G. Ciarlet and J.L. Lions, editors, *Handbook of Numerical Analysis, Volume II: Finite Element Methods (Part 1)*. North-Holland, Amsterdam, 1991.

[10] A.J. Chorin and J.E. Marsden. *A Mathematical Introduction to Fluid Mechanics*. Springer, Berlin–Heidelberg–New York, 1993.

[11] R. Dautray and J.-L. Lions. *Mathematical Analysis and Numerical Methods for Science and Technology. Volume 4: Integral Equations and Numerical Methods*. Springer, Berlin–Heidelberg–New York, 1990.

[12] L.C. Evans. *Partial Differential Equations*. American Mathematical Society, Providence, 1998.

[13] D. Gilbarg and N.S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Springer, Berlin–Heidelberg–New York, 1983 (2nd ed.).

[14] V. Girault and P.-A. Raviart. *Finite Element Methods for Navier-Stokes Equations*. Springer, Berlin–Heidelberg–New York, 1986.

[15] W. Hackbusch. *Elliptic Differential Equations. Theory and Numerical Treatment*. Springer, Berlin–Heidelberg–New York, 1992.

[16] W. Hackbusch. *Iterative Solution of Large Sparse Systems of Equations*. Springer, New York, 1994.

[17] W. Hackbusch. *Multi-Grid Methods and Applications*. Springer, Berlin–Heidelberg–New York, 1985.

[18] L.A. Hageman and D.M. Young. *Applied Iterative Methods*. Academic Press, New York–London–Toronto–Sydney–San Francisco, 1981.

[19] U. Hornung, ed.. *Homogenization and Porous Media*. Springer, New York, 1997.

[20] T. Ikeda. *Maximum Principle in Finite Element Models for Convection–Diffusion Phenomena*. North-Holland, Amsterdam–New York–Oxford, 1983.

[21] C. Johnson. *Numerical Solution of Partial Differential Equations by the Finite Element Method*. Cambridge University Press, Cambridge–New York–New Rochelle–Melbourne–Sydney, 1987.

[22] C.T. Kelley. *Iterative Methods for Linear and Nonlinear Equations*. SIAM, Philadelphia, 1995.

[23] P. Knupp and S. Steinberg. *Fundamentals of Grid Generation*. CRC Press, Boca Raton, 1993.

[24] J.D. Logan. *Transport Modeling in Hydrogeochemical Systems*. Springer, New York–Berlin–Heidelberg, 2001.

[25] J. Nečas. *Les Méthodes Directes en Théorie des Équations Elliptiques*. Masson/Academia, Paris/Prague, 1967.

[26] M. Renardy and R.C. Rogers. *An Introduction to Partial Differential Equations*. Springer, New York, 1993.

[27] H.-G. Roos, M. Stynes, and L. Tobiska. *Numerical Methods for Singularly Perturbed Differential Equations*. Springer, Berlin–Heidelberg–New York, 1996. Springer Series in Computational Mathematics, Vol. 24.

[28] Y. Saad. *Iterative Methods for Sparse Linear Systems*. PWS Publ. Co., Boston, 1996.

[29] D.H. Sattinger. *Topics in Stability and Bifurcation Theory*. Springer, Berlin–Heidelberg–New York, 1973.

[30] J. Stoer. *Introduction to Numerical Analysis*. Springer, Berlin–Heidelberg–New York, 1996 (2nd ed.).

[31] G. Strang and G.J. Fix. *An Analysis of the Finite Element Method*. Wellesley-Cambridge Press, Wellesley, 1997 (3rd ed.).

[32] J.C. Strikwerda. *Finite Difference Schemes and Partial Differential Equations*. Wadsworth & Brooks/Cole, Pacific Grove, 1989.

[33] J.F. Thompson, Z.U.A. Warsi, and C.W. Mastin. *Numerical Grid Generation: Foundations and Applications*. North-Holland, Amsterdam, 1985.

[34] R.S. Varga. *Matrix Iterative Analysis*. Springer, Berlin–Heidelberg–New York, 2000.

[35] R. Verfürth. *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*. Wiley and Teubner, Chichester–New York–Brisbane–Toronto–Singapore and Stuttgart–Leipzig, 1996.

[36] S. Whitaker. *The Method of Volume Averaging*. Kluwer Academic Publishers, Dordrecht, 1998.

[37] J. Wloka. *Partial Differential Equations*. Cambridge University Press, New York, 1987.

