Partial Differential Equations and the Finite Element Method

Partial Differential Equations and the Finite Element Method

Pavel Šolín
The University of Texas at El Paso
Academy of Sciences of the Czech Republic

WILEY-
INTERSCIENCE
A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic format. For information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data:

Šolín, Pavel.
Partial differential equations and the finite element method / Pavel Solin
p. cm.
Includes bibliographical references and index.
ISBN-13 978-0-471-72070-6
ISBN-10 0-471-72070-4 (cloth : acid-free paper)
1. Differential equations, Partial—Numerical solutions. 2. Finite element method. I. Title.

QA377.S65 2005
518'.64—dc22 200548622

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

To Dagmar

CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

Rien ne sert de courir,
il faut partir à point.
Jean de la Fontaine

Many physical processes in nature, whose correct understanding, prediction, and control are important to people, are described by equations that involve physical quantities together with their spatial and temporal rates of change *(partial derivatives)*. Among such processes are the weather, flow of liquids, deformation of solid bodies, heat transfer, chemical reactions, electromagnetics, and many others. Equations involving partial derivatives are called *partial differential equations (PDEs)*. The solutions to these equations are functions, as opposed to standard algebraic equations whose solutions are numbers. For most PDEs we are not able to find their exact solutions, and sometimes we do not even know whether a unique solution exists. For these reasons, in most cases the only way to solve PDEs arising in concrete engineering and scientific problems is to approximate their solutions numerically. Numerical methods for PDEs constitute an indivisible part of modern engineering and science.

The most general and efficient tool for the numerical solution of PDEs is the *Finite element method (FEM)*, which is based on the spatial subdivision of the physical domain into *finite elements* (often triangles or quadrilaterals in 2D and tetrahedra, bricks, or prisms in 3D), where the solution is approximated via a finite set of polynomial *shape functions*. In this way the original problem is transformed into a *discrete problem* for a finite number of unknown coefficients. It is worth mentioning that rather simple shape functions, such as affine or quadratic polynomials, have been used most frequently in the past due to their relatively low implementation cost. Nowadays, higher-order elements are becoming increasingly popular due to their excellent approximation properties and capability to reduce the size of finite element computations significantly.

The higher-order finite element methods, however, require a better knowledge of the underlying mathematics. In particular, the understanding of linear algebra and elementary functional analysis is necessary. In this book we follow the modern trend of building engineering finite element methods upon a solid mathematical foundation, which can be traced in several other recent finite element textbooks, as, e.g., [18] (membrane, beam and plate models), [29] (finite element analysis of shells), or [83] (edge elements for Maxwell's equations).

The contents at a glance

This book is aimed at graduate and Ph.D. students of all disciplines of computational engineering and science. It provides an introduction into the modern theory of partial differential equations, finite element methods, and their applications. The logical beginning of the text lies in Appendix A, which is a course in linear algebra and elementary functional analysis. This chapter is readable with minimum prerequisites and it contains many illustrative examples. Readers who trust their skills in function spaces and linear operators may skip Appendix A, but it will facilitate the study of PDEs and finite element methods to all others significantly.

The core Chapters 1–4 provide an introduction to the theory of PDEs and finite element methods. Chapter 5 is devoted to the numerical solution of ordinary differential equations (ODEs) which arise in the semidiscretization of time-dependent PDEs by the most frequently used *Method of lines (MOL)*. Emphasis is given to higher-order implicit one-step methods. Chapter 6 deals with Hermite and Argyris elements with application to fourth-order problems rooted in the bending of elastic beams and plates. Since the fourth-order problems are less standard than second-order equations, their physical background and derivation are discussed in more detail. Chapter 7 is a newcomer's introduction into computational electromagnetics. Explained are basic laws governing electromagnetics in both their integral and differential forms, material properties, constitutive relations, and interface conditions. Discussed are potentials and problems formulated in terms of potentials, and the time-domain and time-harmonic Maxwell's equations. The concept of Nédélec's *edge elements* for the Maxwell's equations is explained.

Appendix B deals with selected algorithmic and programming issues. We present a universal sparse matrix interface sMatrix which makes it possible to connect multiple sparse matrix solver packages simultaneously to a finite element solver. We mention the advantages of separating the finite element technology from the physics represented by concrete PDEs. Such approach is used in the implementation of a high-performance modular finite element system HERMES. This software is briefly described and applied to several challenging engineering problems formulated in terms of second-order elliptic PDEs and time-harmonic Maxwell's equations. Advantages of higher-order elements are demonstrated.

After studying this introductory text, the reader should be ready to read articles and monographs on advanced topics including a-posteriori error estimation and automatic adaptivity, mixed finite element formulations and saddle point problems, spectral finite element methods, finite element multigrid methods, hierarchic higher-order finite element methods (*hp*-FEM), and others (see, e.g., [9, 23, 69, 105] and [111]). Additional test and homework problems, along with an errata, will be maintained on my home page.

PAVEL ŠOLÍN

El Paso, Texas,
August, 2005

ACKNOWLEDGMENTS

I acknowledge with gratitude the assistance and help of many friends, colleagues and students in the preparation of the manuscript.[1] Tomáš Vejchodský (Academy of Sciences of the Czech Republic) read a significant part of the text and provided me with many corrections and hints that improved its overall quality. Martin Zítka (Charles University, Prague, and UTEP) checked Chapter 2 and made numerous useful observations to various other parts of the text. Invaluable was the expert review of the ODE Chapter 5 by Laurent Jay (University of Iowa). The functional-analytic course in Appendix A was reviewed by Volker John (Universität des Saarlandes, Saarbrücken) from the point of view of a numerical analyst, and by Osvaldo Mendez (UTEP), who is an expert in functional analysis. For numerous corrections to this part of the text I also wish to thank to UTEP's graduate students Svatava Vyvialová and Francisco Ávila.

I am deeply indebted to Prof. Ivo Doležel (Czech Technical University and Academy of Sciences of the Czech Republic), who is a theoretical electrical engineer with lively interest in computational mathematics, for providing me over the years with exciting practical problems to solve. Mainly thanks to him I learned to appreciate the engineer's point of view. The manuscript emerged from handouts, course notes, homeworks, and tests written for students. The students along with their interest and excitement were my main sources of motivation to write this book.

There is no way to express all my gratitude to my wife Dagmar for her support, understanding, and admirable patience during the two years of my work on the manuscript.

P. Š.

[1] The author acknowledges the support of the Czech Science Foundation under the Grant No. 102/05/0629.

Partial Differential Equations and the Finite Element Method
by Pavel Šolín
Copyright © 2006 John Wiley & Sons, Inc.

CHAPTER 1

PARTIAL DIFFERENTIAL EQUATIONS

Many natural processes can be sufficiently well described on the macroscopic level, without taking into account the individual behavior of molecules, atoms, electrons, or other particles. The averaged quantities such as the deformation, density, velocity, pressure, temperature, concentration, or electromagnetic field are governed by partial differential equations (PDEs). These equations serve as a language for the formulation of many engineering and scientific problems. To give a few examples, PDEs are employed to predict and control the static and dynamic properties of constructions, flow of blood in human veins, flow of air past cars and airplanes, weather, thermal inhibition of tumors, heating and melting of metals, cleaning of air and water in urban facilities, burning of gas in vehicle engines, magnetic resonance imaging and computer tomography in medicine, and elsewhere. Most PDEs used in practice only contain the first and second partial derivatives (we call them second-order PDEs).

Chapter 1 provides an overview of basic facts and techniques that are essential for both the qualitative analysis and numerical solution of PDEs. After introducing the classification and mentioning some general properties of second-order equations in Section 1.1, we focus on specific properties of elliptic, parabolic, and hyperbolic PDEs in Sections 1.2–1.4. Indeed, there are important PDEs which are not of second order. To mention at least some of them, in Section 1.5 we discuss first-order hyperbolic problems that are frequently used to model transport processes such as, e.g., inviscid fluid flow. Fourth-order problems rooted in the bending of elastic beams and plates are discussed later in Chapter 6.

1.1 SELECTED GENERAL PROPERTIES

Second-order PDEs (or PDE systems) encountered in physics usually are either elliptic, parabolic, or hyperbolic. Elliptic equations describe a special state of a physical system, which is characterized by the minimum of certain quantity (often energy). Parabolic problems in most cases describe the evolutionary process that leads to a steady state described by an elliptic equation. Hyperbolic equations describe the transport of some physical quantities or information, such as waves. Other types of second-order PDEs are said to be undetermined. In this introductory text we restrict ourselves to linear problems, since nonlinearities induce additional aspects whose understanding requires the knowledge of nonlinear functional analysis.

1.1.1 Classification and examples

Let $\mathcal{O}$ be an open connected set in $\mathbb{R}^n$. A sufficiently general form of a linear second-order PDE in n independent variables $z = (z_1, z_2, \ldots, z_n)^T$ is

$$-\sum_{i,j=1}^{n} \frac{\partial}{\partial z_i}\left(a_{ij}\frac{\partial u}{\partial z_j}\right) + \sum_{i=1}^{n}\left(\frac{\partial}{\partial z_i}(b_i u) + c_i\frac{\partial u}{\partial z_i}\right) + a_0 u = f, \tag{1.1}$$

where $a_{ij} = a_{ij}(z), b_i = b_i(z), c_i = c_i(z), a_0 = a_0(z)$ and $f = f(z)$. For all derivatives to exist in the classical sense, the solution and the coefficients have to satisfy the following regularity requirements: $u \in C^2(\mathcal{O}), a_{ij} \in C^1(\mathcal{O}), b_i \in C^1(\mathcal{O}), c_i \in C^1(\mathcal{O}), a_0 \in C(\mathcal{O})$, $f \in C(\mathcal{O})$. These regularity requirements will be reduced later when the PDE is formulated in the weak sense, and additional conditions will be imposed in order to ensure the existence and uniqueness of solution. If the functions a_{ij}, b_i, c_i, and a_0 are constants, the PDE is said to be with constant coefficients. Since the order of the partial derivatives can be switched for any twice continuously differentiable function u, it is possible to symmetrize the coefficients a_{ij} by defining

$$a_{ij}^{new} := (a_{ij}^{orig} + a_{ji}^{orig})/2$$

and adjusting the other coefficients accordingly so that the equation remains in the form (1.1). This is left to the reader as an exercise. Based on this observation, in the following we always will assume that the coefficient matrix $A(z) = \{a_{ij}\}_{i,j=1}^n$ is symmetric.

Recall that a symmetric $n \times n$ matrix A is said to be positive definite if

$$v^T A v > 0 \qquad \text{for all } 0 \neq v \in \mathbb{R}^n$$

and positive semidefinite if

$$v^T A v \geq 0 \qquad \text{for all } v \in \mathbb{R}^n.$$

Analogously one defines negative definite and negative semidefinite matrices by turning the inequalities. Matrices which do not belong to any of these types are said to be indefinite.

Definition 1.1 (Elliptic, parabolic and hyperbolic equations) *Consider a second-order PDE of the form (1.1) with a symmetric coefficient matrix $A(z) = \{a_{ij}\}_{i,j=1}^n$.*

1. *The equation is said to be* elliptic *at $z \in \mathcal{O}$ if $A(z)$ is positive definite.*

2. *The equation is said to be* parabolic *at $z \in \mathcal{O}$ if $A(z)$ is positive semidefinite, but not positive definite, and the rank of $(A(z), b(z), c(z))$ is equal to n.*

3. *The equation is said to be* hyperbolic *at $z \in \mathcal{O}$ if $A(z)$ has one negative and $n-1$ positive eigenvalues.*

An equation is called elliptic, parabolic, *or* hyperbolic *in the set $\mathcal{O}$ if it is elliptic, parabolic, or hyperbolic everywhere in $\mathcal{O}$, respectively.*

Remark 1.1 (Temporal variable t) *In practice we distinguish between time-dependent and time-independent PDEs. If the equation is time-independent, we put $n = d$ and $z = \boldsymbol{x}$, where d is the spatial dimension and $\boldsymbol{x}$ the spatial variable. This often is the case with elliptic equations. If the quantities in the equation depend on time, which often is the case with parabolic and hyperbolic equations, we put $n = d+1$ and $z = (\boldsymbol{x}, t)$, where t is the temporal variable. In such case the set $\mathcal{O}$ represents some space-time domain. If the spatial part of the space-time domain $\mathcal{O}$ does not change in time, we talk about a space-time cylinder $\Omega \times (0,T)$, where $\Omega \subset \mathbb{R}^d$ and $(0,T)$ is the corresponding time interval.*

Notice that, strictly speaking, the type of the PDE in Definition 1.1 is not invariant under multiplication by -1. For example, the equation

$$-\Delta u = f \qquad \left(\text{where } \Delta = \sum_{i=1}^{3} \frac{\partial^2}{\partial x_i^2} \text{ in } \mathbb{R}^3\right) \tag{1.2}$$

is elliptic everywhere in $\mathbb{R}^3$ since its coefficient matrix $\boldsymbol{A}$ is positive definite,

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

However, the type of the equation

$$\Delta u = -f$$

cannot be determined since its coefficient matrix

$$\boldsymbol{A} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

is negative definite. In such cases it is customary to multiply the equation by (-1) so that Definition 1.1 can be applied. Moreover, notice that Definition 1.1 only applies to second-order PDEs. Later in this text we will discuss two important cases outside of this classification: hyperbolic first-order systems in Section 1.5 and elliptic fourth-order problems in Chapter 6.

Remark 1.2 *Sometimes, linear second-order PDEs are found in a slightly different form*

$$-\sum_{i,j=1}^{n} \tilde{a}_{ij}(z)\frac{\partial^2 u}{\partial z_i \partial z_j} + \sum_{i=1}^{n} \tilde{b}_i(z)\frac{\partial u}{\partial z_i} + \tilde{a}_0(z)u = f(z), \tag{1.3}$$

usually with a symmetric coefficient matrix $\tilde{A}(z) = \{\tilde{a}_{ij}\}_{i,j=1}^{n}$. When transforming (1.3) into the form (1.1), it is easy to see that the matrices $\tilde{A}(z)$ and $A(z)$ are identical, and *thus either one can be used to determine the ellipticity, parabolicity, or hyperbolicity of the problem. Moreover, if the coefficients $\tilde{a}_{ij}$ and $\tilde{b}_i$ are sufficiently smooth, the two forms are equivalent.*

Operator notation It is customary to write elliptic PDEs in a compact form

$$Lu = f,$$

where L defined by

$$Lu = -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right) + \sum_{i=1}^{n}\left(\frac{\partial}{\partial x_i}(b_i u) + c_i\frac{\partial u}{\partial x_i}\right) + a_0 u \tag{1.4}$$

is a second-order elliptic differential operator. The part of L with the highest derivatives,

$$-\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial}{\partial x_j}\right). \tag{1.5}$$

is called the principal (leading) part of L. Most parabolic and hyperbolic equations are motivated in physics, and therefore one of the independent variables usually is the time t. The typical operator form of parabolic equations is

$$\frac{\partial u}{\partial t} + Lu = f, \tag{1.6}$$

where L is an elliptic differential operator. Typical second-order hyperbolic equation can be seen in the form

$$\frac{\partial^2 u}{\partial t^2} + Lu = f. \tag{1.7}$$

where again L is an elliptic differential operator. The following examples show simple elliptic, parabolic, and hyperbolic equations.

■ EXAMPLE 1.1 (Elliptic PDE: Potential equation of electrostatics)

Let the function $\rho \in C(\overline{\Omega})$ represent the electric charge density in some open bounded set $\Omega \subset \mathbb{R}^d$. If the permittivity ϵ is constant in Ω, the distribution of the electric potential φ in Ω is governed by the Poisson equation

$$-\epsilon\Delta\varphi = \rho. \tag{1.8}$$

Notice that (1.8) does not possess a unique solution, since for any solution φ the function $\varphi + C$, where C is an arbitrary constant, also is a solution. In order to yield a well-posed problem, every elliptic equation has to be endowed with suitable boundary conditions. This will be discussed in Section 1.2.

■ **EXAMPLE 1.2 (Parabolic PDE: Heat transfer equation)**

Let $\Omega \subset \mathbb{R}^d$ be an open bounded set and $q \in C(\overline{\Omega})$ the volume density of heat sources in Ω. If the thermal conductivity k, material density ϱ, and specific heat c are constant in Ω, the parabolic equation

$$\frac{\partial \theta}{\partial t} - \frac{k}{\varrho c}\Delta\theta = \frac{q}{\varrho c} \tag{1.9}$$

describes the evolution of the temperature $\theta(\boldsymbol{x}, t)$ in Ω. The steady state of the temperature ($\partial\theta/\partial t = 0$) is described by the corresponding elliptic equation

$$-k\Delta\theta = q.$$

Similarly to the previous case, the solution θ is not determined by (1.9) uniquely. Parabolic equations have to be endowed with both boundary and initial conditions in order to yield a well-posed problem. This will be discussed in Section 1.3.

■ **EXAMPLE 1.3 (Hyperbolic PDE: Wave equation)**

Let $\Omega \subset \mathbb{R}^d$ be an open bounded set. The speed of sound a can be considered constant in Ω if the motion of the air is sufficiently slow. Then the hyperbolic equation

$$\frac{\partial^2 p}{\partial t^2} - a^2\Delta p = 0 \tag{1.10}$$

describes the propagation of sound waves in Ω. Here the unknown function $p(\boldsymbol{x}, t)$ represents the pressure, or its fluctuations around some arbitrary constant equilibrium pressure. Again the function p is not determined by (1.10) uniquely. Hyperbolic equations have to be endowed with both boundary and initial conditions in order to yield a well-posed problem. Definition of boundary conditions for hyperbolic problems is more difficult compared to the elliptic or parabolic case, since generally they depend on the choice of the initial data and on the solution itself. We will return to this issue in Example 1.4 and in more detail in Section 1.5.

1.1.2 Hadamard's well-posedness

The notion of well-posedness of boundary-value problems for partial differential equations was established around 1932 by Jacques Salomon Hadamard.

J.S. Hadamard was a French mathematician who contributed significantly to the analysis of Taylor series and analytic functions of the complex variable, prime number theory, study of matrices and determinants, boundary value problems for partial differential equations, probability theory, Markov chains, several areas of mathematical physics, and education of mathematics.

Definition 1.2 (Hadamard's well-posedness) *A problem is said to be* well-posed *if*

1. *it has a unique solution,*
2. *the solution depends continuously on the given data.*

Otherwise the problem is ill-posed.

Figure 1.1 Jacques Salomon Hadamard (1865–1963).

As the reader may expect, well-posed problems are more pleasant to deal with than the ill-posed ones. The requirement of existence and uniqueness of solution is obvious. The other condition in Definition 1.2 denies well-posedness to problems with unstable solutions. From the point of view of numerical solution of PDEs, the computational domain Ω, boundary and initial conditions, and other parameters are not represented exactly in the computer model. Additional source of error is the finite computer arithmetics. If a problem is well-posed, one has a chance to compute a reasonable approximation of the unique exact solution as long as the data to the problem are approximated reasonably. Such expectation may not be realistic at all if the problem is ill-posed.

The concept of well-posedness deserves to be discussed in more detail. First let us show in Example 1.4 that well-posedness may be violated by endowing a PDE with wrong boundary conditions.

■ EXAMPLE 1.4 (Ill-posedness due to wrong boundary conditions)

Consider an interval $\Omega = (-a, a)$, $a > 0$, and the (inviscid) Burgers' equation

$$\frac{\partial}{\partial t}u(x,t) + u(x,t)\frac{\partial}{\partial x}u(x,t) = 0. \tag{1.11}$$

This equation is endowed with the initial condition

$$u(x,0) = u_0(x) = x, \quad x \in \Omega, \tag{1.12}$$

where u_0 is a function continuous in $(-a, a)$ such that $u_0(\pm a) = \pm a$, and the boundary conditions

$$u(\pm a, t) = \pm a, \quad t > 0. \tag{1.13}$$

The (inviscid) Burgers' equation is an important representant of the class of first-order hyperbolic problems that will be studied in more detail in Section 1.5. In particular, after reading Paragraph 1.5.5 the reader will know that every function $u(x,t)$ that satisfies both equation (1.11) and initial condition (1.12) is constant along the lines

$$x_{x_0}(t) = x_0(t+1), \quad x_0 \in \Omega, \tag{1.14}$$

depicted in Figure 1.2.

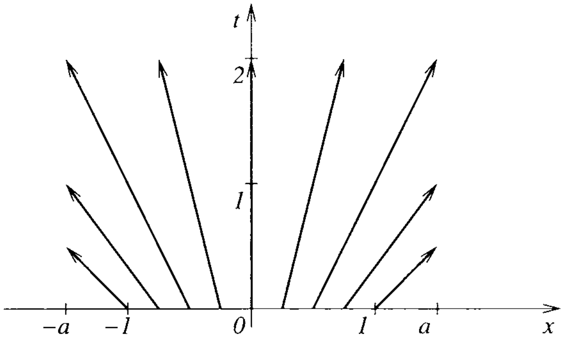


Figure 1.2 Isolines of the solution $u(x,t)$ of Burgers' equation.

It is easy to check the constantness of the solution u along the lines (1.14) by performing the derivative

$$\frac{\mathrm{d}}{\mathrm{d}t}u(x_{x_0}(t),t).$$

From this fact it follows that the solution to (1.11), (1.12) cannot be constant in time at the endpoints of Ω. Hence the problem (1.11), (1.12), (1.13) has no solution.

Some problems are ill-posed because of their very nature, despite their initial and boundary conditions are defined appropriately. This is illustrated in Example 1.5.

■ EXAMPLE 1.5 (Ill-posed problem with unstable solution)

Consider the one-dimensional version of the heat transfer equation (1.9) with normalized coefficients,

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0, \tag{1.15}$$

describing the temperature distribution within a thin slab $\Omega = (0,\pi)$ in the time interval $(0,T)$. We choose an initial temperature distribution $u(x,0) = u_0(x)$ such that $u_0(0) = u_0(\pi) = 0$, fix the temperature at the endpoints to $u(0) = u(\pi) = 0$ and ask about the solution $u(x,t)$ of (1.15) for $t \in (0,T)$. The initial condition $u_0(x)$ can be expressed by means of the Fourier expansion

$$u_0(x) = \sum_{n=1}^{\infty} c_n \sin(nx). \tag{1.16}$$

Thus it is easy to verify that the exact solution $u(x,t)$ has the form

$$u(x,t) = \sum_{n=1}^{\infty} c_n e^{-n^2 t} \sin(nx) \tag{1.17}$$

and hence that

$$u(x,T) = \sum_{n=1}^{\infty} c_n e^{-n^2 T} \sin(nx) \tag{1.18}$$

is the solution corresponding to the time $t = T$. Notice that the coefficients $c_n e^{-n^2 t}$ converge to zero very fast as the time grows, and therefore after a sufficiently long time T the solution will be very close to zero in Ω. Hence, the heat transfer problem evidently is a well-posed in the sense of Hadamard.

Now let us reverse the time by defining a new temporal variable $s = T - t$. The backward heat transfer equation has the form

$$\frac{\partial \hat{u}}{\partial s} + \frac{\partial^2 \hat{u}}{\partial x^2} = 0.$$

We consider an initial condition $\hat{u}_0(x)$ corresponding to $s = 0$, i.e., to $t = T$. Again, $\hat{u}_0(x)$ can be expressed as

$$\hat{u}_0(x) = \sum_{n=1}^{\infty} d_n \sin(nx), \tag{1.19}$$

and the exact solution $\hat{u}(x,s)$ has the form

$$\hat{u}(x,s) = \sum_{n=1}^{\infty} d_n e^{n^2 s} \sin(nx).$$

Notice that now the coefficients $d_n e^{n^2 s}$ are amplified exponentially as the backward temporal variable s grows. This means that the solution of the backward heat transfer equation does not depend continuously on the initial data $\hat{u}_0(x)$, i.e., that the backward problem is ill-posed.

Suppose that we calculate some numerical approximation of the solution $u(x,T)$ for some sufficiently large time T and then use it as the initial condition $\hat{u}_0(x)$ for the backward problem. What we will observe when solving the backward problem is that the solution $\hat{u}(x,s)$ begins to oscillate immediately and the computation ends with a floating point overflow or similar error very soon. Because of the ill-posedness of the backward problem, chances are slim that one can get close to the original initial condition $u_0(x)$ at $s = T$.

Remark 1.3 (Inverse problems) *The ill-posed backward heat transfer equation from Example 1.5 was an inverse problem. There are various types of ill-posed inverse problems: For example, it is an inverse problem to identify suitable initial state and/or parameters for some physical process to obtain a desired final state. Usually, the better-posed the forward problem, the worse the posedness of the inverse problem.*

1.1.3 General existence and uniqueness results

Prior to discussing various aspects of the elliptic, parabolic, and hyperbolic PDEs in Sections 1.2–1.5, we find it useful to mention a few important abstract existence and uniqueness results for general operator equations. Since this paragraph uses some abstract functional analysis, readers who find its contents too difficult may skip it in the first reading and continue with Section 1.2.

In the following we consider a pair of Hilbert spaces V and W, and an equation of the form

$$Lu = f, \tag{1.20}$$

where $L : D(L) \subset V \to W$ is a linear operator and $f \in W$. The existence of solution to (1.20) for any right-hand side $f \in W$ is equivalent to the condition $R(L) = W$, while the uniqueness of solution is equivalent to the condition $N(L) = \{0\}$.

Theorem 1.1 (Hahn–Banach) *Let U be a subspace of a (real or complex) normed space V, and $f \in U'$ a linear form over U. Then there exists an extension $g \in V'$ of f such that $g(u) = f(u)$ for all $u \in U$, moreover satisfying $\|g\|_{V'} = \|f\|_{U'}$.*

Proof: The proof can be found in standard functional-analytic textbooks. See, e.g., [34, 65] and [100]. ■

Theorem 1.1 has important consequences: If $v_0 \in V$ and $f(v_0) = 0$ for all $f \in V'$, then $v_0 = 0$. Further, for any $v_0 \in V$ there exists $f \in V'$ such that $\|f\|_V = 1$ and $f(v_0) = \|v_0\|_V$. The following result is used in the proof of the basic existence theorem: For any two disjoint subsets $A, B \subset V$, where A is compact and B convex, there exists $f \in V'$ and $\gamma \in \mathbb{R}$ such that $f(a) < \gamma < f(b)$ for all $a \in A$ and $b \in B$.

Theorem 1.2 (Basic existence result) *Let V, W be Hilbert spaces and $L : D(L) \subset V \to W$ a bounded linear operator. Then $R(L) = W$ if and only if both $R(L)$ is closed and $R(L)^\perp = \{0\}$.*

Proof: If $R(L) = W$, then obviously $R(L)$ is closed and $R(L)^\perp = \{0\}$. Conversely, assume that $R(L)$ is closed, $R(L)^\perp = \{0\}$ but $R(L) \neq W$. The linearity and boundedness of L implies that $R(L)$ is a closed subspace of W. Let $w \in W \setminus R(L)$. The set $\{w\}$ is compact and the closed set $R(L)$ obviously is convex. By the Hahn–Banach theorem there exists a $w^* \in W'$ such that $(w^*, w) > 0$ and $(w^*, Lv) = 0$ for all $v \in D(L)$. Therefore $0 \neq w^* \in R(L)^\perp$, which is a contradiction. ■

In order to see under what conditions $R(L)$ is closed, let us generalize the notion of continuity by introducing closed operators:

Definition 1.3 (Closed operator) *An operator $T : D(T) \subset V \to W$, where V and W are Banach spaces, is said to be* closed *if for any sequence $\{v_n\}_{n=1}^\infty \subset D(T)$, $v_n \to v$ and $T(v_n) \to w$ imply that $v \in D(T)$ and $w = Tv$.*

It is an easy exercise to show that every continuous operator is closed. However, there are closed operators which are not continuous:

■ **EXAMPLE 1.6 (Closed operator which is not continuous)**

Consider the interval $\Omega = (0, 1) \subset \mathbb{R}$, the Hilbert space $V = L^2(\Omega)$ and the Laplace operator $L : V \to V$, $Lu = -\Delta u = -u''$. This operator is not continuous, since, e.g., $Lv \notin V$ for $v = x^{-1/3} \in V$. We know that the space $C_0^\infty(\Omega)$ is dense in $L^2(\Omega)$ (see Paragraph A.2.10). To show that L is closed in V, for an element $v \in V$ consider some sequence $\{v_n\}_{n=1}^\infty \subset C_0^\infty(\Omega)$ such that $v_n \to v$, and such that the sequence $\{-\Delta v_n\}_{n=1}^\infty$ converges to some $w \in V$. Passing to the limit $n \to \infty$ in the relation

$$\int_\Omega -\Delta v_n \varphi \,\mathrm{d}\boldsymbol{x} = -\int_\Omega v_n \Delta\varphi \,\mathrm{d}\boldsymbol{x} \quad \text{for all } \varphi \in C_0^\infty(\Omega),$$

we obtain

$$\int_\Omega w\varphi \,\mathrm{d}\boldsymbol{x} = -\int_\Omega v\Delta\varphi \,\mathrm{d}\boldsymbol{x} \quad \text{for all } \varphi \in C_0^\infty(\Omega).$$

Therefore $w = -\Delta v$ and the operator L is closed.

Theorem 1.3 (Basic existence and uniqueness result) *Let V, W be Hilbert spaces and $L : D(L) \subset V \to W$ a closed linear operator. Assume that there exists a constant $C > 0$ such that*

$$\|Lv\|_W \geq C\|v\|_V \quad \textit{for all } v \in D(L) \tag{1.21}$$

(this inequality sometimes is called the stability or coercivity estimate). If $R(L)^\perp = \{0\}$, then the operator equation $Lu = f$ has a unique solution.

Proof: First let us verify that $R(L)$ is closed. Let $\{w_n\}_{n=1}^\infty \subset R(L)$ such that $w_n \to w$. Then there is a sequence $\{v_n\}_{n=1}^\infty \subset D(L)$ such that $w_n = Lv_n$. The stability estimate (1.21) implies that $C\|v_n - v_m\|_V \leq \|w_n - w_m\|_W$, which means that $\{v_n\}_{n=1}^\infty$ is a Cauchy sequence in V. Completeness of the Hilbert space V yields existence of a $v \in V$ such that $v_n \to v$. Since L is closed, we obtain $v \in D(L)$ and $w = Lv \in R(L)$. Theorem 1.2 yields the existence of a solution. The uniqueness of the solution follows immediately from the stability estimate (1.21). ■

Now let us introduce the notion of monotonicity and show that strongly monotone linear operators satisfy the stability estimate (1.21):

Definition 1.4 (Monotonicity) *Let V be a Hilbert space and $L \in \mathcal{L}(V, V')$. The operator L is said to be* monotone *if*

$$\langle Lv, v\rangle \geq 0 \quad \textit{for all } v \in V, \tag{1.22}$$

it is strictly monotone *if*

$$\langle Lv, v\rangle > 0 \quad \textit{for all } 0 \neq v \in V, \tag{1.23}$$

and it is strongly monotone *if there exists a constant $C_L > 0$ such that*

$$\langle Lv, v\rangle \geq C_L\|v\|^2 \quad \textit{for all } v \in V. \tag{1.24}$$

For every $u \in V$ the element $Lu \in V'$ is a linear form. The symbol $\langle Lv, v\rangle$, which means the application of Lu to $v \in V$, is called duality pairing.

The notion of monotonicity for linear operators is a special case of a more general definition applicable to nonlinear operators. An operator $T : V \longrightarrow V'$ is said to be monotone if $\langle Tu - Tv, u - v\rangle \geq 0$ for all $u, v \in V$, it is strictly monotone if $\langle Tu - Tv, u - v\rangle > 0$ for all $u, v \in V$, $u \neq v$, and it is strongly monotone if there exists a positive constant C_L such that $\langle Tu - Tv, u - v\rangle \geq C_L\|u - v\|^2$ for all $u, v \in V$. The concept of monotonicity for operators is related to the standard notion of monotonicity of real functions: A function $f : \mathbb{R} \longrightarrow \mathbb{R}$ is monotone if the condition $x_1 < x_2$ implies that $f(x_1) \leq f(x_2)$. The same can be written as the condition $(f(x_1) - f(x_2))(x_1 - x_2) \geq 0$ for all $x_1, x_2 \in \mathbb{R}$.

Lemma 1.1 *Let V be a Hilbert space and $L \in \mathcal{L}(V, V')$ a continuous strongly monotone linear operator. Then there exists a constant $C > 0$ such that L satisfies the stability estimate (1.21).*

Proof: The strong monotonicity condition (1.24) implies

$$C_L\|v\|_V^2 \leq \langle Lv, v\rangle \leq \|Lv\|_{V'}\|v\|_V,$$

which means that

$$C_L\|v\|_V \leq \|Lv\|_{V'}$$

■

The following theorem presents an important abstract existence and uniqueness result for operator equations:

Theorem 1.4 (Existence and uniqueness of solution for strongly monotone operators) *Let V be a Hilbert space, $f \in V'$ and $L \in \mathcal{L}(V, V')$ a strongly monotone linear operator. Then for every $f \in V'$ the operator equation $Lu = f$ has a unique solution $u \in V$.*

Proof: According to Lemma 1.1 the operator L satisfies the stability estimate (1.21). Moreover, if $v \in R(L)^{\perp}$, then $\langle Lv, v\rangle = 0$ and

$$C\|v\|^2 \leq \langle Lv, v\rangle = 0.$$

Hence $R(L)^{\perp} = \{0\}$, and the conclusion follows from Theorem 1.3. ■

1.1.4 Exercises

Exercise 1.1 *Use Definition 1.3 to show that every continuous operator $L : V \longrightarrow W$, where V and W are Banach spaces, is closed.*

Exercise 1.2 *Consider a second-order PDE in the form (1.1) with a nonsymmetric coefficient matrix $A(z)$. Symmetrize the coefficient matrix by defining $\tilde{A} = (A + A^T)/2$. Find out how the remaining coefficients b_i, c_i, and a_0 have to be adjusted so that the equation remains in the form (1.1). Hint: Write $a_{ij} = (a_{ij} + a_{ji})/2 + (a_{ij} - a_{ji})/2$.*

Exercise 1.3 *Consider a second-order PDE in the alternative form (1.3),*

$$-\sum_{i,j=1}^{n} \tilde{a}_{ij}\frac{\partial^2 u}{\partial z_i \partial z_j} + \sum_{i=1}^{n} \tilde{b}_i \frac{\partial u}{\partial z_i} + \tilde{a}_0 u = f.$$

where $\tilde{a}_{ij} = \tilde{a}_{ji}$ *for all* $1 \leq i, j \leq n$.

1. *Turn the equation into the conventional form (1.1),*

$$-\sum_{i,j=1}^{n} \frac{\partial}{\partial z_i}\left(a_{ij}\frac{\partial u}{\partial z_j}\right) + \sum_{i=1}^{n}\left(\frac{\partial}{\partial z_i}(b_i u) + c_i\frac{\partial u}{\partial z_i}\right) + a_0 u = f.$$

2. *Write the relations of the coefficients* a_{ij}, b_i, c_i, a_0 *and* $\tilde{a}_{ij}, \tilde{b}_i, \tilde{c}_i, \tilde{a}_0$.

Exercise 1.4 *Use Definition 1.1 to show that equation (1.8) from Example 1.1 is elliptic.*

Exercise 1.5 *Use Definition 1.1 to show that equation (1.9) from Example 1.2 is parabolic.*

Exercise 1.6 *Use Definition 1.1 to show that equation (1.10) from Example 1.3 is hyperbolic.*

Exercise 1.7 *Verify that the function* $u(x, t)$ *defined in* $(0, \pi)$ *by the relation (1.17) is the solution of the heat-transfer equation (1.15) with the boundary conditions* $u(0, t) = u(\pi, t) = 0$ *for all* $t > 0$.

Exercise 1.8 *In* $\mathbb{R}^3$ *consider the equation*

$$\frac{\partial u}{\partial t} - \left(1 + x_1^2\right)\frac{\partial^2 u}{\partial x_1^2} - \left(1 + x_2^4\right)\frac{\partial^2 u}{\partial x_2^2} - \left(1 + x_3^6\right)\frac{\partial^2 u}{\partial x_3^2} + \sqrt{1 + |\boldsymbol{x}|^2}\frac{\partial u}{\partial x_3} = e^{-|\boldsymbol{x}|}$$

and decide if (and where in $\mathbb{R}^3$*) it is elliptic, parabolic, or hyperbolic.*

Exercise 1.9 *In* $\mathbb{R}^2$ *consider the equation*

$$\frac{\partial^2 u}{\partial t^2} + \left(1 - x_1^2\right)\frac{\partial^2 u}{\partial x_1^2} + \left(1 - x_2^2\right)\frac{\partial^2 u}{\partial x_2^2} - x_1 x_2 \frac{\partial u}{\partial x_1} = \sin(x_1\pi)\cos(x_2\pi).$$

and decide if (and where in $\mathbb{R}^2$*) it is elliptic, parabolic, or hyperbolic.*

Exercise 1.10 *In* $\mathbb{R}^3$ *consider the equation*

$$-\Delta u - 2\frac{\partial^2 u}{\partial x_1 x_2} - 2\frac{\partial^2 u}{\partial x_2 x_3} = f$$

and decide if (and where in $\mathbb{R}^2$*) it is elliptic, parabolic, or hyperbolic.*

Exercise 1.11 *In* $\mathbb{R}^3$ *consider the equation*

$$(1 - x_1^2 - x_2^2)\frac{\partial^2 u}{\partial t^2} - (1 + x_2^2)\frac{\partial^2 u}{\partial x_1^2} - (1 + x_1^4)\frac{\partial^2 u}{\partial x_2^2} + (1 + x_3^2)\frac{\partial u}{\partial x_3} = e^{-|\boldsymbol{x}|^2}$$

and decide if (and where in $\mathbb{R}^3$*) it is elliptic, parabolic, or hyperbolic.*

Exercise 1.12 *In* $\mathbb{R}^2$ *consider the equation*

$$\frac{\partial^2 u}{\partial t^2} - \left(1 - |\boldsymbol{x}|^2\right)\frac{\partial^2 u}{\partial x_1^2} - \left(1 - \frac{|\boldsymbol{x}|^2}{4}\right)\frac{\partial^2 u}{\partial x_2^2} = 0$$

and decide if (or where in $\mathbb{R}^2$*) it is elliptic, parabolic, or hyperbolic.*

1.2 SECOND-ORDER ELLIPTIC PROBLEMS

This section is devoted to the discussion of linear second-order elliptic problems. We begin by deriving the weak formulation of a model problem in Paragraph 1.2.1. Properties of bilinear forms arising in the weak formulation of linear elliptic problems are discussed in Paragraph 1.2.2. In Paragraph 1.2.3 we introduce the Lax–Milgram lemma, which is the basic tool for proving the existence and uniqueness of solution to linear elliptic problems. The weak formulations and solvability analysis of problems involving various types of boundary conditions are discussed in Paragraphs 1.2.5–1.2.8. Abstract energy of elliptic problems, which plays an important role in their numerical solution (error estimation, automatic adaptivity), is introduced in Paragraph 1.2.9. Finally, Paragraph 1.2.10 presents maximum principles for elliptic problems, which are used to prove their well-posedness.

1.2.1 Weak formulation of a model problem

Assume an open bounded set $\Omega \subset \mathbb{R}^d$ with Lipschitz-continuous boundary, and recall the general linear second-order equation (1.1),

$$-\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right) + \sum_{i=1}^{n}\left(\frac{\partial}{\partial x_i}(b_i u) + c_i\frac{\partial u}{\partial x_i}\right) + a_0 u = f, \tag{1.25}$$

where the coefficients and the right-hand side satisfy the regularity assumptions formulated in Paragraph 1.1.1. In this case we put $n = d$. Equation (1.25) is elliptic if the symmetric coefficient matrix $A = \{a_{ij}\}_{i,j=1}^{d}$ is positive definite everywhere in Ω (Definition 1.1).

Consider the model equation

$$-\nabla \cdot (a_1 \nabla u) + a_0 u = f \qquad \text{in } \Omega, \tag{1.26}$$

obtained from (1.25) by assuming $a_{ij}(\boldsymbol{x}) = a_1(\boldsymbol{x})\delta_{ij}$ and $\boldsymbol{b}(\boldsymbol{x}) = \boldsymbol{c}(\boldsymbol{x}) = 0$ in Ω. For the existence and uniqueness of solution we add another important assumption:

$$a_1(\boldsymbol{x}) \geq C_{min} > 0 \quad \text{and} \quad a_0(\boldsymbol{x}) \geq 0 \quad \text{in } \Omega. \tag{1.27}$$

The problem (1.26) is fairly general: Even with $a_0 \equiv 0$ it describes, for example, the following physical processes:

1. Stationary heat transfer (u is the temperature, a_1 is the thermal conductivity, and f are the heat sources),

2. electrostatics (u is the electrostatic potential, a_1 is the dielectric constant, and f is the charge density),

3. transverse deflection of a cable (u is the transverse deflection, a_1 is the axial tension, and f is the transversal load),

4. axial deformation of a bar (u is the axial displacement, $a_1 = EA$ is the product of the elasticity modulus and the cross-sectional area, and f is either the friction or contact force on the surface of the bar),

5. pipe flow (u is the hydrostatic pressure, $a_1 = \pi D^4/128\mu$, D is the diameter, μ is the viscosity and $f = 0$ represents zero flow sources),

6. laminar incompressible flow through a channel under constant pressure gradient (u is the velocity, a_1 is the viscosity, and f is the pressure gradient),
7. porous media flow (u is the fluid head, a_1 is the permeability coefficient, and f is the fluid flux).

To begin with, let (1.26) be endowed with homogeneous Dirichlet boundary conditions

$$u(\boldsymbol{x}) = 0 \quad \text{on } \partial\Omega. \tag{1.28}$$

This type of boundary conditions carries the name of a French mathematician Johann Peter Gustav Lejeune Dirichlet, who made substantial contributions to the solution of Fermat's Last Theorem, theory of polynomial functions, analytic and algebraic number theory, convergence of trigonometric series, and boundary-value problems for harmonic[1] functions.

Figure 1.3 Johann Peter Gustav Lejeune Dirichlet (1805–1859).

Classical solution to the problem (1.26), (1.28) is a function $u \in C^2(\Omega) \cap C(\overline{\Omega})$ satisfying the equation (1.26) everywhere in Ω and fulfilling the boundary condition (1.28) at every $\boldsymbol{x} \in \partial\Omega$. Naturally, one has to assume that $f \in C(\Omega)$. However, neither this nor even stronger requirement $f \in C(\overline{\Omega})$ guarantees the solvability of the problem, for which still stronger smoothness of f is required.