[38] D.M. Young. *Iterative Solution of Large Linear Systems*. Academic Press, New York, 1971.

[39] E. Zeidler. *Nonlinear Functional Analysis and Its Applications. II/A: Linear Monotone Operators*. Springer, Berlin–Heidelberg–New York, 1990.

References: Journal Papers

[40] L. Angermann. Error estimates for the finite-element solution of an elliptic singularly perturbed problem. *IMA J. Numer. Anal.*, 15:161–196, 1995.

[41] T. Apel and M. Dobrowolski. Anisotropic interpolation with applications to the finite element method. *Computing*, 47:277–293, 1992.

[42] D.G. Aronson. The porous medium equation. In: A. Fasano and M. Primicerio, editors, *Nonlinear Diffusion Problems*. Lecture Notes in Mathematics 1224:1–46, 1986.

[43] M. Bause and P. Knabner. Uniform error analysis for Lagrange–Galerkin approximations of convection-dominated problems. *SIAM J. Numer. Anal.*, 39(6):1954–1984, 2002.

[44] R. Becker and R. Rannacher. A feed-back approach to error control in finite element methods: Basic analysis and examples. *East-West J. Numer. Math.*, 4(4):237–264, 1996.

[45] C. Bernardi, Y. Maday, and A.T. Patera. A new nonconforming approach to domain decomposition: the mortar element method. In: H. Brezis and J.-L. Lions, editors, *Nonlinear Partial Differential Equations and Their Applications*. Longman, 1994.

[46] T.D. Blacker and R.J. Meyers. Seams and wedges in plastering: A 3-D hexahedral mesh generation algorithm. *Engineering with Computers*, 9:83–93, 1993.

[47] T.D. Blacker and M.B. Stephenson. Paving: A new approach to automated quadrilateral mesh generation. *Internat. J. Numer. Methods Engrg.*, 32:811–847, 1991.

[48] A. Bowyer. Computing Dirichlet tesselations. *Computer J.*, 24(2):162–166, 1981.

[49] A.N. Brooks and T.J.R. Hughes. Streamline-upwind/Petrov–Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier–Stokes equations. *Comput. Meth. Appl. Mech. Engrg.*, 32:199–259, 1982.

[50] J.C. Cavendish. Automatic triangulation of arbitrary planar domains for the finite element method. *Internat. J. Numer. Methods Engrg.*, 8(4):679–696, 1974.

[51] W.M. Chan and P.G. Buning. Surface grid generation methods for overset grids. *Comput. Fluids*, 24(5):509–522, 1995.

[52] P. Clément. Approximation by finite element functions using local regularization. *RAIRO Anal. Numér.*, 9(R-2):77–84, 1975.

[53] P.C. Hammer and A.H. Stroud. Numerical integration over simplexes and cones. *Math. Tables Aids Comput.*, 10:130–137, 1956.

[54] T.J.R. Hughes, L.P. Franca, and G.M. Hulbert. A new finite element formulation for computational fluid dynamics: VIII. The Galerkin/least-squares method for advective-diffusive equations. *Comput. Meth. Appl. Mech. Engrg.*, 73(2):173–189, 1989.

[55] P. Jamet. Estimation of the interpolation error for quadrilateral finite elements which can degenerate into triangles. *SIAM J. Numer. Anal.*, 14:925–930, 1977.

[56] H. Jin and R. Tanner. Generation of unstructured tetrahedral meshes by advancing front technique. *Internat. J. Numer. Methods Engrg.*, 36:1805–1823, 1993.

[57] P. Knabner and G. Summ. The invertibility of the isoparametric mapping for pyramidal and prismatic finite elements. *Numer. Math.*, 88(4):661–681, 2001.

[58] M. Křížek. On the maximum angle condition for linear tetrahedral elements. *SIAM J. Numer. Anal.*, 29:513–520, 1992.

[59] C.L. Lawson. Software for C^1 surface interpolation. In: J.R. Rice, editor, *Mathematical Software III*, 161–194. Academic Press, New York, 1977.

[60] P. Möller and P. Hansbo. On advancing front mesh generation in three dimensions. *Internat. J. Numer. Methods Engrg.*, 38:3551–3569, 1995.

[61] K.W. Morton, A. Priestley, and E. Süli. Stability of the Lagrange–Galerkin method with non-exact integration. *RAIRO Modél. Math. Anal. Numér.*, 22(4):625–653, 1988.

[62] J. Peraire, M. Vahdati, K. Morgan, and O.C. Zienkiewicz. Adaptive remeshing for compressible flow computations. *J. Comput. Phys.*, 72:449–466, 1987.

[63] S.I. Repin. A posteriori error estimation for approximate solutions of variational problems by duality theory. In: H.G. Bock et al., editors, *Proceedings of ENUMATH 97*, 524–531. World Scientific Publ., Singapore, 1998.

[64] R. Rodríguez. Some remarks on Zienkiewicz–Zhu estimator. *Numer. Meth. PDE*, 10(5):625–635, 1994.

[65] W. Ruge and K. Stueben. Algebraische Mehrgittermethoden. In: S.F. McCormick, editor, *Multigrid Methods*, 73–130. SIAM, Philadelphia, 1987.

[66] R. Schneiders and R. Bünten. Automatic generation of hexahedral finite element meshes. *Computer Aided Geometric Design*, 12:693–707, 1995.

[67] L.R. Scott and S. Zhang. Finite element interpolation of nonsmooth functions satisfying boundary conditions. *Math. Comp.*, 54(190):483–493, 1990.

[68] M.S. Shephard and M.K. Georges. Automatic three-dimensional mesh generation by the finite octree technique. *Internat. J. Numer. Methods Engrg.*, 32:709–749, 1991.

[69] G. Summ. *Quantitative Interpolationsfehlerabschätzungen für Triangulierungen mit allgemeinen Tetraeder- und Hexaederelementen.* Diplomarbeit, Friedrich–Alexander–Universität Erlangen–Nürnberg, 1996. (`http://www.am.uni-erlangen.de/am1/publications/dipl_phd_thesis`)

[70] Ch. Tapp. *Anisotrope Gitter — Generierung und Verfeinerung.* Dissertation, Friedrich–Alexander–Universität Erlangen–Nürnberg, 1999. (`http://www.am.uni-erlangen.de/am1/publications/dipl_phd_thesis`)

[71] D.F. Watson. Computing the n-dimensional Delaunay tesselation with application to Voronoi polytopes. *Computer J.*, 24(2):167–172, 1981.

[72] M.A. Yerry and M.S. Shephard. Automatic three-dimensional mesh generation by the modified-octree technique. *Internat. J. Numer. Methods Engrg.*, 20:1965–1990, 1984.

[73] J.Z. Zhu, O.C. Zienkiewicz, E. Hinton, and J. Wu. A new approach to the development of automatic quadrilateral mesh generation. *Internat. J. Numer. Methods Engrg.*, 32:849–866, 1991.

[74] O.C. Zienkiewicz and J.Z. Zhu. The superconvergent patch recovery and a posteriori error estimates. Parts I,II. *Internat. J. Numer. Methods Engrg.*, 33(7):1331–1364,1365–1382, 1992.

Index

adjoint, 247
adsorption, 12
advancing front method, 179, 180
algorithm
  Arnoldi, 235
  CG, 223
  multigrid iteration, 243
  nested iteration, 253
  Newton's method, 357
algorithmic error, 200
angle condition, 173
angle criterion, 184
anisotropic, 8, 139
ansatz space, 56, 67
  nested, 240
  properties, 67
approximation
  superconvergent, 193
approximation error estimate, 139, 144
  for quadrature rules, 160
  one-dimensional, 137
approximation property, 250
aquifer, 7
Armijo's rule, 357
Arnoldi's method, 235
  algorithm, 235
  modified, 237
artificial diffusion method, 373
assembling, 62
  element-based, 66, 77
  node-based, 66
asymptotically optimal method, 199

Banach space, 404
Banach's fixed-point theorem, 345
barycentric coordinates, 117
basis of eigenvalues
  orthogonal, 300
best approximation error, 70
BICGSTAB method, 238
bifurcation, 363
biharmonic equation, 111
bilinear form, 400
  bounded, 403
  continuous, 93
  definite, 400
  positive, 400
  positive definite, 400
  symmetric, 400
  V-elliptic, 93
  V_h-elliptic, 156
block-Gauss–Seidel method, 211
block-Jacobi method, 211

Bochner integral, 289
boundary, 393
boundary condition, 15
  Dirichlet, 15
  flux, 15
  homogeneous, 15
  inhomogeneous, 15
  mixed, 15
  Neumann, 16
boundary point, 393
boundary value problem, 15
  adjoint, 145
    regular, 145
  weak solution, 107
Bramble–Hilbert lemma, 135
bulk density, 12