Weak formulation In order to reduce the above-mentioned regularity restrictions, we introduce the weak formulation of the problem (1.26), (1.28). The derivation of the weak formulation of (1.26) consists of the following four standard steps:

1. Multiply (1.26) with a test function $v \in C_0^\infty(\Omega)$,

$$-\nabla \cdot (a_1 \nabla u)v + a_0 uv = fv.$$

2. Integrate over Ω,

[1] $\Delta u = 0$

$$-\int_\Omega \nabla \cdot (a_1 \nabla u) v \, \mathrm{d}\boldsymbol{x} + \int_\Omega a_0 u v \, \mathrm{d}\boldsymbol{x} = \int_\Omega f v \, \mathrm{d}\boldsymbol{x}.$$

3. Use the Green's formula (A.80) to reduce the maximum order of the partial derivatives present in the equation. The fact that v vanishes on the boundary $\partial\Omega$ removes the boundary term, and we have

$$\int_\Omega a_1 \nabla u \cdot \nabla v \, \mathrm{d}\boldsymbol{x} + \int_\Omega a_0 u v \, \mathrm{d}\boldsymbol{x} = \int_\Omega f v \, \mathrm{d}\boldsymbol{x}. \tag{1.29}$$

4. Find the largest possible function spaces for u, v, and other functions in (1.29) where all integrals are finite. Originally, identity (1.29) was derived under very strong regularity assumptions $u \in C^2(\Omega) \cap C(\overline{\Omega})$ and $v \in C_0^\infty(\Omega)$. All integrals in (1.29) remain finite when these assumptions are weakened to

$$u, v \in H_0^1(\Omega), \qquad f \in L^2(\Omega), \tag{1.30}$$

where $H_0^1(\Omega)$ is the Sobolev space $W_0^{1,2}(\Omega)$ defined in Section A.4. Similarly the regularity assumptions for the coefficients a_1 and a_0 can be reduced to

$$a_1, a_0 \in L^\infty(\Omega). \tag{1.31}$$

The weak form of the problem (1.26), (1.28) is stated as follows: Given $f \in L^2(\Omega)$, find a function $u \in H_0^1(\Omega)$ such that

$$\int_\Omega a_1 \nabla u \cdot \nabla v + a_0 u v \, \mathrm{d}\boldsymbol{x} = \int_\Omega f v \, \mathrm{d}\boldsymbol{x} \quad \text{for all } v \in H_0^1(\Omega). \tag{1.32}$$

The existence and uniqueness of solution will be discussed in Paragraph 1.2.4.

Let us mention that the assumption $f \in L^2(\Omega)$ can be further weakened to $f \in H^{-1}(\Omega)$, where $H^{-1}(\Omega)$, which is the dual space to $H_0^1(\Omega)$, is larger than $L^2(\Omega)$. Then the integral

$$\int_\Omega f v \, \mathrm{d}\boldsymbol{x}$$

is interpreted as the duality pairing $\langle f, v\rangle$ between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$.

Equivalence of the strong and weak solutions Obviously the classical solution to the problem (1.26), (1.28) also solves the weak formulation (1.32). Conversely, if the weak solution of (1.32) is sufficiently regular, which in this case means $u \in C^2(\Omega) \cap C(\overline{\Omega})$, it also satisfies the classical formulation (1.26), (1.28).

In the language of linear forms Let $V = H_0^1(\Omega)$. We define a bilinear form $a(\cdot,\cdot) : V \times V \to \mathbb{R}$,

$$a(u, v) = \int_\Omega (a_1 \nabla u \cdot \nabla v + a_0 u v) \, \mathrm{d}\boldsymbol{x},$$

and a linear form $l \in V'$,

$$l(v) = \langle l, v\rangle = \int_\Omega fv \, \mathrm{d}x.$$

Then the weak formulation of the problem (1.26), (1.28) reads: Find a function $u \in V$ such that

$$a(u, v) = l(v) \quad \text{for all } v \in V. \tag{1.33}$$

This notation is common in the study of partial differential equations and finite element methods.

1.2.2 Bilinear forms, energy norm, and energetic inner product

In this paragraph we learn more about bilinear forms for elliptic problems, and introduce the notions of energy norm and energetic inner product. Every bilinear form $a : V \times V \longrightarrow \mathbb{R}$ in a Banach space V is associated with a unique linear operator $A : V \longrightarrow V'$ defined by

$$(Au)(v) = \langle Au, v\rangle = a(u, v) \quad \text{for all } u, v \in V. \tag{1.34}$$

Lemma 1.2 *Relation (1.34) defines a one-to-one correspondence between continuous bilinear forms $a : V \times V \longrightarrow \mathbb{R}$ and linear continuous operators $A : V \longrightarrow V'$.*

Proof: If $A \in \mathcal{L}(V, V')$, then the mapping $a : V \times V \longrightarrow \mathbb{R}$ defined by (1.34) is bilinear and bounded,

$$|a(u, v)| \leq \|Au\|_{V'}\|v\|_V \leq \|A\|\|u\|_V\|v\|_V \quad \text{for all } u, v \in V.$$

Conversely, let $a(\cdot, \cdot)$ be a continuous bilinear form on $V \times V$. For any $u \in V$ the map $v \longrightarrow a(u, v)$ defines a continuous linear operator on V. Hence there exists an element $Au \in V'$ such that (1.34) holds. The bilinearity and boundedness of $a(\cdot, \cdot)$ implies the linearity and boundedness of A. ■

Basic properties of bilinear forms in Hilbert spaces are introduced in Definition 1.5 and discussed in Lemma 1.3:

Definition 1.5 *Let V be a real Hilbert space, $a : V \times V \longrightarrow \mathbb{R}$ a bilinear form and $A : V \longrightarrow V'$ a linear operator related to $a(\cdot, \cdot)$ via (1.34). We say that*

1. *a is* bounded *if there exists a constant $C_a > 0$ such that $|a(u, v)| \leq C_a\|u\|\|v\|$ for all $u, v \in V$,*

2. *a is* positive *if $a(v, v) \geq 0$ for all $v \in V$,*

3. *a is* strictly positive *if $a(v, v) > 0$ for all $0 \neq v \in V$,*

4. *a is* V-elliptic (coercive) *if there exists a constant $\tilde{C}_a > 0$ such that $a(v, v) \geq \tilde{C}_a\|v\|_V^2$ for all $v \in V$,*

5. *a is* symmetric *if $a(u, v) = a(v, u)$ for all $u, v \in V$.*

Lemma 1.3 *Under the assumptions of Definition 1.5 it holds:*

1. *The bilinear form a is bounded if and only if the linear operator A is bounded.*
2. *The bilinear form a is positive if and only if the linear operator A is monotone.*
3. *The bilinear form a is strictly positive if and only if the linear operator A is strictly monotone.*
4. *The bilinear form a is V-elliptic if and only if the linear operator A is strongly monotone.*
5. *The bilinear form a is symmetric if and only if the linear operator A is symmetric (i.e., if $\langle Au, v\rangle = \langle Av, u\rangle$ for all $u, v \in V$).*

Proof: Left to the reader as an exercise. ■

Definition 1.6 (Energetic inner product, energy norm) *Let V be a Hilbert space and $a : V \times V \to \mathbb{R}$ a bounded symmetric V-elliptic bilinear form. The bilinear form defines an inner product*

$$(u, v)_e = a(u, v) \tag{1.35}$$

in V, called energetic inner product. *The norm induced by the energetic inner product,*

$$\|u\|_e = \sqrt{(u, u)_e}, \tag{1.36}$$

is called energy norm.

It is easy to verify that $\|\cdot\|_e$ and $(\cdot,\cdot)_e$ fulfill all properties of norm and inner product (use Definitions A.24 and A.41).

Lemma 1.4 *Let V be a Hilbert space and $a : V \times V \to \mathbb{R}$ a bounded symmetric V-elliptic bilinear form. The energy norm induced by a is equivalent to the original norm in V,*

$$C_1\|u\|_V \le \|u\|_e \le C_2\|u\|_V \quad \textit{for all } u \in V, \tag{1.37}$$

where $C_1, C_2 > 0$ are some real constants.

Proof: Left to the reader as an exercise. ■

If the V-elliptic bilinear form $a(\cdot,\cdot)$ is not symmetric, it does not represent an inner product, but still it induces an energy norm. If $a : V \times V \to \mathbb{C}$, then the symmetry requirement $a(u, v) = a(v, u)$ is replaced with the sesquilinearity requirement $a(u, v) = \overline{a(v, u)}$.

Both the energetic inner product $(\cdot,\cdot)_e$ and the energy norm $\|\cdot\|_e$ represent important tools in the error analysis and numerical solution of elliptic PDEs. They are used to derive both a-priori and a-posteriori error estimates, to guide refinement strategies for adaptive finite element methods, and for other purposes. We will return to this topic later, after introducing the finite element discretization in Chapter 2.

1.2.3 The Lax–Milgram lemma

The Lax–Milgram lemma is the basic and most important tool for proving the existence and uniqueness of solution to elliptic problems.

Theorem 1.5 (Lax–Milgram lemma) *Let V be a Hilbert space, $a : V \times V \to \mathbb{R}$ a bounded V-elliptic bilinear form and $l \in V'$. Then there exists a unique solution to the problem*

$$a(u,v) = l(v) \quad \textit{for all } v \in V. \tag{1.38}$$

Remark 1.4 (Lax–Milgram vs. Riesz) *If the bilinear form $a(\cdot,\cdot)$ is symmetric, then the unique solution $u \in V$ of equation (1.38) is nothing else than the unique representant of the linear form $l \in V'$ with respect to the energetic inner product $(\cdot,\cdot)_e = a(\cdot,\cdot)$. In this sense the Lax–Milgram lemma is a special case of the Riesz representation theorem (Theorem A.15).*

Proof: The uniqueness of solution follows immediately from the V-ellipticity of the bilinear form a. We will use Theorem 1.2 to verify the existence. Let $A : V \to V'$ be the linear operator associated with the bilinear form a via (1.34). Then A is bounded and strongly monotone. By $L = \mathcal{J}A : V \to V$ denote the isometric dual mapping from the Riesz theorem,

$$a(u,v) = \langle Au, v\rangle = (\mathcal{J}Au, v) \quad \text{for all } u,v \in V.$$

Recall that $R(L) = V$ if and only if $R(L)$ is closed and $R(L)^{\perp} = \{0\}$. To show that $R(L)$ is closed, let $\{u_n\}_{n=1}^{\infty} \subset R(L)$ be a sequence converging to some function u. Then $u_n = \mathcal{J}Aw_n$ where $\{w_n\}_{n=1}^{\infty} \subset V$. Lemma 1.1 yields the existence of a constant $C > 0$ such that

$$\|u_n - u_m\| = \|\mathcal{J}A(w_n - w_m)\| = \|A(w_n - w_m)\| \geq C\|w_n - w_m\|.$$

Hence $\{w_n\}_{n=1}^{\infty}$ is a Cauchy sequence that has a limit $w \in V$. It holds

$$\|u_n - \mathcal{J}Aw\| = \|\mathcal{J}A(w_n - w)\| = \|A(w_n - w)\| \leq C_a\|w_n - w\| \to 0.$$

Therefore $u = \mathcal{J}Aw \in R(L)$ and $R(L)$ is closed. To prove that $R(L)^{\perp} = \{0\}$, take an arbitrary $u \in R(L)^{\perp}$. Then for any $v \in V$ it is

$$0 = (\mathcal{J}Av, u) = a(v,u).$$

Putting $v = u$, we obtain that the energy norm $\|u\|_e = 0$ and thus that $u = 0$. ■

1.2.4 Unique solvability of the model problem

The existence and uniqueness of solution to the model problem (1.33) can be proved using the Lax–Milgram lemma (Theorem 1.5) under the following assumptions:

Lemma 1.5 *Assume that $a_1(\boldsymbol{x}) \geq C_{min} > 0$ and $a_0(\boldsymbol{x}) \geq 0$ a.e. in Ω. Then the weak problem (1.33) has a unique solution $u \in V$.*

Proof: Since $a_1, a_0 \in L^\infty(\Omega)$, there exists a $C_{max} < \infty$ such that $|a_1(\boldsymbol{x})| \le C_{max}$ and $|a_0(\boldsymbol{x})| \le C_{max}$ a.e. in Ω. Then,

$$|a(u,v)| \le \int_\Omega (a_1|\nabla u \cdot \nabla v| + a_0|uv|)\,\mathrm{d}\boldsymbol{x} \le C_{max}\int_\Omega (|\nabla u \cdot \nabla v| + |uv|)\,\mathrm{d}\boldsymbol{x}. \qquad (1.39)$$

Since $\nabla u, \nabla v \in [L^2(\Omega)]^d$, the Hölder inequality (A.50) yields

$$\int_\Omega |\nabla u \cdot \nabla v|\,\mathrm{d}\boldsymbol{x} \le \left(\int_\Omega |\nabla u|^2\,\mathrm{d}\boldsymbol{x}\right)^{\frac{1}{2}} \left(\int_\Omega |\nabla v|^2\,\mathrm{d}\boldsymbol{x}\right)^{\frac{1}{2}} = |u|_{1,2}|v|_{1,2}. \qquad (1.40)$$

Analogously, for the product $|uv|$ one obtains

$$\int_\Omega |uv|\,\mathrm{d}\boldsymbol{x} \le \left(\int_\Omega u^2\,\mathrm{d}\boldsymbol{x}\right)^{\frac{1}{2}} \left(\int_\Omega v^2\,\mathrm{d}\boldsymbol{x}\right)^{\frac{1}{2}} = \|u\|_{L^2}\|v\|_{L^2}. \qquad (1.41)$$

The norm $\|\cdot\|_{1,2}$ is obtained by adding a nonnegative term to the seminorm $|\cdot|_{1,2}$,

$$|u|_{1,2}|v|_{1,2} \le \|u\|_{1,2}\|v\|_{1,2}. \qquad (1.42)$$

Similarly for the L^2-norm,

$$\|u\|_{L^2}\|v\|_{L^2} \le \|u\|_{1,2}\|v\|_{1,2}. \qquad (1.43)$$

Finally, relations (1.39) to (1.43) together yield

$$|a(u,v)| \le 2C_{max}\|u\|_{1,2}\|v\|_{1,2},$$

which means that the bilinear form is bounded with the constant $C_a = 2C_{max}$. Next let us prove the V-ellipticity of $a(\cdot,\cdot)$. Using the Poincaré–Friedrichs' inequality (Theorem A.26) in the space $V = H_0^1(\Omega)$, together with the nonnegativity of a_0 and strict positivity of a_1, we obtain that there exists a constant $C_{pf} > 0$ such that

$$\begin{aligned} a(v,v) &= \int_\Omega a_1|\nabla v|^2 + a_0 v^2\,\mathrm{d}\boldsymbol{x} \ge \int_\Omega a_1|\nabla v|^2\,\mathrm{d}\boldsymbol{x} \\ &\ge C_{min}\int_\Omega |\nabla v|^2\,\mathrm{d}\boldsymbol{x} = C_{min}|v|_{1,2}^2 \ge C_{min}C_{pf}^2\|v\|_{1,2}^2 \quad \text{for all } v \in V. \end{aligned}$$

Thus the bilinear form $a(\cdot,\cdot)$ is bounded and V-elliptic, and the Lax–Milgram lemma yields the existence and uniqueness of solution for every $f \in L^2(\Omega)$. ■

Discussion of the existence and uniqueness of solution for elliptic operators of the general form (1.25) can be found, e.g., in [93].

1.2.5 Nonhomogeneous Dirichlet boundary conditions

In this paragraph we consider the model equation (1.26) endowed with more general Dirichlet boundary conditions of the form

$$u(\boldsymbol{x}) = g(\boldsymbol{x}) \quad \text{on } \partial\Omega, \tag{1.44}$$

where $g \in C(\partial\Omega)$. For the purpose of the weak formulation we consider a function $G \in C^2(\Omega) \cap C(\overline{\Omega})$ such that $G = g$ on $\partial\Omega$ (the so-called Dirichlet lift of g). Notice that G is not unique, but we will show later that the solution is invariant under its choice. Writing $u = G + U$, the problem (1.26), (1.44) can be reformulated to:

Find $U \in C_0^2(\Omega)$ such that

$$\begin{aligned} -\nabla \cdot [a_1 \nabla(U+G)] + a_0(U+G) &= f \quad \text{in } \Omega, \\ U + G &= g \quad \text{on } \partial\Omega, \end{aligned}$$

or, equivalently,

$$\begin{aligned} -\nabla \cdot (a_1 \nabla U) + a_0 U &= f + \nabla \cdot (a_1 \nabla G) - a_0 G \quad \text{in } \Omega, && (1.45) \\ U &= 0 \quad \text{on } \partial\Omega, && (1.46) \end{aligned}$$

Except for an adjusted right-hand side, this problem is identical to the model problem (1.26), (1.28). We proceed analogously as in Paragraph 1.2.1 to derive its weak formulation:

Find $U \in V = H_0^1(\Omega)$ such that

$$a(U, v) = l(v) \quad \text{for all } v \in V \tag{1.47}$$

with

$$\begin{aligned} a(U, v) &= \int_\Omega (a_1 \nabla U \cdot \nabla v + a_0 U v)\, \mathrm{d}\boldsymbol{x}, \quad v \in V, \\ l(v) &= \int_\Omega (fv - a_1 \nabla G \cdot \nabla v - a_0 G v)\, \mathrm{d}\boldsymbol{x}, \quad v \in V, \end{aligned}$$

This weak formulation is defined under much weaker assumptions on f, g, and G. In particular, we can assume that $f \in L^2(\Omega)$ and $G \in H^1(\Omega)$ with the trace $g \in H^{\frac{1}{2}}(\partial\Omega)$.

We have seen in Paragraph 1.2.4 that the bilinear form $a(\cdot, \cdot)$ is bounded and V-elliptic. In other words, the Lax–Milgram lemma yields the existence and uniqueness of solution to (1.47) for every Dirichlet lift G.

Independence of the solution $u = U + G$ ***on the Dirichlet lift*** G***:*** Assume that $U_1 + G_1 = u_1 \in H^1(\Omega)$ and $U_2 + G_2 = u_2 \in H^1(\Omega)$ are two weak solutions. By (1.47) the difference $u_1 - u_2 \in V = H_0^1(\Omega)$ satisfies

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

Taking $u_1 - u_2$ for v and using the V-ellipticity of the bilinear form a, we obtain

$$0 = a(u_1 - u_2, u_1 - u_2) \geq C_{el} \|u_1 - u_2\|_V^2.$$

This means that

$$\|u_1 - u_2\|_V = 0,$$

i.e., that $u_1 = u_2$ a.e. in Ω.

1.2.6 Neumann boundary conditions

Consider the model equation (1.26) with Neumann boundary conditions of the form

$$\frac{\partial u}{\partial \boldsymbol{\nu}} = g \quad \text{on } \partial\Omega, \tag{1.48}$$

where $g \in C(\partial\Omega)$. This time we have to strengthen the positivity assumption on the coefficient a_0 to

$$a_0(\boldsymbol{x}) \geq \hat{C}_{min} > 0 \quad \text{in } \Omega. \tag{1.49}$$

The weak formulation of the problem (1.26), (1.48) is derived as follows: Assume that $u \in C^\infty(\Omega) \cap C^1(\overline{\Omega})$. Multiply (1.26) with a test function $v \in C^\infty(\Omega) \cap C^1(\overline{\Omega})$, integrate over Ω, and use the Green's theorem to reduce the maximum order of the partial derivatives. The boundary integrals do not vanish as they did in the homogeneous Dirichlet case, and we get an extra boundary term,

$$\int_\Omega (a_1 \nabla u \cdot \nabla v + a_0 uv)\, \mathrm{d}\boldsymbol{x} - \int_{\partial\Omega} a_1 \frac{\partial u}{\partial \boldsymbol{\nu}} v \, \mathrm{d}\boldsymbol{S} = \int_\Omega fv \, \mathrm{d}\boldsymbol{x}.$$

Here $\boldsymbol{\nu}$ is the unit outer normal vector to $\partial\Omega$ and $\partial u/\partial\boldsymbol{\nu} = \nabla u \cdot \boldsymbol{\nu}$. Substituting the boundary condition (1.48) into the boundary integral, and weakening the regularity assumptions, we obtain the following weak formulation:

Given $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$, find $u \in V = H^1(\Omega)$ such that

$$\int_\Omega (a_1 \nabla u \cdot \nabla v + a_0 uv)\, \mathrm{d}\boldsymbol{x} = \int_\Omega fv \, \mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} a_1 gv \, \mathrm{d}\boldsymbol{S} \quad \text{for all } v \in V.$$

Stated in the language of linear forms, one has to find a function $u \in V$ such that

$$a(u, v) = l(v) \quad \text{for all } v \in V, \tag{1.50}$$

where

$$\begin{aligned} a(u, v) &= \int_\Omega a_1 \nabla u \cdot \nabla v + a_0 uv \, \mathrm{d}\boldsymbol{x} \quad &&\text{for all } u, v \in V, \\ l(v) &= \int_\Omega fv \, \mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} a_1 gv \, \mathrm{d}\boldsymbol{S} \quad &&\text{for all } v \in V. \end{aligned}$$

Notice that although the bilinear form $a(\cdot, \cdot)$ is given by the same formula as in the case of Dirichlet boundary conditions, it is different since the space V changed.

The boundedness of the bilinear form $a(\cdot, \cdot)$ in $V \times V$ can be shown analogously to the proof of Lemma 1.5. Notice, however, that one cannot use the Poincaré–Friedrichs' inequality to prove the V-ellipticity of $a(\cdot, \cdot)$, since now the solution is not zero on the boundary. Here the additional assumption (1.49) comes into the play, and we obtain

$$a(v, v) \geq \min(C_{min}, \hat{C}_{min}) \|v\|_V^2.$$

The Lax–Milgram lemma guarantees that the problem (1.50) has a unique solution $u \in V$.

Remark 1.5 (Neumann problem without the assumption (1.49)) *The assumption (1.49) guarantees the presence of a nonzero L^2-term in the bilinear form. Without this term, neither the classical nor the weak formulation has a unique solution in Sobolev spaces. For example, if u is a solution of $-\Delta u = f$ with Neumann boundary conditions, then also $u + C$, where C is an arbitrary constant, is a solution. Let us formulate this problem in the weak sense:*

Find $u \in H^1(\Omega)$ such that

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx + \int_{\partial\Omega} g v \, dS \quad \textit{for all } v \in H^1(\Omega). \tag{1.51}$$

Using the test function $v = 1 \in H^1(\Omega)$, one finds that a necessary condition for (1.51) to have a solution at all is

$$\int_\Omega f \, dx + \int_{\partial\Omega} g \, dS = 0. \tag{1.52}$$

It follows from a deeper analysis in the quotient space $H^1(\Omega)/\mathbb{R}$ that condition (1.52) is sufficient for the existence and uniqueness of solution in $H^1(\Omega)/\mathbb{R}$ (see, e.g., [6]).

Remark 1.6 (Essential and natural boundary conditions) *Dirichlet boundary conditions are sometimes called essential since they essentially influence the weak formulation: They determine the function space in which the solution is sought. On the other hand, Neumann boundary conditions do not influence the function space and can be naturally incorporated into the boundary integrals. Therefore they are called natural.*

1.2.7 Newton (Robin) boundary conditions

Another frequently used type of natural boundary conditions involves a combination of function values and normal derivatives. Consider the model equation (1.26) equipped with such boundary conditions,

$$\begin{aligned} -\nabla \cdot (a_1 \nabla u) + a_0 u &= f \quad \text{in } \Omega, && (1.53)\\ c_1 u + c_2 \frac{\partial u}{\partial \nu} &= g \quad \text{on } \partial\Omega, && (1.54) \end{aligned}$$

where $f \in C(\Omega)$, $g \in C(\partial\Omega)$, and $c_1, c_2 \in C(\partial\Omega)$ are such that $c_1 c_2 > 0$ and $0 < \epsilon \le |c_2|$ on $\partial\Omega$. The positivity assumptions (1.27) and (1.49) on the coefficients a_0, a_1 apply.

For a sufficiently regular function $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$, the weak identity

$$\int_\Omega a_1 \nabla u \cdot \nabla v + a_0 u v \, \mathrm{d}x - \int_{\partial\Omega} a_1 \frac{\partial u}{\partial \nu} v \, \mathrm{d}S = \int_\Omega f v \, \mathrm{d}x$$

is derived analogously to the Neumann case. Using the boundary condition (1.54), we obtain the following weak formulation:

Given $f \in L^2(\Omega)$, $g \in L^2(\partial\Omega)$, and $a_0, a_1 \in L^\infty(\Omega)$, find $u \in V = H^1(\Omega)$ such that

$$\int_\Omega a_1 \nabla u \cdot \nabla v + a_0 u v \, \mathrm{d}x + \int_{\partial\Omega} \frac{a_1 c_1}{c_2} u v \, \mathrm{d}S = \int_\Omega f v \, \mathrm{d}x + \int_{\partial\Omega} \frac{a_1 g}{c_2} v \, \mathrm{d}S \quad \text{for all } v \in V.$$

In other words, it is our task to find $u \in V$ such that

$$a(u,v) = l(v) \quad \text{for all } v \in V, \tag{1.55}$$

where

$$\begin{aligned} a(u,v) &= \int_\Omega a_1 \nabla u \cdot \nabla v + a_0 uv \, \mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} \frac{a_1 c_1}{c_2} uv \, \mathrm{d}\boldsymbol{S} \quad \text{for all } u, v \in V, \\ l(v) &= \int_\Omega fv \, \mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} \frac{a_1 g}{c_2} v \, \mathrm{d}\boldsymbol{S} \quad \text{for all } v \in V. \end{aligned}$$

Since the bilinear form $a(\cdot,\cdot)$ is both bounded and V-elliptic (use Theorem A.28), the Lax–Milgram lemma implies that problem (1.55) has a unique solution $u \in V$.

1.2.8 Combining essential and natural boundary conditions

What remains to be discussed is the combination of essential and natural boundary conditions. Let us choose, for example, the Dirichlet and Neumann conditions for this purpose. Hence, let the boundary $\partial\Omega$ be split into two nonempty disjoint open parts Γ_D and Γ_N, and consider the problem

$$\begin{aligned} -\nabla \cdot (a_1 \nabla u) + a_0 u &= f \quad \text{in } \Omega, &(1.56)\\ u &= g_D \quad \text{on } \Gamma_D, &(1.57)\\ \frac{\partial u}{\partial \boldsymbol{\nu}} &= g_N \quad \text{on } \Gamma_N. &(1.58) \end{aligned}$$

The weak formulation is derived as follows: First extend the function $g_D \in C(\Gamma_D)$ to the rest of the boundary $\partial\Omega$ by introducing a function $\tilde{g}_D \in C(\partial\Omega)$ such that $\tilde{g}_D \equiv g_D$ on Γ_D. The nonuniqueness of this extension is not going to cause any problems. Next find some Dirichlet lift $G \in C^2(\Omega) \cap C(\overline{\Omega})$ of $\tilde{g}_D$ (i.e., $G \equiv \tilde{g}_D$ on $\partial\Omega$). The solution u is sought in the form $u = U + G$ analogously to the pure Dirichlet case. The equations

$$\begin{aligned} -\nabla \cdot [a_1 \nabla (U+G)] + a_0(U+G) &= f \quad \text{in } \Omega, &(1.59)\\ (U+G) &= g_D \quad \text{on } \Gamma_D, &(1.60)\\ \frac{\partial (U+G)}{\partial \boldsymbol{\nu}} &= g_N \quad \text{on } \Gamma_N, &(1.61) \end{aligned}$$

yield

$$\begin{aligned} -\nabla \cdot (a_1 \nabla U) + a_0 U &= f + \nabla \cdot (a_1 \nabla G) - a_0 G \quad \text{in } \Omega, &(1.62)\\ U &= 0 \quad \text{on } \Gamma_D, &(1.63)\\ \frac{\partial (U+G)}{\partial \boldsymbol{\nu}} &= g_N \quad \text{on } \Gamma_N. &(1.64) \end{aligned}$$

The appropriate space for the function U is

$$V = \{u \in H^1(\Omega);\ u = 0 \text{ on } \Gamma_D\}. \tag{1.65}$$

Applying the standard procedure that we went through several times, we arrive at the weak identity

$$\int_\Omega (a_1 \nabla U \cdot \nabla v + a_0 U v) \, \mathrm{d}\boldsymbol{x}$$
$$= \int_\Omega (fv - a_1 \nabla G \cdot \nabla v - a_0 G v) \, \mathrm{d}\boldsymbol{x} + \int_{\Gamma_N} \left(a_1 \frac{\partial (U+G)}{\partial \boldsymbol{\nu}} v \right) \mathrm{d}S \quad \text{for all } v \in V.$$

Using the Neumann boundary condition (1.64) on Γ_N, we finally obtain the following weak problem:

Find a function U in the space V such that

$$a(U, v) = l(v) \quad \text{for all } v \in V, \tag{1.66}$$

where

$$\begin{aligned} a(U, v) &= \int_\Omega (a_1 \nabla U \cdot \nabla v + a_0 U v) \, \mathrm{d}\boldsymbol{x}, \quad U, v \in V, \qquad (1.67) \\ l(v) &= \int_\Omega (fv - a_1 \nabla G \cdot \nabla v - a_0 G v) \, \mathrm{d}\boldsymbol{x} + \int_{\Gamma_N} a_1 g_N v \, \mathrm{d}S \quad \text{for all } v \in V. \end{aligned}$$

The bilinear form $a(\cdot, \cdot)$ is bounded and V-elliptic (the proof is analogous to Paragraph 1.2.4). The Poincaré–Friedrichs' inequality holds in V due to the zero boundary condition for U on Γ_D (see Remark A.8). Therefore the Lax–Milgram lemma implies that problem (1.66) has a unique solution $U \in V$. As usual, the final solution satisfying both the essential and natural boundary conditions is $u = U + G$.

1.2.9 Energy of elliptic problems

It was mentioned in Paragraph 1.1.1 that elliptic problems usually describe some equilibrium or minimum-energy state of a system. In this paragraph we introduce the explicit form of the abstract energy, at least for symmetric problems. The most important numerical scheme based on the minimization of the abstract energy, the Ritz method, will be discussed later in Chapter 2.

Theorem 1.6 *Let V be a linear space, $a : V \times V \longrightarrow \mathbb{R}$ a symmetric V-elliptic bilinear form and $l \in V'$. Then the functional of abstract energy,*

$$E(v) = \frac{1}{2} a(v, v) - l(v), \tag{1.68}$$

attains its minimum in V at an element $u \in V$ if and only if

$$a(u, v) = l(v) \quad \textit{for all } v \in V. \tag{1.69}$$

Moreover, the minimizer $u \in V$ is unique.

Proof: Let (1.69) hold. Then

$$\begin{aligned} E(u + tv) &= \frac{1}{2} a(u + tv, u + tv) - l(u + tv) \\ &= E(u) + t(a(u, v) - l(v)) + \frac{1}{2} t^2 a(v, v) \qquad (1.70) \end{aligned}$$

for all $u, v \in V$ and $t \in \mathbb{R}$. If $u \in V$ satisfies (1.69), then the last equation with $t = 1$ implies

$$E(u+v) = E(u) + \frac{1}{2}a(v,v) > E(u) \qquad \text{for all } 0 \neq v \in V.$$

Thus $u \in V$ is a unique minimizer of (1.68).

Conversely, if E has a minimum at $u \in V$, then for every $v \in V$ the derivative of the quadratic function $\phi(t) = E(u + tv)$ must vanish at $t = 0$. By (1.70),

$$0 = \phi'(0) = a(u,v) - l(v),$$

and (1.69) holds. ■

Another interesting theoretical application of the energy-minimization concept is an alternative proof of the Lax–Milgram lemma for symmetric elliptic problems in convex sets:

Theorem 1.7 (Lax–Milgram lemma for convex sets) *Let W be a closed convex set in a Hilbert space V and $a : V \times V \to \mathbb{R}$ a bounded V-elliptic bilinear form. Then for every $l \in V'$ there exists a unique $u \in W$ such that $E(u) = \inf\{E(v);\ v \in W\}$, where*

$$E(v) = \frac{1}{2}a(v,v) - l(v).$$

Proof: The functional E is bounded from below since

$$E(v) \geq \frac{1}{2}C_a\|v\|^2 - \|l\|\|v\| = \frac{1}{2C_a}(C_a\|v\| - \|l\|)^2 - \frac{\|l\|^2}{2C_a} \geq -\frac{\|l\|^2}{2C_a}.$$

Let $e_0 = \inf\{E(v);\ v \in W\}$ and let $\{v_n\}_{n=1}^{\infty}$ be a minimizing sequence, i.e.,

$$\lim_{n\to\infty} E(v_n) = e_0.$$

Then

$$\begin{aligned} C_a\|v_n - v_m\|^2 &\leq a(v_n - v_m, v_n - v_m) \\ &= 2a(v_n, v_n) + 2a(v_m, v_m) - a(v_n + v_m, v_n + v_m) \\ &= 4E(v_n) + 4E(v_m) - 8E\left(\frac{v_n + v_m}{2}\right) \\ &\leq 4E(v_n) + 4E(v_m) - 8e_0, \end{aligned}$$

where $\frac{1}{2}(v_n + v_m) \in W$ thanks to the convexity of W. Now $E(v_n), E(v_m) \to e_0$ implies $\|v_n - v_m\| \to 0$ as $n, m \to \infty$. Thus $\{v_n\}_{n=1}^{\infty}$ is a Cauchy sequence in V and there exists a limit $u \in V$, $v_n \to u$. Since W is closed, we also have $u \in W$. The continuity of E implies

$$E(u) = \lim_{n\to\infty} E(v_n) = \inf_{v \in W} E(v).$$

Let us show that the solution $u \in W$ is unique. Suppose that both u_1 and u_2 are solutions. Clearly the sequence $u_1, u_2, u_1, u_2, \ldots$ is a minimizing sequence. Above we saw that every minimizing sequence has to be a Cauchy sequence. Thus $u_1 = u_2$. ■

1.2.10 Maximum principles and well-posedness

Another important aspect of elliptic problems is the existence of maximum principles. We find it useful to present several of them here and illustrate how they imply the well-posedness of elliptic problems. The counterpart of the maximum principles on the numerical level are the discrete maximum principles (see, e.g., [11, 14, 19, 31, 57, 67] and [112]), which find particularly important application in problems where physically nonnegative quantities like the temperature, density, or concentration are computed.

Theorem 1.8 (Basic maximum principle) *Consider an open bounded set $\Omega \subset \mathbb{R}^d$ and a symmetric elliptic operator of the form*

$$Lu = -\sum_{i,j=1}^{d} a_{ij}(\boldsymbol{x})\frac{\partial^2 u}{\partial x_i \partial x_j}, \tag{1.71}$$

where $a_{ij} \in C(\Omega)$. Let $u \in C^2(\Omega) \cap C(\overline{\Omega})$ be the solution of the equation $Lu = f$, where $f \in C(\Omega)$ and

$$f \leq 0 \quad \mathit{in}\ \Omega.$$

Then the maximum of u in $\overline{\Omega}$ is attained on the boundary $\partial\Omega$. Furthermore it holds that if the maximum is attained at an interior point of Ω, then the function u is constant.

This result remains true under less restrictive assumptions on the coefficients a_{ij}.

Proof: Recall that L is elliptic if the coefficient matrix $A(\boldsymbol{x}) = \{a_{ij}\}_{i,j=1}^d$ is positive definite in Ω. First we carry out the proof under a stronger assumption that $f < 0$ in Ω. Suppose that there exists some $\tilde{\boldsymbol{x}} \in \Omega$ such that

$$u(\tilde{\boldsymbol{x}}) = \sup_{\boldsymbol{x}\in\Omega} u(\boldsymbol{x}) > \sup_{\boldsymbol{x}\in\partial\Omega} u(\boldsymbol{x}). \tag{1.72}$$

Since $A(\tilde{\boldsymbol{x}}) = \{a_{ij}(\tilde{\boldsymbol{x}})\}_{i,j=1}^d$ is symmetric and positive definite, it is diagonalizable and has positive real eigenvalues $\lambda_1(\tilde{\boldsymbol{x}}), \lambda_2(\tilde{\boldsymbol{x}}), \ldots, \lambda_d(\tilde{\boldsymbol{x}})$. Thus there exists a nonsingular $d \times d$ matrix C such that

$$\Lambda = C^{-1}A(\tilde{\boldsymbol{x}})C,$$

where $\Lambda = \mathrm{diag}(\lambda_1(\tilde{\boldsymbol{x}}), \lambda_2(\tilde{\boldsymbol{x}}), \ldots, \lambda_d(\tilde{\boldsymbol{x}}))$. In a new coordinate system defined by

$$\boldsymbol{\xi} = \boldsymbol{\xi}(\boldsymbol{x}) = C\boldsymbol{x}$$

we have that

$$\begin{aligned} 0 &> f(\tilde{\boldsymbol{x}}) = (Lu)(\tilde{\boldsymbol{x}}) \\ &= -\sum_{i,j=1}^{d} (C^{-1}A(\tilde{\boldsymbol{x}})C)_{ij}\frac{\partial^2 u}{\partial \xi_i \partial \xi_j}(\tilde{\boldsymbol{x}}) \\ &= -\sum_{i=1}^{d} \lambda_i(\tilde{\boldsymbol{x}})\frac{\partial^2 u}{\partial \xi_i^2}(\tilde{\boldsymbol{x}}), \end{aligned} \tag{1.73}$$

which is a contradiction since $\lambda_i(\tilde{\boldsymbol{x}}) > 0$ for all $1 \leq i \leq d$, and $\tilde{\boldsymbol{x}} \in \Omega \setminus \partial\Omega$ is a maximum point of u, meaning that

$$\frac{\partial^2 u}{\partial \xi_i^2}(\tilde{\boldsymbol{x}}) \leq 0 \quad \text{for all } 1 \leq i \leq d.$$

Next let us prove the result for the weaker assumption $f \leq 0$ in Ω. Again, suppose that there exists some $\tilde{\boldsymbol{x}} \in \Omega$ satisfying (1.72). Consider the function

$$h(\boldsymbol{x}) = \sum_{i=1}^{d} (x_i - \tilde{x}_i)^2.$$

Since the maximum point $\tilde{\boldsymbol{x}}$ of u lies in the interior of Ω and $h(\boldsymbol{x})$ is bounded in Ω, for a sufficiently small $\beta > 0$ the function $w(\boldsymbol{x}) = u(\boldsymbol{x}) + \beta h(\boldsymbol{x})$ attains its maximum at some interior point $\boldsymbol{x}_0 \in \Omega$. Since

$$\frac{\partial^2 h}{\partial x_i \partial x_j}(\boldsymbol{x}) = 2\delta_{ij} \qquad \text{for all } \boldsymbol{x} \in \Omega,$$

we have

$$(Lw)(\boldsymbol{x}) = (Lu)(\boldsymbol{x}) + \beta(Lh)(\boldsymbol{x}) = f(\boldsymbol{x}) - 2\beta \sum_{i=1}^{d} a_{ii}(\boldsymbol{x}) = \tilde{f}(\boldsymbol{x}) < 0 \quad \text{in } \Omega.$$

Thus we can apply the result of the first part of the proof. ■

■ EXAMPLE 1.7 (Maximum principle)

Consider an open bounded set $\Omega = (-1, 1)^2 \subset \mathbb{R}^2$ and the Poisson equation

$$-\Delta u = -4 \qquad \text{in } \Omega \tag{1.74}$$

($L = -\Delta$ is obtained from (1.71) putting $a_{ij} = \delta_{ij}$). The solution u has the form

$$u(x_1, x_2) = x_1^2 + x_2^2 + C,$$

where $C \in \mathbb{R}$ is an arbitrary constant to be determined from the boundary conditions. Since $f \leq 0$ in Ω, the maximum principle (Theorem 1.8) implies that u attains its maximum on the boundary $\partial\Omega$. This indeed is true, as shown in Figure 1.4.

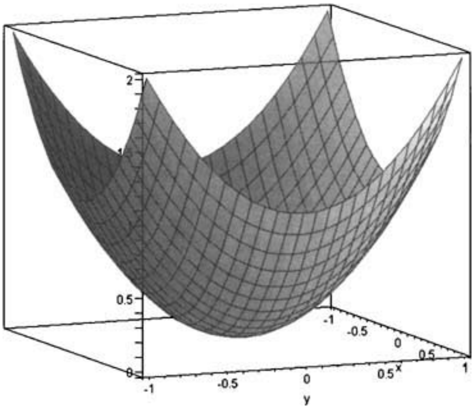


Figure 1.4 Maximum principle for the Poisson equation in 2D.

Immediate consequences of the maximum principle are the minimum principle, comparison principle, and the continuous dependence of the solution on boundary and initial data. Most of these results are straightforward consequences of Theorem 1.8. We encourage the reader to perform the proofs using the hints given.

Corollary 1.1 (Minimum principle) *Let $\Omega \subset \mathbb{R}^d$ be an open bounded set and L an elliptic operator of the form (1.71). If $Lu = f \geq 0$ in Ω, then u attains its minimum on the boundary $\partial\Omega$.*

Proof: Apply Theorem 1.8 to $\tilde{u} := -u$. ■

Corollary 1.2 (Comparison principle) *Let $\Omega \subset \mathbb{R}^d$ be an open bounded set and L an elliptic operator of the form (1.71). Suppose that functions $u, v \in C^2(\Omega) \cap C(\overline{\Omega})$ solve the equations $Lu = f_u$ and $Lv = f_v$, respectively, and*

$$\begin{aligned} f_u &\leq f_v \quad \textit{in } \Omega, \\ u &\leq v \quad \textit{on } \partial\Omega. \end{aligned}$$

Then $u \leq v$ in Ω.

Proof: Apply the minimum principle to $w := v - u$. ■

Corollary 1.3 (Continuous dependence on boundary data) *Let $\Omega \subset \mathbb{R}^d$ be an open bounded set and L an elliptic operator of the form (1.71). Suppose that u_1 and u_2 solve the equation $Lu = f$ with different Dirichlet boundary data. Then*

$$\sup_{\boldsymbol{x}\in\Omega} |u_1(\boldsymbol{x}) - u_2(\boldsymbol{x})| = \sup_{\boldsymbol{x}\in\partial\Omega} |u_1(\boldsymbol{x}) - u_2(\boldsymbol{x})|.$$

Proof: The function $w = u_1 - u_2$ satisfies the homogeneous equation $Lw = 0$ in Ω. Apply both the maximum and minimum principles to obtain the result. ■

Before introducing the continuous dependence of solution on the right-hand side, we need to define the notion of uniform ellipticity:

Definition 1.7 (Uniform ellipticity) *A linear elliptic operator L of the form (1.4) is said to be* uniformly elliptic *in an open set $\Omega \subset \mathbb{R}^d$ if there exists a constant $C_A > 0$ such that*

$$\boldsymbol{\xi}^T A(\boldsymbol{x})\boldsymbol{\xi} \geq C_A \|\boldsymbol{\xi}\|^2 \quad \textit{for all } \boldsymbol{\xi} \in \mathbb{R}^d,$$

and all $\boldsymbol{x} \in \Omega$, where $A(\boldsymbol{x})$ is the corresponding coefficient matrix.

Corollary 1.4 (Continuous dependence on the right-hand side) *Let $\Omega \subset \mathbb{R}^d$ be an open bounded set and L an elliptic operator of the form (1.71). Moreover, assume that L is uniformly elliptic in Ω. Then there exists a constant C only depending on the set Ω and the uniform ellipticity constant C_A, such that*

$$|u(\boldsymbol{x})| \leq \sup_{\boldsymbol{y}\in\partial\Omega} |u(\boldsymbol{y})| + C \sup_{\boldsymbol{y}\in\Omega} |f(\boldsymbol{y})| \tag{1.75}$$

for all $\boldsymbol{x} \in \Omega$.

Proof: Since Ω is bounded, it is contained in some open ball $B(0, r)$. Let

$$w(\boldsymbol{x}) = r^2 - \sum_{i=1}^{d} x_i^2.$$

Clearly $0 \le w \le r^2$ in Ω. Since

$$\frac{\partial^2 w}{\partial x_i \partial x_j} = -2\delta_{ij},$$

it is $Lw \ge 2dC_A$, where C_A is the uniform ellipticity constant of L. Let

$$v(\boldsymbol{x}) := \sup_{\boldsymbol{y}\in\partial\Omega} |u(\boldsymbol{y})| + w(\boldsymbol{x})\frac{1}{2dC_A} \sup_{\boldsymbol{y}\in\partial\Omega} |Lu(\boldsymbol{y})|.$$

Then $Lv \ge |Lu|$ in Ω and $v \ge |u|$ on $\partial\Omega$. The comparison principle implies that $-v(\boldsymbol{x}) \le u(\boldsymbol{x}) \le v(\boldsymbol{x})$ in Ω. Since $w \le r^2$, (1.75) holds with $C = r^2/(2dC_A)$. ■

Corollary 1.5 (Elliptic operator with a Helmholtz term) *Consider an elliptic operator L of the form*

$$Lu = -\sum_{i,j=1}^{d} a_{ij}(\boldsymbol{x})\frac{\partial^2 u}{\partial x_i \partial x_j} + a_0(\boldsymbol{x})u$$

with $a_0(\boldsymbol{x}) \ge 0$ in Ω. Then $Lu \le 0$ in Ω implies that

$$\sup_{\boldsymbol{x}\in\Omega} u(\boldsymbol{x}) \le \max\{0, \sup_{\boldsymbol{x}\in\partial\Omega} u(\boldsymbol{x})\}.$$

Proof: Without loss of generality, let $\boldsymbol{x}_0 \in \Omega$ be such that

$$u(\boldsymbol{x}_0) = \sup_{\boldsymbol{y}\in\Omega} u(\boldsymbol{y}) > 0.$$

Then $(Lu)(\boldsymbol{x}_0) - a_0(\boldsymbol{x}_0)u(\boldsymbol{x}_0) \le (Lu)(\boldsymbol{x}_0) \le 0$, and the principal part $Lu - a_0 u$ defines an elliptic operator of the form (1.71). The conclusion follows from Theorem 1.8. ■

1.2.11 Exercises

Exercise 1.13 *Show that the bilinear form $a(\cdot,\cdot)$ from (1.55) is bounded and V-elliptic.*

Exercise 1.14 *Show that relation (1.35) in Lemma 1.4 defines an inner product. Further show that the energy norm (1.36) induced by this inner product satisfies the relation (1.37) (i.e., that it is equivalent to the norm $\|\cdot\|_V$).*

Exercise 1.15 *Let $\Omega \subset \mathbb{R}^d$ be an open bounded set with Lipschitz-continuous boundary. Let the boundary $\partial\Omega$ be split into two nonempty disjoint open parts Γ_N and Γ_D such that $\overline{\Gamma}_N \cup \overline{\Gamma}_D = \partial\Omega$. Consider boundary data (real functions) g_N, g_D defined on Γ_N and Γ_D, respectively. Write the weak formulation of the boundary value problem for the Poisson equation*

$$-\Delta u = f,$$

equipped with boundary conditions

$$\frac{\partial u}{\partial \nu}(\boldsymbol{x}) + u(\boldsymbol{x}) = g_N(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Gamma_N,$$

and

$$u(\boldsymbol{x}) = g_D(\boldsymbol{x}), \qquad \boldsymbol{x} \in \Gamma_D,$$

where f is a real-valued load function defined in Ω. Identify the largest function spaces where the solution u as well as the test functions v and data g_N, g_D, and f must lie in order that all integrals in the weak formulation be defined.

Exercise 1.16 *Prove Corollary 1.1.*

Exercise 1.17 *Prove Corollary 1.2.*

1.3 SECOND-ORDER PARABOLIC PROBLEMS

Next let us turn our attention to linear parabolic problems (the notion of parabolicity was introduced in Definition 1.1). Let $\Omega \subset \mathbb{R}^d$ be an open set with Lipschitz-continuous boundary. We will study a class of linear parabolic equations

$$\frac{\partial u}{\partial t} + Lu = f \qquad \text{in } \Omega, \tag{1.76}$$

where t is the time, $u = u(\boldsymbol{x}, t)$, $f = f(\boldsymbol{x}, t)$ and L is an elliptic operator of the form (1.1) with time-independent coefficients. The equation (1.76) is considered in a space-time cylinder $Q_T = \Omega \times (0, T)$, where $T > 0$.

1.3.1 Initial and boundary conditions

Boundary conditions for parabolic problems are analogous to the elliptic case: Dirichlet, Neumann, Newton, and combined (see Section 1.2). For simplicity, let us denote them by

$$(Bu)(\boldsymbol{x}, t) = g(\boldsymbol{x}, t) \qquad \text{for all } (\boldsymbol{x}, t) \in \partial\Omega \times (0, T). \tag{1.77}$$

Parabolic problems describe evolutionary processes, and thus one needs to provide an initial condition of the form

$$u(\boldsymbol{x}, 0) = u_0(\boldsymbol{x}) \qquad \text{for all } \boldsymbol{x} \in \Omega. \tag{1.78}$$

If the problem is considered in the classical sense, then the initial condition $u_0(\boldsymbol{x})$ must moreover satisfy the boundary conditions (this is known as compatibility condition).

1.3.2 Weak formulation

At every time instant the solution is sought in a closed subspace $V \subset H^1(\Omega)$ such that $H_0^1(\Omega) \subset V$. The form of the space V depends on the boundary conditions analogously to the elliptic case (see Paragraphs 1.2.5–1.2.8).

For the analysis of existence and uniqueness of solution we need to introduce function spaces and norms for time-dependent functions:

Definition 1.8 *First by $L^q(0,T;W^{k,p}(\Omega))$ we denote the space*

$$L^q(0,T;W^{k,p}(\Omega)) = \quad \{u : (0,T) \to W^{k,p}(\Omega); \\ u \textit{ is measurable and } \int_0^T \|u(t)\|^q_{k,p,\Omega}\,\mathrm{d}t < \infty\},$$

endowed with the norm

$$\|u\|_{L^q(0,T;W^{k,p}(\Omega))} = \left(\int_0^T \|u(t)\|^q_{k,p,\Omega}\,\mathrm{d}t\right)^{\frac{1}{q}}. \tag{1.79}$$

The symbol $u(t)$ stands for a function of $\boldsymbol{x}$ such that $u(t) : \boldsymbol{x} \longrightarrow u(\boldsymbol{x},t)$. Further we define the space

$$C([0,T];L^p(\Omega)) = \{u : [0,T] \to L^p(\Omega);\ \|u(t)\|_{p,\Omega} \textit{ is continuous in } [0,T]\}. \tag{1.80}$$

Analogously we use the $W^{k,p}$-norm in Ω to define the space

$$C([0,T];W^{k,p}(\Omega)) = \{u : [0,T] \to W^{k,p}(\Omega);\ \|u(t)\|_{k,p,\Omega} \textit{ is continuous in } [0,T]\}. \tag{1.81}$$

Weak formulation The weak formulation of parabolic problems is derived using a procedure analogous to elliptic equations. For example, in the case of homogeneous Dirichlet boundary conditions the weak formulation of the problem (1.76), (1.77), (1.78) reads:

Given $f \in L^2(Q_T)$ and $u_0 \in V = H^1_0(\Omega)$, find $u \in L^2(0,T;V) \cap C([0,T];L^2(\Omega))$ such that

$$\frac{\mathrm{d}}{\mathrm{d}t}(u(t),v)_{L^2} + a(u(t),v) = (f(t),v)_{L^2} \quad \text{for all } v \in V,\ t \in (0,T), \tag{1.82}$$

$$u(0) = u_0, \tag{1.83}$$

where the bilinear form $a(\cdot,\cdot)$ corresponds to the elliptic operator (1.1),

$$a(u,v) = \int_\Omega \left[\sum_{i,j=1}^d a_{ij}\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i} - \sum_{i=1}^d \left(b_i u\frac{\partial v}{\partial x_i} - c_i v\frac{\partial u}{\partial x_i}\right) + a_0 uv\right]\mathrm{d}\boldsymbol{x}. \tag{1.84}$$

The other types of boundary conditions are handled analogously to elliptic problems.

1.3.3 Existence and uniqueness of solution

Since the difficulty of the proof of existence and uniqueness of solution to the problem (1.82), (1.83) exceeds the scope of this text, we restrict ourselves to formulating the principal theoretical result, and providing appropriate references.

We need to introduce the notion of weak coercivity of the form $a(u, v)$ in the space V: There exist two constants $c_{1,2} > 0$ and $c_2 \geq 0$ such that

$$a(u,u) + c_2\|u\|_{L^2}^2 \geq c_{1,2}\|u\|_{W^{1,2}}^2 \qquad \text{for all } u \in V. \tag{1.85}$$

If the form $a(u, v)$ is V-elliptic (coercive), then this inequality holds with $c_2 = 0$. This is the case, for example, for the heat transfer equation $\partial u/\partial t - \Delta u = f$ with homogeneous Dirichlet boundary conditions. In general, condition (1.85) is satisfied for all types of boundary value problems we deal with, provided that all coefficients a_{ij}, b_i, c_i, and a_0 of the operator (1.4) belong to $L^\infty(\Omega)$.

Before introducing the existence and uniqueness theorem, let us show an interesting trick that turns the weakly coercive bilinear form $a(\cdot,\cdot)$ into a coercive one. Applying the substitution

$$\tilde{u}(\boldsymbol{x}, t) = e^{-c_2 t}u(\boldsymbol{x}, t),$$

equation (1.76) comes over to the form

$$\frac{\partial \tilde{u}}{\partial t} + L\tilde{u} + c_2\tilde{u} = e^{-c_2 t} f \quad \text{in } Q_T.$$

Defining $\tilde{f} := e^{-c_2 t} f$ and $\tilde{L} := (L + c_2 I)$, where I stands for the identity operator, the equation returns to the form (1.76). However, if the original bilinear form $a(u, v)$ is weakly coercive, the bilinear form $\tilde{a}(u, v) = a(u, v) + c_2(u, v)$ is coercive. This technique is used in the analysis of parabolic PDEs quite frequently. Now let us formulate the promised existence and uniqueness result:

Theorem 1.9 (Existence and uniqueness of solution) *Let the bilinear form $a(\cdot,\cdot)$ be continuous in $V \times V$ and weakly coercive. Given $f \in L^2(Q_T)$ and $u_0 \in V$, there exists a unique solution $u \in L^2(0, T; V) \cap C([0, T]; L^2(\Omega))$ to the system (1.82), (1.83). Moreover, $\partial u/\partial t \in L^2(0, T; V')$ and the energy estimate*

$$\max_{t\in[0,T]} \|u(t)\|_{L^2}^2 + c_{1,2}\int_0^T \|u(t)\|_{W^{1,2}}^2 \leq \|u(0)\|_{L^2}^2 + \frac{1}{c_{1,2}}\int_0^T \|f(t)\|_2^2 \tag{1.86}$$

holds.

Proof: See, e.g., [93], pages 366 to 369. ■

1.3.4 Exercises

Exercise 1.18 *Let $Q_T = \Omega \times (0, T)$, where $\Omega \subset \mathbb{R}^d$ is an open bounded set with Lipschitz-continuous boundary. Consider the heat-transfer equation*

$$\frac{\partial u}{\partial t} - \Delta u = f \quad \text{in } \Omega, \tag{1.87}$$

$f \in L^2(Q_T)$, equipped with some initial condition $u(\boldsymbol{x}, 0) = u_0(\boldsymbol{x})$, $u_0 \in H^1(\Omega)$, and Neumann boundary conditions

$$\frac{\partial u}{\partial \boldsymbol{\nu}} = g \quad on\ \partial\Omega, \tag{1.88}$$

$g \in C(\partial\Omega)$.

1. *What is the space V in this case?*

2. *Verify in detail all assumptions ofTheorem 1.9 and use it to show the unique solvability of this problem.*

3. *Consider the elliptic problem $-\Delta u = f$ in Ω, which is the stationary version of equation (1.87), equipped with the pure Neumann boundary conditions (1.88). Does this problem have a unique solution?*

4. *Explain the difference between the V-ellipticity condition (Definition 1.5) and condition (1.85). What does this difference imply for the unique solvability of elliptic and parabolic problems?*

1.4 SECOND-ORDER HYPERBOLIC PROBLEMS

In this section we study linear second-order hyperbolic problems. A model equation with appropriate boundary and initial conditions is formulated in Paragraph 1.4.1. In Paragraph 1.4.2 we derive its weak formulation and present a basic existence and uniqueness result. In Paragraph 1.4.3 we show the link between the second-order hyperbolic equations and first-order hyperbolic systems.

1.4.1 Initial and boundary conditions

The notion of hyperbolicity was first introduced in Definition 1.1. Consider the model equation

$$\frac{\partial^2 u}{\partial t^2} + Lu = f, \tag{1.89}$$

where L is an elliptic operator of the form

$$L = \sum_{i,j=1}^{d} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial}{\partial x_j}\right) \tag{1.90}$$

with time-independent coefficients. We are interested in solving equation (1.89) in a space-time cylinder $Q_T = \Omega \times (0,T)$, where $\Omega \subset \mathbb{R}^d$ is some open bounded set with Lipschitz-continuous boundary, and $T > 0$.

Let the boundary $\partial\Omega$ be split into two open parts $\Gamma_D, \Gamma_N \subset \partial\Omega$ such that $\Gamma_D \cap \Gamma_N = \emptyset$ and $\overline{\Gamma}_D \cup \overline{\Gamma}_N = \partial\Omega$. We prescribe a Dirichlet boundary condition

$$u(\boldsymbol{x},t) = g_D(\boldsymbol{x},t) \qquad \text{for all } (\boldsymbol{x},t) \in \Gamma_D \times (0,T), \tag{1.91}$$

and a Neumann boundary condition

$$\frac{\partial u}{\partial \boldsymbol{\nu}_L}(\boldsymbol{x},t) = g_N(\boldsymbol{x},t) \qquad \text{for all } (\boldsymbol{x},t) \in \Gamma_N \times (0,T). \tag{1.92}$$

Here

$$\frac{\partial u}{\partial \boldsymbol{\nu}_L} = \sum_{i,j=1}^{d} a_{ij} \frac{\partial u}{\partial x_j} n_i$$

is the conormal derivative to $\partial\Omega$, $\boldsymbol{\nu} = (n_1, n_2, \ldots, n_d)^T$ being the unit outer normal vector to $\partial\Omega$.

Since the equation is of second-order in time, one has to prescribe initial boundary conditions for both the function values,

$$u(\boldsymbol{x},0) = u_0(\boldsymbol{x}) \qquad \text{for all } \boldsymbol{x} \in \Omega, \tag{1.93}$$

and the temporal derivative,

$$\frac{\partial u}{\partial t}(\boldsymbol{x},0) = u_1(\boldsymbol{x}) \qquad \text{for all } \boldsymbol{x} \in \Omega. \tag{1.94}$$

1.4.2 Weak formulation and unique solvability

To avoid complications related to the Dirichlet lift, for simplicity consider homogeneous boundary conditions on $\partial\Omega$. Then $V = H_0^1(\Omega)$, and the weak formulation of the problem (1.89)–(1.94) reads:

Given some right-hand side $f \in L^2(Q_T)$ and initial conditions $u_0 \in V$ and $u_1 \in L^2(\Omega)$, find a function $u \in C([0,T];V) \cap C^1([0,T];L^2(\Omega))$ such that

$$\begin{aligned} \frac{\mathrm{d}^2}{\mathrm{d}t^2}(u(t),v)_{L^2} + a(u(t),v) &= (f(t),v)_{L^2} \qquad \text{for all } v \in V,\ t \in (0,T), && (1.95)\\ u(0) &= u_0, && (1.96)\\ \frac{\mathrm{d}u}{\mathrm{d}t}(0) &= u_1, && (1.97) \end{aligned}$$

where the bilinear form $a(\cdot,\cdot)$ corresponds to the elliptic operator (1.90).

Theorem 1.10 *Under the above assumptions on the data, the problem (1.95)–(1.97) has a unique solution.*

Proof: The technicality of the proof exceeds the scope of this text. We refer the reader, e.g., to [79] and [94]. ■

1.4.3 The wave equation

Sometimes it is practical to abbreviate the notation for partial derivatives using a subscript, for example, $\partial u/\partial x = u_x$, $\partial u/\partial t = u_t$, $\partial^2 u/\partial x^2 = u_{xx}$, etc. We shall take advantage of this notation in what follows. One of the simplest examples of a second-order hyperbolic equation is the one-dimensional wave equation

$$u_{tt} = c^2 u_{xx}, \tag{1.98}$$

to be satisfied for all $(x,t) \in \mathbb{R} \times (0,T)$. The positive constant $c > 0$ is the wave speed. The equation (1.98) does not require boundary conditions since it is defined in $\mathbb{R}$, but it has to be supplemented with some initial conditions of the form

$$\begin{aligned} u(x,0) &= u_0(x), \\ u_t(x,0) &= u_1(x). \end{aligned} \tag{1.99}$$

Using the substitution

$$v = u_x \quad \text{and} \quad w = u_t,$$

the equation (1.98) comes over to a system of two first-order equations

$$\begin{pmatrix} v_t \\ w_t \end{pmatrix} + \begin{pmatrix} -w_x \\ -c^2 v_x \end{pmatrix} = \mathbf{0},$$

which can be written in the matrix form

$$\begin{pmatrix} v_t \\ w_t \end{pmatrix} + \begin{pmatrix} 0 & -1 \\ -c^2 & 0 \end{pmatrix} \begin{pmatrix} v_x \\ w_x \end{pmatrix} = \begin{pmatrix} v_t \\ w_t \end{pmatrix} + \boldsymbol{A} \begin{pmatrix} v_x \\ w_x \end{pmatrix} = \mathbf{0}. \tag{1.100}$$

The initial conditions to (1.100) are

$$\begin{aligned} v(x,0) &= u_0'(x), \\ w(x,0) &= u_1(x). \end{aligned} \tag{1.101}$$

This problem belongs to the class of first-order hyperbolic conservation laws that will be studied in Section 1.5. There the reader will learn how to derive the exact solution to (1.98), (1.99) in the form

$$u(x,t) = \frac{1}{2}\left[u_0(x-ct) + u_0(x+ct) - \frac{1}{c}U_1(x-ct) + \frac{1}{c}U_1(x+ct)\right], \tag{1.102}$$

where $U_1(x)$ is a primitive function to $u_1(x)$.

1.4.4 Exercises

Exercise 1.19 *Can equation (1.89), when equipped with a Neumann boundary condition on the whole boundary $\partial\Omega$, have a unique solution? How would this change in the stationary case $Lu = f$?*

Exercise 1.20 *Calculate the eigenvalues and eigenvectors of the matrix $\boldsymbol{A}$ in (1.100).*

Exercise 1.21 *Verify that the function $u(x,t)$ defined in (1.102) is the exact solution of the 1D wave equation (1.98) with the initial conditions (1.99).*

1.5 FIRST-ORDER HYPERBOLIC PROBLEMS

This section is devoted to first-order hyperbolic problems of the form

$$\frac{\partial}{\partial t}\boldsymbol{u}(x,t) + \operatorname{div}\boldsymbol{f}(\boldsymbol{u}(x,t)) = 0. \tag{1.103}$$

These equations differ from the previously studied second-order PDEs significantly and methods other than FEM are usually used for their numerical solution. PDEs of the form (1.103) are referred to as conservation laws, and they play an important role in the continuum mechanics and fluid dynamics.

The (generally nonlinear) flux function $\boldsymbol{f} = (\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_d)^T$, where d is the spatial dimension, consists of d directional fluxes $\boldsymbol{f}_i : \mathbb{R}^m \to \mathbb{R}^m$ that describe the transport of the solution in the axial directions x_i. The equation (1.103) is equipped with an initial condition

$$\boldsymbol{u}(\boldsymbol{x},0) = \boldsymbol{u}_0(\boldsymbol{x}).$$

Boundary conditions are not required if the problem is stated in $\Omega = \mathbb{R}^d$, otherwise suitable conditions on the boundary have to be imposed. An example of a conservation law are the Euler equations of compressible inviscid flow, which consist of the law of conservation of mass (continuity equation), law of conservation of momentum (Euler momentum equations), and the law of conservation of energy. For the analysis and numerical solution of the compressible Euler equations see, e.g., [52] and the references therein.

After a brief general introduction in Paragraph 1.5.1 we begin with the study of scalar and vector-valued linear conservation laws in one spatial dimension. Due to the existence of characteristics, the solutions of conservation laws have a unique structure. Characteristics are space-time curves that distribute the information from the initial and boundary conditions through the space-time cylinder $Q_T = \Omega \times (0,T)$. We will define and study the characteristics in Paragraph 1.5.2, and consequently utilize them to construct the exact solutions to a general one-dimensional linear first-order system in Paragraph 1.5.3.

Exciting things happen when the flux function $\boldsymbol{f}$ is nonlinear. Nonlinear hyperbolic systems exhibit discontinuous solutions, a feature unknown in elliptic and parabolic problems. The discontinuities, which may arise at finite times and even in problems with infinitely smooth initial and boundary data, banish the solution from Sobolev spaces and pose serious difficulties to both the analysis and numerical solution of hyperbolic problems. In Paragraph 1.5.5 we exploit the characteristics introduced in Paragraph 1.5.2 to understand the mechanism of creation of discontinuities in solutions to nonlinear hyperbolic problems.

1.5.1 Conservation laws

In one spatial dimension the conservation law (1.103) takes the form

$$\frac{\partial}{\partial t}\boldsymbol{u}(x,t) + \frac{\partial}{\partial x}\boldsymbol{f}(\boldsymbol{u}(x,t)) = 0, \tag{1.104}$$

where $\boldsymbol{f} : \mathbb{R}^m \to \mathbb{R}^m$ is the flux function and $\boldsymbol{u} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^m$ is m-dimensional vector of conserved quantities (state variables) such as, e.g., the mass, momentum or energy. When we say that a quantity $\boldsymbol{u}(x,t)$ is conserved, we mean that all its components satisfy

$$\int_{\mathbb{R}} u_i(x,t)\,\mathrm{d}x = \mathrm{const}_i, \tag{1.105}$$

or,

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}} u_i(x,t)\,\mathrm{d}x = 0. \tag{1.106}$$

Notice that while satisfying (1.106), the functions u_i themselves may change in time. Moreover notice that (1.104) implies (1.106).

Definition 1.9 (Cauchy problem) *By* Cauchy problem *we mean the pure initial-value problem where one requires that (1.104) holds for all $x \in \mathbb{R}$ and all $t \geq 0$. In this case one has to specify the initial condition only,*

$$\boldsymbol{u}(x,0) = \boldsymbol{u}_0(x), \quad x \in \mathbb{R}.$$

Of particular interest are conservation laws (1.104) which are hyperbolic:

Definition 1.10 (Hyperbolicity) *The system (1.104) is said to be* hyperbolic *if the flux function $\boldsymbol{f}$ is continuously differentiable and the $m \times m$ Jacobi matrix $D\boldsymbol{f}/D\boldsymbol{u}$ is diagonalizable and has real eigenvalues only.*

Recall that a square $m \times m$ matrix is diagonalizable if and only if it is similar to a diagonal matrix (Definition A.20). It is worth mentioning that the first-order system (1.100) associated with the second-order wave equation (1.98) was a hyperbolic conservation law: The flux function was linear, $\boldsymbol{f}(\boldsymbol{u}) = \boldsymbol{A}\boldsymbol{u}$, and the eigenvalues of its Jacobi matrix $D\boldsymbol{f}/D\boldsymbol{u} = \boldsymbol{A}$ were real numbers $\pm c$.

More generally, in $\mathbb{R}^d$ the conservation law (1.103) takes the form

$$\frac{\partial}{\partial t}\boldsymbol{u}(\boldsymbol{x},t) + \sum_{i=1}^{3} \frac{\partial}{\partial x_i}\boldsymbol{f}_i(\boldsymbol{u}(\boldsymbol{x},t)) = 0, \tag{1.107}$$

where $\boldsymbol{u} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^m$, and $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_d : \mathbb{R}^m \to \mathbb{R}^m$ are flux functions in the directions $x_1, \ldots, x_d$. Equation (1.107) is said to be hyperbolic if every linear combination of the Jacobi matrices

$$\sum_{i=1}^{d} a_i \frac{D\boldsymbol{f}_i}{D\boldsymbol{u}}, \tag{1.108}$$

where $a_i \in \mathbb{R}$ are arbitrary constants, is diagonalizable and has real eigenvalues only.

The Reynolds' transport theorem Conservation laws come from physics, where in most cases they are stated in integral form. For example, the law of mass conservation in fluids holds in the integral form

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\sigma(t)} \varrho(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x} = 0. \tag{1.109}$$

where $\sigma(t)$ is an arbitrary control volume. Control volume is a volume of fluid that is formed by the same particles at all times, and the integral of the density ϱ over $\sigma(t)$ yields the mass of $\sigma(t)$.

Since the integral formulations of conservation laws are very difficult to handle numerically, it is customary to use the Reynolds' transport theorem to convert them into PDEs. For a general density function $\mathcal{D}(\boldsymbol{x}, t)$ and under suitable regularity assumptions (see, e.g., [52]) the Reynolds' transport theorem says

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\sigma(t)} \mathcal{D}\,\mathrm{d}\boldsymbol{x} = \int_{\sigma(t)} \left(\frac{\partial \mathcal{D}}{\partial t} + \operatorname{div}(\mathcal{D}\boldsymbol{v})\right)\mathrm{d}\boldsymbol{x}, \tag{1.110}$$

where $\boldsymbol{v}(\boldsymbol{x}, t)$ is the fluid velocity. Applying (1.110)–(1.109) with $\mathcal{D} = \varrho$, we obtain

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}\int_{\sigma(t)} \varrho(\boldsymbol{x}, t)\,\mathrm{d}\boldsymbol{x} = \int_{\sigma(t)} \left(\frac{\partial \varrho}{\partial t} + \operatorname{div}(\varrho\boldsymbol{v})\right)\mathrm{d}\boldsymbol{x}. \tag{1.111}$$

Since the control volume $\sigma(t) \subset \Omega$ in (1.111) is arbitrary, the standard localization theorem says that the integrand has to be zero almost everywhere in Ω. Thus (1.111) yields the continuity equation,

$$\frac{\partial \varrho}{\partial t} + \operatorname{div}(\varrho\boldsymbol{v}) = 0 \qquad \text{a.e. in } Q_T = \Omega \times (0, T). \tag{1.112}$$

The localization theorem is intuitively clear and its proof straightforward. In particular, if the function ϱ is continuous, (1.112) holds everywhere in Q_T. For $\varrho \in H^1(\Omega)$ one proceeds by the density argument (see the end of Paragraph A.2.10).

Standard difficulties related to conservation laws The transformation of an integral equation to a PDE is not an equivalent operation. Usually the PDE is less general, undefined on discontinuities (shocks) where the integral form holds. Therefore one has to go back to the integral equation and derive suitable jump conditions to hold at the discontinuities and incorporate them back into the weak formulation of the PDE.

The weak solution usually admits more solutions than the unique physically admissible solution corresponding to the integral form of the conservation law. Therefore one has to impose some selection principle that excludes nonphysical solutions. For fluid dynamics problems one can appeal the second law of thermodynamics which states that the entropy is not decreasing. In particular, as molecules of a fluid pass through a shock, their entropy must increase. It turns out that this condition is sufficient to reliably distinguish between physically correct and incorrect discontinuities. Generally, such conditions are called entropy conditions.

1.5.2 Characteristics

The existence of characteristics (characteristic curves) is a unique aspect of hyperbolic PDEs. These space-time curves determine how the values of the initial and boundary conditions are distributed through the space-time cylinder $Q_T = \Omega \times (0, T)$.

To begin with, consider a constant $a \in \mathbb{R}$ and the Cauchy problem for a scalar hyperbolic equation with the linear flux function $f(u) = au$,

$$u_t + au_x = 0 \qquad \text{for all } x \in \mathbb{R},\ t > 0. \tag{1.113}$$

equipped with the initial condition

$$u(x,0) = u_0(x) \qquad \text{for all } x \in \mathbb{R}. \tag{1.114}$$

Definition 1.11 (Characteristics) *Characteristic curve of equation (1.113), passing through the point $(x_0, 0)$, $x_0 \in \mathbb{R}$, is the graph of the solution of the ordinary differential equation*

$$\begin{aligned} x'(t) &= a \quad \text{for all } t > 0, \\ x(0) &= x_0. \end{aligned} \tag{1.115}$$

Lemma 1.6 *The solution of (1.113), (1.114) is constant along the characteristics $x(t)$, and thus it is fully determined by the initial data,*

$$u(x,t) = u_0(x - at). \tag{1.116}$$

Proof: Since $a \in \mathbb{R}$ is constant, by (1.115) the characteristics are straight lines,

$$x(t) = at + x_0.$$

Consider the solution along these lines, $u(x(t), t)$, and take its derivative in time. Using the original equation (1.113), we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t} u(at + x_0, t) = a \frac{\partial u}{\partial x}(x(t), t) + \frac{\partial}{\partial t} u(x(t), t) = 0.$$

For an arbitrary $(x, t) \in \mathbb{R} \times (0, T)$, the characteristics $x(t)$ passing through this point intersects with the real axis at $x(0) = x - at$, where it takes the value $u(x,t) = u(x - at, 0) = u_0(x - at)$. ■

Remark 1.7 (Equation (1.113) describes "flow") *Equation (1.113) does not generate any new information, it only shifts the initial condition u_0 in time. The initial condition moves to the right if $a > 0$ and to the left if $a < 0$. In the degenerated case of $a = 0$ the equation reduces to $\partial u / \partial t = 0$, i.e., the solution is constant in time, which is compatible with the fact that the characteristics have the form $x(t) = x_0$.*

1.5.3 Exact solution to linear first-order systems

The next natural step to take is to analyze linear vector-valued problems in one spatial dimension. Hence, for $m \geq 1$ consider the hyperbolic conservation law (1.104) with a linear flux function $\boldsymbol{f}(\boldsymbol{u}) = \boldsymbol{A}\boldsymbol{u}$,

$$\boldsymbol{u}_t(x,t) + \boldsymbol{A}\boldsymbol{u}_x(x,t) = 0, \tag{1.117}$$

$$\boldsymbol{u}(x,0) = \boldsymbol{u}_0(x), \tag{1.118}$$

where $\boldsymbol{u} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^m$ and $\boldsymbol{A} \in \mathbb{R}^m \times \mathbb{R}^m$ is a constant matrix. By the hyperbolicity of the problem the matrix $\boldsymbol{A}$ is diagonalizable with real eigenvalues, i.e., there exists a nonsingular $m \times m$ matrix $\boldsymbol{R}$ such that

$$\boldsymbol{A} = \boldsymbol{R}\Lambda\boldsymbol{R}^{-1}. \tag{1.119}$$

Here $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_m)$ is a diagonal eigenvalue matrix, and it is worth mentioning that the matrix $\boldsymbol{R}$ contains the right eigenvectors of $\boldsymbol{A}$ in its columns. Thus for the columns of $\boldsymbol{R}$ we have

$$\boldsymbol{A}\boldsymbol{r}_i = \lambda_i \boldsymbol{r}_i \qquad \text{for all } 1 \le i \le m.$$

Let us introduce the notion of strict hyperbolicity for reference:

Definition 1.12 (Strictly hyperbolic system) *The system (1.117), (1.118) is called* strictly hyperbolic *if the eigenvalues λ_i, $1 \le i \le m$, are distinct.*

Characteristic variables One can solve (1.117), (1.118) by switching to characteristic variables

$$\boldsymbol{v} = \boldsymbol{R}^{-1}\boldsymbol{u}.$$

Multiplying (1.117) by $\boldsymbol{R}^{-1}$ and using (1.119), one obtains

$$\boldsymbol{R}^{-1}\boldsymbol{u}_t + \boldsymbol{\Lambda}\boldsymbol{R}^{-1}\boldsymbol{u}_x = \boldsymbol{0},$$

which further yields

$$\boldsymbol{v}_t + \boldsymbol{\Lambda}\boldsymbol{v}_x = 0. \tag{1.120}$$

By the diagonality of $\boldsymbol{\Lambda}$, this is a system of m independent linear advection equations for the components of $\boldsymbol{v}$,

$$\begin{aligned} (v_i)_t + \lambda_i(v_i)_x &= 0, \\ v_i(0) &= v_{0,i}, \end{aligned}$$

$i = 1, 2, \dots, m$. The initial condition for v_i is the ith component of the vector $\boldsymbol{R}^{-1}\boldsymbol{u}_0$. Using what we learned in Paragraph 1.5.2, for each $1 \le i \le m$ the solution is

$$v_i(x,t) = v_i(x - \lambda_i t, 0) = v_{0,i}(x - \lambda_i t).$$

The solution $\boldsymbol{u}$ is finally recovered using the relation

$$\boldsymbol{u}(x,t) = \boldsymbol{R}\boldsymbol{v}(x,t) = \sum_{i=1}^{m} v_i(x,t)\boldsymbol{r}_i, \tag{1.121}$$

which yields

$$\boldsymbol{u}(x,t) = \sum_{i=1}^{m} v_{0,i}(x - \lambda_i t)\boldsymbol{r}_i. \tag{1.122}$$

Simple waves The solution (1.122) is the superposition of m independently advected linear waves. The ith wave has the form

$$v_i(x,0)\boldsymbol{r}_i,$$

and propagates at the wave speed λ_i.

1.5.4 Riemann problem

The solution of the Riemann problem plays an important role in the design of finite volume methods for the approximate solution of nonlinear conservation laws.

Figure 1.5 Georg Friedrich Bernhard Riemann (1826–1866).

G.F.B. Riemann was a German mathematician who, besides other important achievements, introduced topological methods into the theory of complex functions, studied the representation of functions by trigonometric series, and established new foundations of geometry which were used later in relativity and cosmology. The Riemann hypothesis, related to the prime number theory, remains one of the most famous unsolved problems of modern mathematics.

Consider the one-dimensional linear hyperbolic equation (1.117),

$$\boldsymbol{u}_t + \boldsymbol{A}\boldsymbol{u}_x = 0, \tag{1.123}$$

with a piecewise-constant initial condition consisting of two different states $\boldsymbol{u}_L, \boldsymbol{u}_R \in \mathbb{R}^m$ on the negative and positive half of the real line, respectively,

$$\boldsymbol{u}(x,0) = \begin{cases} \boldsymbol{u}_L & x \leq 0, \\ \boldsymbol{u}_R & x > 0. \end{cases} \tag{1.124}$$

For simplicity we assume that the problem (1.123) is strictly hyperbolic. This means that the matrix $\boldsymbol{A}$ has m eigenvalues which are real and distinct. They can be denoted as follows,

$$\lambda_1 < \lambda_2 < \ldots < \lambda_m.$$

Exact solution in characteristic variables Recall that the exact solution to (1.123) is given by (1.122). We can simplify the situation by expressing the initial states $\boldsymbol{u}_L$ and $\boldsymbol{u}_R$ in terms of eigenvectors of the matrix $\boldsymbol{A}$,

$$\boldsymbol{u}_L = \sum_{i=1}^{m} \alpha_i \boldsymbol{r}_i, \quad \boldsymbol{u}_R = \sum_{i=1}^{m} \beta_i \boldsymbol{r}_i.$$

Then

$$v_i(x,0) = \begin{cases} \alpha_i & x \leq 0, \\ \beta_i & x > 0, \end{cases}$$

and the problem (1.123) decouples into m independent scalar Riemann problems

$$\begin{aligned}(v_i)_t + \lambda_i (v_i)_x &= 0, \\ v_i(x,0) &= \begin{cases} \alpha_i & x \le 0, \\ \beta_i & x > 0. \end{cases}\end{aligned} \tag{1.125}$$

For ith scalar problem, the initial discontinuity $[\beta_i - \alpha_i]$ at $x = 0$ propagates into the space-time domain along the characteristics $x_i(t) = \lambda_i t$, as illustrated in Figure 1.6.

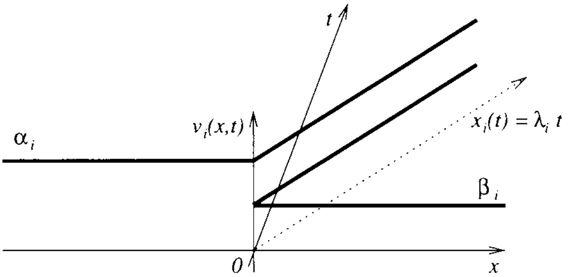


Figure 1.6 Propagation of the jump $[\beta_i - \alpha_i]$ in the ith characteristic variable $v_i(x,t)$ along the ith zero characteristics $x_i(t) = \lambda_i t$.)

Solution at $x = 0$ Finite volume schemes are based on the value of the solution $\boldsymbol{u}(0,t)$, which is constant in time. It is defined if $\lambda_i \neq 0$ for all i (i.e., if no jump is propagated along the temporal axis $x = 0$). It is easy to see that the characteristic variable v_i satisfies

$$v_i(0,t) = \begin{cases} \alpha_i & \lambda_i \ge 0, \\ \beta_i & \lambda_i < 0. \end{cases}$$

Equation (1.121) then yields

$$\boldsymbol{u}(0,t) = \boldsymbol{R}\boldsymbol{v}(0,t) = \sum_{i=1}^{m} v_i(0,t)\boldsymbol{r}_i.$$

Let the first m_0 eigenvalues λ_i be negative and the rest positive. Then the exact solution at $x = 0$ can be expressed as

$$\boldsymbol{u}(0,t) = \sum_{i=1}^{m_0} \beta_i \boldsymbol{r}_i + \sum_{i=m_0+1}^{m} \alpha_i \boldsymbol{r}_i.$$

An important quantity is the (also time-independent) value of $\boldsymbol{A}\boldsymbol{u}(0,t)$ that represents the linear flux across the interface $x = 0$,

$$\begin{aligned}\boldsymbol{A}\boldsymbol{u}(0,t) &= \boldsymbol{A}\sum_{i=1}^{m_0} \beta_i \boldsymbol{r}_i + \boldsymbol{A} \sum_{i=m_0+1}^{m} \alpha_i \boldsymbol{r}_i \\ &= \sum_{i=1}^{m_0} \beta_i \lambda_i \boldsymbol{r}_i + \sum_{i=m_0+1}^{m} \alpha_i \lambda_i \boldsymbol{r}_i \\ &= \sum_{i=1}^{m} \beta_i \lambda_i^- \boldsymbol{r}_i + \sum_{i=1}^{m} \alpha_i \lambda_i^+ \boldsymbol{r}_i \\ &= \boldsymbol{A}^- \boldsymbol{u}_R + \boldsymbol{A}^+ \boldsymbol{u}_L.\end{aligned} \tag{1.126}$$

Here

$$\begin{aligned}\lambda_i^- &= \min(\lambda_i, 0),\\ \lambda_i^+ &= \max(\lambda_i, 0).\end{aligned}$$

The matrices $\boldsymbol{A}^-$, $\boldsymbol{A}^+$ are the negative and positive parts of the matrix $\boldsymbol{A}$, defined using the decomposition $\boldsymbol{A} = \boldsymbol{R}\boldsymbol{\Lambda}\boldsymbol{R}^{-1}$ and $\lambda_i = \lambda_i^- + \lambda_i^+$, as

$$\begin{aligned}\boldsymbol{A}^- &= \boldsymbol{R}\boldsymbol{\Lambda}^-\boldsymbol{R}^{-1},\\ \boldsymbol{A}^+ &= \boldsymbol{R}\boldsymbol{\Lambda}^+\boldsymbol{R}^{-1}.\end{aligned}$$

Here $\boldsymbol{\Lambda}^- = \mathrm{diag}(\lambda_1^-, \lambda_2^-, \ldots, \lambda_m^-)$ and $\boldsymbol{\Lambda}^+ = \mathrm{diag}(\lambda_1^+, \lambda_2^+, \ldots, \lambda_m^+)$. Obviously, $\boldsymbol{A} = \boldsymbol{A}^- + \boldsymbol{A}^+$. Analogously we define the absolute value of the matrix $\boldsymbol{A}$, $|\boldsymbol{A}| = \boldsymbol{A}^+ - \boldsymbol{A}^- = \boldsymbol{R}|\boldsymbol{\Lambda}|\boldsymbol{R}^{-1}$, where $|\boldsymbol{\Lambda}| = \mathrm{diag}(|\lambda_1|, |\lambda_2|, \ldots, |\lambda_m|)$.

Application to nonlinear conservation laws The matrices $\boldsymbol{A}^+$, $\boldsymbol{A}^-$, $|\boldsymbol{A}|$ are used by several popular finite volume schemes for the solution of nonlinear hyperbolic conservation laws, including the compressible Euler equations. The basic idea of the approximation consists in the linearization of the nonlinear flux functions (their replacement with their Jacobi matrices) and consequent application of the above-described procedure for the linear Riemann problem. The approximation of the time-independent value $\boldsymbol{A}\boldsymbol{u}(0, t)$ plays a key role in the finite volume schemes. Let us stop the comment at this point, since the finite volume method lies beyond the scope of this text. There is a vast literature devoted to this topic. We refer the reader, e.g., to [52, 54, 77] and [78].