Cantor's function, 53
capillary pressure, 10
Cauchy sequence, 404
Cauchy–Schwarz inequality, 400
CG method, 221
  algorithm, 223
  error reduction, 224
  with preconditioning, 228
CGNE method, 235
CGNR method, 234
characteristics, 388
Chebyshev polynomial, 225
Cholesky decomposition, 84
  incomplete, 231
  modified
    incomplete, 232
chord method, 354
circle criterion, 184
closure, 393
coarse grid correction, 242, 243
coefficient, 16
collocation method, 68
collocation point, 68
column sum criterion
  strict, 398
comparison principle, 40, 328
completion, 404
complexity, 88
component, 5
condition number, 209, 397
  spectral, 398
conjugate, 219
conjugate gradient, *see* CG
connectivity condition, 173
conormal, 16
conormal derivative, 98
conservative form, 14
conservativity
  discrete global, 278
consistency, 28
consistency error, 28, 156
constitutive relationship, 7
continuation method, 357, 363
continuity, 402
continuous problem, 21
  approximation, 21
contraction, 402
contraction number, 199
control domain, 257
control volume, 257
convection
  forced, 5, 12
  natural, 5
convection-diffusion equation, 12
convection-dominated, 268
convective part, 12
convergence, 27
  global, 343
  linear, 343
  local, 343
  quadratic, 343
  superlinear, 343
  with order of convergence p, 343
  with respect to a norm, 401
correction, 201
Crank-Nicolson method, 313
cut-off strategy, 187
Cuthill–McKee method, 89

Darcy velocity, 7
Darcy's law, 8
decomposition
  regular, 232
definiteness, 400
degree of freedom, 62, 115, 120
Delaunay triangulation, 178, 263
dense, 96, 288, 404
density, 7
derivative
  generalized, 53
  material, 388

weak, 53, 289
diagonal field, 362
diagonal scaling, 230
diagonal swap, 181
difference quotient, 23
backward, 23
forward, 23
symmetric, 23
differential equation
convection-dominated, 12, 368
degenerate, 9
elliptic, 17
homogeneous, 16
hyperbolic, 17
inhomogeneous, 16
linear, 16
nonlinear, 16
order, 16
parabolic, 17
quasilinear, 16
semilinear, 16, 360
type of, 17
differential equation model
instationary, 8
linear, 8
stationary, 8
diffusion, 5
diffusive mass flux, 11
diffusive part, 12
Dirichlet domain, 262
Dirichlet problem
solvability, 104
discrete problem, 21
discretization, 21
five-point stencil, 24
upwind, 372
discretization approach, 55
discretization parameter, 21
divergence, 20
divergence form, 14
domain, 19, 394
C^l, 407
C^k-, 96
C^∞-, 96
Lipschitz, 96, 407
strongly, 407
domain of (absolute) stability, 317
Donald diagram, 265
dual problem, 194
duality argument, 145

edge swap, 181
eigenfunction, 285
eigenvalue, 285, 291, 394
eigenvector, 291, 394
element, 57
isoparametric, 122, 169
element stiffness matrix, 78
element-node table, 74
ellipticity
uniform, 100
embedding, 403
$H^k(\Omega)$ in $C(\bar{\Omega})$, 99
empty sphere criterion, 178
energy norm, 218
energy norm estimates, 132
energy scalar product, 217
equidistribution strategy, 187
error, 201
error equation, 68, 242
error estimate
a priori, 131, 185
anisotropic, 144
error estimator
a posteriori, 186
asymptotically exact, 187
efficient, 186
reliable, 186
residual, 188
dual-weighted, 194
robust, 187
error level
relative, 199
Euler method
explicit, 313
implicit, 313
extensive quantity, 7
extrapolation factor, 215
extrapolation method, 215

face, 123
family of triangulations
quasi-uniform, 165
regular, 138
Fick's law, 11
fill-in, 85
finite difference method, 17, 24
finite element, 115, 116