1.5.5 Nonlinear flux and shock formation

To illustrate the mechanism of the creation of discontinuities in nonlinear first-order hyperbolic problems, consider a nonlinear analogy to (1.113), (1.114),

$$u_t(x,t) + [f(u(x,t))]_x = 0 \quad \text{for all } x \in \mathbb{R},\ t > 0, \tag{1.127}$$

$$u(x,0) = u_0(x), \tag{1.128}$$

where the flux function $f : \mathbb{R} \to \mathbb{R}$ is once continuously differentiable. For demonstration purposes let us pick the function

$$f(u) = \frac{1}{2}u^2.$$

This choice leads to Burgers' equation (1.11),

$$u_t(x,t) + u(x,t)u_x(x,t) = 0. \tag{1.129}$$

The characteristics of equation (1.127) are defined as

$$x'(t) = \frac{\mathrm{d}f}{\mathrm{d}u}(u(x(t),t)) = u(x(t),t), \tag{1.130}$$

$$x(0) = x_0.$$

Using (1.130) and (1.129), it is easy to verify that the solution $u(x(t),t)$ along these characteristics is constant,

$$\frac{\mathrm{d}}{\mathrm{d}t}u(x(t),t) = \frac{\partial u}{\partial x}(x(t),t)\underbrace{u(x(t),t)}_{x'(t)} + \frac{\partial}{\partial t}u(x(t),t) = 0.$$

Since $x'(t) = u(x(t), t)$ is the slope of the characteristics and $u(x(t), t)$ is constant, also in this case the characteristics are straight lines. A characteristic curve passing through $(x_0, 0)$ has the slope $u(x_0, 0) = u_0(x_0)$. When two different characteristic curves, carrying two different values of the solution on them, intersect, a discontinuity (shock) is born. This is illustrated in Figure 1.7.

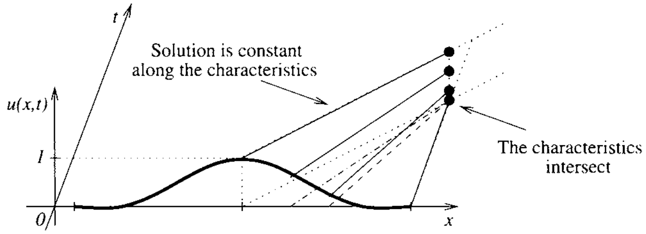

t
Solution is constant along the characteristics
u(x,t)
1
0
x
The characteristics intersect

Figure 1.7 Formation of shock in the solution $u(x, t)$ of Burgers' equation.

Nonlinear hyperbolic problems constitute a more or less autonomous field in applied mathematics, and there is a wide class of literature dedicated to both their theoretical and computational aspects. See the literature listed at the end of the previous paragraph and references therein.

1.5.6 Exercises

Exercise 1.22 *Under sufficient regularity conditions for the flux $\boldsymbol{f}$ and the solution $\boldsymbol{u}$, show that every solution $\boldsymbol{u}$ of (1.104) is conserved in time, i.e., it satisfies condition (1.106). Hint: Integrate (1.104) over $\mathbb{R}$, use the fundamental theorem of calculus and decay conditions for functions integrable in $\mathbb{R}$.*

Exercise 1.23 (Exact solution to the wave equation) *Calculate the eigenvectors of the matrix $\boldsymbol{A}$ defined in (1.100). Use the characteristic variables to construct the exact solution (1.102) of the linear first-order hyperbolic system (1.100), (1.101).*

Exercise 1.24 *Prove a simplified version of the localization theorem: Let $\Omega \subset \mathbb{R}^d$ be an open bounded set. Let $f \in C(\overline{\Omega})$. Let*

$$\int_{\sigma} f\, dx = 0$$

be valid for all open bounded sets $\sigma \subset \Omega$. Then f is zero everywhere in Ω.

Exercise 1.25 *Consider a linear hyperbolic problem of the form (1.117), (1.118) with the flux function $\boldsymbol{f}(\boldsymbol{u}) = \boldsymbol{A}\boldsymbol{u}$, where the matrix $\boldsymbol{A}$ has the form*

$$\boldsymbol{A} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

Consider a general initial condition $\boldsymbol{u}(x, 0) = \boldsymbol{u}_0(x)$ for all $x \in \mathbb{R}$. Write the exact solution to this problem.

CHAPTER 2

CONTINUOUS ELEMENTS FOR 1D PROBLEMS

After reviewing the basic theory of partial differential equations in Chapter 1, let us now introduce the Galerkin method and its important special case, the Finite element method.

2.1 THE GENERAL FRAMEWORK

Let V be a Hilbert space, $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ a bilinear form (coming, e.g., from the weak formulation of a PDE) and $l \in V'$ (representing, e.g., the right-hand side of a PDE). It is our task to find $u \in V$ such that

$$a(u,v) = l(v) \quad \text{for all } v \in V. \tag{2.1}$$

We assume that the bilinear form $a(\cdot,\cdot)$ is bounded and V-elliptic, i.e., that there exist constants $C_b, C_{el} > 0$ such that

$$|a(u,v)| \le C_b \|u\|_V \|v\|_V \quad \text{for all } u,v \in V, \tag{2.2}$$

and

$$a(v,v) \ge C_{el} \|v\|_V^2 \quad \text{for all } v \in V. \tag{2.3}$$

Recall that the weak problem (2.1) has a unique solution by the Lax–Milgram lemma (Theorem 1.5).

2.1.1 The Galerkin method

Problem (2.1) was stated in an infinitely-dimensional space V. Therefore, its exact solution, as a "function of infinitely many unknown parameters", is impossible to find in general. The finite-dimensional (numerical) approximation of such problems was first studied systematically by Boris Grigorievich Galerkin.

Figure 2.1 Boris Grigorievich Galerkin (1871–1945).

B.G. Galerkin was a Russian mathematician who became famous for his results related to thin elastic plates, numerical solution of partial differential equations, and investigation of the stress in dams and breast walls with trapezoidal profile. His work found many industrial applications, including the construction of large dams and hydroelectric power stations.

The Galerkin method, which he first published in 1915, is based on a sequence of finite-dimensional subspaces $\{V_n\}_{n=1}^{\infty} \subset V$, $V_n \subset V_{n+1}$, that fill the space V in the limit. In each finite-dimensional space V_n problem (2.1) is solved exactly. It can be shown that under suitable assumptions the sequence of the approximate solutions $\{u_n\}_{n=1}^{\infty}$, $u_n \in V_n$, converges to the exact solution of problem (2.1).

Let $\{V_n\}_{n=1}^{\infty} \subset V$ be a sequence of subspaces of V such that

$$\overline{\bigcup_{i=1}^{\infty} V_n} = V, \tag{2.4}$$

where $V_n \subset V_{n+1} \subset V$ and $\dim(V_n) = N_n < \infty$ for all $n = 1, 2, \ldots$. Every finite-dimensional subspace of a Hilbert space is closed and therefore a Hilbert space (see Remark A.5). The Galerkin approximate problem usually is called discrete problem.

Discrete problem Find a solution $u_n \in V_n$, satisfying

$$a(u_n, v) = l(v) \quad \text{for all } v \in V_n. \tag{2.5}$$

Lemma 2.1 (Unique solvability) *Problem (2.5) has a unique solution $u_n \in V_n$.*

Proof: The form $a(\cdot,\cdot)$, restricted to $V_n \times V_n$, obviously remains bilinear, bounded, and V_n-elliptic. The linear form $l(v)$, restricted to V_n, remains linear, and therefore $l \in V_n'$. Thus the assumptions of the Lax–Milgram lemma (Theorem 1.5) are fulfilled and there exists a unique solution to (2.5). ■

The solution $u_n \in V_n$ to the discrete problem (2.5) can be found explicitly thanks to the fact that the space V_n has a finite basis $\{v_n\}_{n=1}^{N_n}$. The solution u_n can be written as a linear combination of these basis functions with unknown coefficients,

$$u_n = \sum_{j=1}^{N_n} y_j v_j. \tag{2.6}$$

Substituting (2.6) into (2.5), one obtains

$$a\left(\sum_{j=1}^{n} y_j v_j, v\right) = l(v) \quad \text{for all } v \in V_n. \tag{2.7}$$

The linearity of $a(\cdot, \cdot)$ in its first component yields

$$\sum_{j=1}^{N_n} a\left(v_j, v\right) y_j = l(v) \quad \text{for all } v \in V_n. \tag{2.8}$$

Substituting the basis functions $v_1, v_2, \ldots, v_{N_n}$ for v in (2.8), we obtain

$$\sum_{j=1}^{N_n} a\left(v_j, v_i\right) y_j = l(v_i), \quad i = 1, 2, \ldots, N_n. \tag{2.9}$$

It is worth taking a moment to see that (2.8) and (2.9) are equivalent: The implication from (2.8) to (2.9) is easy since every basis function $v_i \in V$ is a special case of a general $v \in V$. Conversely, an arbitrary $v \in V_n$ can be written as a linear combination

$$v = \sum_{i=1}^{N_n} \beta_i v_i.$$

Multiplying the ith equation in (2.9) with β_i and using the linearity of the forms a and l, we obtain

$$\sum_{j=1}^{N_n} a\left(v_j, \beta_i v_i\right) y_j = l(\beta_i v_i) \quad i = 1, 2, \ldots, N_n.$$

Summing up these equations over all i, we see that

$$\sum_{j=1}^{N_n} a\left(v_j, \sum_{i=1}^{N_n} \beta_i v_i\right) y_j = l\left(\sum_{i=1}^{N_n} \beta_i v_i\right),$$

i.e.,

$$\sum_{j=1}^{N_n} a\left(v_j, v\right) y_j = l(v),$$

Next let us define the stiffness matrix

$$\boldsymbol{S}_n = \{s_{ij}\}_{i,j=1}^{N_n}, \qquad s_{ij} = a(v_j, v_i), \tag{2.10}$$

the load vector

$$\boldsymbol{F}_n = \{f_i\}_{i=1}^{N_n}, \qquad f_i = l(v_i), \tag{2.11}$$

and the unknown coefficient vector

$$\boldsymbol{Y}_n = \{y_i\}_{i=1}^{N_n}. \tag{2.12}$$

Then the system of linear algebraic equations (2.9) can be written in a matrix form

$$\boldsymbol{S}_n \boldsymbol{Y}_n = \boldsymbol{F}_n. \tag{2.13}$$

In order to show the invertibility of the matrix $\boldsymbol{S}_n$, let us prove its positive definiteness first:

Lemma 2.2 (Positive definiteness of $\boldsymbol{S}_n$) *Let V_n, $dim(V_n) = N_n < \infty$ be a Hilbert space and $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ a bilinear V-elliptic form. Then the stiffness matrix $\boldsymbol{S}_n$ of the discrete problem (2.13) is positive definite.*

Proof: It is our aim to show that $\hat{\boldsymbol{Y}}^T \boldsymbol{S} \hat{\boldsymbol{Y}} > 0$ for all $0 \neq \hat{\boldsymbol{Y}} \in \mathbb{R}^{N_n}$. Thus take an arbitrary $\hat{\boldsymbol{Y}} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_{N_n})^T$ and define the vector

$$\hat{v} = \sum_{i=1}^{N_n} \hat{y}_i v_i,$$

where $\{v_1, v_2, \dots, v_{N_n}\}$ is some basis in V_n. By the V-ellipticity of the form $a(\cdot,\cdot)$ it is

$$\begin{aligned} \hat{\boldsymbol{Y}}^T \boldsymbol{S}_n \hat{\boldsymbol{Y}} &= \sum_{i=1}^{n} \sum_{j=1}^{n} \hat{y}_i s_{ji} \hat{y}_j \\ &= a\left(\sum_{i=1}^{n} \hat{y}_i v_i, \sum_{j=1}^{n} \hat{y}_j v_j \right) = a(\hat{v}, \hat{v}) \geq C_{el} \|\hat{v}\|_V^2 > 0, \end{aligned}$$

which was to be shown. ■

Corollary 2.1 (Invertibility of $\boldsymbol{S}_n$) *The stiffness matrix $\boldsymbol{S}_n$ of the discrete problem (2.13) is nonsingular.*

Proof: This fact follows immediately from the existence and uniqueness of the solution $u_n \in V_n$ (Lemma 2.1). Alternatively, let us assume that $\boldsymbol{S}_n$ is singular. Then there exists a nontrivial vector $\boldsymbol{Y}_0 \in \mathbb{R}^{N_n}$ such that $\boldsymbol{S}_n \boldsymbol{Y}_0 = 0$. Then necessarily $\boldsymbol{Y}_0^T \boldsymbol{S}_n \boldsymbol{Y}_0 = 0$, which is a contradiction with the positive definiteness of $\boldsymbol{S}_n$ (Lemma 2.2). ■

Thus we conclude that the system of linear algebraic equations (2.13) has a unique solution $\boldsymbol{Y}_n$ that defines a unique solution $u_n \in V_n$ of (2.5) via (2.6).

Now let us interrupt the discussion of the Galerkin method for a moment and introduce an important concept of orthogonality of error for elliptic problems and Céa's lemma in Paragraph 2.1.2. The convergence proof for the Galerkin sequence will be presented as a simple consequence of these results in Paragraph 2.1.3.

2.1.2 Orthogonality of error and Céa's lemma

The error $e_n = u - u_n$ of the solution to the discrete problem (2.9) exhibits the following orthogonality property:

Lemma 2.3 (Orthogonality of error for elliptic problems) *Let $u \in V$ be the exact solution of the continuous problem (2.1) and u_n the exact solution of the discrete problem (2.5). Then the error $e_n = u - u_n$ satisfies*

$$a(u - u_n, v) = 0 \quad \textit{for all } v \in V_n. \tag{2.14}$$

Proof: Subtract (2.5) from (2.1) restricted to $V_n \subset V$. ■

Remark 2.1 (Geometrical interpretation) *If the bilinear form $a(\cdot,\cdot)$ is symmetric, it induces an energetic inner product*

$$(u, v)_e = a(u, v).$$

It follows from (2.14) that

$$(e_n, v)_e = 0 \quad \textit{for all } v \in V_n,$$

i.e., that the error of the Galerkin approximation $e_n = u - u_n$ is orthogonal to the Galerkin subspace V_n in the energetic inner product. Hence the approximate solution $u_n \in V_n$ is an orthogonal projection of the exact solution $u \in V$ onto the Galerkin subspace V_n in the energetic inner product, and thus it is the nearest element in the space V_n to the exact solution u in the energy norm,

$$\|u - u_n\|_e = \inf_{v \in V_n} \|u - v\|_e. \tag{2.15}$$

Next let us introduce Céa's lemma, which establishes the relation between the error of the approximation $e_n = u - u_n$ and the interpolation properties of the subspace V_n, using the continuity and V-ellipticity constants C_b, C_{el} of the bilinear form $a(\cdot,\cdot)$.

Theorem 2.1 (Céa's lemma) *Let V be a Hilbert space, $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ a bilinear bounded V-elliptic form and $l \in V'$. Let $u \in V$ be the solution of problem (2.1). Further, let V_n be a subspace of V and $u_n \in V_n$ the solution of the Galerkin approximation (2.5). Let C_b, C_{el} be the continuity and V-ellipticity constants of the form $a(\cdot,\cdot)$. Then*

$$\|u - u_n\|_V \leq \frac{C_b}{C_{el}} \inf_{v \in V_n} \|u - v\|_V.$$

Proof: Using relation (2.14), we obtain that

$$\begin{aligned} a(u - u_n, u - u_n) &= a(u - u_n, u - v) - a(u - u_n, u_n - v) \\ &= a(u - u_n, u - v) \end{aligned}$$

for an arbitrary $v \in V_n$. By the V-ellipticity of the bilinear form $a(\cdot,\cdot)$ we have

$$a(u - u_n, u - u_n) \geq C_{el} \|u - u_n\|_V^2. \tag{2.16}$$

The boundedness of $a(\cdot,\cdot)$ yields

$$a(u-u_n,u-u_n) \le C_b\|u-u_n\|_V\|u-v\|_V \quad \text{for all } v \in V_n. \tag{2.17}$$

Putting relations (2.16) and (2.17) together, we obtain

$$\|u-u_n\|_V \le \frac{C_b}{C_{el}}\|u-v\|_V \quad \text{for all } v \in V_n,$$

which was to be shown. ■

Theorem 2.1 was first proved by Céa [27] in 1964 for the symmetric case and extended to the nonsymmetric case four years later in [13].

Remark 2.2 *Céa's lemma states that the approximation error $e_n = u - u_n$ depends on the choice of the Galerkin subspace V_n, but it does not depend on the choice of its basis. Therefore, when working with finite element methods for elliptic problems, one should think in terms of function spaces rather than in terms of concrete basis functions. Also numerical results should stay independent of a concrete choice of finite element basis functions.*

Where the choice of the basis matters is the condition number of the stiffness matrix $\boldsymbol{S}_n$, which influences the performance of iterative matrix solvers. This issue will be discussed in more detail in Paragraph 2.5.2.

2.1.3 Convergence of the Galerkin method

The convergence of the Galerkin method for elliptic problems is a simple consequence of Céa's lemma (Theorem 2.1).

Theorem 2.2 *Let V be a Hilbert space and $V_1 \subset V_2 \subset \ldots \subset V$ a sequence of its finite dimensional subspaces such that (2.4),*

$$\overline{\bigcup_{n=1}^{\infty} V_n} = V. \tag{2.18}$$

Let $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ be a bounded bilinear V-elliptic form and $l \in V'$. Then

$$\lim_{n\to\infty} \|u-u_n\|_V = 0,$$

i.e., the Galerkin method for problem (2.1) converges.

Proof: Given the exact solution $u \in V$ of (2.1), by (2.18) it is possible to find some sequence $\{v_n\}_{n=1}^{\infty}$ such that $v_n \in V_n$ for every $n = 1,2,\ldots$ and

$$\lim_{n\to\infty} \|u-v_n\|_V = 0. \tag{2.19}$$

Lemma 2.5 yields the existence and uniqueness of a solution $u_n \in V_n$ of the discrete problem (2.5) for every $n \ge 1$. By Céa's lemma,

$$\|u-u_n\|_V \le \frac{C_b}{C_{el}} \inf_{v\in V_n} \|u-v\|_V \le \frac{C_b}{C_{el}}\|u-v_n\|_V \quad \text{for all } n = 1,2,\ldots$$

By (2.19) we conclude that

$$\lim_{n\to\infty} \|u - u_n\|_V = 0,$$

which was to be shown. ■

2.1.4 Ritz method for symmetric problems

We have shown in Paragraph 1.2.9 that for a symmetric bounded bilinear V-elliptic form $a(\cdot,\cdot)$ problem (2.1) is equivalent to a minimization problem for the abstract energy functional (1.68),

$$E(v) = \frac{1}{2}a(v, v) - l(v),$$

in the space V. On the discrete level, it follows from Theorem 1.6 that the discrete problem (2.5) is equivalent to a discrete minimization problem of minimizing $E(v)$ in the finite-dimensional subspace V_n. The equivalence of the Galerkin and Ritz methods in the symmetric case is the reason why sometimes the Galerkin method is referred to as the Ritz–Galerkin method.

2.1.5 Exercises

Exercise 2.1 *Prove that every symmetric bilinear form $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ yields a symmetric stiffness matrix S. Hint: A bilinear form $a(\cdot,\cdot) : V \times W \to \mathbb{R}$ can only be symmetric if $V = W$ (see Definition 1.5).*

Exercise 2.2 *Prove the equivalence of the Galerkin and Ritz methods in the symmetric case, stated in Paragraph 2.1.4:*

Let V be a linear space, $V_n \subset V$ its subspace, $a : V \times V \to \mathbb{R}$ a symmetric V-elliptic bilinear form, and $l \in V'$. Show that the abstract energy

$$E(v) = \frac{1}{2}a(v, v) - l(v)$$

attains its minimum over V_n at $u_n \in V_n$ if and only if

$$a(u_n, v) = l(v) \quad \textit{for all } v \in V_n.$$

Proceed similarly to the proof of Theorem 1.6.

2.2 LOWEST-ORDER ELEMENTS

Let $\Omega \subset \mathbb{R}^d$, where d is the spatial dimension, be an open bounded set. If the Hilbert space V consists of functions defined in Ω and the Galerkin subspaces $V_n \subset V$ comprise piecewise-polynomial functions, the Galerkin method is called the Finite element method (FEM). Let us begin with the exposition of the simplest case of piecewise-affine elements in one spatial dimension.

2.2.1 Model problem

Consider the model equation (1.26),

$$-\nabla \cdot (a_1 \nabla u) + a_0 u = -(a_1 u')' + a_0 u = f, \tag{2.20}$$

where $f \in L^2(\Omega)$, in a bounded interval $\Omega = (a, b) \subset \mathbb{R}$, equipped with the homogeneous Dirichlet boundary conditions (1.28). The weak formulation of this problem (see Paragraph 1.2.1) takes place in the Sobolev space

$$V = H_0^1(\Omega).$$

We assume that the coefficient functions $a_1, a_0 \in L^\infty(\Omega)$ satisfy the unique solvability assumptions (1.27),

$$a_1(x) \geq C_{min} > 0, \qquad a_0(x) \geq 0 \quad \text{a.e. in } \Omega.$$

At the beginning let a_1 and a_0 be constants and assume a simple load function of the form

$$f(x) = 1 \quad \text{in } \Omega. \tag{2.21}$$

The model problem reads: Find a function $u \in V$ satisfying

$$a(u, v) = l(v) \quad \text{for all } v \in V, \tag{2.22}$$

where the bilinear form $a : V \times V \to \mathbb{R}$ is given by

$$a(u, v) = \int_\Omega a_1 \nabla u \cdot \nabla v + a_0 uv \, \mathrm{d}x = \int_\Omega a_1 u'(x) v'(x) + a_0 u(x) v(x) \, \mathrm{d}x$$

and the linear form $l \in V'$ is defined by

$$l(v) = \langle l, v \rangle = \int_\Omega f v \, \mathrm{d}x.$$

2.2.2 Finite-dimensional subspace $V_n \subset V$

The Galerkin procedure assumes a sequence of finite-dimensional subspaces

$$V_1 \subset V_2 \subset \ldots \subset V$$

of the infinite-dimensional Hilbert space V, satisfying (2.18),

$$\overline{\bigcup_n V_n} = V. \tag{2.23}$$

Let $n \geq 1$ be a natural number. Consider a partition

$$a = x_0^{(n)} < x_1^{(n)} \ldots < x_{M_n}^{(n)} = b$$

of the interval $\Omega = (a, b)$, and define the finite element mesh

$$\mathcal{T}_n = \{K_1^{(n)}, K_2^{(n)}, \dots, K_{M_n}^{(n)}\}.$$

The open intervals

$$K_i^{(n)} = \left(x_{i-1}^{(n)}, x_i^{(n)}\right)$$

are called finite elements, and the value

$$h(n) = \max_{1 \le i \le M_n} (x_i^{(n)} - x_{i-1}^{(n)})$$

is said to be the mesh diameter.

Remark 2.3 *For historical reasons the subscript $h = h(n)$ is often used instead of the subscript n to distinguish between the Galerkin subspaces (the mesh diameter h is closely related to the approximation error for lowest-order methods). Since*

$$V_{h(n)} \to V \text{ as } h(n) \to 0,$$

the limit $n \to \infty$ can be replaced with the limit $h \to 0$ in the Galerkin procedure. The Galerkin method itself remains unchanged.

In the case of piecewise-affine elements, the Galerkin subspace $V_n \subset V$ consists of continuous functions that are affine polynomials in every domain $K_i^{(n)} \in \mathcal{T}_h$. One defines

$$V_n = \left\{v \in V;\ v|_{K_i^{(n)}} \in P^1\left(K_i^{(n)}\right) \text{ for all } i = 1, 2, \dots, M_n\right\}. \tag{2.24}$$

Recall that functions in the space $H^1(\Omega)$ in one spatial dimension are continuous (Examples A.53 and A.54). Therefore one refers to the finite elements in this space as to continuous elements.

2.2.3 Piecewise-affine basis functions

While the Galerkin method assumes an arbitrary basis of the space V_n, the finite element method (up to rare exceptions) prefers basis functions with small supports, so that as many of them as possible are disjoint. When the supports of v_i and v_j are disjoint, the stiffness matrix entry $s_{ij} = a(v_j, v_i)$ is zero. Matrices with few nonzero entries are called sparse, and in comparison with dense matrices they are much easier to store in the computer memory and to solve numerically. More about sparse matrices and their properties will be said in Paragraph 2.5.1.

The most convenient basis of the space V_n in the piecewise-affine one-dimensional case consists of $M_n - 1$ continuous "hat functions" v_j of the form

$$v_i(x) = \begin{cases} (x - x_{i-1})\dfrac{1}{x_i - x_{i-1}}, & x \in \overline{K}_i, \\[2ex] (x_{i+1} - x)\dfrac{1}{x_{i+1} - x_i}, & x \in \overline{K}_{i+1}, \\[2ex] 0, & \text{elsewhere in } (a, b), \end{cases} \tag{2.25}$$

$i = 1, 2, \ldots, M_n - 1$. The basis functions v_i satisfy

$$v_i(x_j) = \delta_{ij},$$

where x_j are the grid points and δ_{ij} the Kronecker delta. The support of each v_i, formed by the pair of elements K_i and K_{i+1}, is minimal. It is easy to see that $\dim(V_n) = M_n - 1$. The hat functions are shown in Figure 2.2.

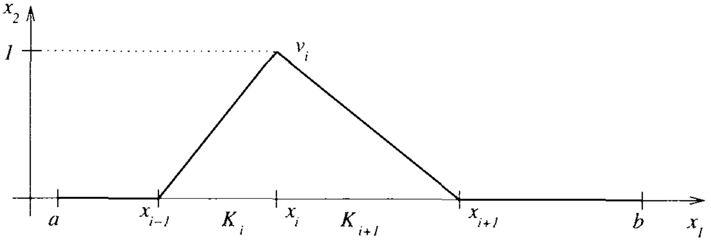


Figure 2.2 Example of a basis function v_i of the space V_n.

Using the basis functions (2.25), the solution u_n to the discrete problem (2.9) can be written in the form (2.6),

$$u_n(x) = \sum_{i=1}^{M_n-1} y_i v_i(x), \tag{2.26}$$

where y_i are unknown real coefficients.

2.2.4 The system of linear algebraic equations

Now we are in position to construct the linear algebraic system (2.13),

$$\boldsymbol{S}_n \boldsymbol{Y}_n = \boldsymbol{F}_n. \tag{2.27}$$

By h_i let us denote the length of the element $K_i = (x_{i-1}, x_i)$. According to (2.10), the stiffness matrix $\boldsymbol{S}_n = \{s_{ij}\}_{i,j=1}^{M_n-1}$ has the entries

$$s_{ij} = a(v_j, v_i) = \int_\Omega a_1 v_j'(x) v_i'(x) + a_0 v_j(x) v_i(x) \,\mathrm{d}x.$$

Using (2.25), one obtains

$$s_{ij} = \begin{cases} -\dfrac{a_1}{h_i} + a_0 \dfrac{h_i}{6}, & j = i-1, \\[2ex] a_1\left(\dfrac{1}{h_i} + \dfrac{1}{h_{i+1}}\right) + a_0\left(\dfrac{h_i}{3} + \dfrac{h_{i+1}}{3}\right), & j = i, \\[2ex] -\dfrac{a_1}{h_{i+1}} + a_0 \dfrac{h_{i+1}}{6}, & j = i+1, \\[2ex] 0, & \text{otherwise}. \end{cases} \tag{2.28}$$

With (2.21), the load vector $F_n = \{f_i\}_{i=1}^{M_n-1}$ has the components

$$f_i = l(v_i) = \int_{K_i \cup K_{i+1}} v_i(x)\,\mathrm{d}x = \frac{1}{2}h_i + \frac{1}{2}h_{i+1} = \frac{1}{2}(h_i + h_{i+1}). \tag{2.29}$$

Hence the system of linear algebraic equations (2.27) has a tridiagonal form illustrated in Figure 2.3.

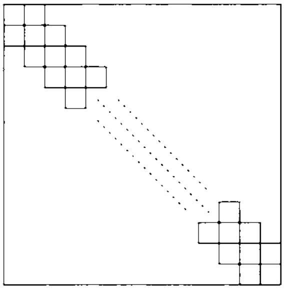


Figure 2.3 Tridiagonal stiffness matrix S_n for piecewise-affine approximations in 1D.

It follows from Corollary 2.1 that the stiffness matrix S_n is invertible, and thus there exists a unique vector Y_n containing the coefficients of the approximate solution $u_n \in V_n$. This model situation is particularly interesting, since the linear algebraic system (2.27) can be written down very easily, and even solved exactly with some effort (when all elements have the same length). This is left to the reader as an exercise. However, the reader should be aware of the fact that in practice, computer programs have to be written for both the assembly and solution of the linear algebraic system (2.27). Let us discuss the assembling algorithm in the next paragraph.

2.2.5 Element-by-element assembling procedure

The ith row in the linear algebraic system (2.27) corresponds to the ith test function $v_i^{(n)} \in V_n$, which is associated with the ith grid point $x_i^{(n)}$. Therefore it seems natural to write the assembling algorithm as a loop over all internal grid vertices $x_1^{(n)}, x_2^{(n)}, \ldots, x_{M_n-1}^{(n)}$:

Algorithm 2.1 (Vertex-by-vertex scheme)

```
//Contributions corresponding to the grid vertex $x_1^{(n)}$:
$s_{1,1} = a_1(1/h_1 + 1/h_2) + a_0(h_1/3 + h_2/3)$;
$s_{1,2} = -a_1/h_2 + a_0h_2/6$;
$f_1 = (h_1 + h_2)/2$;
//Contributions corresponding to the grid vertices
$x_2^{(n)}, x_3^{(n)}, \ldots, x_{M_n-2}^{(n)}$:
for $i = 2, 3, \ldots, M_n - 2$ do {
   $s_{i,i-1} = -a_1/h_i + a_0h_i/6$;
   $s_{ii} = a_1(1/h_i + 1/h_{i+1}) + a_0(h_i/3 + h_{i+1}/3)$;
   $s_{i,i+1} = -a_1/h_{i+1} + a_0h_{i+1}/6$;
   $f_i = (h_i + h_{i+1})/2$;
}
```

```
//Contributions corresponding to the grid vertex $x^{(n)}_{M_n-1}$:
$s_{M_n-1,M_n-2} = -a_1/h_{M_n-1} + a_0 h_{M_n-1}/6$;
$s_{M_n-1,M_n-1} = a_1(1/h_{M_n-1} + 1/h_{M_n}) + a_0(h_{M_n-1}/3 + h_{M_n}/3)$;
$f_{M_n-1} = (h_{M_n-1} + h_{M_n})/2$;
```

However, the vertex scheme is difficult to work with in higher spatial dimensions and to extend to higher-order finite element methods. In particular, higher-order finite elements come with basis functions associated with vertices, edges, faces (in 3D only), and element interiors, and thus the vertices lose their unique role in the assembling procedure. From the point of view of future extensions it is better to assemble the linear system (2.27) in an element-by-element fashion:

Algorithm 2.2 (Element-by-element scheme)

```
Set the stiffness matrix $S_n$ zero.
Set the load vector $F_n$ zero.
//Contributions corresponding to the element $K^{(n)}_1$:
$s_{1,1} = s_{1,1} + a_1/h_1 + a_0 h_1/3$;
$f_1 = f_1 + h_1/2$;
//Contributions corresponding to the elements $K^{(n)}_2, K^{(n)}_3, \ldots, K^{(n)}_{M_n-1}$:
for $i = 2, 3, \ldots, M_n - 1$ do {
    $s_{i-1,i-1} = s_{i-1,i-1} + a_1/h_i + a_0 h_i/3$;
    $s_{i-1,i} = s_{i-1,i} - a_1/h_i + a_0 h_i/6$;
    $s_{i,i-1} = s_{i,i-1} - a_1/h_i + a_0 h_i/6$;
    $s_{ii} = s_{ii} + a_1/h_i + a_0 h_i/3$;
    $f_{i-1} = f_{i-1} + h_i/2$;
    $f_i = f_i + h_i/2$;
}
//Contributions corresponding to the element $K^{(n)}_{M_n}$:
$s_{M_n-1,M_n-1} = s_{M_n-1,M_n-1} + a_1/h_{M_n} + a_0 h_{M_n}/3$;
$f_{M_n-1} = f_{M_n-1} + h_{M_n}/2$;
```

It is left to the reader as an exercise to verify that indeed Algorithms 2.1 and 2.2 yield the same system of linear algebraic equations. We shall use element-by-element algorithms similar to Algorithm 2.2 in the following.

2.2.6 Refinement and convergence

In Paragraphs 2.2.2–2.2.5 we discussed one step of the Galerkin method only: The construction of the approximate solution $u_n \in V_n$ in a given finite-dimensional space $V_n \subset V$. To accomplish the Galerkin procedure, we need a sequence of subspaces $V_1 \subset V_2 \subset \ldots \subset V$ such that $V_n \to V$.

Assume a space V_n associated with a mesh $\mathcal{T}_n$. The next mesh $\mathcal{T}_{n+1}$ can be defined, for example, by halving all intervals $K^{(n)}_i$. Then we have the diameter $h(n+1) = h(n)/2$ and $M_{n+1} = 2M_n$. Clearly the space V_{n+1} of continuous piecewise-affine functions on the refined mesh $\mathcal{T}_{n+1}$ satisfies

$$V_n \subset V_{n+1} \subset V, \tag{2.30}$$

and when the refinements are repeated, one obtains a sequence of spaces satisfying (2.23),

$$\overline{\bigcup_n V_n} = V. \tag{2.31}$$

The discretization procedure is performed on each mesh

$$\mathcal{T}_1, \mathcal{T}_2, \ldots,$$

and one obtains the desired Galerkin sequence of approximate solutions $\{u_n\}_{n=1}^{\infty} \subset V$. Theorem 2.2 yields the convergence to the exact solution u of the continuous problem,

$$\lim_{n \to \infty} \|u_n - u\|_V = 0.$$

A-posteriori error estimation The above-described approach to mesh refinement is not very practical, since the number of elements and consequently the number of unknowns grow exponentially. In practice one needs to use more sophisticated adaptive strategies that refine the mesh only where the error $u - u_n$ is largest. Since the exact solution u is not known, the a-posteriori error estimation (i.e., error estimation based on the values of the computed approximation u_n) comes into the play. The Galerkin subspaces $V_n \subset V$ are constructed in such a way that the distance $\mathrm{dist}(u, V_n) = \inf_{v \in V_n} \|u - v\|_V$ is minimized most efficiently as the number of unknowns $N_n = \dim(V_n)$ is increased.

2.2.7 Exercises

Exercise 2.3 *Assume problem (2.22) with $a_1 = 1$ and $a_0 = 0$, and an equidistant partition of the domain $\Omega = (-1, 1)$ with $M_n \geq 2$ elements (i.e., $h_1 = h_2 = \ldots = h_{M_n} = 2/M_n$).*

1. *Calculate the exact solution $u \in C^2(\Omega) \cap C(\overline{\Omega})$.*
2. *Solve analytically the linear system (2.27) defined via formulae (2.28) and (2.29).*
3. *Take the limit $n \to \infty$ to see that $u_n \to u$.*

Exercise 2.4 *Show that Algorithms 2.1 and 2.2 yield the same system of linear algebraic equations for piecewise-affine approximations.*

Exercise 2.5 *Verify in detail the inclusions (2.30) for the case that the next one-dimensional mesh $\mathcal{T}_{n+1}$ is obtained by halving all intervals in the current mesh $\mathcal{T}_n$.*

Exercise 2.6 *Consider problem (2.22) with $a_1 = 1$ and $a_0 = 0$ in the interval $\Omega = (-1, 1)$ on equidistant meshes. Take the number of elements $M_n \geq 2$ as an input parameter. Construct the system of linear algebraic equations (2.27) using Algorithm 2.2. Write an appropriate Gauss elimination algorithm for the tridiagonal stiffness matrix S_n. Your output will be the H^1-seminorm $OUT(M_n) = |u_n - u|_{1,2}$, where u_n is the Galerkin approximation and*

$$u(x) = \frac{1 - x^2}{2}.$$

Let $N_n = M_n - 1$ be the number of unknowns. Produce a graph of values $[N_n, OUT(M_n)]$ for $M_1 = 2$, $M_2 = 4$, $M_3 = 8$, $M_4 = 16$, $M_5 = 32$, $M_6 = 64$, $M_7 = 128$, $M_7 = 256$,

$M_8 = 512$, $M_9 = 1024$, $M_{10} = 2048$. *Use both the decimal and decimal-logarithmic scales. Try to read parameters (constants) from an input file and write results into an output data file. What will be the limit of* $OUT(M_n)$ *for* $M_n \to \infty$? *Hint: Use an appropriate theorem that states that, and verify that its assumptions are satisfied.*

Exercise 2.7 *Consider problem (2.22) with* $a_1 = 1$ *and* $a_0 = 0$, *and a right-hand side* $f(x) = 4 - 6x$,

$$-u''(x) = 4 - 6x \quad \textit{in}\ \Omega = (a,b) \subset \mathbb{R}, \tag{2.32}$$

equipped with the boundary conditions

$$u(x) = 0 \quad \textit{on}\ \partial\Omega. \tag{2.33}$$

Suppose that Ω *is covered with a finite element mesh containing* $M \geq 2$ *equally-spaced affine elements.*

1. *First find an exact solution* $u \in C^2(\Omega) \cap C(\overline{\Omega})$ *of (2.32), (2.33). Hint: Perform integration to eliminate the derivatives. Use the boundary conditions to calculate constants that will appear.*
2. *Write the weak formulation and explain why there exists a unique solution.*
3. *Write the discrete problem and explain why there exists a unique solution.*
4. *What is the minimum order of accuracy of numerical quadrature that should be used for the discretization?*
5. *Write a computer code that constructs the stiffness matrix* $\boldsymbol{S}$ *and load vector* $\boldsymbol{F}$, *and that solves the system of linear algebraic equations. The numbers* $a, b \in \mathbb{R}$ *and* M *will be input parameters. The output will be the graph containing both the approximate solution* u_h *and the exact solution* u *(in whatever form you prefer). Other output parameters will be the* H^1*-norm of error,* $OUT_1(M) = \|u - u_h\|_{1,2}$, *and the* H^1*-seminorm of error,* $OUT_2(M) = |u - u_h|_{1,2}$.
6. *Consider input parameters* $a = 0$, $b = 1$. *For* $M = 2, 5, 10, 50, 100$ *produce graphs containing the pair of functions* u, u_h.
7. *Run the code for* $M = 2, 3, 5, 10, 30, 50, 100, 150, 200, 300, 500$ *and produce convergence curves in* H^1*-norm and* H^1*-seminorm (i.e., graphs of values* $[N, OUT_1(M)]$ *and* $[N, OUT_2(M)]$, *respectively, where* $N = M - 1$ *is the number of unknowns).*
8. *Explain why the* H^1*-norm is equivalent to* H^1*-seminorm for problem (2.32), (2.33).*
9. *Guess the algebraic order of convergence of the method, i.e., a positive integer number* α *satisfying*
$$0 < \mathrm{const} = \lim_{M\to\infty} \frac{\|u - u_h\|_{1,2}}{h^\alpha} < \infty,$$
where $h = (b-a)/M$. *Hint: Use the sequence of values* $\|u - u_h\|_{1,2}$ *you have for* $M = 2, 3, 5, 10, 30, 50, 100, 150, 200, 300, 500$. *Construct three sequences corresponding to* $\alpha = 0$, $\alpha = 1$, *and* $\alpha = 2$. *See which one converges to zero, which one diverges and which one converges to a nonzero finite number.*
10. *After validating the code on a simple example with known exact solution (this should become your standard first step whenever implementing a new numerical scheme),*

now you can use it to solve a more difficult problem whose exact solution is not known. Exchange the load function f for

$$\tilde{f}(x) = -\arctan(x - 1/2)\cos(\pi x/2), \quad x \in (0, 1).$$

Present a plot of the numerical solution u_h that is "optically identical" with the unknown exact solution. Hint: Refine the meshes and observe the shape of u_h. Stop when u_h "does not change anymore".

2.3 HIGHER-ORDER NUMERICAL QUADRATURE

The explicit form of the stiffness matrix $\boldsymbol{S}_n$ and the load vector $\boldsymbol{F}_n$, shown in Section 2.2, should not make the reader think that the integrals in the finite element method are calculated on the paper. In reality the load function $f \in L^2(\Omega)$ may be nonpolynomial or even defined via tabulated data. In general, the right-hand side integrals of the form

$$l(v_i) = \int_\Omega f(x)v_i(x)\,\mathrm{d}x$$

cannot be calculated exactly. Usually it is not a bad idea to use a numerical quadrature for the stiffness matrix entries as well. As we will see in a moment, the Gaussian quadrature rules are exact for polynomials up to certain degree that depends on the quality of the quadrature rule. Therefore it is convenient to evaluate numerically even integrals that could be calculated on the paper. One more pro of the numerical quadrature is that when the basis functions in the code change, the values of the integrals are updated automatically. Among the wide scale of existing numerical quadrature methods (see, e.g., [111]) we prefer the Gaussian quadrature rules for their high efficiency.

The derivation and basic properties of these rules are discussed in Paragraph 2.3.1. In Paragraph 2.3.2 we present a few tables with the integration points and weights for practical implementation, and in Paragraph 2.3.3 some approaches to adaptive numerical quadrature are described.

2.3.1 Gaussian quadrature rules

A class of highly efficient quadrature rules was invented by a German mathematician Carl Friedrich Gauss.

C.F. Gauss achieved a large amount of fundamental results in algebra and geometry, number theory, mathematical statistics, approximate integration, differential geometry, geodesy, theoretical astronomy, and other fields. For example, he proved mathematically that the Earth has two different magnetic poles and used the Laplace equation to locate the magnetic south pole.

The k-point Gaussian quadrature rule in the interval $K_a = (-1, 1)$ has the form

$$\int_{-1}^{1} g(\xi)\mathrm{d}\xi \approx \sum_{i=1}^{k} w_{k,i}\, g(\xi_{k,i}), \tag{2.34}$$

where g is a bounded continuous function, $\xi_{k,i} \in (-1, 1), i = 1, 2, \ldots, k$, are the integration points, and $w_{k,i} \in \mathbb{R}$ are the integration weights. The integration weights have to satisfy

Figure 2.4 Carl Friedrich Gauss (1777–1855).

$$\sum_{i=1}^{k} w_{k,i} = 2,$$

so that the rule (2.34) is exact for constants. If the points and weights are chosen carefully, the formula (2.34) can be exact for polynomials up to certain degree $q > 0$. The Gaussian rules are designed to maximize the degree q for a given number of points k:

For some $k \geq 1$ we have k unknown integration points $\xi_{k,i}$ and k unknown integration weights $w_{k,i}$, $i = 1, 2, \ldots, k$. Thus we need $2k$ suitable equations to solve for these unknowns. These equations can be created by inserting, for example, the $2k$ linearly independent monomials $1, x, x^2, \ldots, x^{2k-1}$ into (2.34). This yields a system of nonlinear algebraic equations

$$\begin{aligned}
\sum_{i=1}^{k} w_{k,i} &= \int_{-1}^{1} 1 \mathrm{d}\xi, \\
\sum_{i=1}^{k} w_{k,i}\xi_{k,i}^{1} &= \int_{-1}^{1} \xi^{1} \mathrm{d}\xi, \\
&\vdots \\
\sum_{i=1}^{k} w_{k,i}\xi_{k,i}^{2k-1} &= \int_{-1}^{1} \xi^{2k-1} \mathrm{d}\xi.
\end{aligned} \tag{2.35}$$

After solving (2.35) for the unknown points and weights, the Gaussian integration rule (2.34) is ready. Since it integrates exactly all functions of a basis of the space $P^{2k-1}(-1,1)$, it is easy to see that it is exact for all polynomials in $P^{2k-1}(-1,1)$. We say that the k-point Gaussian quadrature rule has the order of accuracy $2k-1$. For higher k the solution of the nonlinear systems is much easier with the Legendre polynomials $L_0, L_1, \ldots, L_{2k-1}$ than with the monomials $1, x, \ldots, x^{2k-1}$.

Remark 2.4 (Existence and uniqueness of the points and weights) *Since the algebraic system (2.35) is nonlinear, the existence and uniqueness of its solution is not obvious.*

Actually, the nonuniqueness of the integration points and weights is a difficult open problem in the design of Gaussian quadrature rules in 2D and 3D. In one spatial dimension it can be shown that the integration points for the k-point rule (2.34) are the roots of the Legendre polynomial L_k. With k known integration points the nonlinear system (2.35) reduces to a system of k linear algebraic equations for the weights. The analysis leads even further: The weights $w_{k,i}$ for the k-point rule (2.34) have the form

$$w_{k,i} = \frac{2}{(1-\xi_{k,i}^2)L_k'(\xi)^2}, \quad i = 1, \dots, k. \tag{2.36}$$

Quadrature on arbitrary intervals

Let $K = (x_{i-1}, x_i) \subset \mathbb{R}$ be an arbitrary interval. It is easy to calculate the coefficients $c_1, c_2 \in \mathbb{R}$ of an affine map $x_K : K_a \to K$,

$$\begin{aligned} x_K(\xi) &= c_1 + c_2\xi, \\ x_K(-1) &= x_{i-1}, \\ x_K(1) &= x_i. \end{aligned} \tag{2.37}$$

It follows from $x_{i-1} < x_i$ that $c_2 > 0$. The new integration points $\tilde{\xi}_{k,i} \in K$ are then defined as

$$\tilde{\xi}_{k,i} = x_K(\xi_{k,i}), \quad i = 1, 2, \dots, k.$$

The integration weights $\tilde{w}_{k,i}$ are obtained via the Substitution Theorem (see, e.g., [99]),

$$\int_K g(x)\,\mathrm{d}x = \int_{K_a} |J_K(\xi)| (g \circ x_K)(\xi)\,\mathrm{d}\xi, \tag{2.38}$$

which yields $\tilde{w}_{k,i} = J_K w_{k,i} = c_2 w_{k,i}$ (recall that the constant Jacobian J_K of the affine map x_K is positive).

2.3.2 Selected quadrature constants

Let us list the integration points and weights for a few selected k-point Gaussian rules in the reference interval $K_a = (-1, 1)$ in Tables 2.1–2.5. Since the integration points are symmetric with respect to zero, only the positive ones are listed. Symmetric integration points have identical weights, k stands for their total number. Numerous 1D, 2D, and 3D Gaussian quadrature rules up to the order of accuracy $p = 20$ are available on the CD-ROM accompanying [111].

Table 2.1 Gaussian quadrature on K_a, order $2k - 1 = 3$.

Point #	± ξ-Coordinate	Weight
1.	0.57735 02691 89625 76450 91488	1.00000 00000 00000 00000 00000

Table 2.2 Gaussian quadrature on K_a, order $2k - 1 = 5$.

Point #	$\pm$ ξ-Coordinate	Weight
1.	0.00000 00000 00000 00000 00000	0.88888 88888 88888 88888 88889
2.	0.77459 66692 41483 37703 58531	0.55555 55555 55555 55555 55556

Table 2.3 Gaussian quadrature on K_a, order $2k - 1 = 7$.

Point #	$\pm$ ξ-Coordinate	Weight
1.	0.33998 10435 84856 26480 26658	0.65214 51548 62546 14262 69361
2.	0.86113 63115 94052 57522 39465	0.34785 48451 37453 85737 30639

Table 2.4 Gaussian quadrature on K_a, order $2k - 1 = 9$.

Point #	$\pm$ ξ-Coordinate	Weight
1.	0.00000 00000 00000 00000 00000	0.56888 88888 88888 88888 88889
2.	0.53846 93101 05683 09103 63144	0.47862 86704 99366 46804 12915
3.	0.90617 98459 38663 99279 76269	0.23692 68850 56189 08751 42640

Table 2.5 Gaussian quadrature on K_a, order $2k - 1 = 11$.

Point #	$\pm$ ξ-Coordinate	Weight
1.	0.23861 91860 83196 90863 05017	0.46791 39345 72691 04738 98703
2.	0.66120 93864 66264 51366 13996	0.36076 15730 48138 60756 98335
3.	0.93246 95142 03152 02781 23016	0.17132 44923 79170 34504 02961

2.3.3 Adaptive quadrature

In some situations even high-order Gaussian quadrature rules fail or deliver unacceptable errors. This may happen, for example, if the integrated function is discontinuous or oscillates. A possible remedy is to apply some suitable adaptive quadrature algorithm in critical elements. These algorithms usually are not very difficult to implement and can improve the accuracy and reliability of the numerical quadrature significantly.

Let us begin with introducing a basic prototype of an adaptive quadrature algorithm, and perform a few numerical tests. We assume an elementary higher-order Gaussian quadrature procedure

```
double Gauss(double a, double b);
```

that integrates numerically some given function f in the interval (a, b). The following recursive algorithm uses the procedure `Gauss(a,b)` to perform adaptive quadrature. The adaptivity consists in recursive halving of intervals where an error indicator exceeds some given tolerance. The error indicator used is based on the relative difference between the approximation over the whole interval (a, b), and the sum of the approximations in the half-intervals $(a, (a+b)/2)$ and $((a+b)/2, b)$,

$$ERR_{rel} = \frac{\mathtt{Gauss(a,(a+b)/2)} + \mathtt{Gauss((a+b)/2,b)} - \mathtt{Gauss(a,b)}}{\mathtt{Gauss(a,(a+b)/2)} + \mathtt{Gauss((a+b)/2,b)}}.$$

Algorithm 2.3 (Adaptive quadrature in 1D)

```
double ZERO = 1e-12;
double QuadAdapt(double a, double b, double TOL) {
  double L = Gauss(a, 0.5*(a+b));
  double R = Gauss(0.5*(a+b), b);
  double LR = Gauss(a, b);
  if(fabs(L+R) < ZERO) return 0;
  double rel_err = fabs((L+R-LR)/(L+R));
  if(rel_err < TOL) {
    return L + R;
  }
  else {
    L = QuadAdapt(a, 0.5*(a+b), TOL);
    R = QuadAdapt(0.5*(a+b), b, TOL);
    return L + R;
  }
}
```

The adaptive process in (a, b) stops as soon as the approximate integral over (a, b) is sufficiently close to the sum of the approximate integrals over its two subintervals $(a, (a+b)/2)$ and $((a+b)/2, b)$. One can formulate various other stopping criteria. Let us stress, however, that the parameter TOL has no direct relation to the true relative error

$$e_{rel} = \frac{\left|\mathtt{Gauss(a,b)} - \int_a^b f(x)\,\mathrm{d}x\right|}{\left|\int_a^b f(x)\,\mathrm{d}x\right|}.$$

Performance of Algorithm 2.3 is illustrated in the next example.

■ EXAMPLE 2.1 (Adaptive quadrature in 1D)

For testing purposes consider the anisotropically behaved function

$$f(x) = \frac{10}{1 + x^2}, \tag{2.39}$$

defined in the interval $(a_0.b_0) = (0, 10)$. The function f is depicted in Figure 2.5.

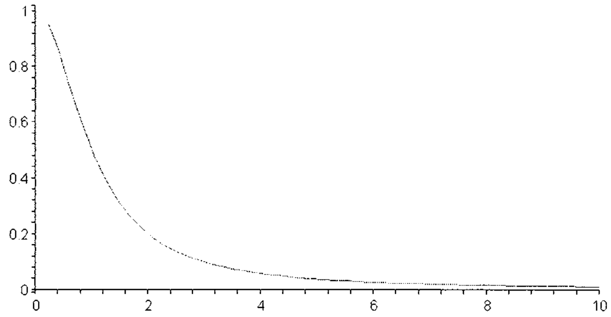


Figure 2.5 Benchmark function f for adaptive numerical quadrature.

The knowledge of the primitive function to f,

$$F(x) = 10 \arctan(x),$$

allows us to evaluate the quadrature error exactly.

First let us investigate the role of the order of accuracy of the elementary quadrature routine Gauss(a,b) on the performance of Algorithm 2.3. Figure 2.6 shows the convergence of the adaptive quadrature when the quadrature routine Gauss(a,b) is third-, fifth- and seventh-order accurate. The horizontal axis represents the final number of integration points in the interval $(a_0.b_0)$, and the vertical axis the true relative error of the approximate quadrature in decimal-logarithmic scale.

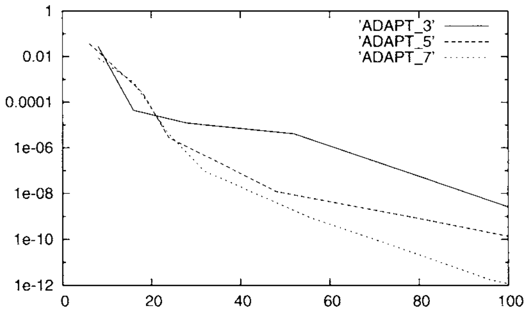


Figure 2.6 Performance of Algorithm 2.3 using the Gaussian quadrature procedure Gauss(a,b) with two, three, and four integration points, respectively.

Next, Figure 2.7 compares the convergence of the adaptive seventh-order Gaussian quadrature to the convergence of a nonadaptive seventh-order Gaussian quadrature scheme based on equidistant subdivisions. One can see that the adaptive procedure performs more efficiently.

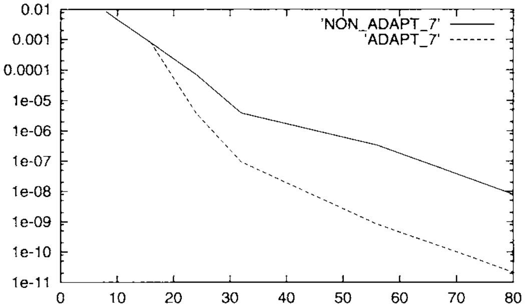


Figure 2.7 Comparison of adaptive and nonadaptive quadrature.

2.3.4 Exercises

Exercise 2.8 *Calculate the coefficients* $c_1, c_2 \in \mathbb{R}$ *of the affine map* x_K *from (2.37).*

Exercise 2.9 *Use Legendre polynomials* $L_0, L_1, \ldots, L_6$ *constructed in Example A.44 to calculate integration points and weights for the Gaussian quadrature rule (2.34) in the interval* $K_a = (-1, 1)$ *for* $k = 3$.

Exercise 2.10 *Let us see the superiority of higher-order Gaussian quadrature rules over the classical trapezoidal rule. Consider, for example, the function* $g(x) = \sin(x)$ *in the interval* $\Omega = (0, \pi)$ *(or some other nonoscillatory continuous function of your choice).*

1. *Calculate the integral* $I_{ex} = \int_0^\pi g(x)\, dx$.

2. *Calculate a series of approximate integrals* I_M *using the trapezoidal rule with equidistant subdivisions of* Ω *into* $M = 2, 5, 10, 20, 50, 100, 200, 500$ *elements. Plot the corresponding convergence curve: Put the number of integration points on the horizontal axis and the error* $|I - I_M|$ *on the vertical one. Use decimal-logarithmic scale.*

3. *Produce an analogous convergence curve for the third-order Gaussian quadrature (Table 2.1). Use equidistant subdivisions with* $M = 1, 2, 5, 10, 50, 100, 250$ *elements.*

4. *At last produce a convergence curve for the fifth-order Gaussian quadrature (Table 2.2). Use equidistant subdivisions with* $M = 1, 2, 5, 10, 50, 100$ *elements.*

5. *Use the convergence curves to compare the efficiency of these three quadrature schemes.*

Exercise 2.11 *Rewrite Algorithm 2.3 in a nonrecursive manner and implement it. Hint: Reserve a sufficiently large array to store the integration subinterval data. Enumerate the integration subintervals at all refinement levels in a suitable unique way (e.g., row-wise in the refinement tree). Link these indices uniquely to positions in the global array. Compare the CPU performance for various values of* $TOL > 0$, *using the function* $f(x)$ *from (2.39).*

2.4 HIGHER-ORDER ELEMENTS

In Section 2.2 we constructed a Galerkin sequence $V_1 \subset V_2 \ldots$ in the Sobolev space V by subdividing selected mesh elements into subelements of the same polynomial degree (h-refinement). Sometimes, much faster convergence can be achieved by increasing the polynomial degree of the elements instead (p-refinement). Such approach usually is more efficient in elements where the solution is very smooth, without singularities, oscillations, or boundary/internal layers. An illustrative example is given in the next paragraph.

2.4.1 Motivation problem

In this paragraph we compare the performance of two simple finite element schemes with (a) two piecewise-affine elements and (b) one quadratic element. Consider the Poisson equation

$$-u''(x) = f(x) \qquad \text{in } \Omega = (-1, 1), \tag{2.40}$$

where $f(x) = \pi^2 \cos(\pi x/2)/4$, equipped with homogeneous Dirichlet boundary conditions. The weak formulation of problem (2.40) reads: Find $u \in V = H_0^1(-1, 1)$ such that

$$\int_{-1}^{1} u'(x)v'(x)\,\mathrm{d}x = \int_{-1}^{1} f(x)v(x)\,\mathrm{d}x \qquad \text{for all } v \in V. \tag{2.41}$$

The exact solution to (2.40) [and (2.41)], has the form

$$u(x) = \cos\left(\frac{\pi x}{2}\right).$$

First let us cover Ω with a pair of affine elements $(-1, 0)$ and $(0, 1)$. The corresponding finite element space V_h is generated by a single piecewise-affine function v_h, defined as $v_h(x) = x + 1$ in $(-1, 0]$ and $v_h(x) = 1 - x$ in $[0, 1)$. The approximate solution $u_h \in V_h$ has the form $u_h(x) = y_1 v_h(x)$, where y_1 is an unknown coefficient. After substituting u_h for u and v_h for v in (2.41), we obtain a single linear algebraic equation for y_1 whose solution is $y_1 = 1$. The functions u and u_h are shown in Figure 2.8.

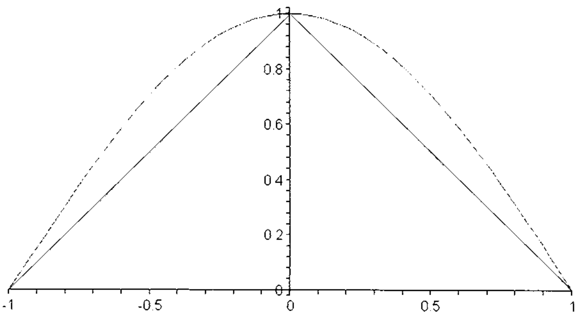


Figure 2.8 Exact solution u and piecewise-affine approximation u_h.

It is left to the reader as an exercise to verify that the approximation error in H^1-seminorm (which by the Poincaré–Friedrich's inequality is equivalent to the full H^1-norm in the space V) is

$$|u - u_h|_{1,2} = \left(\int_{-1}^{1} [u'(x) - u_h'(x)]^2 \, \mathrm{d}x \right)^{\frac{1}{2}} \approx 0.683667.$$

Next assume a single quadratic element $(-1, 1)$. We can choose, for example, the function $v_p(x) = 1 - x^2$ to be the basis of the corresponding finite element space V_p. The approximate solution has the form $u_p(x) = \tilde{y}_1 v_p(x)$. After substituting u_p for u and v_p for v in (2.41), we calculate that $\tilde{y}_1 = 3/\pi$. The functions u and u_p are depicted in Figure 2.9.

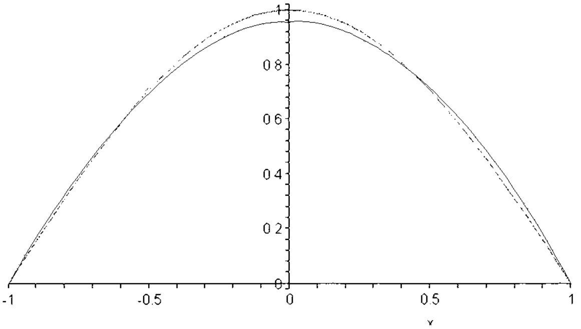


Figure 2.9 Exact solution u and quadratic approximation u_p.

The approximation error $|u - u_p|_{1,2} \approx 0.20275$ is less than 30% of $|u - u_h|_{1,2}$. The next step is left to the reader as an exercise: Use (a) four equally-long piecewise-affine elements and (b) one quartic ($p = 4$) element. The number of unknowns in each case is three. The error in the quartic case is less than 2.5% of the error of the piecewise-affine approximation.

This indicates that smooth functions are better approximated by means of large higher-order elements. On the other hand, less regular functions can be approximated more efficiently on smaller piecewise low-degree elements. The ultimately best Galerkin sequences $V_1 \subset V_2 \ldots \subset V$ can be obtained by combining appropriately the spatial subdivision of elements with the selection of suitable polynomial degrees in the subelements (hp-adaptivity). See, e.g., [111] and the references therein.

2.4.2 Affine concept: reference domain and reference maps

The affine concept of finite elements is closely related to the element-by-element assembling procedure. It is particularly suitable for higher-order finite element discretizations. The basic idea is to define a single set of shape functions on some suitable reference domain, say, $K_a = (-1, 1)$. For each element K_i in the mesh we define an affine reference map $x_{K_i} : K_a \to K_i$ (Paragraph 2.3.2), and use it to transfer the shape functions from K_a to K_i. In this way one obtains the desired finite element basis in the physical mesh.

In addition, the weak formulation is transformed from K_i to K_a via the maps x_{K_i}, and in the end all computational work is done on the reference domain. This approach also is efficient from the point of view of computer memory, since the numerical quadrature data are stored on the reference domain only. The shape functions and their partial derivatives can be stored via their values at integration points in the reference domain.

Model problem Let us stay with the model problem (1.26), (1.28) in a bounded interval $\Omega = (a, b) \subset \mathbb{R}$. Recall the weak formulation from Paragraph 1.2.1: We seek a function $u \in V = H_0^1(\Omega)$, such that

$$a(u, v) = l(v) \quad \text{for all } v \in V, \tag{2.42}$$

where

$$a(u, v) = \int_\Omega a_1 \nabla u \cdot \nabla v + a_0 uv \, \mathrm{d}x = \int_\Omega a_1 u'(x)v'(x) + a_0 u(x)v(x) \, \mathrm{d}x,$$

and

$$l(v) = \langle l, v \rangle = \int_\Omega fv \, \mathrm{d}x,$$

$f \in L^2(\Omega)$. The coefficient functions $a_1, a_0 \in L^\infty(\Omega)$, $a_1(x) \geq C_{min} > 0$ and $a_0(x) \geq 0$ a.e. in Ω, are assumed constant.

Finite element space Let the interval Ω be covered with a mesh $\mathcal{T}_{h,p} = \{K_1, K_2, \ldots, K_M\}$ where the elements K_m carry arbitrary polynomial degrees $1 \leq p_m$, $m = 1, 2, \ldots, M$. For each element $K_m = (x_{m-1}, x_m)$, $m = 1, 2, \ldots, M$ we define an affine reference map (2.37) of the form

$$x_{K_m}(\xi) = c_1^{(m)} + c_2^{(m)}\xi, \tag{2.43}$$

where

$$c_1^{(m)} = \frac{x_{m-1} + x_m}{2}, \quad c_2^{(m)} = J_{K_m} = \frac{x_m - x_{m-1}}{2}.$$

The space $V_{h,p}$ has the form

$$V_{h,p} = \{v \in V;\ v|_{K_m} \in P^{p_m}(K_m) \text{ for all } m = 1, 2, \ldots, M\}, \tag{2.44}$$

or, equivalently,

$$V_{h,p} = \{v \in V;\ v|_{K_m} \circ x_{K_m} \in P^{p_m}(K_a) \text{ for all } m = 1, 2, \ldots, M\}. \tag{2.45}$$

Here $(f \circ g)(x) \equiv f(g(x))$. The dimension of the space $V_{h,p}$ is

$$N = \dim(V_{h,p}) = \underbrace{M - 1}_{\text{first-order part}} + \underbrace{\sum_{m=1}^{M} (p_m - 1)}_{\text{higher-order part}} = -1 + \sum_{m=1}^{M} p_m. \tag{2.46}$$

Discrete problem The discrete problem (2.5) reads: Find a function $u_{h,p} \in V_{h,p}$, such that

$$a(u_{h,p}, v_{h,p}) = l(v_{hp}) \quad \text{for all } v_{h,p} \in V_{h,p}.$$

Consider some basis $\{v_1, v_2, \ldots, v_N\} \subset V_{h,p}$ (a concrete basis will be presented in Paragraph 2.4.7). When expressing as usual

$$u_{h,p} = \sum_{j=1}^{N} y_j v_j,$$

we obtain

$$\sum_{j=1}^{n} y_j a(v_j, v_i) = l(v_i) \qquad i = 1, 2, \ldots, N, \tag{2.47}$$

or

$$\boldsymbol{S}\boldsymbol{Y} = \boldsymbol{F}. \tag{2.48}$$

For future reference let us rewrite (2.47) into a sum over all elements K_m, $m = 1, 2, \ldots, M$,

$$\sum_{m=1}^{M}\sum_{j=1}^{N} y_j \int_{K_m} a_1 v_j'(x) v_i'(x) + a_0 v_j(x) v_i(x)\,\mathrm{d}x = \sum_{m=1}^{M} \int_{K_m} f(x) v_i(x)\,\mathrm{d}x, \tag{2.49}$$

$i = 1, 2, \ldots, N$.

2.4.3 Transformation of weak forms to the reference domain

Next let us transform the integrals in the weak formulation (2.49) from the mesh elements $K_m \in \mathcal{T}_{h,p}$ to the reference domain $K_a = (-1, 1)$, using the reference maps (2.43):

Transformation of function values The transformation of the approximate solution $u_{h,p}$ is simple:

$$\tilde{u}_{h,p}^{(m)}(\xi) \equiv (u_{h,p} \circ x_{K_m})(\xi) = u_{h,p}(x_{K_m}(\xi)). \tag{2.50}$$

Transformation of derivatives One has to be more careful when transforming derivatives. The chain rule yields

$$[\tilde{u}_{h,p}^{(m)}(\xi)]' = (u_{h,p} \circ x_{K_m})'(\xi) = u_{h,p}'(x)|_{x = x_{K_m}(\xi)} J_{K_m}(\xi). \tag{2.51}$$

This means that

$$u_{h,p}'(x) = \frac{1}{J_{K_m}(\xi)} [\tilde{u}_{h,p}^{(m)}]'(\xi), \tag{2.52}$$

i.e., the derivative at a reference point $\xi \in K_a$ is obtained by dividing the derivative of $u_{h,p}$ at its image $x = x_{K_m}(\xi) \in K_m$ by the constant Jacobian $0 \neq J_{K_m}$.

Transformation of integrals from (2.49) to K_a The test functions $v_{h,p}$ and their derivatives are transformed in the same way. Using the Substitution Theorem, it is easy to conclude that

$$\begin{aligned} &\int_{K_m} a_1 u_{h,p}'(x) v_{h,p}'(x) + a_0 u_{h,p}(x) v_{h,p}(x)\,\mathrm{d}x \\ &= \int_{K_a} \frac{a_1}{J_{K_m}} [\tilde{u}_{h,p}^{(m)}]'(\xi) [\tilde{v}_{h,p}^{(m)}]'(\xi) + a_0 J_{K_m} [\tilde{u}_{h,p}^{(m)}](\xi) [\tilde{v}_{h,p}^{(m)}](\xi)\,\mathrm{d}\xi, \end{aligned} \tag{2.53}$$

for all $m = 1, 2, \ldots, M$. The right-hand side transforms as

$$\int_{K_m} f(x) v_{h,p}(x)\,\mathrm{d}x = \int_{K_a} J_{K_m} \tilde{f}^{(m)}(\xi) \tilde{v}_{h,p}^{(m)}(\xi)\,\mathrm{d}\xi, \tag{2.54}$$

where $\tilde{f}^{(m)}(\xi) = (f \circ x_{K_m})(\xi)$.

2.4.4 Higher-order Lagrange nodal shape functions

The construction of suitable shape functions [basis in the polynomial space $P^{p_m}(K_a)$ on the reference domain] matters: With a wrong choice the linear system $\boldsymbol{SY} = \boldsymbol{F}$ will pose a serious problem to both iterative and direct matrix solvers. This issue will be discussed in more detail in Section 2.5. Now we will introduce the nodal and hierarchic approaches to the construction of suitable basis functions. Let us begin with the nodal basis, which is based on the idea of the Lagrange interpolation.

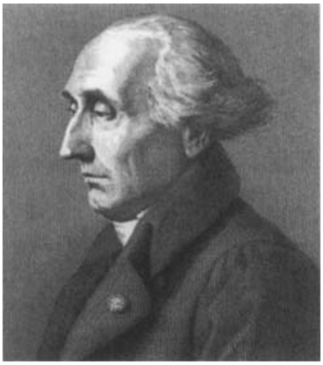


Figure 2.10 Joseph–Louis Lagrange (1736–1813).

J.–L. Lagrange was a French mathematician who was largely self-taught. In spite of that, he influenced numerous fields of mathematics. His work covers a variety of topics including algebra, number theory, mathematical probability, theoretical astronomy, and others. It is assumed that one of his greatest contributions is his transformation of mechanics into a mathematical framework based on differential equations.

The $p_m + 1$ Lagrange nodal shape functions $\theta_1, \theta_2, \ldots, \theta_{p_m+1} \in P^{p_m}(K_a)$ are associated with an equal number of pairwise-distinct nodal points,

$$-1 = y_1 < y_2 < \ldots < y_{p_m+1} = 1, \tag{2.55}$$

via the standard Lagrange interpolation condition

$$\theta_j(y_k) = \delta_{jk}. \tag{2.56}$$

Exploiting the Lagrange interpolation polynomial (A.75), condition (2.56) yields the explicit formulae of the Lagrange nodal shape functions,

$$\theta_i(\xi) = \prod_{1 \le j \le p_m+1, j \ne i} \frac{(\xi - y_j)}{(y_i - y_j)}, \qquad i = 1, 2, \ldots, p_m + 1. \tag{2.57}$$

Obviously all of these functions are polynomials of the degree p_m. In particular, for piecewise-affine approximations ($p_m = 1$) the nodal points $y_1 = -1$ and $y_2 = 1$ yield the pair of affine shape functions

$$\theta_1(\xi) = \frac{1 - \xi}{2}, \quad \theta_2(\xi) = \frac{\xi + 1}{2}. \tag{2.58}$$

Our first idea might be to distribute the nodal points in $K_a = (-1, 1)$ equidistantly. However, the equidistant points are known to be notoriously bad from both the conditioning and interpolation points of view. In practice we prefer more sophisticated point sets.

2.4.5 Chebyshev and Gauss–Lobatto nodal points

Among the best known points for the construction of higher-order nodal elements are the Chebyshev and Gauss–Lobatto points.

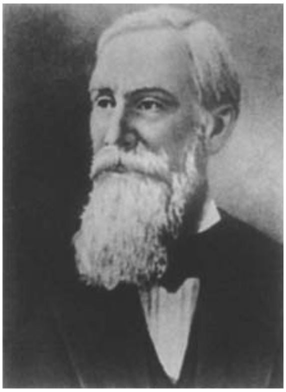


Figure 2.11 Pafnuty Lvovich Chebyshev (1821–1894).

P.L. Chebyshev was a Russian mathematician who made famous contributions to the analysis of the Taylor series, number theory, theory of integrals, mathematical probability, and other fields of mathematics. He introduced his polynomials in 1854 and developed a general theory of orthogonal polynomials. He is assumed to be one of the founders of modern approximation theory.

For a polynomial degree $p > 1$, the $p + 1$ Chebyshev points in the reference interval K_a are defined by

$$y_j = \cos\left(\frac{\pi(j-1)}{p}\right), \quad j = 1, 2, \ldots, p+1. \tag{2.59}$$

The Gauss–Lobatto points are the roots of the function

$$(1 - x^2)L_p'(x), \tag{2.60}$$

where $L_p(x)$ is the pth Legendre polynomial. There is no explicit formula for these points, but they have been tabulated (see, e.g., the companion CD-ROM accompanying [111]). Figure 2.12 shows that the Gauss–Lobatto and Chebyshev points are very similar. The numerical practice confirms that also the properties of the corresponding Lagrange nodal shape functions are analogous. This issue will be addressed in more detail in Paragraph 2.5.3.

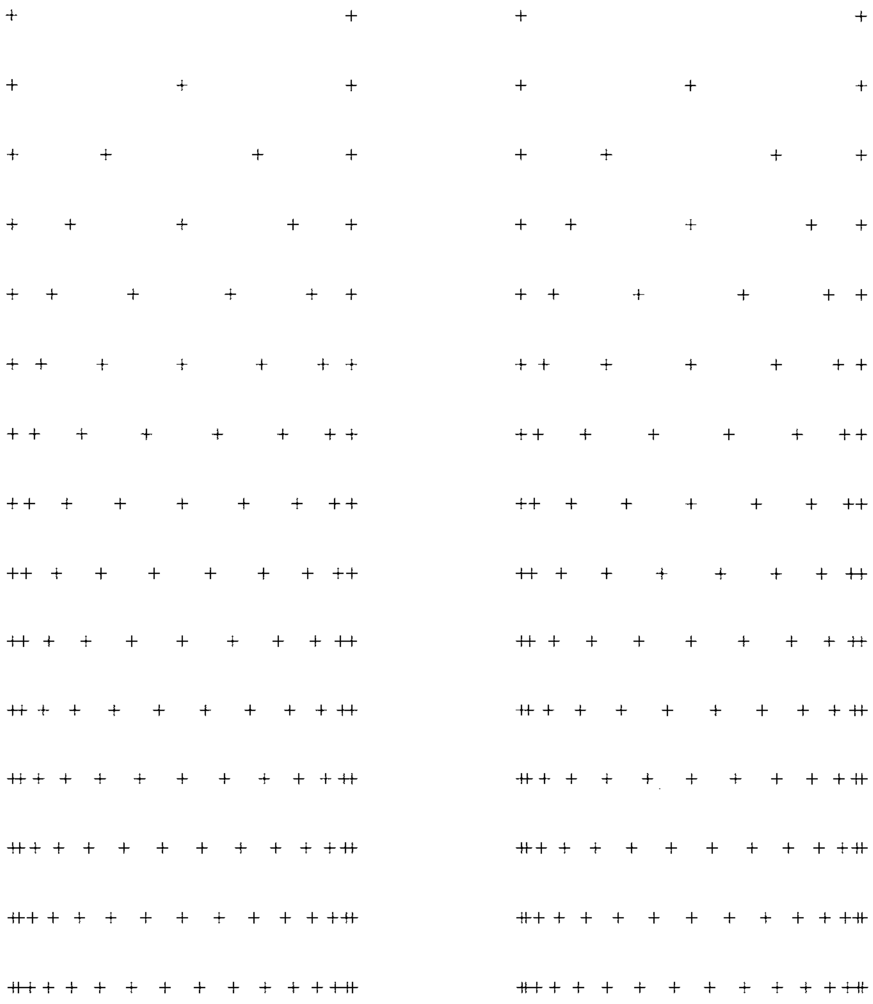


Figure 2.12 The Gauss–Lobatto (left) and Chebyshev points (right) for $p = 1, 2, \ldots, 15$.

Let us show a few examples of higher-order Lagrange nodal shape functions built on the Gauss–Lobatto points. These functions are presented in the quadratic, cubic, quartic and quintic cases in Figures 2.13–2.16. Shape functions associated with vertices are called vertex functions, and remaining shape functions (that vanish at ± 1) are said to be bubble functions.

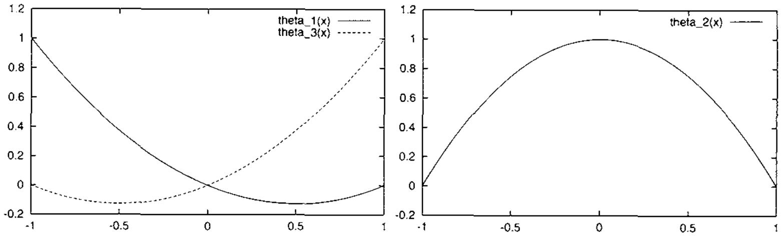


Figure 2.13 Quadratic Lagrange–Gauss–Lobatto shape functions; vertex functions θ_1, θ_3 (left) and the bubble function θ_2 (right).

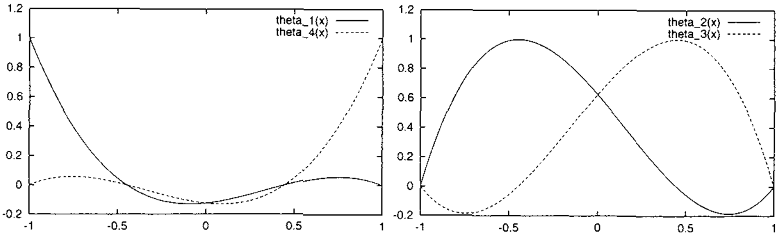


Figure 2.14 Cubic Lagrange–Gauss–Lobatto nodal shape functions; vertex functions θ_1, θ_4 (left) and bubble functions θ_2, θ_3 (right).

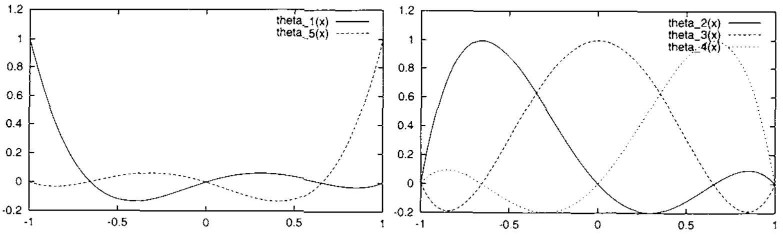


Figure 2.15 Quartic Lagrange–Gauss–Lobatto nodal shape functions; vertex functions θ_1, θ_5 (left) and bubble functions $\theta_2, \theta_3, \theta_4$ (right).

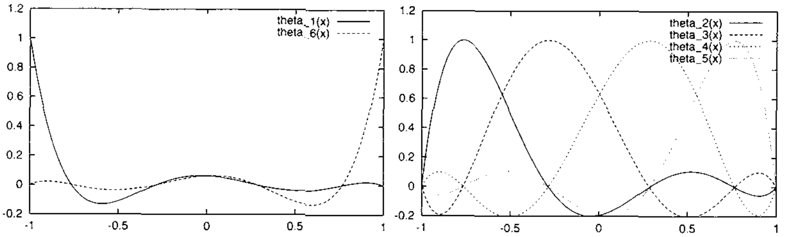


Figure 2.16 Quintic Lagrange–Gauss–Lobatto nodal shape functions; vertex functions θ_1, θ_6 (left) and bubble functions $\theta_2, \theta_3, \ldots, \theta_5$ (right).

2.4.6 Higher-order Lobatto hierarchic shape functions

An alternative way of constructing a suitable basis of the polynomial space $P^{p_m}(K_a)$ is to use hierarchic shape functions. The idea of the hierarchic approach is as follows: When a set of shape functions

$$\mathcal{B}_{p_m} = \{\theta_1, \theta_2, \dots, \theta_{p_m}\}$$

forms a basis of the polynomial space $P^{p_m}(K_a)$, the basis of the next space $P^{p_m+1}(K_a)$ is defined by adding a new shape function to the basis $\mathcal{B}_{p_m}$,

$$\mathcal{B}_{p_m+1} = \mathcal{B}_{p_m} \cup \{\theta_{p_m+1}\}.$$

The lowest-order basis, $\mathcal{B}_1 = \{(1-\xi)/2, (1+\xi)/2\}$, is identical to the Lagrange basis (2.58) for piecewise-affine approximations.

Currently, among the best known hierarchic shape functions for elliptic problems in 1D are the Lobatto polynomials,

$$\begin{aligned} l_0(\xi) &= \frac{1-\xi}{2}, \quad l_1(\xi) = \frac{1+\xi}{2}, \\ l_k(\xi) &= \int_{-1}^{\xi} L_{k-1}(\zeta)\,\mathrm{d}\zeta, \quad 2 \le k. \end{aligned} \tag{2.61}$$

Here, L_k are the (normalized) Legendre polynomials, $\|L_k\|_{L^2(-1,1)} = 1$ for all $k \geq 0$, that are constructed in Example A.44. It is easy to see that $l_k(-1) = 0$, $k = 2, 3, \dots$. The orthogonality of the Legendre polynomials further yields that

$$l_k(1) = \int_{-1}^{1} L_{k-1}(\xi)\,\mathrm{d}\xi = \int_{-1}^{1} L_{k-1}(\xi)L_0(\xi)\,\mathrm{d}\xi = 0 \qquad \text{for all } 2 \le k. \tag{2.62}$$

Evidently the functions $l_0, l_1, l_2, \dots, l_{p_m}$ constitute a basis in the space $P^{p_m}(K_a)$. Their optimality for the discretization of the Laplace operator follows from their orthonormality in the H_0^1-inner product $(u, v) = \int_{-1}^{1} u'(x)v'(x)\,\mathrm{d}x$. More about this will be said in Paragraph 2.5.3.

Several Lobatto hierarchic shape functions are shown below for reference, and they are depicted in Figures 2.17–2.21:

$$\begin{aligned} l_2(\xi) &= \frac{1}{2}\sqrt{\frac{3}{2}}(\xi^2 - 1), \\ l_3(\xi) &= \frac{1}{2}\sqrt{\frac{5}{2}}(\xi^2 - 1)\xi, \\ l_4(\xi) &= \frac{1}{8}\sqrt{\frac{7}{2}}(\xi^2 - 1)(5\xi^2 - 1), \\ l_5(\xi) &= \frac{1}{8}\sqrt{\frac{9}{2}}(\xi^2 - 1)(7\xi^2 - 3)\xi, \\ l_6(\xi) &= \frac{1}{16}\sqrt{\frac{11}{2}}(\xi^2 - 1)(21\xi^4 - 14\xi^2 + 1), \\ l_7(\xi) &= \frac{1}{16}\sqrt{\frac{13}{2}}(\xi^2 - 1)(33\xi^4 - 30\xi^2 + 5)\xi, \\ l_8(\xi) &= \frac{1}{128}\sqrt{\frac{15}{2}}(\xi^2 - 1)(429\xi^6 - 495\xi^4 + 135\xi^2 - 5), \end{aligned} \tag{2.63}$$

$$\begin{aligned} l_9(\xi) &= \frac{1}{128}\sqrt{\frac{17}{2}}(\xi^2-1)(715\xi^6-1001\xi^4+385\xi^2-35)\xi, \\ l_{10}(\xi) &= \frac{1}{256}\sqrt{\frac{19}{2}}(\xi^2-1)(2431\xi^8-4004\xi^6+2002\xi^4-308\xi^2+7). \end{aligned}$$

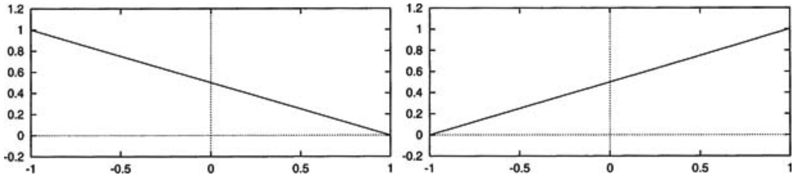


Figure 2.17 Lowest-order Lobatto shape functions l_0, l_1.

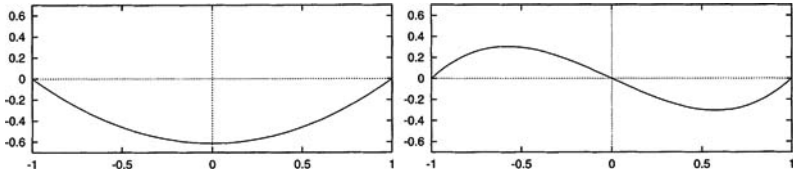


Figure 2.18 H_0^1-orthonormal (Lobatto) hierarchic shape functions l_2, l_3.

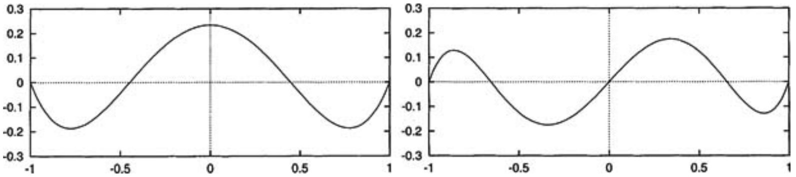


Figure 2.19 H_0^1-orthonormal (Lobatto) hierarchic shape functions l_4, l_5.

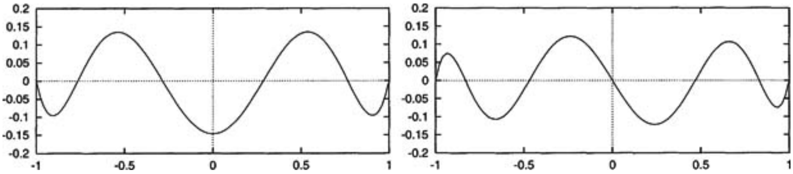


Figure 2.20 H_0^1-orthonormal (Lobatto) hierarchic shape functions l_6, l_7.

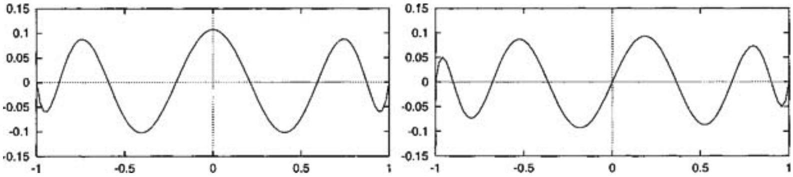


Figure 2.21 H_0^1-orthonormal (Lobatto) hierarchic shape functions l_8, l_9.

2.4.7 Constructing basis of the space $V_{h,p}$

With the shape functions and reference maps in hand, we can define the basis functions of the space $V_{h,p}$:

Lagrange nodal basis Let us begin with the Lagrange basis functions. In this case it is customary to assume a uniform polynomial degree p in all elements. As indicated in Paragraph 2.4.4, the Lagrange shape functions can be split into the vertex and bubble functions. The basis functions of the space $V_{h,p}$ inherit the same structure. A vertex basis function v_i represents the value at the grid vertex x_i, $1 \le i \le M - 1$, and it is zero in Ω except for the elements adjacent to x_i:

$$v_i(x) = \begin{cases} (\theta_{p+1} \circ x_{K_i}^{-1})(x), & x \in K_i, \\ (\theta_1 \circ x_{K_{i+1}}^{-1})(x), & x \in K_{i+1}. \end{cases} \tag{2.64}$$

The reader does not have to worry about the inverse reference maps in (2.64), since in the element-by-element procedure they are never evaluated explicitly. This will be explained in detail in Paragraph 2.4.9.