C^1-, 115, 127
affine equivalent, 122
Bogner–Fox–Schmit rectangle, 127
C^0-, 115
cubic ansatz on simplex, 121
cubic Hermite ansatz on simplex, 126
d-polynomial ansatz on cuboid, 123
equivalent, 122
Hermite, 126
Lagrange, 115, 126
linear, 57
linear ansatz on simplex, 119
quadratic ansatz on simplex, 120
simplicial, 117
finite element code
assembling, 176
kernel, 176
post-processor, 176
finite element discretization
conforming, 114
condition, 115
nonconforming, 114
finite element method, 18
characterization, 67
convergence rate, 131
maximum principle, 175
mortar, 180
finite volume method, 18
cell-centred, 258
cell-vertex, 258
node-centred, 258
semidiscrete, 297
five-point stencil, 24
fixed point, 342
fixed-point iteration, 200, 344
consistent, 200
convergence theorem, 201
fluid, 5
Fourier coefficient, 287
Fourier expansion, 287
Friedrichs–Keller triangulation, 64
frontal method, 87
full discretization, 293
full upwind method, 373
function
almost everywhere vanishing, 393
continuous, 407
essentially bounded, 405
Lebesgue integrable, 407
measurable, 393
piecewise continuous, 48
support, 394
functional, 403
functional matrix, 348
functions
equal almost everywhere, 393

Galerkin method, 56
stability, 69
unique solvability, 63
Galerkin product, 248
Galerkin/least squares–FEM, 377
Gauss's divergence theorem, 14, 47, 266
Gauss–Seidel method, 204
convergence, 204, 205
symmetric, 211
Gaussian elimination, 82
generating function, 316
GMRES method, 235
truncated, 238
with restart, 238
gradient, 20
gradient method, 218
error reduction, 219
gradient recovery, 192
graph
dual, 263
grid
chimera, 180
combined, 180
hierarchically structured, 180
logically structured, 177
overset, 180
structured, 176
in the strict sense, 176
in the wider sense, 177
unstructured, 177
grid adaptation, 187
grid coarsening, 183
grid function, 24
grid point, 21, 22
close to the boundary, 24, 327
far from the boundary, 24, 327
neighbour, 23

harmonic, 31

heat equation, 9
Hermite element, 126
Hessenberg matrix, 398
Hilbert space, 404
homogenization, 6
hydraulic conductivity, 8

IC factorization, 231
ill-posedness, 16
ILU factorization, 231
  existence, 232
ILU iteration, 231
inequality
  of Kantorovich, 218
  Friedrichs', 105
  inverse, 376
  of Poincaré, 71
inflow boundary, 108
inhomogeneity, 15
initial condition, 15
initial-boundary value problem, 15
inner product
  on $H^1(\Omega)$, 54
integral form, 14
integration by parts, 97
interior, 394
interpolation
  local, 58
interpolation error estimate, 138, 144
  one-dimensional , 136
interpolation operator, 132
interpolation problem
  local, 120
isotropic, 8
iteration
  inner, 355
  outer, 355
iteration matrix, 200
iterative method, 342

Jacobi matrix, 348
Jacobi's method, 203
  convergence, 204, 205
jump, 189
jump condition, 14

Krylov (sub)space, 222
Krylov subspace
  method, 223, 233

L_0-matrix, 399
L-matrix, 399
Lagrange element, 115, 126
Lagrange–Galerkin method, 387
Lagrangian coordinate, 387
Lanczos biorthogonalization, 238
Langmuir model, 12
Laplace equation, 9
Laplace operator, 20
lemma
  Bramble–Hilbert, 135
  Céa's, 70
  first of Strang, 155
lexicographic, 25
linear convergence, 199
Lipschitz constant, 402
Lipschitz continuity, 402
load vector, 62
LU factorization, 82
  incomplete, 231

M-matrix, 41, 399
macroscale, 6
mapping
  bounded, 402
  continuous, 402
  contractive, 402
  linear, 402
  Lipschitz continuous, 402
mass action law, 11
mass average mixture velocity, 7
mass lumping, 314, 365
mass matrix, 163, 296, 298
mass source density, 7
matrix
  band, 84
  bandwidth, 84
  consistently ordered, 213
  Hessenberg, 398
  hull, 84
  inverse monotone, 41
  irreducible, 399
  L_0-, 399
  L-, 399
  LU factorizable, 82
  M-, 399
  monotone, 399
  of monotone type, 399
  pattern, 231