Notice that for $p = 1$ relation (2.64) yields the "hat functions" (2.25). An example of a quadratic Lagrange nodal vertex basis function is shown in Figure 2.22.

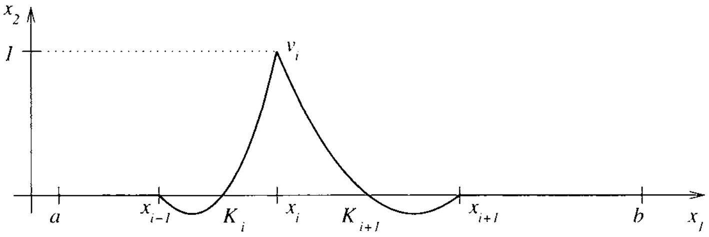


Figure 2.22 A quadratic Lagrange nodal vertex basis function.

The bubble functions are local to element interiors. There are $p-1$ bubble basis functions per every element $K_m \in \mathcal{T}_{h,p}$, defined as

$$(\theta_2 \circ x_{K_m}^{-1})(x),\ (\theta_3 \circ x_{K_m}^{-1})(x), \ldots, (\theta_p \circ x_{K_m}^{-1})(x). \tag{2.65}$$

It is easy to verify that the $M - 1$ vertex functions (2.64) together with the $\sum_{m=1}^{M}(p_m - 1)$ bubble functions (2.65) constitute a basis of the space $V_{h,p}$ defined in (2.44).

Lobatto hierarchic basis In this case we allow for different polynomial degrees $1 \le p_m = p(K_m)$ in the mesh, $1 \le m \le M$. For each interior grid vertex x_i there is one vertex function v_i, which is identical to the piecewise-affine Lagrange vertex function (independently of p_m),

$$v_i(x) = \begin{cases} (l_1 \circ x_{K_i}^{-1})(x), & x \in K_i, \\ (l_0 \circ x_{K_{i+1}}^{-1})(x), & x \in K_{i+1}. \end{cases} \tag{2.66}$$

The quadratic and higher-order hierarchic shape functions are bubble functions, given by

$$(l_2 \circ x_{K_m}^{-1})(x),\ (l_3 \circ x_{K_m}^{-1})(x), \dots, (l_{p_m} \circ x_{K_m}^{-1})(x). \tag{2.67}$$

Also in this case it is easy to verify that the functions (2.66) and (2.67) together constitute a basis of the space $V_{h,p}$.

2.4.8 Data structures

The rest of this section is devoted to the implementation of higher-order finite elements in one spatial dimension. As the reader expects, the implementation of the nodal and hierarchic elements is done in different ways. We choose the hierarchic case for illustration.

To begin with, recall the model problem (2.20) with homogeneous Dirichlet boundary conditions. In this case $V = H_0^1(\Omega)$, and the approximate solution $u_{h,p}$ is sought in the space $V_{h,p} \subset V$ of continuous, piecewise polynomial functions (2.44),

$$V_{h,p} = \{v \in V;\ v|_{K_m} \in P^{p_m}(K_m) \text{ for all } m = 1, 2, \dots, M\}.$$

The dimension of $V_{h,p}$, which at the same time is the number of unknowns, was calculated in (2.46),

$$N = \dim(V_{h,p}) = -1 + \sum_{i=1}^{M} p_m.$$

Some remarks are in order before we introduce concrete data structures and algorithms. Generally, data structures differ from implementation to implementation. A safe way to avoid criticism for the complicatedness or inoptimality of one's data structures and algorithms is not to expose them. On the other hand, the presentation of the data structures and algorithms may be of considerable help to beginners. Therefore let us try to be concrete, without claiming that our data structures or algorithms are optimal.

Element data structure Choose a reasonable upper bound MAXP for the highest polynomial degree in the mesh $\mathcal{T}_{h,p}$. A basic Element data structure can be defined as follows:

```
struct {
  int p;               //polynomial degree of element
  int vert_dir[2];     //vertex Dirichlet flags
  int vert_dof[2];     //vertex connectivity array
  int *bubb_dof;       //bubble connectivity array (length MAXP-1)
  ...
} Element;
```

This amount of information per element is superfluous. However, let us keep a data structure that can most naturally be extended into two and three spatial dimensions. The Dirichlet flags Elem[m].vert_dir[j], $j = 1, 2$, have the following meaning: Elem[m].vert_dir[1] = 0 if the left vertex of $K_m = (x_{m-1}, x_m)$ is unconstrained by a Dirichlet boundary condition, and Elem[m].vert_dir[1] = 1 otherwise. The flag Elem[m].vert_dir[2] is related to the right vertex of K_m in the same way.

Unique enumeration of shape and basis functions The element-by-element assembling algorithm relies on the vertex and bubble connectivity arrays vert_dof and bubb_dof, that for every element $K_m \in \mathcal{T}_{h,p}$ link the global indices $1, 2, \dots, N$ of all basis functions of the space $V_{h,p}$, whose support includes K_m, to the local indices $1, 2, \ldots, p_m+1$ of the corresponding shape functions on the reference domain K_a.

First one has to enumerate the N basis functions of the space $V_{h,p}$ in a unique way. For the sake of compatibility with piecewise-affine approximations, it is reasonable to first enumerate all vertex functions analogously to the lowest-order element case in Paragraph 2.2.3. After that, higher-order basis functions can be enumerated in an element-by-element fashion, always from quadratic to the highest degree p_m on the element K_m. In the Lagrange nodal case, where all bubble functions have the same polynomial degree, it is natural to sort them according to the ordering of the nodal points.

Element connectivity arrays The values of the Dirichlet lift $G(x)$ at the endpoints of $\Omega = (a, b)$, only nonzero in the case of nonhomogeneous Dirichlet boundary conditions, are stored in a global array `double DIR_BC_ARRAY[2]` $= \{G(a), G(b)\}$. The variable `Elem[m].vert_dof[1]` contains either

- a positive index i of a vertex basis function v_i of the space $V_{h,p}$ associated with the left vertex of the element K_m (if the vertex is unconstrained, i.e., `Elem[m].vert_dir[1] == 0`),
- or `-1`, so that $G(a) =$ `DIR_BC_ARRAY[-Elem[m].vert_dof[1]]` (if `Elem[m].vert_dir[1] == 1`).

Analogously one defines `Elem[m].vert_dof[2]` for the right vertex of the element K_m. If `Elem[m].vert_dir[2] == 1`, then `Elem[m].vert_dof[2] == -2`. The bubble functions are always unconstrained, and the value `Elem[m].bubb_dof[j], j = 1,2,...,Elem[m].p-1`, contains the index of the bubble basis function of the polynomial degree $j + 1$ associated with the element K_m.

The construction of the connectivity arrays always represents a considerable part of the total programming work. In two dimensions these are the Algorithms 4.1, 4.3, 4.4 and 4.5. In one dimension the connectivity algorithm may look as follows:

Algorithm 2.4 (Enumeration of DOF)

```
count := 1;
//Visiting vertex basis functions on the element $K_1$:
if (Elem[1].vert_dir[1] == 1) then Elem[1].vert_dof[1] := -1;
else {
  Elem[1].vert_dof[1] := count;
  count := count + 1;
}
Elem[1].vert_dof[2] := count;
//Visiting vertex basis functions on interior elements $K_2, K_3, \ldots, K_{M-1}$:
for m = 2,3,...,M-1 do {
  Elem[m].vert_dof[1] := count;
  count := count + 1;
  Elem[m].vert_dof[2] := count;
}
//Visiting vertex basis functions on the element $K_M$:
Elem[M].vert_dof[1] := count;
count := count + 1;
if (Elem[M].vert_dir[2] == 1) then {
  Elem[M].vert_dof[2] := -2;
}
else {
  Elem[M].vert_dof[2] := count;
  count := count + 1;
}
//Visiting bubble basis functions on all elements:
for m = 1,2,...,M do {
  for j = 1,2,...,Elem[m].p-1 do {
    Elem[m].bubb_dof[j] := count;
    count := count + 1;
 }
}
```

More about the implementation of nonhomogeneous boundary conditions will be said in Paragraph 2.6.

■ EXAMPLE 2.2

Consider a mesh $\mathcal{T}_{h,p}$ consisting of three elements K_1, K_2, and K_3 of the polynomial degrees $p_1 = 3$, $p_2 = 4$ and $p_3 = 2$, and the model problem (2.20) with homogeneous Dirichlet boundary conditions. In this case the connectivity Algorithm 2.4 obtains the following input data:

```
Elem[1].p = 3;
Elem[1].vert_dir = {1,0};
Elem[2].p = 4;
Elem[2].vert_dir = {0,0};
Elem[3].p = 2;
Elem[3].vert_dir = {0,1};
```

The resulting element connectivity arrays have the form

```
Elem[1].vert_dof = {-1,1};
Elem[1].bubb_dof = {3,4};
Elem[2].vert_dof = {1,2};
Elem[2].bubb_dof = {5,6,7};
Elem[3].vert_dof = {2,-2};
Elem[3].bubb_dof = {8};
```

Next let us present the assembling procedure.

2.4.9 Assembling algorithm

In the following we distinguish between two situations:

1. The differential operator L in the equation $Lu = f$ does not explicitly depend on space or time. This is the case when all coefficients a_{ij}, b_i, c_i and a_0 in (1.4) are constant. For example, the operators

$$Lu = -\Delta u,\ Lu = -\Delta u + \frac{\partial u}{\partial x},\ Lu = -\Delta u + \frac{\partial u}{\partial x} + u$$

belong to this category, and so does the general operator L in (2.20) if $a_1 > 0$ and $a_0 \geq 0$ are constant.

2. The differential operator L does explicitly depend on space or time, as, for example, the operators

$$Lu = \frac{-\Delta u}{1+x^2} + u,\ Lu = -\Delta u + (e^{-t^2})\frac{\partial u}{\partial x},\ Lu = -\Delta u + \sin(x)u.$$

In the former case it is possible to avoid repeated numerical integration on every element and assemble the global stiffness matrix S efficiently by means of precomputed prototype integrals calculated on the reference domain K_a. The integrals present in the weak formulation of a concrete problem determine which constants have to be precomputed. For example, problem (2.20) with constant coefficients requires the $L^2(K_a)$-products of the first derivatives of the shape functions (master element stiffness integrals. MESI) and, if $a_0 \neq 0$, then also the $L^2(K_a)$-products of the shape functions themselves (master element mass integrals, MEMI), In one dimension these constants can be organized in the form of square matrices.

If we denote the maximum polynomial degree in the mesh by p_{max} and consider some set of shape functions $\varphi_1, \varphi_2, \dots, \varphi_{p_{max}+1} \in P^{p_{max}}(K_a)$, the master element stiffness matrix S_{K_a} of problem (2.20) has the form

$$S_{K_a} = \{\hat{s}_{ij}\}_{i,j=1}^{p_{max}+1} = \left\{ \int_{K_a} \varphi_i'(\xi)\varphi_j'(\xi)\,\mathrm{d}\xi \right\}_{i,j=1}^{p_{max}+1}. \tag{2.68}$$

The master element mass matrix M_{K_a} is defined as

$$M_{K_a} = \{\hat{m}_{ij}\}_{i,j=1}^{p_{max}+1} = \left\{ \int_{K_a} \varphi_i(\xi)\varphi_j(\xi)\,\mathrm{d}\xi \right\}_{i,j=1}^{p_{max}+1}. \tag{2.69}$$

The only information about the reference map x_{K_m} that is needed on every element $K_m \in \mathcal{T}_{h,p}$ in the assembling algorithm is its Jacobian. Therefore, for each element K_m we introduce one more constant, `Elem[m].jac` $:= |J_{K_m}|$. The assembling procedure for model problem (2.20) with homogeneous Dirichlet boundary conditions can be written as follows.

Algorithm 2.5 (Assembling algorithm)

```
//Calculate the dimension of the space $V_{h,p}$:
N := -1;
for m = 1,2,...,M do N := N + Elem[m].p;
  //Calculate the master element stiffness integrals MESI:
  //(Use sufficiently accurate Gaussian quadrature to obtain exact results)
  for i = 1,2,...,MAXP+1 do {
    for j= 1,2,...,MAXP+1 do {
      MESI[i][j] := $\int_{-1}^{1} \varphi_i'(x)\varphi_j'(x)\,\mathrm{d}x$;
    }
  }
  //Calculate the master element mass integrals MEMI:
  for i = 1,2,...,MAXP+1 do {
    for j= 1,2,...,MAXP+1 do {
      MEMI[i][j] := $\int_{-1}^{1} \varphi_i(x)\varphi_j(x)\,\mathrm{d}x$;
    }
  }
  //Calculate the value of Elem[m].jac for all elements $K_m$, $m = 1,2,\dots,M$:
  for m = 1,2,...,M do Elem[m].jac := $(x_m - x_{m-1})/2$;
  //Set the stiffness matrix $S$ zero:
  for i = 1,2,...,N do for j = 1,2,...,N do S[i][j] := 0;
  //Set the right-hand side vector $F$ zero:
  for i = 1,2,...,N do F[i] := 0;
  //Element loop:
  for m = 1,2,...,M do {
      //Loop over vertex test functions:
      for i = 1,2 do {
        //If $>$ -1, this is index of a test function $v_{m_1} \in V_{h,p}$,
        //i.e., row position in $S$:
        m1 := Elem[m].vert_dof[i];
        //Loop over vertex basis functions:
        if (m1 > -1) then for j = 1,2 do {
        //If $>$ -1, this is index of a basis function $v_{m_2} \in V_{h,p}$,
        //i.e., column position in $S$:
        m2 := Elem[m].vert_dof[j];
        if (m2 > -1) then
          S[m1][m2] := S[m1][m2] + a1*MESI[i][j]/Elem[m].jac
          + a0*Elem[m].jac*MEMI[i][j];
      } //End of inner loop over vertex functions
      //Loop over bubble basis functions:
      for j = 1,2,...,Elem[m].p-1 do {
        m2 := Elem[m].bubb_dof[j];
        if (m2 > -1) then
          S[m1][m2] := S[m1][m2] + a1*MESI[i][j+2]/Elem[m].jac
          + a0*Elem[m].jac*MEMI[i][j+2];
      } //End of inner loop over bubble functions
      //Contribution of the vertex test function $v_{m_1}$
      //to the right-hand side $F$:
      if (m1 > -1) then F[m1] := F[m1] + $\int_{K_a} |J_{K_m}|\tilde{f}^{(m)}(\xi)\varphi_i(\xi)\,\mathrm{d}\xi$;
    } //End of outer loop over vertex functions
    //Loop over bubble test functions:
    for i = 1,2,...,Elem[m].p-1 do {
      m1 := Elem[m].bubb_dof[i];
      //Loop over vertex basis functions:
      if (m1 > -1) then for j = 1,2 do {
        m2 := Elem[m].vert_dof[j];
        if (m2 > -1) then
          S[m1][m2] := S[m1][m2] + a1*MESI[i+2][j]/Elem[m].jac
          + a0*Elem[m].jac*MEMI[i+2][j];
      } //End of inner loop over vertex functions
      //Loop over bubble basis functions:
      if (m1 > -1) then for j = 1,2,...,Elem[m].p-1 do {
        m2 := Elem[m].bubb_dof[j];
        if (m2 > -1) then
          S[m1][m2] := S[m1][m2] + a1*MESI[i+2][j+2]/Elem[m].jac
          + a0*Elem[m].jac*MEMI[i+2][j+2];
      } //End of inner loop over bubble functions
      //Contribution of the bubble test function $v_{m_1}$
      //to the right-hand side $F$:
      if (m1 > -1) then F[m1] := F[m1] + $\int_{K_a} |J_{K_m}|\tilde{f}^{(m)}(\xi)\varphi_i(\xi)\,\mathrm{d}\xi$;
    } //End of outer loop over bubble functions
  } //End of element loop
```

In Algorithm 2.5 we used the notation $\tilde{f}^{(m)}(\xi) = f(x_{K_m}(\xi))$. If the operator L is space- or time-dependent (for example, if the coefficient functions a_1 and a_0 in the model problem (2.20) are not constant), the precomputed MESI and MEMI arrays cannot be used. Instead, appropriate numerical quadrature must be performed each time the MESI or MEMI arrays in Algorithm 2.5 are accessed.

Efficient implementation of Algorithm 2.5 For the sake of transparency, significant portion of Algorithm 2.5 (the application of a given test function to all vertex and bubble basis functions) was repeated two times with minor changes. This part of the code can be moved to a separate subroutine. Moreover, it is not necessary to store the full Elem[m].bubb_dof array of the length `Elem[m].p-1`, since according to the enumeration of the bubble shape functions (Paragraph 2.4.8) it holds

```
Elem[m].bubb_dof[2] = Elem[m].bubb_dof[1] + 1;
Elem[m].bubb_dof[3] = Elem[m].bubb_dof[1] + 2;
...
Elem[m].bubb_dof[Elem[m].p] = Elem[m].bubb_dof[1] + Elem[m].p-1;
```

2.4.10 Exercises

Exercise 2.12 *Verify in detail the inclusions $V_1 \subset V_{2,h} \subset V$ and $V_1 \subset V_{2,p} \subset V$ for the spaces defined in Paragraph 2.4.1.*

Exercise 2.13 *Consider the Poisson problem $-u'' = f$, $f \in L^2(\Omega)$, in a bounded interval $\Omega = (a,b) \subset \mathbb{R}$. Suppose that Ω is covered with a finite element mesh $\mathcal{T}_{h,p}$ containing $M \geq 2$ elements of polynomial degrees $1 \leq p_1, p_2, \ldots, p_M$. Consider (A) homogeneous Dirichlet boundary conditions on $\partial\Omega$, (B) nonhomogeneous Dirichlet boundary conditions on $\partial\Omega$, (C) a nonhomogeneous Dirichlet boundary condition at a and a Neumann boundary condition at b.*

1. *Write the weak formulation of these problems.*

2. *Use the Lax–Milgram lemma to show that in each case there exists a unique solution.*

3. *Write the discrete problems.*

4. *How many unknowns has the discrete problem in each case?*

Exercise 2.14 *Consider the Helmholtz equation $-u'' + u = f$, $f \in L^2(\Omega)$, with homogeneous Dirichlet boundary conditions $u(a) = u(b) = 0$ in a bounded interval $\Omega = (a,b) \subset \mathbb{R}$. Let $a < x_{i-1} < x_i < b$ be a pair of neighboring grid points, $K_m = (x_{i-1}, x_i)$ and $K_a = (-1,1)$.*

1. *Write the weak formulation of this problem.*

2. *Write the affine map $x_{K_m} : K_a \to (x_{i-1}, x_i)$.*

3. *Transform the weak formulation from the interval K_m to the reference interval K_a.*

Exercise 2.15 *Consider the reference interval $K_a = (-1,1)$ and $p = 4$.*

1. *Write explicit formulae for the Lobatto hierarchic shape functions $l_0, l_1, \ldots, l_4$.*

2. *Consider equidistant nodal points $-1 = y_1 < \ldots < y_5 = 1$. Write the Lagrange nodal shape functions $\theta_1, \theta_2, \ldots, \theta_5$ such that $\theta_i(y_j) = \delta_{ij}$, $1 \leq i,j \leq 5$.*

3. *Write master element stiffness matrices $S^{(h)}_{K_a}$ and $S^{(n)}_{K_a}$ for the Poisson equation, corresponding to the above two sets of shape functions.*

4. *In each case calculate the condition number of the 3×3 block corresponding to bubble functions (use, e.g., Matlab).*

5. *Which condition number is greater and what is the implication for the performance of iterative matrix solvers?*

Exercise 2.16 *Again consider the reference interval $K_a = (-1, 1)$, polynomial degree p and $p+1$ distinct nodal points $-1 = y_1 < \ldots < y_{p+1} = 1$. Show that all Lagrange nodal shape functions $\theta_1, \theta_2, \ldots, \theta_{p+1}$ from (2.57) must be polynomials of the degree at least p.*

Exercise 2.17 *Consider the problem from Exercise 2.7 with the load function $f(x) = 4-6x$ and an equidistant mesh $\mathcal{T}_{h,p}$ with M quadratic elements.*

1. *Write formulae for the affine reference maps $x_{K_m} : K_a \rightarrow K_m$.*

2. *Transform the weak formulation to the reference domain $K_a = (-1, 1)$. Write the integrals explicitly.*

3. *Perform a suitable unique enumeration of the basis functions and write the element connectivity arrays.*

4. *Write the 3×3 master element stiffness matrix (2.68) for the Lobatto hierarchic shape functions l_0, l_1, l_2.*

5. *Implement a finite element discretization using Algorithm 2.5.*

6. *Produce plots of u and u_h for $M = 2, 5, 10, 50$.*

7. *Consider $M = 2, 3, 5, 10, 30, 50, 100, 150, 200, 300$ and produce convergence curve in H^1-seminorm (be careful to put the correct number of unknowns on the horizontal axis).*

8. *Compare with the H^1-seminorm curve for piecewise-affine approximation from Exercise 2.7. Was the piecewise-affine or the piecewise-quadratic scheme more efficient? Why?*

9. *Again guess the algebraic order of convergence α of the method. Compare it with the value of α obtained in Exercise 2.7.*

Exercise 2.18 *Extend your code from Exercise 2.17 to finite elements of arbitrary polynomial degrees $1 \le p_m = p(K_m) \le 5$, $i = 1, 2, \ldots, M$.*

1. *Read the polynomial degrees $p_m = p(K_m)$ together with all other input parameters from an input data file.*

2. *Write the 6×6 master element stiffness matrix S_{K_a} for the Lobatto hierarchic shape functions $l_0, l_1, \ldots, l_5$.*

3. *When evaluating integrals of polynomial expressions, make sure to use Gaussian quadrature data of an appropriate order of accuracy.*

4. *Calculate the exact solution for the cubic load function $f = -50x(1-x)^2$.*

5. *Present results of suitable convergence tests proving that the code works correctly.*

2.5 THE SPARSE STIFFNESS MATRIX

As we mentioned in Paragraph 2.2.3, the finite element method prefers basis functions with small and possibly nonoverlapping supports. Then almost all entries in the stiffness matrix S are zero, which is convenient for the computation. Matrices with this property are said to be sparse. The question of efficient storage and operation with large sparse matrices is essential.

With $N = 100,000$ unknowns, which is a moderate number in practical applications, a full $N \times N$ stiffness matrix S in double precision arithmetics would consume 80 GB of computer memory. Hence, disregarding the well-known fact that the Gaussian elimination procedure is unstable on large systems, the storage argument alone calls for a much more economical treatment.

There is extensive literature on the numerical solution of sparse systems of linear algebraic equations (see, e.g., [18] and the references therein), and vast resources of concrete program packages are available on the web. Most of the solvers are sufficiently robust and user friendly, so that the reader can use them without any problems after fitting their more or less standard input format.

2.5.1 Compressed sparse row (CSR) data format

One of the most frequently used data formats for sparse matrices is the Compressed Sparse Row (CSR) format. Let N be the dimension of the stiffness matrix S and by NNZ denote the number of nonzero entries in S. The CSR representation of S consists of three arrays:

1. Array A of length NNZ: This is a real-valued array containing all nonzero entries of the matrix S listed from the left to the right, starting with the first and ending with the last row.
2. Array IA of length $N + 1$: This is an integer array, $IA[1] = 1$. $IA[k + 1] = IA[k] + nnz_k$. where nnz_k is the number of nonzero entries in the kth row.
3. Array JA of length NNZ: This is an integer array containing the row positions of all entries of array A.

Sometimes one uses an analogous Compressed Sparse Column (CSC) sparse matrix format.

2.5.2 Condition number

The reader knows from Paragraph 2.1.1 that every symmetric V-elliptic bilinear form $a(\cdot, \cdot) : V \times V \to \mathbb{R}$ leads to a symmetric positive definite stiffness matrix S. All eigenvalues are then positive real numbers (see, e.g., [100]). It is well known that iterative solvers perform better on matrices where the ratio of the largest and smallest eigenvalue $\lambda_{max}/\lambda_{min}$ is close to one – such matrices are called well-conditioned. Figure 2.23 illustrates the convergence history of a standard iterative matrix solver (an incompletely LU-preconditioned conjugate gradient method) on two matrices of the same size and sparsity structure, but different condition numbers.

Before introducing the condition number of a nonsingular matrix in Definition 2.2, let us define the spectrum and spectral radius:

Definition 2.1 (Spectrum, spectral radius) *Let M be a square matrix. By $\sigma(M)$ we denote the* spectrum *(set of all eigenvalues) of the matrix M. The* spectral radius *$\rho(M)$ is defined as*

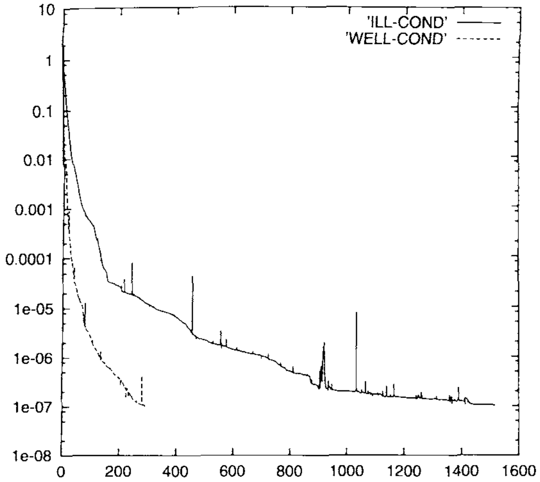


Figure 2.23 Performance of an iterative matrix solver on two differently conditioned matrices of the same size and sparsity structure. The horizontal axis represents the number of iterations and the vertical one shows the norm of the residuum of the approximate solution.

$$\rho(\boldsymbol{M}) = \max_{\lambda\in\sigma(\boldsymbol{M})} |\lambda|.$$

Definition 2.2 (Condition number) *Let $\boldsymbol{M}$ be a nonsingular $n \times n$ matrix. The product*

$$\kappa(\boldsymbol{M}) = \|\boldsymbol{M}\|\|\boldsymbol{M}^{-1}\|,$$

where $\|\,.\,\|$ is some matrix norm, is called condition number *of the matrix $\boldsymbol{M}$ (with respect to the norm $\|\,.\,\|$).*

One may use, for example, the standard Frobenius norm

$$\|\boldsymbol{M}\| = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}^2},$$

or the spectral norm

$$\|\boldsymbol{M}\|_* = \max_{\boldsymbol{x}\neq 0} \frac{\|\boldsymbol{M}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \sqrt{\rho(\boldsymbol{M}\boldsymbol{M}^T)},$$

where $\|\boldsymbol{M}\boldsymbol{x}\|$ is the Euclidean norm in $\mathbb{R}^n$. The spectral (Todd) condition number

$$\kappa^*(\boldsymbol{M}) = \|\boldsymbol{M}\|_*\|\boldsymbol{M}^{-1}\|_* = \frac{\max_{\lambda\in\sigma(\boldsymbol{M})} |\lambda|}{\min_{\lambda\in\sigma(\boldsymbol{M})} |\lambda|} \tag{2.70}$$

has the minimum property

$$1 \leq \kappa^*(\boldsymbol{M}) \leq \kappa(\boldsymbol{M}),$$

where $\kappa(\boldsymbol{M})$ is a condition number induced by any other matrix norm.

Clearly, for every symmetric positive definite matrix $\boldsymbol{S}$, the spectral condition number $\kappa(\boldsymbol{S}) = \kappa^*(\boldsymbol{S})$ can be written as

$$1 \leq \kappa(\boldsymbol{S}) = \frac{\lambda_{max}}{\lambda_{min}}.$$

The following aspects influence the condition number of the stiffness matrix $\boldsymbol{S}$ significantly:

1. the discretized differential operator,

2. quality of the mesh,

3. the set of shape functions.

In practice the differential operator is given, and the mesh can be optimized outside of the finite element solver. Therefore let us look at the last aspect in more detail.

2.5.3 Conditioning of shape functions

The simplest comparison of the quality of different sets of higher-order shape functions can be done using a one-element mesh, equipped with appropriate boundary conditions so that the discrete problem has a unique solution. Such test, of course, does not cover the influence of the geometrical structure of the entire mesh, but still the results usually provide a valuable information.

For model problem (2.20) let us consider a one-element mesh $K_a = (-1, 1)$ equipped with homogeneous Dirichlet boundary conditions. The corresponding stiffness matrix $\boldsymbol{S}_0$ is obtained by leaving out of the master element stiffness matrix $\boldsymbol{S}_{K_a}$ all rows and columns corresponding to the vertex shape functions. The mass matrix $\boldsymbol{M}_0$ is obtained analogously from the master element mass matrix $\boldsymbol{M}_{K_a}$. The next example compares the quality of the Lagrange nodal and Lobatto hierarchic shape functions.

■ **EXAMPLE 2.3 (Comparison of Lagrange and Lobatto shape functions)**

Figures 2.24 and 2.25 show the condition number of the stiffness and mass matrices for the Lagrange nodal shape functions on the equidistant, Gauss–Lobatto and Chebyshev nodal points, and for the Lobatto hierarchic shape functions. The horizontal axis represents the polynomial degree $p = 2, 3, \ldots, 10$.

The Lagrange nodal shape functions on equidistant points cause an exponential growth of the condition number of both the stiffness and mass matrices, which indicates that these shape functions should be avoided. It is clear from Figure 2.25 that the Chebyshev and Gauss–Lobatto points are a better choice for Lagrange nodal elements. The Lobatto hierarchic bubble functions perform best: they are orthogonal in the H_0^1-product, which makes them optimal for the discretization of the Laplace operator in one dimension.

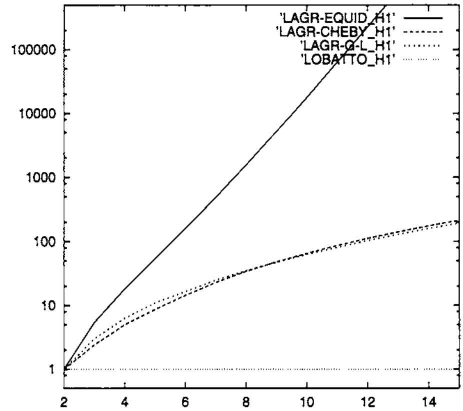


Figure 2.24 Conditioning of various types of shape functions in the $H_0^1(K_a)$-inner product (condition number of the matrix $\boldsymbol{S}_0$).

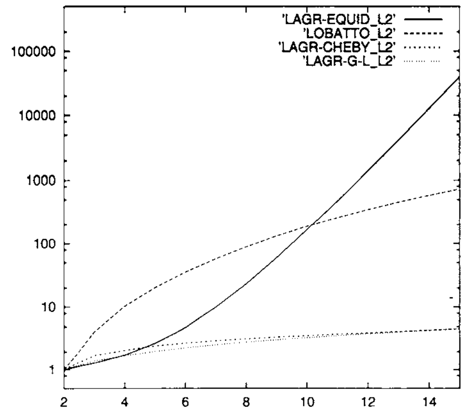


Figure 2.25 Conditioning of various types of shape functions in the $L^2(K_a)$-inner product (condition number of the matrix $\boldsymbol{M}_0$).

Regarding the more complex model problem (2.22), the Lobatto hierarchic shape functions will perform well as long as $a_0 << a_1$. Otherwise their worse conditioning in the L^2-product becomes important, and for $a_0 >> a_1$ the Lagrange shape functions on the Gauss–Lobatto and Chebyshev points may yield a better-conditioned discrete problem. Let us close this paragraph with a lemma that is useful for practical implementation:

Lemma 2.4 *The spectral condition number of a symmetric stiffness matrix $\boldsymbol{S}$ does not depend on the enumeration of the basis functions of the space $V_{h,p}$.*

Proof: Consider a permutation that exchanges the indices of a pair of basis functions v_k and v_l. It follows from Definition A.17 that the new stiffness matrix $\tilde{\boldsymbol{S}}$ has the same set of eigenvalues. The new eigenvectors are obtained from the original ones by exchanging their kth and lth components. ■

2.5.4 Stiffness matrix for the Lobatto shape functions

Let us have a closer look at the the sparsity structure of the stiffness matrix S obtained in the discretization of the Laplace operator by means of the Lobatto hierarchic shape functions. It follows from the L^2-orthogonality of the Legendre polynomials that

$$\int_{-1}^{1} l_i'(x) l_j'(x)\,\mathrm{d}x = \int_{-1}^{1} L_{i-1}(x) L_{j-1}(x)\,\mathrm{d}x = \delta_{ij}, \quad \text{for all } 2 \le i, j. \tag{2.71}$$

Moreover, we have

$$\int_{-1}^{1} l_0'(x) l_j'(x)\,\mathrm{d}x = \int_{-1}^{1} l_1'(x) l_j'(x)\,\mathrm{d}x = 0, \quad \text{for all } 2 \le j. \tag{2.72}$$

Therefore the master element stiffness matrix S_{K_a} itself is sparse,

$$S_{K_a} = \begin{pmatrix} 1/2 & -1/2 & 0 & 0 & \dots & 0 \\ -1/2 & 1/2 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & & & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{pmatrix}. \tag{2.73}$$

Due to (2.73), the global stiffness matrix S corresponding to the Lobatto hierarchic shape functions has a particularly nice block-diagonal sparse structure shown in Figure 2.26.

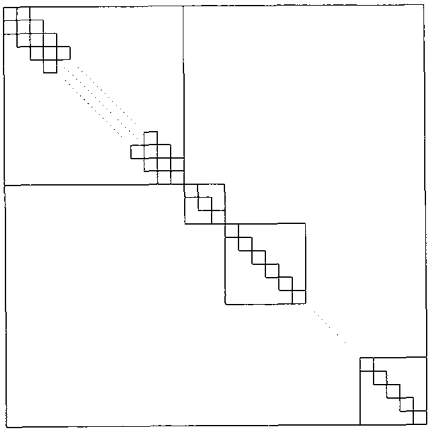


Figure 2.26 Sparsity structure of the stiffness matrix for the Laplace operator discretized by means of the Lobatto hierarchic shape functions.

The number of blocks is $M+1$, where M is the number of elements in the mesh $\mathcal{T}_{h,p}$. The $(M-1)\times(M-1)$ block in the upper left corner corresponds to the piecewise-affine basis functions $v_1, v_2, \dots, v_{M-1}$ – this block is identical to the tridiagonal stiffness matrix (2.28) corresponding to the piecewise-affine case (Paragraph 2.2.4). The remaining M diagonal blocks of the type $(p_1-1)\times(p_1-1), (p_2-1)\times(p_2-1), \ldots, (p_M-1)\times(p_M-1)$ correspond to higher-order bubble basis functions associated with each element $K_1, K_2, \ldots, K_M$, respectively. This structure is given by the enumeration of the basis functions of the space $V_{h,p}$ (see Paragraph 2.4.8).

If the Lobatto hierarchic shape functions were replaced with the Lagrange or other nonorthogonal shape functions, additional nonzero off-diagonal entries would appear in the stiffness matrix, and its condition number would rise.

2.5.5 Exercises

Exercise 2.19 *Show that for every symmetric positive definite matrix S the spectral condition number (2.70) satisfies*

$$\kappa^*(S) = \|S\|_*\|S^{-1}\|_* = \frac{\lambda_{max}}{\lambda_{min}}.$$

Exercise 2.20 *Consider a nonsingular $N \times N$ matrix S, and a matrix $\overline{S}$ obtained by switching the kth and lth row and the kth and lth column in S, $1 \le k, l \le N$, $k \neq l$. Show that the matrices S and $\overline{S}$ have the same set of eigenvalues.*

Exercise 2.21 *Use the result of Exercise 2.20 to prove that the condition number of the (symmetric positive definite) stiffness matrix S, obtained from the discretization of a V-elliptic operator L, does not depend on the enumeration of the basis functions of the space $V_{h,p}$.*

Exercise 2.22 *Write a computer code that turns a sparse matrix represented as an array $S[i][j]$, $1 \le i,j \le N$, into the CSR sparse matrix format. The numbers $1 \le N, NNZ$ are input parameters. Assume that exactly NNZ entries in the array $S[\cdot][\cdot]$ are nonzero.*

2.6 IMPLEMENTING NONHOMOGENEOUS BOUNDARY CONDITIONS

The implementation of various types of boundary conditions closely follows the discussion in Paragraphs 1.2.5, 1.2.6, and 1.2.7. Let us begin with the nonhomogeneous Dirichlet case.

2.6.1 Dirichlet boundary conditions

According to Paragraph 1.2.5, any problem with nonhomogeneous Dirichlet boundary conditions can be treated as a homogeneous Dirichlet problem with an adjusted right-hand side. Let us stay with the model equation (2.20),

$$-\nabla \cdot (a_1 \nabla u) + a_0 u = -(a_1 u')' + a_0 u = f,$$

$a_1 > 0$, $a_0 \ge 0$, $f \in L^2(\Omega)$, in a bounded domain $\Omega = (a,b) \subset \mathbb{R}$, but consider the nonhomogeneous Dirichlet boundary conditions

$$\begin{aligned} u(a) &= g_a, \\ u(b) &= g_b, \end{aligned} \tag{2.74}$$

where $g_a, g_b \in \mathbb{R}$. Recall that the solution u is sought in the form

$$u = U + G, \tag{2.75}$$

where $G \in H^1(\Omega)$ is a Dirichlet lift such that $G(a) = g_a$, $G(b) = g_b$, and the new unknown function $U \in V = H_0^1(\Omega)$. The task is to find a function $U \in V$ satisfying the weak formulation (1.47),

$$a(U, v) = l(v) \quad \text{for all } v \in V \tag{2.76}$$

with

$$\begin{aligned} a(U, v) &= \int_\Omega a_1 U'(x)v'(x) + a_0 U(x)v(x)\,\mathrm{d}x, \quad U, v \in V, \\ l(v) &= \int_\Omega f(x)v(x) - a_1 G'(x)v'(x) - a_0 G(x)v(x)\,\mathrm{d}x, \quad v \in V. \end{aligned} \tag{2.77}$$

Choice of the Dirichlet lift When using the Lobatto hierarchic elements, define G as a continuous piecewise-affine function that vanishes in all interior elements (Figure 2.27). In the case of the Lagrange nodal elements choose, for example, a piecewise pth-degree polynomial function G that vanishes in all interior elements, and that in the elements K_1 and K_M coincides with the appropriate Lagrange functions $g_a\theta_1 \circ x_{K_1}^{-1}$ and $g_b\theta_{p+1} \circ x_{K_M}^{-1}$, respectively.

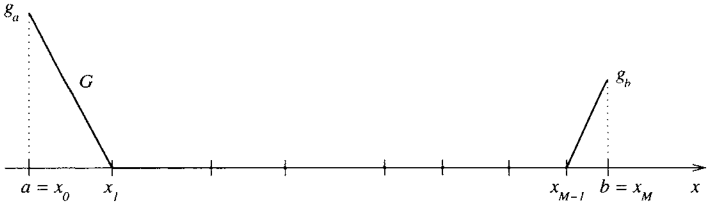


Figure 2.27 Typical piecewise-affine Dirichlet lift G.

Implementation When using the Lobatto hierarchic shape functions, the Dirichlet lift G transforms from the mesh element K_m to the reference interval K_a as follows:

$$(G \circ x_{K_m})(\xi) = \begin{cases} 0, & 2 \le m \le M - 1, \\ g_a l_0(\xi), & m = 1, \\ g_b l_1(\xi), & m = M. \end{cases} \tag{2.78}$$

The case of the Lagrange nodal shape functions is analogous,

$$(G \circ x_{K_m})(\xi) = \begin{cases} 0, & 2 \le m \le M - 1, \\ g_a \theta_1(\xi), & m = 1, \\ g_b \theta_{p+1}(\xi), & m = M. \end{cases}$$

The values of the Dirichlet lift G at the endpoints of $\Omega = (a, b)$ may be stored as described in Paragraph 2.4.8,

$$\begin{aligned}\mathrm{DIR\_BDY\_ARRAY}[1] &:= g_a;\\ \mathrm{DIR\_BDY\_ARRAY}[2] &:= g_b;\end{aligned}$$

Algorithm 2.5 needs to be changed as follows: Whenever a contribution to the stiffness matrix $\boldsymbol{S}$ is made, a new contribution to the load vector $\boldsymbol{F}$ appears. For example, the portion of the code

```
...
//Loop over vertex basis functions:
if (m1 > -1) then for j = 1,2 do {
  m2 := Elem[m].vert_dof[j];
  if (m2 > -1) then
    S[m1][m2] := S[m1][m2] + MESI[i][j]/Elem[m].jac;
    + a0*Elem[m].jac*MEMI[i][j];
} //End of inner loop over vertex functions
...
```

needs to be changed to

```
...
//Loop over vertex basis functions:
if (m1 > -1) then for j = 1,2 do {
  m2 := Elem[m].vert_dof[j];
  if (m2 > -1) then
    S[m1][m2] := S[m1][m2] + MESI[i][j]/Elem[m].jac
    + a0*Elem[m].jac*MEMI[i][j];
  else
    F[m1] := F[m1] - DIR_BDY_ARRAY[-m2]*a1*MESI[i][j]/Elem[m].jac
    - DIR_BDY_ARRAY[-m2]*a0*Elem[m].jac*MEMI[i][j];
} //End of inner loop over vertex functions
...
```

and so on. The stiffness matrix $\boldsymbol{S}$ is the same as with homogeneous Dirichlet boundary conditions.

2.6.2 Combination of essential and natural conditions

Since the incorporation of Neumann or Newton boundary conditions occurs exactly as described in Paragraphs 1.2.6 and 1.2.7, let us discuss in more detail the case when essential and natural boundary conditions are combined. Consider the model equation (2.20) in a bounded domain $\Omega = (a, b) \subset \mathbb{R}$ with the boundary conditions

$$\begin{aligned}\frac{\partial u}{\partial \nu}(a) = -u'(a) &= g_a,\\ u(b) &= g_b,\end{aligned} \tag{2.79}$$

where $g_a, g_b \in \mathbb{R}$. The solution u is sought in the form $u = U + G$, where $G \in H^1(\Omega)$ is a Dirichlet lift satisfying $G(b) = g_b$, and the new unknown function U lies in the space V defined in (1.65),

$$V = \{v \in H^1(\Omega);\ v(b) = 0\}.$$

The weak formulation (1.66) then reads

$$a(U, v) = l(v) \quad \text{for all } v \in V,$$

where

$$\begin{aligned} a(u, v) &= \int_\Omega a_1 U'(x)v'(x) + a_0 U(x)v(x)\,\mathrm{d}x, \quad U, v \in V, \\ l(v) &= \int_\Omega f(x)v(x) - a_1 G'(x)v'(x) - a_0 G(x)v(x)\,\mathrm{d}x + g_a(a)v(a), \quad v \in V. \end{aligned}$$

The Dirichlet lift G is defined analogously to the case with nonhomogeneous Dirichlet boundary conditions, but now it vanishes also at the endpoint where the natural boundary condition is prescribed (Figure 2.28).

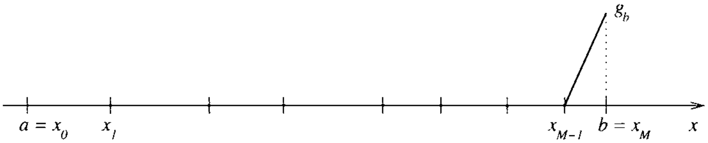


Figure 2.28 Dirichlet lift for combined boundary conditions (2.79).

2.6.3 Exercises

Exercise 2.23 *Extend the code from Exercise 2.18 to nonhomogeneous Dirichlet boundary conditions*

$$\begin{aligned} u(a) &= g_a, \\ u(b) &= g_b, \end{aligned}$$

where $g_a, g_b \in \mathbb{R}$ are additional input parameters.

1. *For the new boundary conditions recalculate the exact solution u of the problem $-u'' = f$ using the cubic load function f from Exercise 2.18.*

2. *Choose $a = 0$, $b = 1$, $g_a = 1/2$, $g_b = 1$.*

3. *For $M = 10$ elements which are (A) linear, (B) quadratic, (C) cubic, (D) fourth-order, (E) fifth-order, produce plots of the error $e_{h,p} = u - u_{h,p}$. Plot all the curves together in one figure using decimal-logarithmic scale.*

Exercise 2.24 *Extend your code from Exercise 2.23 to nonequidistant meshes.*

1. *Read the number of elements M together with the coordinates of the grid points $a = x_0 < x_1 < \ldots < x_M = b$ and the polynomial degrees $p_1, p_2, \ldots, p_M$ together with the other input parameters from an input data file.*

2. *Verify that the code is correct.*

2.7 INTERPOLATION ON FINITE ELEMENTS

Assume some restricted set of functions C (such as, for example, polynomial, piecewise-polynomial or trigonometric polynomial functions) in a linear space V and a function $g \in V$ that does not belong to C. The prototype approximation problem is to find a suitable function $g_c \in C$ (approximation of g) such that g_c is in some sense close to g. The measure of the quality of the approximation (abstract distance of g_c from g), can be defined as an error estimate, the norm $\|g - g_c\|_V$ if the space V is normed, or it can be defined otherwise. By best approximation one means an approximation that minimizes this distance.

Approximation becomes interpolation when the sought function $g_c \in C$ has to satisfy some additional constraints. These conditions are formulated generally as

$$L_i(g_c) = b_i, \quad i = 1, 2, \ldots, N_c, \tag{2.80}$$

where $L_i : V \rightarrow \mathbb{R}$ are linearly independent linear forms in V' and $b_1, b_2, \ldots, b_{N_c}$ some given constants.

For example, in the traditional Lagrange interpolation one requires the approximation g_c to coincide with the original function g at some points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{N_c} \in \Omega$ via the choice

$$L_i(g_c) = g_c(\boldsymbol{x}_i),$$

and defining the constants b_i in (2.80) as

$$b_i = g(\boldsymbol{x}_i).$$

There are many natural questions related to the approximation and interpolation: What assumptions have to be put on V, C and g to ensure the existence and uniqueness of the best approximation? What conditions must the linear forms L_i obey to guarantee a unique solution of the interpolation problem? What can be said about the error of the approximation/interpolation?

The analysis is highly nontrivial in the general setting of a basic linear or normed space V and a general subset $C \subset V$. However, the good news is that all important assumptions on the space V, the set C, and the function g, developed in the framework of the abstract Approximation Theory, are fulfilled automatically when V is a Hilbert space and C its closed subspace.

2.7.1 The Hilbert space setting

Let $V = V(\Omega)$ be a Hilbert space corresponding to the solved problem, $a(\cdot,\cdot) : V \times V \rightarrow \mathbb{R}$ a bounded V-elliptic bilinear form, $l \in V'$, and $V_{h,p}$ a finite-dimensional subspace of V determined by the finite element mesh $\mathcal{T}_{h,p}$. Consider the continuous problem (2.1),

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

and the discrete problem (2.5),

$$a(u_{h,p}, v) = l(v) \quad \text{for all } v \in V_{h,p}.$$

According to Céa's lemma (Theorem 2.1, Paragraph 2.1.2), the discretization error $\|u - u_{h,p}\|_V$ is bounded by the interpolation properties of the subspace $V_{h,p} \subset V$ and the continuity and V-ellipticity constants C_b, C_{el} of the bilinear form $a(\cdot,\cdot)$,

$$\|u - u_{h,p}\|_V \le \frac{C_b}{C_{el}} \inf_{v \in V_{h,p}} \|u - v\|_V = \frac{C_b}{C_{el}} \mathrm{dist}(u, V_{h,p})_V.$$

Hence the interpolation properties of the space $V_{h,p}$ are largely responsible for the final form of the error estimate.

In practice we always have a concrete interpolation operator $P : V \longrightarrow V_{h,p}$ that obviously satisfies

$$\mathrm{dist}(u, V_{h,p})_V \le \|u - Pu\|_V.$$

Hence, for a sufficiently regular function $u \in V$ it is our aim to estimate the interpolation error $\|u - Pu\|_V$ using some parameters of the mesh $\mathcal{T}_{h,p}$ as well as the amount of regularity of the function u. A typical interpolation error estimate has the form

$$\|u - Pu\|_V \le C(u)h^\alpha,$$

where $h = \max_i h_i$ is the mesh diameter and $C(u)$ is a constant depending on the amount of regularity of the function u. In addition to its application in error analysis, interpolation also finds practical use in the finite element technology, when a given function $g \in V$ needs to be represented by a sufficiently close function $g_{h,p} \in V_{h,p}$. Problems of this type are encountered in the finite element solution of evolutionary problems (to be discussed in Chapter 5), as well as in multigrid methods, automatic hp-adaptivity, and numerous other situations.

2.7.2 Best interpolant

In the Hilbert space setting the question of existence and uniqueness of the best approximation is trivial. Since $V_{h,p} \subset V$ is finite-dimensional and therefore automatically closed, according to Lemma A.39 the nearest representant of a function $g \in V$ in the norm $\|\cdot\|_V$ is its unique orthogonal projection $g_{h,p} = Pg \in V_{h,p}$. Theorem A.14 implies that the orthogonal projection P is defined uniquely via the condition

$$(g - g_{h,p}, v_{h,p})_V = 0 \qquad \text{for all } v_{h,p} \in V_{h,p}. \tag{2.81}$$

With some basis $\{v_1, v_2, \ldots, v_N\} \subset V_{h,p}$, (2.81) can be rewritten equivalently as

$$(g - g_{h,p}, v_i)_V = 0 \qquad \text{for all } i = 1, 2, \ldots, N. \tag{2.82}$$

Expressing

$$g_{h,p} = \sum_{j=1}^{N} y_j v_j \tag{2.83}$$

and substituting into (2.82), one obtains a system of linear algebraic equations

$$\sum_{j=1}^{N} y_j(v_j, v_i)_V = (g, v_i)_V, \quad i = 1, 2, \ldots, N, \tag{2.84}$$

for the unknown coefficients $y_1, y_2, \ldots, y_N$.

■ EXAMPLE 2.4

Consider a domain $\Omega = (-1, 1)$, covered with a finite element mesh $\mathcal{T}_{h,p} = \{K_1, K_2\}$ consisting of affine elements $K_1 = (-1, 0)$ and $K_2 = (0, 1)$. Assume the space $V = H_0^1(-1, 1)$ related to some problem with homogeneous Dirichlet boundary conditions. The finite element subspace $V_{h,p}$ is one-dimensional, defined as

$$V_{h,p} = \{v \in V;\ v|_{K_m} \in P^1(K_m),\ i = 1, 2\}.$$

Let us construct the best approximation $g_{h,p} \in V_{h,p}$ of the function $g(x) = 1 - x^4 \in V$. In other words, we are looking for a function $g_{h,p} \in V_{h,p}$ such that

$$\mathrm{dist}(g, g_{h,p}) = \mathrm{dist}(g, V_{h,p}). \tag{2.85}$$

The linear system (2.84) reduces to a single equation, which yields the best approximation $g_{h,p}$,

$$g_{h,p}(x) = \begin{cases} \frac{11}{10}(1 + x), & x \in K_1, \\ \frac{11}{10}(1 - x), & x \in K_2, \end{cases}$$

depicted in Figure 2.29.

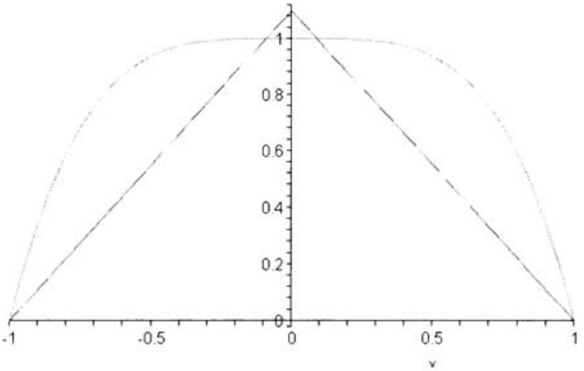


Figure 2.29 Best approximation $g_{h,p} \in V_{h,p}$ of the function $g \in V$.

Notice that the best approximation $g_{h,p}$ does not coincide with the function g at the grid point $x = 0$.

In some cases the construction of the best approximation may be too demanding from the practical point of view, since the cost of the calculation of $g_{h,p}$ is similar to the cost of solution of the global finite element problem. In such cases the only possibility is to abandon the optimality requirement (2.85) and find some less expensive interpolant. The first natural choice is to perform the orthogonal projection locally in elements.

2.7.3 Projection-based interpolant

Piecewise-affine case In the simplest case when all elements $K_1, K_2, \dots, K_M$ are affine, the continuity requirement implies that the projection-based interpolant $g_{h,p} \in V_{h,p}$ be defined as the usual piecewise-affine vertex interpolant,

$$g_{h,p}(x_j) = g^v_{h,p}(x_j) = g(x_j), \quad j = 0, 1, \dots, M, \tag{2.86}$$

where $g_{h,p}|_{K_m} \in P^1(K_m)$ for all $K_m \in \mathcal{T}_{h,p}$, as illustrated in Figure 2.30.

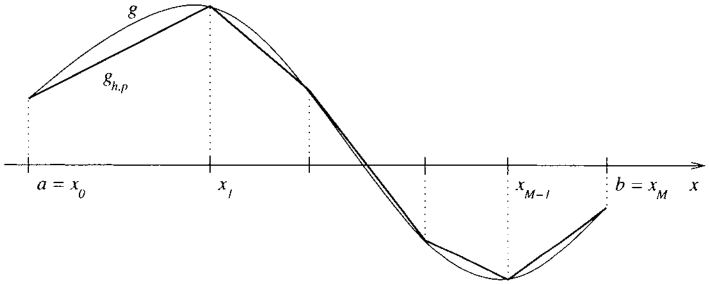


Figure 2.30 Projection-based interpolation reduces to the usual piecewise-affine Lagrange interpolation on piecewise-affine elements.

Higher-order case On a general higher-order finite element mesh $\mathcal{T}_{h,p}$, as the reader may guess, the interpolation problem is decoupled by subtracting the piecewise-affine vertex interpolant $g^v_{h,p}$ from the interpolated function g. The function $g - g^v_{h,p}$ vanishes at all grid points, and can be projected locally onto the polynomial spaces

$$P_0^{p_m}(K_m) = \{v \in H_0^1(K_m);\ v \in P^{p_m}(K_m)\}.$$

In this way one calculates the bubble interpolant $g^b_{h,p}$. The resulting interpolant $g_{h,p}$ is then obtained as a sum of the vertex and bubble parts,

$$g_{h,p} = g^v_{h,p} + g^b_{h.p} \tag{2.87}$$

Since we are in $H_0^1(K_m)$, either the full $H^1(K_m)$-norm or the equivalent $H^1(K_m)$-seminorm can be used. The fact that the standard vertex interpolation is combined with the orthogonal projection on higher-order subspaces is why one speaks about projection-based interpolation.

Choose, for example, the H^1-seminorm for the projection part. Then the associated inner product has the form

$$(\phi, \psi)_{H_0^1(K_m)} = \int_{K_m} \phi'(x)\psi'(x)\,\mathrm{d}x. \tag{2.88}$$

The orthogonality condition that determines $g_{h,p}^b$ is

$$((g - g_{h,p}^v) - g_{h,p}^b, v)_{H_0^1(K_m)} = 0 \quad \text{for all } v \in P_0^{p_m}(K_m). \tag{2.89}$$

This equivalent to

$$((g - g_{h,p}^v) - g_{h,p}^b, \vartheta_k^{(m)})_{H_0^1(K_m)} = 0 \quad k = 2, 3, \dots, p_m, \tag{2.90}$$

where $\vartheta_k^{(m)}$, $k = 2, 3, \dots, p_m$, is a suitable basis of $P_0^{p_m}(K_m)$. Utilizing the Lobatto bubble shape functions (2.63) and the reference maps (2.37), this basis has the form

$$\begin{aligned} \vartheta_2^{(m)}(x) &= l_2(x_{K_m}^{-1}(x)), \\ \vartheta_3^{(m)}(x) &= l_3(x_{K_m}^{-1}(x)), \\ &\vdots \\ \vartheta_{p_m}^{(m)}(x) &= l_{p_m}(x_{K_m}^{-1}(x)). \end{aligned} \tag{2.91}$$

Expressing now

$$g_{h,p}^b|_{K_m} = \sum_{r=2}^{p_m} \alpha_r^{(m)} \vartheta_r^{(m)},$$

and inserting this linear combination into (2.90), one obtains on K_m a system of $p_m - 1$ linear algebraic equations,

$$\sum_{r=2}^{p_m} \alpha_r^{(m)} \int_{K_m} \left(\vartheta_r^{(m)}\right)' \left(\vartheta_k^{(m)}\right)' \mathrm{d}x = \int_{K_m} (g - g_{h,p}^v)' \left(\vartheta_k^{(m)}\right)' \mathrm{d}x, \quad k = 2, 3, \dots, p_m, \tag{2.92}$$

for the unknown coefficients $\alpha_r^{(m)}$. By Substitution Theorem, (2.92) attains on the reference domain K_a a simple form

$$\sum_{r=2}^{p_m} \alpha_r^{(m)} \underbrace{\int_{K_a} l_r'(\xi) l_k'(\xi)\,\mathrm{d}\xi}_{\delta_{rk}} = \int_{K_a} \left(\tilde{g}^{(m)} - \tilde{g}_{h,p}^{v(m)}\right)'(\xi)\, l_k'(\xi)\,\mathrm{d}\xi, \quad k = 2, 3, \dots, p_m, \tag{2.93}$$

which by the orthogonality of the Lobatto bubble functions yields

$$\alpha_k^{(m)} = \int_{K_a} \left(\tilde{g}^{(m)} - \tilde{g}_{h,p}^{v(m)}\right)'(\xi)\, l_k'(\xi)\,\mathrm{d}\xi, \quad k = 2, 3, \dots, p_m. \tag{2.94}$$

Here, $\tilde{g}^{(m)}(\xi) = g(x_{K_m}(\xi))$ and $\tilde{g}_{h,p}^{v(m)}(\xi) = (g_{h,p}^v(x_{K_m}(\xi))$ is $l_0(\xi)g(x_{m-1}) + l_1(\xi)g(x_m)$. The orthogonality of the Lobatto shape functions is once more advantageous here. If one used the Lagrange nodal bubble shape functions $\theta_2, \theta_3, \dots, \theta_{p_m}$ from (2.57) instead, the simplification (2.94) would not have taken place, and a linear algebraic system of the form

(2.93) would have to be solved on every element K_m with $p_m \geq 2$. The projection problem (2.94) is illustrated in Figure 2.31.

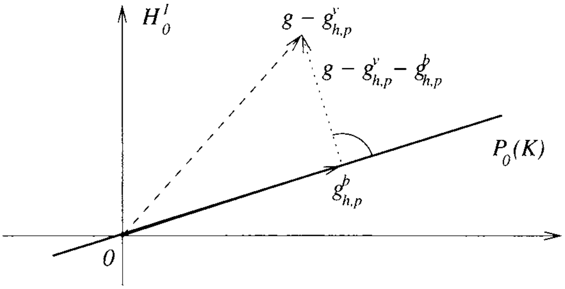


Figure 2.31 Graphical interpretation of the projection problem (2.94).

Lemma 2.5 (Local optimality of the projection-based interpolant) *Let $\Omega = (a, b) \subset \mathbb{R}$ be covered with a finite element mesh $\mathcal{T}_{h,p}$ consisting of M finite elements $K_m = (x_{m-1}, x_m)$ equipped with the polynomial degrees $1 \leq p_m = p(K_m)$. Let $g \in H^1(\Omega) \cap C(\overline{\Omega})$, $g_{h,p} \in V_{h,p}$ its projection-based interpolant (2.87) and $\tilde{g}_{h,p} \in V_{h,p}$ an arbitrary other interpolant satisfying $\tilde{g}_{h,p}(x_j) = g(x_j)$ for all $j = 0, 1, \ldots, M$. Then*

$$|g - g_{h,p}|_{1,2,K_m} \leq |g - \tilde{g}_{h,p}|_{1,2,K_m} \quad \text{for all } m = 1, 2, \ldots, M, \tag{2.95}$$

and consequently

$$|g - g_{h,p}|_{1,2,\Omega} \leq |g - \tilde{g}_{h,p}|_{1,2,\Omega}. \tag{2.96}$$

If the bubble interpolant $g^b_{h,p}$ is calculated using the full H^1-product $(\cdot, \cdot)_{1,2}$ instead of (2.88), the inequalities (2.95) and (2.96) hold with the full H^1-norm $\|\cdot\|_{1,2}$.

Proof: The fact that the bubble interpolant $g^b_{h,p}$ is defined as the orthogonal projection of $g - g^v_{h,p} \in H^1_0(K_m)$ onto $P_0^{p_m}(K_m)$ implies that

$$\begin{aligned} |g - g_{h,p}|_{1,2,K_m} &= |(g - g^v_{h,p}) - g^b_{h,p}|_{1,2,K_m} \\ &= \min_{w \in P_0^{p_m}(K_m)} |(g - g^v_{h,p}) - w|_{1,2,K_m} \\ &\leq |(g - g^v_{h,p}) - (\tilde{g}_{h,p} - g^v_{h,p})|_{1,2,K_m} \\ &= |g - \tilde{g}_{h,p}|_{1,2,K_m}. \end{aligned}$$

The integral $|g - g_{h,p}|^2_{1,2,\Omega}$ can be written as a sum

$$|g - g_{h,p}|^2_{1,2,\Omega} = \sum_{m=1}^{M} |g - g_{h,p}|^2_{1,2,K_m}.$$

Inequality (2.95) finally yields

$$\sum_{i=1}^{M} |g - g_{h,p}|^2_{1,2,K_m} \leq \sum_{i=1}^{M} |g - \tilde{g}_{h,p}|^2_{1,2,K_m} = |g - \tilde{g}_{h,p}|^2_{1,2,\Omega}.$$

Things work in the same way when (2.88) is replaced with the H^1-product $(\cdot, \cdot)_{1,2}$. ■

Let us close this paragraph by mentioning that the projection-based interpolation is significantly more efficient than the full projection from Paragraph 2.7.2. The cost of the local optimality on the elements $K_m, i = 1, 2, \ldots, M$, is one numerical integration over K_a in (2.94) when the orthogonal Lobatto hierarchic shape functions are used, or in the worst case (with a set of nonorthogonal shape functions) the solution of M systems of $p_m - 1$ linear algebraic equations of the form (2.93).

2.7.4 Nodal interpolant

The last important interpolation technique is the Lagrange nodal interpolation, which is based on the evaluation of (a) the interpolated function at a given set of nodal points and (b) a suitable set of interpolation polynomials. Depending on the selection of the nodal points (such as, e.g., equidistant, Chebyshev, Gauss–Lobatto, Fekete, or other points), one obtains various variants of the general Lagrange interpolation methods, which produce different interpolants.

By Lemma 2.5, all Lagrange interpolants are equally or less accurate than the projection-based interpolant (2.87). On the other hand, their explicit nature with no system of linear equations solved makes them extremely efficient. The Lagrange interpolation is a special case of nodal interpolation on general nodal elements, which will be discussed in detail in Chapter 3. In particular, the question of optimal interpolation points in 2D will be addressed in Paragraphs 4.3.1 and 4.3.4.

Although the Lagrange interpolation is natural for Lagrange nodal elements and the projection-based interpolation for Lobatto hierarchic elements, the projection-based interpolation can be performed on Lagrange nodal elements and vice versa.

Interpolation conditions Consider an interval $K_m = (x_{m-1}, x_m) \subset \Omega \subset \mathbb{R}$ and a set of Lagrange nodal points $x_{m-1} = \tilde{y}_1^{(m)} < \tilde{y}_2^{(m)} < \ldots < \tilde{y}_{p_m+1}^{(m)} = x_m$. Using the reference maps $x_{K_m} : K_a \to K_m$ from (2.37), define the corresponding points in the reference domain $K_a = (-1, 1)$ as $y_j = x_{K_m}^{-1}(\tilde{y}_j^{(m)})$. On the element K_m, the interpolation conditions

$$g_{h,p}\left(\tilde{y}_j^{(m)}\right) = g\left(\tilde{y}_j^{(m)}\right) \quad \text{for all } 1 \le j \le p_m + 1, \ g_{h,p} \in V_{h,p},$$

are equivalent to

$$(g_{h,p} \circ x_{K_m})(y_j) = (g \circ x_{K_m})(y_j) \quad \text{for all } 1 \le j \le p_m + 1, \ g_{h,p} \circ x_{K_m} \in P^{p_m}(K_a).$$

Hence, the interpolation can be performed elementwise on the reference domain K_a. In practice a unique set of Lagrange nodal points is defined on the reference domain and used for all elements.

A basic result related to the accuracy of the Lagrange interpolation in the maximum norm is formulated in the following lemma.

Lemma 2.6 (Error of the Lagrange interpolation) *Let* $-1 = y_1 < y_2 < \ldots < y_{p+1} = 1$ *and* $g \in C^{p+1}(\overline{K_a})$. *Consider the Lagrange interpolant*

$$g_{h,p}(x) = \sum_{i=1}^{p+1} \left(\prod_{j \neq i} \frac{x - y_j}{y_i - y_j} \right) g(y_i). \tag{2.97}$$

There exists a ξ_y, $\min\{-1,x\} \le \xi_y \le \max\{x,1\}$, *such that*

$$g(x) - g_{h,p}(x) = \frac{\prod_{i=1}^{p+1}(x-y_i)}{(p+1)!} g^{(p+1)}(\xi_y). \tag{2.98}$$

Proof: The result obviously holds if $x = y_i$. Hence suppose $x \neq y_i$ for all $1 \le i \le p+1$, and denote

$$e(x) = g(x) - g_{h,p}(x).$$

The function

$$\sigma(t) = e(t) - \frac{\prod_{i=1}^{p+1}(t-y_i)}{\prod_{i=1}^{p+1}(x-y_i)} e(x)$$

has $p+2$ distinct roots $t = x$ and $t = y_i$, $1 \le i \le p+1$. The Mean Value Theorem implies that $\sigma'(t)$ has $p+1$ distinct roots. Applying the Mean Value Theorem to higher derivatives of σ, we find that $\sigma^{(p+1)}(t)$ has a single root $\xi_y \in (\min\{-1,x\}, \max\{x,1\})$, satisfying

$$0 = \sigma^{(p+1)}(\xi_y) = g^{(p+1)}(\xi_y) - \frac{(p+1)!}{\prod_{i=1}^{p+1}(x-y_i)} e(x),$$

and (2.98) follows. ■

The function $\beta_p(x) = \prod_{i=1}^{p+1}(x-y_i)$ in (2.98) is the only way the distribution of the nodal points influences the distribution of the interpolation error. Compare with the projection-based interpolation from Paragraph 2.7.3, where the interpolation error was independent of the concrete representation of the polynomial space. Let us look at $\beta_p(x)$ for equidistributed nodal points in Figure 2.32.

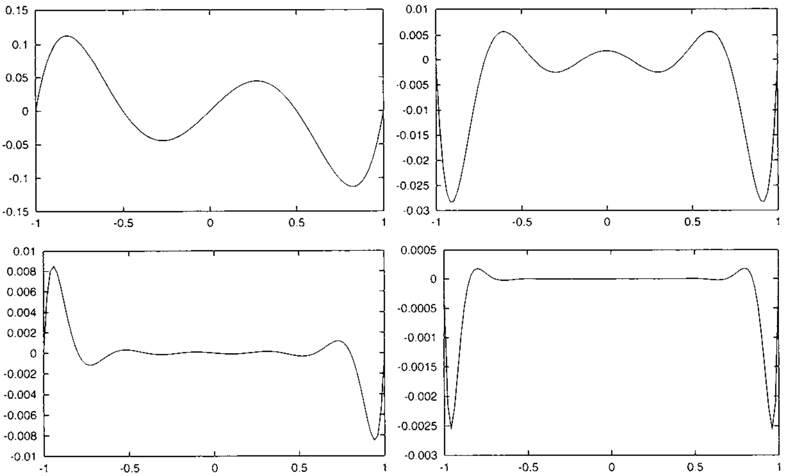


Figure 2.32 Error factor $\beta_p(x)$ for equidistributed nodal points, $p = 4, 7, 10$, and 13.

From these plots it is clear that the behavior of the error $e(x) = g(x) - g_{h,p}(x)$ is significantly worse near the endpoints than in the interior. The Lagrange interpolation with equidistributed nodal points is known to be notoriously bad. In his famous example from 1901, Carl Runge shows that the sequence of Lagrange interpolants $g_{h,p}$ with equidistributed nodal points diverges for otherwise a very nice function $g(x) = 1/(1+25x^2)$ in the interval $(-1,1)$ as $p \to \infty$ (for details see, e.g., [62]).

Chebyshev interpolant The Lagrange interpolant (2.97) based on the nodal points (2.59) is called Chebyshev interpolant. The error factors β_p for the Chebyshev interpolation with $p = 4, 7, 10$ and 13 are shown in Figure 2.33. Compare with Figure 2.32, and notice the different scales.

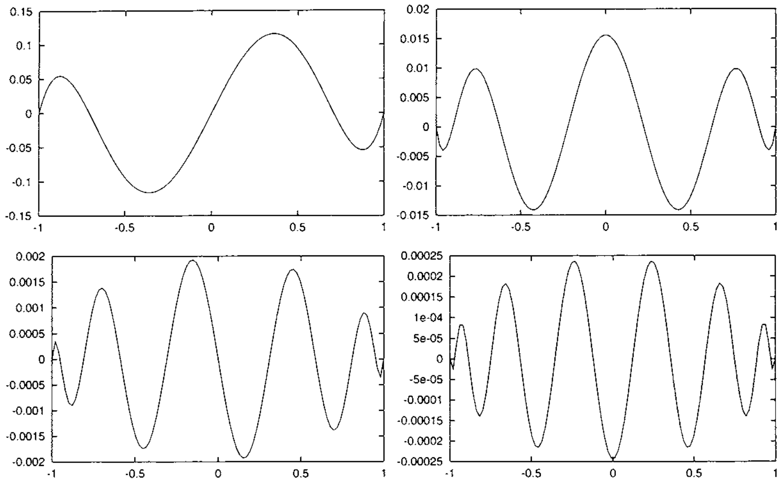


Figure 2.33 Error factor $\beta_p(x)$ for Chebyshev nodal points, $p = 4, 7, 10$ and 13.

Before introducing a basic Chebyshev interpolation error estimate, we need the weighted L^2-space

$$L^2_w(K_a) = \{v \in L^2(K_a);\ v \text{ is measurable and } \|v\|_{2,w} < \infty\}$$

with

$$\|v\|^2_{2,w} = \int_{-1}^{1} |v(x)|^2 w(x)\,\mathrm{d}x, \tag{2.99}$$

where $w(x) = 1/\sqrt{1-x^2}$ is the Chebyshev weight function. The norm (2.99) induces an inner product

$$(u, v)_w = \int_{-1}^{1} u(x)v(x)w(x)\,\mathrm{d}x$$

on $L^2_w \times L^2_w$. Further define a weighted Sobolev space

$$H^s_w(K_a) = \left\{v \in L^2_w(K_a);\ v^{(k)} \in L^2_w \text{ for all } k = 1, 2, \ldots, s\right\}$$

with the norm

$$\|v\|_{s,2,w} = \left(\sum_{k=0}^{s} \|v^{(k)}\|^2_{2,w}\right)^{\frac{1}{2}}.$$

Here $v^{(k)}$ denotes the kth weak derivative of v.

Theorem 2.3 (Chebyshev interpolation error estimate) *Let $u \in H^s_w(K_a)$ for some $s \geq 1$. Let $P_N u$ be the Chebyshev interpolant based on the $N + 1$ Lagrange nodal points (2.59). Then there exists a constant C independent of u such that*

$$\|u - P_N u\|_{2,w} \leq CN^{-s}\|u\|_{s,2,w}.$$

Proof: See, e.g., [90]. ■

Among nodal interpolation schemes, Chebyshev interpolation is very popular due to its accuracy. More details can be found, e.g., in [5] and [62].

2.7.5 Exercises

Exercise 2.25 *Consider a bounded interval $(a, b) \subset \mathbb{R}$ and $p + 1$ distinct points $a = y_1 < y_2 < \ldots < y_{p+1} = b$. Consider two polynomials $f, g \in P^p(a, b)$ such that*

$$f(y_j) = g(y_j) \quad \textit{for all } j = 1, 2, \ldots, p + 1.$$

Prove that necessarily $f = g$. Do not use the explicit formula of the Lagrange interpolation polynomial. Use the maximum number of roots of a polynomial instead.

Exercise 2.26 *In 1901, without the help of a computer, Carl Runge presented a famous example of a divergent series of Lagrange interpolation polynomials on equidistant meshes.*

Consider a function

$$g(x) = \frac{1}{1 + 25x^2}, \quad x \in K_a.$$

Construct the Lagrange interpolation polynomials

$$g_{h,p}(x) = \sum_{j=1}^{p+1} g(y_j)\theta_j^{(p)}(x),$$

where $\theta_j^{(p)}$ are the Lagrange nodal shape functions (2.57) corresponding to $p+1$ equidistant points $-1 = y_1 < y_2 < \ldots < y_{p+1} = 1$. Present plots of $g, g_{h,p}$ together with the H^1-seminorm of the error $g - g_{h,p}$ for $p = 2, 4, 6, 8$, and 10. You can use a computer.

Exercise 2.27 *Consider Exercise 2.26 with $p + 1$ Chebyshev nodal points (2.59). Again present plots of $g, g_{h,p}$ together with the H^1-seminorm of the error $g - g_{h,p}$ for $p = 2, 4, 6, 8$ and 10. Compare with the results of Exercise 2.26.*

Exercise 2.28 *At last consider Exercise 2.26 with the projection-based interpolation from Paragraph 2.7.3 instead of Lagrange nodal interpolation. Present plots of the projection-based interpolants $g_{h,p}$ of the function g in the spaces $P^p(K_a)$, where $p = 2, 4$ and 6. Use mesh containing a single element $K_a = (-1, 1)$. In all three cases calculate the H^1-seminorm of the error $g - g_{h,p}$. Compare with the results of Exercises 2.26 and 2.27.*

CHAPTER 3

GENERAL CONCEPT OF NODAL ELEMENTS

The reader knows from Chapter 2 the nodal and hierarchic concepts in the FEM. In the following we discuss in detail the nodal concept, which is both historically older and more suitable for an introduction. The strong side of nodal elements is their extremely general definition of degrees of freedom via linear forms, which allows for a very fast interpolation and makes these elements applicable to a large variety of problems in various spaces of functions.

3.1 THE NODAL FINITE ELEMENT

Let us return for a moment to the one-dimensional Lagrange nodal element $K = (a, b)$ of the degree p, equipped with $p + 1$ nodal points $a = y_1 < y_2 < \ldots < y_{p+1} = b$. The corresponding polynomial space on the element is $P = P^p(K)$. For every nodal point y_j, one can define a mapping

$$L_j : g \in P \to g(y_j) \in \mathbb{R}. \tag{3.1}$$

This functional is linear since

$$L_j(g + \tilde{g}) = (g + \tilde{g})(y_j) = g(y_j) + \tilde{g}(y_j) = L_j(g) + L_j(\tilde{g})$$

and

$$L_j(\alpha g) = \alpha g(y_j) = \alpha L_j(g),$$

for all $g, \tilde{g} \in P$ and all $\alpha \in \mathbb{R}$. Hence L_j are linear forms and they belong to the dual space P'. The number of the linear forms L_j is equal to the dimension $\dim(P) = p + 1$.

In Paragraph 2.4.4 we designed a basis of $P^p(K)$ consisting of Lagrange nodal shape functions $\theta_1, \theta_2, \ldots, \theta_{p+1}$ that satisfied the delta property (2.56),

$$\theta_i(y_j) = L_j(\theta_i) = \delta_{ij}. \tag{3.2}$$

Here δ_{ij} is the Kronecker delta ($\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise). General nodal elements are defined by replacing the interval K with a general bounded domain $K \in \mathbb{R}^d$ and by extending the Lagrange linear forms (3.1) to general linear forms

$$L_j : P \to \mathbb{R}.$$

A classical book on nodal elements is [30].

Definition 3.1 (Nodal finite element) Nodal finite element *is a triad* $\mathcal{K} = (K, P, \Sigma)$, *where*

- *K is a bounded domain in $\mathbb{R}^d$ with a Lipschitz-continuous boundary,*
- *P is a space of polynomials on K of the dimension dim$(P) = N_P$,*
- *$\Sigma = \{L_1, L_2, \ldots, L_{N_P}\}$ is a set of linear forms*

$$L_i : P \to \mathbb{R}, \quad i = 1, 2, \ldots, N_P.$$

The elements of Σ are called degrees of freedom (DOF).

In most cases it is clear from the context what finite element $\mathcal{K}$ is associated with a domain K. Then the set K itself often is called finite element, as we did in the previous chapter, and as we shall occasionally do in what follows.

3.1.1 Unisolvency and nodal basis

The one-dimensional Lagrange nodal points $y_1, y_2, \ldots, y_{p+1}$ were chosen pairwise distinct in Paragraph 2.4.4 in order to ensure that on every element K, any set of $p + 1$ given numbers $g_1, g_2, \ldots, g_{p+1}$ identifies a unique polynomial $g \in P^p(K)$ with the property $g(y_1) = g_1, g(y_2) = g_2, \ldots, g(y_{p+1}) = g_{p+1}$.

This requirement guarantees that the vector of computed coefficients,

$$\boldsymbol{Y} = (y_1, y_2, \ldots, y_N)^T = \boldsymbol{S}^{-1}\boldsymbol{F},$$

where $\boldsymbol{S}$ is the stiffness matrix and $\boldsymbol{F}$ the load vector, identifies a unique piecewise-polynomial function

$$u_{h,p} = \sum_{i=1}^{N} y_i v_i \in V_{h,p}.$$

Here $v_1, v_2, \ldots, v_N$ is a basis of the space $V_{h,p}$, and $u_{h,p}$ is the solution to the discrete problem. In the context of general nodal elements, the generalization of this property is called unisolvency:

Definition 3.2 (Unisolvency) *A nodal finite element (K, P, Σ) is said to be* unisolvent *if for every polynomial $g \in P$ it holds*

$$L_1(g) = L_2(g) = \ldots = L_{N_P}(g) = 0 \quad \Rightarrow \quad g = 0.$$

Lemma 3.1 *Let (K, P, Σ) be a unisolvent nodal element. Given any set of numbers $\{g_1, g_2, \ldots, g_{N_P}\} \in \mathbb{R}^{N_P}$, where $N_P = dim(P)$, there exists a unique polynomial $g \in P$ such that*

$$L_1(g) = g_1,\ L_2(g) = g_2, \ldots, L_{N_P}(g) = g_{N_P}. \tag{3.3}$$

Proof: Left to the reader as an exercise. ■

Unisolvency is characterized by a generalization of the delta property (3.2):

Definition 3.3 (Nodal basis) *Let (K, P, Σ), $dim(P) = N_P$, be a nodal finite element. We say that a set of functions $\mathcal{B} = \{\theta_1, \theta_2, \ldots, \theta_{N_P}\} \subset P$ is a* nodal basis *of P if it satisfies*

$$L_i(\theta_j) = \delta_{ij} \quad \textit{for all } 1 \le i, j \le N_P. \tag{3.4}$$

The functions θ_i are usually called nodal shape functions.

Theorem 3.1 (Characterization of unisolvency) *Consider a nodal finite element (K, P, Σ), $dim(P) = N_P$. The finite element is unisolvent if and only if there exists a unique nodal basis $\mathcal{B} = \{\theta_1, \theta_2, \ldots, \theta_{N_P}\} \subset P$.*

Proof: First let us consider a unisolvent finite element and construct a unique nodal basis $\mathcal{B} = \{\theta_1, \theta_2, \ldots, \theta_{N_P}\}$. Begin with any basis $\{g_1, g_2, \ldots, g_{N_P}\} \subset P$. Express each sought function θ_j, $j = 1, \ldots, N_P$, as

$$\theta_j = \sum_{k=1}^{N_P} a_{kj} g_k. \tag{3.5}$$

Condition (3.4) implies

$$\delta_{ij} = L_i(\theta_j) = L_i\left(\sum_{k=1}^{N_P} a_{kj} g_k\right) = \sum_{k=1}^{N_P} L_i(g_k) a_{kj}, \quad 1 \le i, j \le N_P, \tag{3.6}$$

which yields a system of N_P linear equations for each j. Putting together the N_P linear systems related to $\theta_1, \theta_2, \ldots, \theta_{N_P}$, one obtains a matrix equation

$$\boldsymbol{L}\boldsymbol{A} = \boldsymbol{I}. \tag{3.7}$$

Here $\boldsymbol{L} = \{L_i(g_k)\}_{i,k=1}^{N_P}$ is a generalized Vandermonde matrix, $\boldsymbol{I}$ the identity matrix, and the matrix $\boldsymbol{A} = \{a_{kj}\}_{k,j=1}^{N_P}$ contains in its columns the unknown coefficients of the functions $\theta_1, \theta_2, \ldots, \theta_{N_P}$, respectively.

Let us verify that the matrix $\boldsymbol{L}$ is invertible: If the columns of $\boldsymbol{L}$ were linearly dependent, there would exist a nontrivial set of coefficients $b_1, b_2, \ldots, b_{N_P}$ such that

$$0 = \sum_{k=1}^{N_P} b_k L_i(g_k) = L_i\left(\sum_{k=1}^{N_P} b_k g_k\right) \quad \text{for all } i = 1, 2, \ldots, N_P. \tag{3.8}$$

However, this is in contradiction with the unisolvency assumption. Therefore $\boldsymbol{L}$ is nonsingular and the functions $\theta_1, \theta_2, \ldots, \theta_{N_P}$ form a basis in P.

Conversely, let $\mathcal{B} = \{\theta_1, \theta_2, \ldots, \theta_{N_P}\}$ be a nodal basis of the space P. Assume that

$$L_1(g) = L_2(g) = \ldots = L_{N_P}(g) = 0$$

for some function $g \in P$. Express

$$g = \sum_{j=1}^{N_P} \gamma_j \theta_j.$$

Since

$$0 = L_i(g) = L_i\left(\sum_{j=1}^{N_P} \gamma_j \theta_j\right) = \gamma_i \quad \text{for all } i = 1, 2, \ldots, N_P,$$

we conclude that $g = 0$ and thus the finite element is unisolvent. ■

3.1.2 Checking unisolvency

Theorem 3.1 describes how to check the unisolvency of an arbitrary nodal finite element (K, P, Σ):

- Consider an arbitrary basis $\{g_1, g_2, \ldots, g_{N_P}\} \subset P$.

- Construct the generalized Vandermonde matrix

$$\boldsymbol{L} = \{L_i(g_k)\}_{i,k=1}^{N_P}.$$

- If $\boldsymbol{L}$ is invertible, then the element is unisolvent, and moreover $\boldsymbol{L}^{-1}$ has in its jth column the coefficients a_{kj}, $k = 1, 2, \ldots, N_P$, which define the jth nodal basis function θ_j via (3.5).

- If the matrix $\boldsymbol{L}$ is not invertible, then the element is not unisolvent.

■ EXAMPLE 3.1 (A nonunisolvent element)

Usually one deals with unisolvent finite elements. Therefore, let us show at least one example of a nodal finite element which is not unisolvent. Consider the polynomial space

$$Q^1(K_q) = \text{span}\{1, \xi_1, \xi_2, \xi_1\xi_2\}$$

in the square domain $K_q = (-1,1)^2$. The set Σ comprises four linear forms $L_i : Q^1(K_q) \to \mathbb{R}$, associated with function values at the edge midpoints $[-1,0]$, $[1,0]$, $[0,-1]$, and $[0,1]$,

$$\begin{aligned} L_1(g) &= g(-1,0), \\ L_2(g) &= g(1,0), \\ L_3(g) &= g(0,-1), \\ L_4(g) &= g(0,1), \end{aligned}$$

as shown in Figure 3.1.

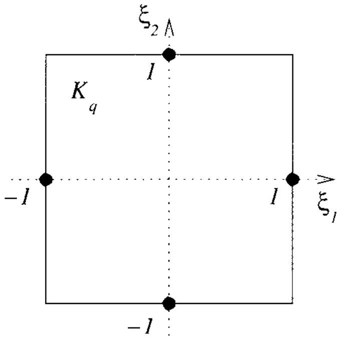


Figure 3.1 Nonunisolvent nodal finite element consisting of a square domain K, polynomial space $Q^1(K_q) = \text{span}\{1, \xi_1, \xi_2, \xi_1\xi_2\}$ and linear forms associated with the values at edge midpoints.

The generalized Vandermonde matrix $\boldsymbol{L} = \{L_i(g_j)\}_{i,j=1}^4$ corresponding to the functions $g_1(\boldsymbol{\xi}) = 1$, $g_2(\boldsymbol{\xi}) = \xi_1$, $g_3(\boldsymbol{\xi}) = \xi_2$ and $g_3(\boldsymbol{\xi}) = \xi_1\xi_2$,

$$\boldsymbol{L} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix},$$

is singular.

3.2 EXAMPLE: LOWEST-ORDER Q^1- AND P^1-ELEMENTS

The unisolvency of the nodal finite element from Example 3.1 can be fixed by replacing the edge midpoints with vertices. In this way one obtains the basic and most frequently used lowest-order element for H^1-problems on quadrilateral meshes in 2D: the Q^1-element.