positive definite, 394
profile, 84
reducible, 399
row bandwidth, 84
row diagonally dominant
strictly, 398
weakly, 399
sparse, 25, 82, 198
symmetric, 394
triangular
lower, 398
upper, 398
matrix norm
compatible, 396
induced, 397
mutually consistent, 396
submultiplicative, 396
subordinate, 397
matrix polynomial, 394
matrix-dependent, 248
max-min-angle property, 179
maximum angle condition, 144
maximum column sum, 396
maximum principle
strong, 36, 39, 329
weak, 36, 39, 329
maximum row sum, 396
mechanical dispersion, 11
mesh width, 21
method
advancing front, 179, 180
algebraic multigrid, 240
Arnoldi's , 235
artificial diffusion, 373
asymptotically optimal, 199
BICGSTAB, 238
block-Gauss–Seidel, 211
block-Jacobi, 211
CG, 221
classical Ritz–Galerkin, 67
collocation, 68
consistent, 28
convergence, 27
Crank-Nicolson, 313
Cuthill–McKee, 89
reverse, 90
Euler explicit, 313
Euler implicit, 313
extrapolation, 215
finite difference, 24
full upwind, 373
Galerkin, 56
Gauss–Seidel, 204
GMRES, 235
iterative, 342
Jacobi's, 203
Krylov subspace, 223, 233
Lagrange–Galerkin, 387
linear stationary, 200
mehrstellen, 30
moving front, 179
multiblock, 180
multigrid, 243
Newton's, 349
of bisection, 182
stage number of, 182
one-step, 316
one-step-theta, 312
overlay, 177
PCG, 228, 229
r-, 181
relaxation, 207
Richardson, 206
Ritz, 56
Rothe's, 294
semi-iterative, 215
SOR, 210
SSOR, 211
streamline upwind Petrov–Galerkin, 375
streamline-diffusion, 377
method of conjugate directions, 219
method of lines
horizontal, 294
vertical, 293
method of simultaneous displacements, 203
method of successive displacements, 204
MIC decomposition, 232
micro scale, 5
minimum angle condition, 141
minimum principle, 36
mobility, 10
molecular diffusivity, 11
monotonicity
inverse, 41, 280
monotonicity test, 357

moving front method, 179
multi-index, 53, 394
  length, 53, 394
  order, 53, 394
multiblock method, 180
multigrid iteration, 243
  algorithm, 243
multigrid method, 243
  algebraic, 240

neighbour, 38
nested iteration, 200, 252
  algorithm, 253
Neumann's lemma, 398
Newton's method, 349
  algorithm, 357
  damped, 357
  inexact, 355
  simplified, 353
nodal basis, 61, 125
nodal value, 58
node, 57, 115
  adjacent, 127
  degree, 89
  neighbour, 63, 89, 211
norm, 400
  discrete L^2-, 27
  equivalence of, 401
  Euclidean, 395
  Frobenius, 396
  induced by a scalar product, 400
  ℓ_p-, 395
  matrix, 395
  maximum, 395
  maximum , 27
  maximum column sum, 396
  maximum row sum, 396
  of an operator, 403
  spectral, 397
  streamline-diffusion, 378
  stronger, 401
  total, 396
  vector, 395
  ε-weighted, 374
normal derivative, 98
normal equations, 234
normed space
  complete, 404
norms
  equivalent, 395
numbering
  columnwise, 25
  rowwise, 25

octree technique, 177
one-step method, 316
  A-stable, 317
    strongly, 319
  L-stable, 319
  nonexpansive, 316
  stable, 320
one-step-theta method, 312
operator, 403
operator norm, 403
order of consistency, 28
order of convergence, 27
orthogonal, 401
orthogonality of the error, 68
outer unit normal, 14, 97
outflow boundary, 108
overlay method, 177
overrelaxation, 209
overshooting, 371

parabolic boundary, 325
parallelogram identity, 400
Parseval's identity, 292
particle velocity, 7
partition, 256
partition of unity, 407
PCG
  method, 228, 229
Péclet number
  global, 12, 368
  grid, 372
  local, 269
permeability, 8
perturbation lemma, 398
phase, 5
  immiscible, 7
phase average
  extrinsic, 6
  intrinsic, 6
k-phase flow, 5
$(k+1)$-phase system, 5
piezometric head, 8
point
  boundary, 40

close to the boundary, 40
far from the boundary, 40
Poisson equation, 8
Dirichlet problem, 19
polynomial
characteristic, 395
matrix, 394
pore scale, 5
pore space, 5
porosity, 6
porous medium, 5
porous medium equation, 9
preconditioner, 227
preconditioning, 207, 227
from the left, 227
from the right, 227
preprocessor, 176
pressure
global, 10
principle of virtual work, 49
projection
elliptic, 303, 304
prolongation, 246, 247
canonical, 246
pyramidal function, 62

quadrature points, 80
quadrature rule, 80, 151
accuracy, 152
Gauss–(Legendre), 153
integration points, 151
nodal, 152
trapezoidal rule, 66, 80, 153
weights, 151
quadtree technique, 177

range, 343
reaction
homogeneous, 13
inhomogeneous, 11
surface, 11
recovery operator, 193
red mblack ordering, 212
reduction strategy, 187
reference element, 58
standard simplicial, 117
refinement
iterative, 231
red/green, 181
relative permeability, 9
relaxation method, 207
relaxation parameter, 207
representative elementary volume, 6
residual, 188, 189, 201, 244
inner, 355
restriction, 248
canonical, 247
Richards equation, 10
Richardson method, 206
optimal relaxation parameter, 208
Ritz method, 56
Ritz projection, 304
Ritz–Galerkin method
classical, 67
root of equation, 342
Rothe's method, 294
row sum criterion
strict, 204, 398
weak, 205, 399
2:1-rule, 181

saturated, 10
saturated-unsaturated flow, 10
saturation, 7
saturation concentration, 12
scalar product, 400
energy, 217
Euclidean, 401
semi-iterative method, 215
semidiscrete problem, 295
semidiscretization, 293
seminorm, 400, 406
separation of variables, 285
set
closed, 393, 402
connected, 394
convex, 394
open, 394
set of measure zero, 393
shape function, 59
cubic ansatz on simplex, 121
d-polynomial ansatz on cube, 123
linear ansatz on simplex, 120
quadratic ansatz on simplex, 121
simplex
barycentre, 119
degenerate, 117
face, 117

regular d-, 117
sliver element, 179
smoothing
barycentric, 181
Laplacian, 181
weighted barycentric, 181
smoothing property, 239, 250
smoothing step, 178, 242
a posteriori, 243
a priori, 243
smoothness requirements, 20
Sobolev space, 54, 94
solid matrix, 5
solute concentration, 11
solution
classical, 21
of an (initial-) boundary value problem, 17
variational, 49
weak, 49, 290
uniqueness, 51
solvent, 5
SOR method, 210, 213
convergence, 212
optimal relaxation parameter, 213
sorbed concentration, 12
source term, 14
space
normed, 400
space-time cylinder, 15
bottom, 15
lateral surface, 15
spectral norm, 397
spectral radius, 395
spectrum, 395
split preconditioning, 228
SSOR method, 211
stability function, 316
stability properties, 36
stable, 28
static condensation, 128
stationary point, 217
step size, 21
stiffness matrix, 62, 296, 298
element entries, 76
streamline upwind Petrov–Galerkin method, 375
streamline-diffusion method, 377
streamline-diffusion norm, 378
superposition principle, 16
surface coordinate, 119
system of equations
positive real, 233

test function, 47
theorem
of Aubin and Nitsche, 145
of Kahan, 212
of Lax–Milgram, 93
of Ostrowski and Reich, 212
of Poincaré, 71
Trace, 96
Thiessen polygon, 262
three-term recursion, 234
time level, 312
time step, 312
tortuosity factor, 11
trace, 97
transformation
compatible, 134
isoparametric, 168
transformation formula, 137
transmission condition, 34
triangle inequality, 400
triangulation, 56, 114
anisotropic, 140
conforming, 56, 125
element, 114
properties, 114
refinement, 76
truncation error, 28
two-grid iteration, 242
algorithm, 242

underrelaxation, 209
unsaturated, 10
upscaling, 6
upwind discretization, 372
upwinding
exponential, 269
full, 269

V-cycle, 244
V-elliptic, 69
variation of constants, 286
variational equation, 49
equivalence to minimization problem, 50

solvability, 93
viscosity, 8
volume averaging, 6
volumetric fluid velocity, 7
volumetric water content, 11
Voronoi diagram, 262
Voronoi polygon, 262
Voronoi tesselation, 178
Voronoi vertex, 262
degenerate, 262
regular, 262

W-cycle, 244
water pressure, 8
weight, 30, 80
well-posedness, 16
Wigner–Seitz cell, 262

Z^2–estimate, 192
zero of function f, 342