3.2.1 Q^1-element

The reference square domain $K_q = (-1,1)^2$ is endowed with the polynomial space

$$Q^1(K_q) = \text{span}\{1, \xi_1, \xi_2, \xi_1\xi_2\}. \tag{3.9}$$

The set of degrees of freedom $\Sigma_q = \{L_1, L_2, \ldots, L_4\}$ consists of the linear forms $L_i : Q^1(K_q) \to \mathbb{R}$,

$$\begin{aligned} L_1(g) &= g(\boldsymbol{v}_1), \\ L_2(g) &= g(\boldsymbol{v}_2), \\ L_3(g) &= g(\boldsymbol{v}_3), \\ L_4(g) &= g(\boldsymbol{v}_4), \end{aligned} \tag{3.10}$$

as illustrated in Figure 3.2.

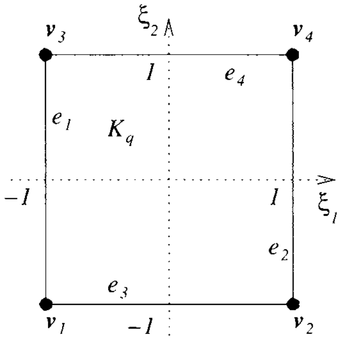


Figure 3.2 Q^1-element on the reference domain K_q.

Lemma 3.2 *The finite element $(K_q, Q^1(K_q), \Sigma_q)$ is unisolvent, and the nodal basis of the space $Q^1(K_q)$ consists of the biaffine shape functions*

$$\begin{aligned} \varphi_q^{v_1}(\boldsymbol{\xi}) &= \frac{(1-\xi_1)(1-\xi_2)}{4}, \\ \varphi_q^{v_2}(\boldsymbol{\xi}) &= \frac{(1+\xi_1)(1-\xi_2)}{4}, \\ \varphi_q^{v_3}(\boldsymbol{\xi}) &= \frac{(1-\xi_1)(1+\xi_2)}{4}, \\ \varphi_q^{v_4}(\boldsymbol{\xi}) &= \frac{(1+\xi_1)(1+\xi_2)}{4}. \end{aligned} \tag{3.11}$$

Proof: Since the generalized Vandermonde matrix

$$\boldsymbol{L} = \{L_i(g_k)\}_{i,k=1}^{N_P} = \begin{pmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

is nonsingular, the element is unisolvent. According to Theorem 3.1, the nodal basis (3.11) is obtained by inverting the matrix L. It is easy to verify that the nodal shape functions satisfy the delta property (3.4). ■

Since $K_q = K_a \times K_a$, the nodal shape functions (3.11) are Cartesian products of the one-dimensional lowest-order Lagrange shape functions θ_1 and θ_2 (same as l_0 and l_1 in the Lobatto hierarchic case):

$$\begin{aligned} \varphi_q^{v_1}(\boldsymbol{\xi}) &= \theta_1(\xi_1)\theta_1(\xi_2) = l_0(\xi_1)l_0(\xi_2), \\ \varphi_q^{v_2}(\boldsymbol{\xi}) &= \theta_2(\xi_1)\theta_1(\xi_2) = l_1(\xi_1)l_0(\xi_2), \\ \varphi_q^{v_3}(\boldsymbol{\xi}) &= \theta_1(\xi_1)\theta_2(\xi_2) = l_0(\xi_1)l_1(\xi_2), \\ \varphi_q^{v_4}(\boldsymbol{\xi}) &= \theta_2(\xi_1)\theta_2(\xi_2) = l_1(\xi_1)l_1(\xi_2). \end{aligned} \tag{3.12}$$

Q^1-element on a convex quadrilateral K Consider an arbitrary convex quadrilateral domain $K \subset \mathbb{R}^2$ with straight edges $s_1, s_2, \ldots, s_4$, illustrated in Figure 3.3.

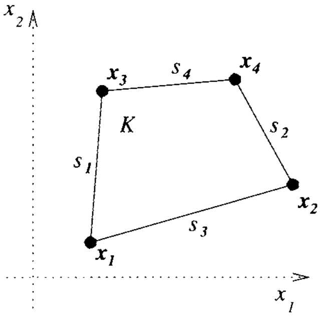


Figure 3.3 Q^1-element on a quadrilateral domain $K \subset \mathbb{R}^2$.

The Q^1-element on K is defined using the Q^1-element on the reference square domain K_q and a suitable reference map $\boldsymbol{x}_K : K_q \to K$. A natural choice is the isoparametric map, defined as a linear combination of the nodal shape functions (3.11) with the coordinates of the vertices $\boldsymbol{x}_i$,

$$\boldsymbol{x}_K(\boldsymbol{\xi}) = \sum_{i=1}^{4} \boldsymbol{x}_i \varphi_q^{v_i}(\boldsymbol{\xi}). \tag{3.13}$$

Since the Substitution Theorem is involved in the finite element discretization (this will be discussed in Chapter 4), the reference map $\boldsymbol{x}_K(\boldsymbol{\xi})$ must be a bijection. However, the question of invertibility of (3.13) is not trivial. We will study this topic in Paragraph 3.2.3.

Remark 3.1 *Elements where the same shape functions are used for the approximation and for the construction of the reference maps, like in this case, are called isoparametric. The map $\boldsymbol{x}_K$ is called isoparametric reference map. The coefficients $\boldsymbol{x}_i$ are called geometrical degrees of freedom (GDOF). The map (3.13) can be generalized to quadrilateral elements with curved edges by adding terms corresponding to higher-order shape functions (see, e.g., [111]).*

Proposition 3.1 *The isoparametric reference map (3.13) satisfies*

$$\boldsymbol{x}_K(\boldsymbol{v}_i) = \boldsymbol{x}_i \qquad \textit{for all } i = 1, 2, \ldots, 4, \tag{3.14}$$

where $\boldsymbol{v}_i$ are the vertices of the reference square domain K_q, and

$$\boldsymbol{x}_K(e_i) = s_i \qquad \textit{for all } i = 1, 2, \ldots, 4, \tag{3.15}$$

where e_i are the edges of K_q.

Proof: The relation (3.14) follows from the delta property (3.4),

$$\varphi_q^{v_i}(\boldsymbol{v}_j) = \delta_{ij} \qquad \text{for all } 1 \le i, j \le 4,$$

and (3.15) holds due to affinity of the shape functions $\varphi_q^{v_1}, \varphi_q^{v_2}, \ldots, \varphi_q^{v_4}$ on the edges of the reference domain K_q. ■

The design of the finite element $(K, Q^1(K), \Sigma_K)$ in the sense of Definition 3.1 is accomplished by defining the space

$$Q^1(K) = \{q \circ \boldsymbol{x}_K^{-1};\ g \in Q^1(K_q)\}, \tag{3.16}$$

and the set Σ_K consisting of four degrees of freedom $L_i^{(K)} : Q^1(K) \to \mathbb{R}$,

$$\begin{aligned} L_1^{(K)}(g) &= g(\boldsymbol{x}_1), \\ L_2^{(K)}(g) &= g(\boldsymbol{x}_2), \\ L_3^{(K)}(g) &= g(\boldsymbol{x}_3), \\ L_4^{(K)}(g) &= g(\boldsymbol{x}_4). \end{aligned} \tag{3.17}$$

Proposition 3.2 *The finite element $(K, Q^1(K), \Sigma_K)$ is unisolvent, and the shape functions*

$$\varphi_K^{v_i}(\boldsymbol{x}) = (\varphi_q^{v_i} \circ \boldsymbol{x}_K^{-1})(\boldsymbol{x}), \qquad 1 \le i \le 4, \tag{3.18}$$

constitute a unique nodal basis of the space (3.16).

Proof: This is left to the reader as an exercise. ■

Remark 3.2 *Notice that the inverse of a polynomial map generally is not polynomial (e.g., x^2 vs. $\sqrt{x}$), and therefore it is not obvious whether $Q^1(K)$ is a polynomial space or not. This will be discussed in Paragraph 3.2.3.*

3.2.2 P^1-element

The natural counterpart of the Q^1-element on triangular meshes is the P^1-element, sometimes called Courant triangle in honor of Richard Courant, a former assistant to David Hilbert. R. Courant first used a numerical scheme that we would call the Finite element method in 1943 to solve a torsion problem. His work was based on his previous results with Hurwitz and Hilbert in the 1920s. R. Courant was forced to leave Europe during the World War II. At the New York University he founded a new Institute of Mathematical Sciences, which since 1964 carries his name. The name "Finite element method" appeared in the 1960s.

Figure 3.4 Richard Courant (1888–1972).

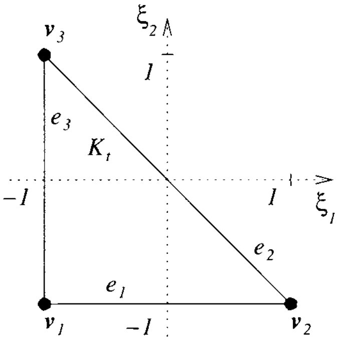


Figure 3.5 P^1-element on the reference domain K_t with the nodal points at its vertices.

Consider the triangular reference domain K_t shown in Figure 3.5. Alternative reference domains may be used, but K_t has certain advantages which will be discussed later.

The domain K_t is equipped with the polynomial space

$$P^1(K_t) = \text{span}\{1, \xi_1, \xi_2\}.$$

The set of degrees of freedom Σ_t contains the linear forms $L_i : P^1(K_t) \to \mathbb{R}$,

$$\begin{aligned} L_1(g) &= g(\boldsymbol{v}_1), \\ L_2(g) &= g(\boldsymbol{v}_2), \\ L_3(g) &= g(\boldsymbol{v}_3). \end{aligned} \tag{3.19}$$

The element $(K_t, P^1(K_t), \Sigma_t)$ is evidently unisolvent and the corresponding nodal basis consists of three affine functions

$$\begin{aligned}\varphi_t^{v_1}(\boldsymbol{\xi}) &= -\frac{\xi_1 + \xi_2}{2}, \\ \varphi_t^{v_2}(\boldsymbol{\xi}) &= \frac{1+\xi_1}{2}, \\ \varphi_t^{v_3}(\boldsymbol{\xi}) &= \frac{1+\xi_2}{2}.\end{aligned} \tag{3.20}$$

It is easy to verify that these shape functions meet the delta property (3.4) with the linear forms (3.19).

P^1-element on an arbitrary triangle K Next consider an arbitrary triangular domain $K \subset \mathbb{R}^2$ with the vertices $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$ and straight edges s_1, s_2, s_3, as shown in Figure 3.6.

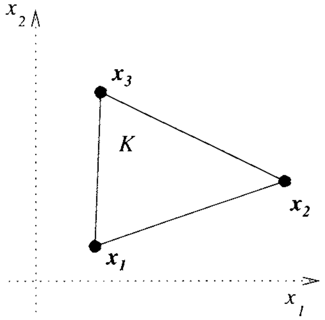


Figure 3.6 P^1-element on a triangular domain $K \subset \mathbb{R}^2$ with straight edges.

The isoparametric reference map $\boldsymbol{x}_K : K_t \to K$ is defined analogously to (3.13),

$$\boldsymbol{x}_K(\boldsymbol{\xi}) = \sum_{i=1}^{3} \boldsymbol{x}_i \varphi_t^{v_i}(\boldsymbol{\xi}), \tag{3.21}$$

where $\varphi_t^{v_i}$ are the nodal basis functions (3.20).

Proposition 3.3 *For any nondegenerate triangle $K \subset \mathbb{R}^2$, the isoparametric reference map $\boldsymbol{x}_K$ is invertible, and the inverse map $\boldsymbol{x}_K^{-1} : K \to K_t$ is affine.*

Proof: Since the map $\boldsymbol{x}_K$ is affine and the triangle K nondegenerate, the Jacobian J_K is a nonzero constant. Therefore also the Jacobian of the inverse map, J_K^{-1}, is a nonzero constant. This means that the inverse map $\boldsymbol{x}_K^{-1}$ is affine. ■

Proposition 3.3 yields that the space

$$P^1(K) = \{q \circ \boldsymbol{x}_K^{-1};\ g \in P^1(K_t)\} \tag{3.22}$$

is polynomial. The definition of the finite element $(K, P^1(K), \Sigma_K)$ is accomplished by defining the set of degrees of freedom Σ_K using the linear forms $L_i^{(K)} : P^1(K) \to \mathbb{R}$,

$$\begin{aligned} L_1^{(K)}(g) &= g(\boldsymbol{x}_1), \\ L_2^{(K)}(g) &= g(\boldsymbol{x}_2), \\ L_3^{(K)}(g) &= g(\boldsymbol{x}_3). \end{aligned} \tag{3.23}$$

Proposition 3.4 *The shape functions*

$$\varphi_{t,K}^{v_i}(\boldsymbol{x}) = (\varphi_t^{v_i} \circ \boldsymbol{x}_K^{-1})(\boldsymbol{x}), \quad 1 \le i \le 3, \tag{3.24}$$

constitute the unique nodal basis of the space (3.22), satisfying the delta property (3.4) with the degrees of freedom (3.23).

Proof: This follows easily from Definition 3.3. ■

The application of the Q^1- and P^1-elements to the discretization of two-dimensional problems formulated in the Sobolev space H^1 will be described in Section 4.1.

3.2.3 Invertibility of the quadrilateral reference map $\boldsymbol{x}_K$

The invertibility of reference maps for nonsimplicial elements always is a nontrivial issue in the finite element analysis. The question of invertibility of triaffine hexahedral reference maps, for example, has not been completely resolved yet. The situation is simpler in the quadrilateral case, where it is known that the Jacobian J_K of the isoparametric reference map (3.13) is nonzero in K_q if and only if the domain K is nondegenerate and convex. To our best knowledge, this result was first proved in [113]. Let us present a slightly different version of the proof here.

Lemma 3.3 *The Jacobian $J_K(\boldsymbol{\xi})$ of the biaffine isoparametric reference map (3.13) is an affine function. In particular, its minimum over K_q is attained at one of the vertices $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_4$.*

Proof: Let the vertices of the mesh quadrilateral K be denoted by $\boldsymbol{x}_1 = (x_1, y_1), \boldsymbol{x}_2 = (x_2, y_2), \ldots, \boldsymbol{x}_4 = (x_4, y_4)$ (in harmony with Figure 3.3). Use the functions (3.12) to write the isoparametric reference map (3.13) in the form

$$\begin{aligned} \boldsymbol{x}_K\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} &= \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} l_0(\xi_1)l_0(\xi_2) + \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} l_1(\xi_1)l_0(\xi_2) \\ &+ \begin{pmatrix} x_3 \\ y_3 \end{pmatrix} l_0(\xi_1)l_1(\xi_2) + \begin{pmatrix} x_4 \\ y_4 \end{pmatrix} l_1(\xi_1)l_1(\xi_2). \end{aligned}$$

Further in agreement with Figure 3.3 denote $(u_1, v_1) := \boldsymbol{x}_3 - \boldsymbol{x}_1, (u_2, v_2) := \boldsymbol{x}_4 - \boldsymbol{x}_2, (u_3, v_3) := \boldsymbol{x}_2 - \boldsymbol{x}_1, (u_4, v_4) := \boldsymbol{x}_4 - \boldsymbol{x}_3$. Recall the lowest-order one-dimensional Lobatto shape functions $l_0(\xi) = (1-\xi)/2$ and $l_1(\xi) = (1+\xi)/2$, and use the identity $l_1(\xi) + l_0(\xi) = 1$ to calculate

$$\boldsymbol{x}_K\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} u_4 - u_3 \\ v_4 - v_3 \end{pmatrix} l_0(\xi_1)l_0(\xi_2) - \begin{pmatrix} u_4 \\ v_4 \end{pmatrix} l_0(\xi_1) - \begin{pmatrix} u_2 \\ v_2 \end{pmatrix} l_0(\xi_2) + \begin{pmatrix} x_4 \\ y_4 \end{pmatrix}.$$

The Jacobi matrix of $\boldsymbol{x}_K$ has the form

$$\frac{D\boldsymbol{x}_K}{D\boldsymbol{\xi}}(\boldsymbol{\xi}) = -\frac{1}{2}\begin{pmatrix} (u_4 - u_3)l_0(\xi_2) - u_4 & (u_4 - u_3)l_0(\xi_1) - u_2 \\ (v_4 - v_3)l_0(\xi_2) - v_4 & (v_4 - v_3)l_0(\xi_1) - v_2 \end{pmatrix},$$

which means that its determinant,

$$J_K(\boldsymbol{\xi}) = \frac{(u_4v_2 - u_2v_4)l_1(\xi_2) + (u_3v_2 - u_2v_3)l_0(\xi_2) + (u_4v_3 - u_3v_4)l_0(\xi_1)}{4}, \quad (3.25)$$

is an affine function. ■

Theorem 3.2 *Let $K \subset \mathbb{R}^2$ be a nondegenerate quadrilateral with straight edges. We assume the ordering of the vertices shown in Figure 3.3. The Jacobian $J_K(\boldsymbol{\xi})$ of the isoparametric reference map (3.13) is positive in the reference domain K_q if and only if K is convex.*

Proof: Rewrite (3.25) using the cross-products of the edge vectors s_2, s_3 and s_4 as

$$J_K(\boldsymbol{\xi}) = \frac{(s_4 \times s_2)l_1(\xi_2) + (s_3 \times s_2)l_0(\xi_2) + (s_4 \times s_3)l_0(\xi_1)}{4}.$$

At the vertices $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_4$ the Jacobian J_K attains the values

$$\begin{aligned} J_K(\boldsymbol{v}_1) &= \frac{1}{4}(s_3 \times s_2) + \frac{1}{4}(s_4 \times s_3) = \frac{1}{4}(s_3 \times s_1), \\ J_K(\boldsymbol{v}_2) &= \frac{1}{4}(s_3 \times s_2), \\ J_K(\boldsymbol{v}_3) &= \frac{1}{4}(s_4 \times s_2) + \frac{1}{4}(s_4 \times s_3) = \frac{1}{4}(s_4 \times s_1), \\ J_K(\boldsymbol{v}_4) &= \frac{1}{4}(s_4 \times s_2). \end{aligned}$$

By Lemma 3.3, the minimum of $J_K(\boldsymbol{\xi})$ in K_q is one of these four values. All of them are positive if and only if the mesh quadrilateral K is convex. ■

Corollary 3.1 *Even for a convex quadrilateral element $K \in \mathcal{T}_{h,p}$ with straight edges, the inverse $\boldsymbol{x}_K^{-1}$ of the isoparametric biaffine reference map (3.13) generally is not polynomial. In particular, the space $Q^1(K)$ defined in (3.16) is not a polynomial space.*

This can be seen after expressing the inverse map explicitly, via a formula containing square roots (see, e.g., [113]). In special cases, when K is the Cartesian product of two intervals, both the reference map $\boldsymbol{x}_K$ and its inverse $\boldsymbol{x}_K^{-1}$ are affine.

3.3 INTERPOLATION ON NODAL ELEMENTS

The interpolation on finite elements is a procedure that takes a function $g \in V(\Omega_h)$ and produces its suitable piecewise-polynomial representant in the finite element space $g_{h,p} \in V_{h,p}(\Omega_h)$. Here by Ω_h we mean a suitable open polygonal domain that approximates the domain Ω for the purposes of the finite element discretization (more about this will be said in Chapter 4). Various interpolation techniques with different quality and computational cost can be used, ranging from the fastest fully explicit interpolation to the full orthogonal projection where a system of N linear algebraic equations, $N = \dim(V_{h,p})$, is solved (see Section 2.7). Among these approaches, the fully explicit interpolation is typical for nodal elements.

3.3.1 Local nodal interpolant

Recall the Lagrange nodal interpolation on the one-dimensional element (K, P, Σ) from Paragraph 2.7.4,

$$g_p = \sum_{i=1}^{N_P} g(y_i)\theta_i, \tag{3.26}$$

where $K = K_a = (-1, 1)$, $P = P^p(K_a)$, $g \in H^1(K_a)$, $g_p \in P$ and θ_i are the Lagrange nodal shape functions (2.57) forming a basis of P. Using the linear forms (3.1), the interpolant (3.26) can be written as

$$g_p = \sum_{i=1}^{N_P} L_i(g)\theta_i. \tag{3.27}$$

This is a bridge that extends the one-dimensional Lagrange interpolation to the interpolation on general nodal elements:

Definition 3.4 (Local nodal interpolant) *Let $\mathcal{B} = \{\theta_1, \theta_2, \ldots, \theta_{N_P}\}$ be the unique nodal basis of a unisolvent finite element (K, P, Σ). Let $g \in V$, where $P \subset V$, be a function for which the values $L_1(g), L_2(g), \ldots, L_{N_P}(g)$ are defined. Then the* local nodal interpolant *is defined as*

$$\mathcal{I}_K(g) = \sum_{i=1}^{N_P} L_i(g)\theta_i. \tag{3.28}$$

Remark 3.3

1. *The requirement that all the values $L_1(g), L_2(g), \ldots, L_{N_P}(g)$ be defined is important. Since the linear forms L_i are defined for polynomials from the finite element space P only (see Definition 3.1), there exist functions outside of P that cannot be interpolated.*

2. *Further, notice that it follows from the linearity of the forms L_i that the interpolation operator $\mathcal{I}_K : V \to P$ is linear.*

Next let us discuss basic properties of nodal interpolants:

Proposition 3.5 *Let (K, P, Σ) be a unisolvent nodal finite element and $\mathcal{I}_K(g)$ the nodal interpolant of a function $g \in V$, $P \subset V$. Then*

$$L_i(\mathcal{I}_K(g)) = L_i(g), \quad 1 \le i \le N_P.$$

Proof: It follows immediately from Definition 3.4 and (3.4) that

$$L_i\left(\sum_{j=1}^{N_P} L_j(g)\theta_j\right) = \sum_{j=1}^{N_P} L_j(g)L_i(\theta_j) = L_i(g).$$

■

Proposition 3.6 *Let (K, P, Σ) be a unisolvent nodal finite element. The nodal interpolation operator $\mathcal{I}_K$ is idempotent,*

$$\mathcal{I}_K^2 = \mathcal{I}_K.$$

Proof: It follows immediately from Proposition 3.5 that

$$\mathcal{I}_K(g) = g \qquad \text{for all } g \in P.$$

Let $P \subset V$. For all $g \in V$ such that $\mathcal{I}_K(g)$ is defined, we have

$$\mathcal{I}_K(\underbrace{\mathcal{I}_K(g)}_{\in P}) = \mathcal{I}_K(g),$$

which had to be shown. ■

■ **EXAMPLE 3.2**

Consider the Q^1-element on the reference square domain $K_q = (-1, 1)^2$ and the function

$$g(\boldsymbol{x}) = (x_1 - 1)^2(x_1 + 1) - 2x_1x_2(x_2 + 1) \quad \in H^1(K_q).$$

The values $L_1(g), L_2(g), \ldots, L_4(g)$ (function values of g in the corners of K_q) are defined. Hence the nodal interpolant $\mathcal{I}(g)$ exists, and using the nodal basis functions $\varphi_q^{v_i}$ from (3.11), we obtain

$$\mathcal{I}(g) = g(-1,-1)\varphi_q^{v_1}(\boldsymbol{x}) + g(1,-1)\varphi_q^{v_2}(\boldsymbol{x}) + g(-1,1)\varphi_q^{v_3}(\boldsymbol{x}) + g(1,1)\varphi_q^{v_4}(\boldsymbol{x}).$$

The functions g and $\mathcal{I}(g)$ are depicted in Figure 3.7.

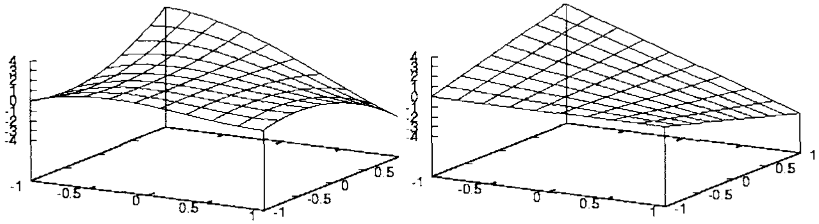


Figure 3.7 The interpolated function $g \in H^1(K_q)$ and the nodal interpolant $\mathcal{I}(g) \in Q^1(K_q)$.

3.3.2 Global interpolant and conformity

The form of the local nodal interpolant determines the conformity of finite element meshes consisting of such elements to the space of functions where the underlying PDE is solved. Before discussing the conformity, we have to be more specific about the shape of the meshes we consider:

Definition 3.5 (Regular mesh) *Let $\Omega_h \subset \mathbb{R}^d$ be an open bounded domain with polygonal boundary, and $\mathcal{T}_{h,p}$ a partition of Ω_h into a finite number of open polygonal subdomains $K_1, K_2, \ldots, K_M$, such that*

$$\bigcup_{i=1}^{M} \overline{K_i} = \overline{\Omega}_h$$

and $K_i \cap K_j = \emptyset$ if $i \neq j$. A two-dimensional finite element mesh $\mathcal{T}_{h,p}$ is said to be regular *if every nonempty intersection of $\overline{K}_i \cap \overline{K}_j$, $i \neq j$, can either be a whole shared edge or a single shared vertex. In 3D this holds analogously with faces, edges and vertices.*

In the following we assume a regular finite element mesh $\mathcal{T}_{h,p}$ consisting of unisolvent nodal finite elements $K_1, K_2, \ldots, K_M$,

$$P_i(K_i) \subset V(\Omega_h)|_{K_i} \qquad \text{for all } i = 1, 2, \ldots, M.$$

Definition 3.6 (Global nodal interpolant) *The global nodal interpolant $\mathcal{I}(g)$ of a function $g \in V(\Omega_h)$ is defined as*

$$\mathcal{I}(g)|_{K_i} \equiv \mathcal{I}_{K_i}(g) \qquad \textit{for all } i = 1, 2, \ldots, M,$$

where $\mathcal{I}_{K_i}$ are (local) nodal interpolants corresponding to the finite elements K_1, K_2, ..., K_M.

The global nodal interpolant is obtained by constructing the local nodal interpolants separately in all elements in the mesh. Since the local interpolation procedures are decoupled, one can expect that the implication

$$g \in V \quad \Rightarrow \quad \mathcal{I}(g) \in V \tag{3.29}$$

may not always hold. This is illustrated in the following example.

■ EXAMPLE 3.3

Consider a pair of adjacent piecewise-affine elements $K_1 = (-1, 0)$ and $K_2 = (0, 1)$. For completeness let us mention that the corresponding polynomial spaces are $P_1 = P^1(K_1)$ and $P_2 = P^1(K_2)$, and the sets of degrees of freedom Σ_1 and Σ_2 comprise the linear forms

$$\begin{array}{rcl} L_1^{(1)}(g) & = & g(-1), \\ L_2^{(1)}(g) & = & g(0), \\ L_1^{(2)}(g) & = & g(0), \\ L_2^{(2)}(g) & = & g(1). \end{array}$$

respectively. It is easy to calculate (or see) that $\mathcal{B}_1 = \{-x, x+1\}$ and $\mathcal{B}_2 = \{1-x, x\}$ are the unique nodal bases of the elements K_1 and K_2.

Let us construct, for example, the global interpolant $\mathcal{I}(g)$ of the function

$$g(x) = x^3 \quad \in V,$$

where $V = H^1(-1, 1)$. The local interpolants in the elements K_1 and K_2 have the form

$$\mathcal{I}_{K_1}(g) = g(-1)(-x) + g(0)(x+1) = x$$

and

$$\mathcal{I}_{K_2}(g) = g(0)(1-x) + g(1)(x) = x,$$

respectively, and thus the global interpolant $\mathcal{I}(g) = x \in V$. The situation is depicted in Figure 3.8.

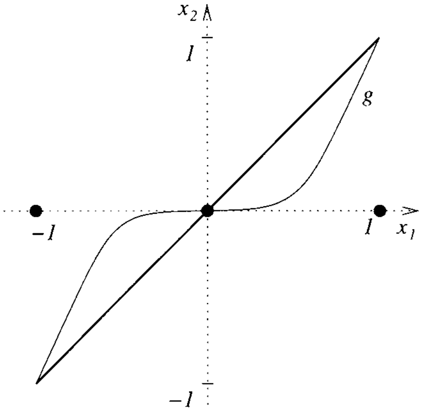


Figure 3.8 The implication (3.29) holds: the global interpolant $\mathcal{I}(g)$ lies in V.

Next define another pair of nodal elements with the same domains and polynomial spaces, but change the linear forms to

$$\begin{aligned} L_1^{(1)}(g) &= g(-2/3),\\ L_2^{(1)}(g) &= g(-1/3),\\ L_1^{(2)}(g) &= g(1/3),\\ L_2^{(2)}(g) &= g(2/3), \end{aligned}$$

respectively. The new nodal bases are $\mathcal{B}_1 = \{-3x - 1, 3x + 2\}$ and $\mathcal{B}_2 = \{2 - 3x, 3x - 1\}$. It can easily be calculated that the new local interpolants of the function $g(x) = x^3$ have the form

$$\mathcal{I}_{K_1}(g) = g(-2/3)(-3x - 1) + g(-1/3)(3x + 2) = \frac{7}{9}x + \frac{2}{9}$$

and

$$\mathcal{I}_{K_2}(g) = g(1/3)(2 - 3x) + g(2/3)(3x - 1) = \frac{7}{9}x - \frac{2}{9}.$$

In this case the global interpolant $\mathcal{I}(g)$ is discontinuous and thus it does not lie in V. This is depicted in Figure 3.9.

Example 3.3 suggests that condition (3.29) is important for nodal elements. Since point values of functions are generally not defined in Sobolev spaces (see Section A.4.4), it is practical to weaken condition (3.29) to hold in their dense subspaces only:

Definition 3.7 (Conformity of finite elements) *A finite element mesh $\mathcal{T}_{h,p}$ is said to be* conforming *to the space V if there exists a dense subspace $W \subset V$ such that*

$$\mathcal{I}(g) \in V \quad \textit{for all } g \in W. \tag{3.30}$$

Recall, for example, that the space of continuous functions $C(\overline{\Omega}_h)$ is dense in the Sobolev space $H^1(\Omega_h)$.

Remark 3.4 *Conforming elements are used more frequently than the nonconforming ones, since they better fit into the Galerkin framework. However, in special applications such as*

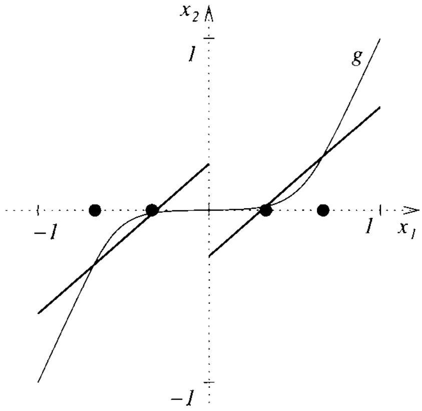


Figure 3.9 The global interpolant does not lie in the space V.

capturing of discontinuities or satisfying divergence or other constraints, nonconforming elements may perform better than the conforming ones. The Discontinuous Galerkin (DG) methods, for example, nowadays are very popular in computational PDEs.

3.3.3 Conformity to the Sobolev space H^1

Conformity requirements of the Sobolev space H^1 are formulated in the following lemma:

Lemma 3.4 (Conformity requirements of the space $H^1(\Omega_h)$) *Consider a bounded domain $\Omega_h \subset \mathbb{R}^d$ covered with a finite element mesh $\mathcal{T}_{h,p}$. A function $v : \Omega_h \to \mathbb{R}$ belongs to $H^1(\Omega_h)$ if and only if*

1. *$v|_K \in H^1(K)$ for each element $K \in \mathcal{T}_{h,p}$,*

2. *for each common face $f = \overline{K}_1 \cap \overline{K}_2$, $K_1, K_2 \in \mathcal{T}_{h,p}$ the trace of $v|_{K_1}$ and $v|_{K_2}$ on f is the same.*

Proof: For this proof we need to review some terminology related to weak derivatives (Paragraph A.4.2): By $\mathcal{D}(\Omega_h)$ we denote the space of infinitely smooth functions with compact support in Ω_h (distributions over Ω_h),

$$\mathcal{D}(\Omega_h) = \{\varphi \in C_0^\infty(\Omega_h);\ \mathrm{supp}(\varphi) \subset \Omega_h\},$$

where the support of a function $\varphi : \Omega_h \to \mathbb{R}$ is defined by

$$\mathrm{supp}(\varphi) = \overline{\{\boldsymbol{x} \in \Omega_h;\ \varphi(\boldsymbol{x}) \neq 0\}}.$$

Recall that since Ω_h is open and $\mathrm{supp}(\varphi)$ closed, the support cannot touch the boundary $\partial\Omega_h$. In other words, there must be a belt along the boundary $\partial\Omega_h$ where φ vanishes. We use the symbol $D^j(v)$ for $\partial v/\partial x_j$ in the sense of distributions (see Definition A.56).

Using *1.*, define the functions $w_j \in L^2(\Omega_h)$, $j = 1, 2, \ldots, d$ as

$$w_j|_K = D^j(v|_K)$$

for all $K \in \mathcal{T}_{h,p}$. We will show that $v \in H^1(\Omega_h)$ by verifying that $w_j = D^j v$.

Using Green's theorem (Theorem A.29), for every $\varphi \in \mathcal{D}(\Omega_h)$ we obtain

$$\int_{\Omega_h} w_j\varphi = \sum_{K\in\mathcal{T}_{h,p}} \int_K w_j\varphi = -\sum_K \int_K (v|_K) D^j\varphi + \sum_K \int_{\partial K} v|_K \varphi \boldsymbol{\nu}_{K,j},$$

where $\boldsymbol{\nu}_K$ is the outer unit normal vector to ∂K. Since φ vanishes on $\partial\Omega_h$ and $\boldsymbol{\nu}_{K_1} = -\boldsymbol{\nu}_{K_2} = \boldsymbol{\nu}$ on the common face f, by *2.* we have

$$\begin{aligned}\int_{\Omega_h} w_j\varphi &= -\int_{\Omega_h} vD^j\varphi + \sum_{f, f=\overline{K}_1\cap\overline{K}_2, K_1,K_2\in\mathcal{T}_{h,p}} \int_f (v|_{K_1} - v|_{K_2})\varphi\boldsymbol{\nu}_j \\ &= -\int_{\Omega_h} vD^j\varphi,\end{aligned}$$

and thus $w_j = D^j v$.

Conversely, if we assume that $v \in H^1(\Omega_h)$, it follows at once that *1.* holds. Using further $w_j = D^j v$, we obtain that

$$\sum_{f, f=\overline{K}_1\cap\overline{K}_2, K_1,K_2\in\mathcal{T}_{h,p}} \int_f (v|_{K_1} - v|_{K_2})\varphi\boldsymbol{\nu}_j = 0$$

for all $\varphi \in \mathcal{D}(\Omega_h)$, $j = 1, 2, \ldots, d$. Hence, *2.* is satisfied. ■

The conformity of meshes consisting of Lagrange Q^1- and P^1-elements to the Sobolev space H^1 will be discussed in Chapter 4. Let us close this chapter with the discussion of equivalence of nodal elements.

3.4 EQUIVALENCE OF NODAL ELEMENTS

Let us return to the one-dimensional Lagrange nodal elements for a moment again. Assume the reference domain $K_a = (-1, 1)$ and $p + 1$ disjoint points $-1 = y_1 < y_2 < \ldots < y_{p+1} = 1$. The polynomial space has the form $P = P^p(K_a)$ and the degrees of freedom are defined as $L_i(g) = g(y_i)$, $i = 1, 2, \ldots, p + 1$ for all $g \in P$.

Consider another interval $\tilde{K} \subset \mathbb{R}$ connected with K_a through the affine reference map (2.37), $x_{\tilde{K}} : K_a \to \tilde{K}$. We construct the Lagrange element $(\tilde{K}, \tilde{P}, \tilde{\Sigma})$ by defining new nodal points $\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_{p+1} \in \tilde{K}$, $\tilde{y}_i = x_{\tilde{K}}(y_i)$, new polynomial space $\tilde{P}$,

$$\tilde{P} = \{g \circ x_{\tilde{K}}^{-1};\ g \in P\},$$

and a new set of degrees of freedom $\tilde{\Sigma}$,

$$\tilde{L}_i(\tilde{g}) = \tilde{g}(\tilde{y}_i) \text{ for all } \tilde{g} \in \tilde{P}.$$

The affinity of the map $x_{\tilde{K}}$ implies that $\tilde{P} = P^p(\tilde{K})$. The degrees of freedom are invariant under the map $\Phi : P \to \tilde{P}$,

$$\Phi(g) = g \circ x_{\tilde{K}}^{-1},$$

in the sense that $g(y_i) = \tilde{g}(x_{\tilde{K}}(y_i))$. This means that

$$L_i(g) = \tilde{L}_i(\Phi(g)) \quad \text{for all } g \in P, \tag{3.31}$$

and we say that the Lagrange elements are equivalent under the map Φ. Since the underlying reference map $x_{\tilde{K}}$ is affine, sometimes one talks about affine-equivalence. The preservation of the degrees of freedom (3.31) allowed us to perform the complete finite element discretization on the reference interval K_a in Chapter 2. The next definition extends the notion of equivalence of one-dimensional Lagrange elements to general nodal elements:

Definition 3.8 (Equivalence of nodal elements) *Assume a pair of nodal finite elements* (K, P, Σ), $\Sigma = \{L_1, L_2, \ldots, L_{N_P}\}$ *and* $(\tilde{K}, \tilde{P}, \tilde{\Sigma})$, $\tilde{\Sigma} = \{\tilde{L}_1, \tilde{L}_2, \ldots, \tilde{L}_{N_P}\}$. *Let* $\Phi : P \longrightarrow \tilde{P}$ *be a bijection. We say that the elements are* equivalent *if*

$$\tilde{P} = \Phi(P), \tag{3.32}$$

and if the degrees of freedom satisfy

$$L_i(g) = \tilde{L}_i(\Phi(g)) \quad \textit{for all } g \in P \textit{ and } i = 1, 2, \ldots, N_P. \tag{3.33}$$

Notice that condition (3.32) includes the existence of a spatial bijection $x_{\tilde{K}} : K \longrightarrow \tilde{K}$.

■ EXAMPLE 3.4 (Equivalence of Lagrange elements on simplices in $\mathbb{R}^d$)

Let $(K, P^p(K), \Sigma)$ and $(\tilde{K}, P^p(\tilde{K}), \tilde{\Sigma})$ be a pair of unisolvent Lagrange nodal elements on simplices (i.e., K and $\tilde{K}$ are intervals in 1D, triangles in 2D or tetrahedra in 3D). Then there exists a bijective affine map $x_{\tilde{K}} : K \longrightarrow \tilde{K}$. The elements are equivalent if and only if the nodal points in K and $\tilde{K}$ are compatible under $x_{\tilde{K}}$.

■ EXAMPLE 3.5 (Elements containing DOF associated with derivatives)

Elements containing derivatives as DOF usually are not equivalent. Let us demonstrate this using a simple one-dimensional example. Consider two bounded intervals $K = (a, b)$ and $\tilde{K} = (\tilde{a}, \tilde{b})$ of different lengths $|K|$ and $|\tilde{K}|$. Define a nodal element (K, P, Σ) using the space $P = P^1(K)$ and the degrees of freedom

$$L_1(g) = g(a), \qquad L_2(g) = g'(b), \qquad g \in P.$$

Analogously the nodal element $(\tilde{K}, \tilde{P}, \tilde{\Sigma})$ is equipped with the space $\tilde{P} = P^1(\tilde{K})$ and the degrees of freedom

$$\tilde{L}_1(\tilde{g}) = \tilde{g}(\tilde{a}), \qquad \tilde{L}_2(\tilde{g}) = \tilde{g}'(\tilde{b}), \qquad g \in \tilde{P}.$$

The situation is depicted in Figure 3.10.

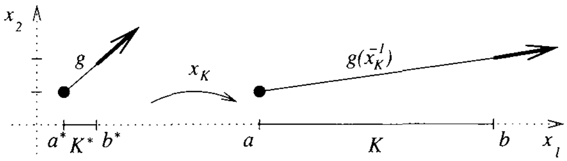


Figure 3.10 A pair of one-dimensional finite elements which are not equivalent: the black circles stand for DOF associated with function values and the arrows indicate DOF related to the derivatives.

Let $x_K : K \longrightarrow \tilde{K}$ be an affine reference map. Because of the presence of the Lagrange degrees of freedom L_1 and $\tilde{L}_1$ the linear operator $\Phi : P \longrightarrow \tilde{P}$ from Definition 3.8 must have the form

$$\Phi(g) = g \circ x_K^{-1}.$$

Then $L_1(g) = \tilde{L}_1(\Phi(g))$ for all $g \in P$. However, for the degrees of freedom L_2 and $\tilde{L}_2$ it holds

$$\tilde{L}_2(\Phi(g)) = \tilde{g}'(\tilde{b}) = \frac{|K|}{|\tilde{K}|} g'(b) \neq g'(b) = L_2(g),$$

and thus these elements are not equivalent.

3.5 EXERCISES

Exercise 3.1 *Prove Lemma 3.1.*

Exercise 3.2 *Prove Proposition 3.2.*

Exercise 3.3 *Prove Proposition 3.3.*

Exercise 3.4 *Prove Proposition 3.4.*

Exercise 3.5 *Write the explicit form of the affine inverse map $\boldsymbol{x}_K^{-1}$ from Proposition 3.3.*

Exercise 3.6 *Construct the explicit formula for the inverse of the biaffine map $\boldsymbol{x}_K : K_q \longrightarrow K$, corresponding to a convex quadrilateral with straight edges, whose nonpolynomial character was discussed in Corollary 3.1.*

Exercise 3.7 *Consider a regular finite element mesh $\mathcal{T}_{h,p}$ over a bounded domain $\Omega_h \subset \mathbb{R}^2$, consisting of a family of Q^1-elements constructed using the master Q^1-element on the reference domain K_q and the reference maps (3.13).*

1. *Is the finite element mesh conforming to the space $H^1(\Omega_h)$? Show in detail.*

2. *Show that Q^1-elements are equivalent under the map (3.13).*

Exercise 3.8 *Consider a Lagrange P^p-element on the reference triangular domain K_t. The polynomial space P is defined as $P = \mathrm{span}\{x_1^i x_2^j;\ 0 \le i + j \le p\}$, $p \ge 1$, and the set of degrees of freedom Σ consists of $N_P = (p+1)(p+2)/2$ linear forms $L_{kl}(g) = g(\boldsymbol{y}_{kl})$, where the N_P nodal points $\boldsymbol{y}_{kl}$ are defined by*

$$\boldsymbol{y}_{kl} = (-1 + 2k/p, -1 + 2l/p) \quad k = 0, 1, \ldots, p;\ l = 0, 1, p - k.$$

1. *Check the unisolvency of this finite element.*

2. *Construct the corresponding nodal basis.*

3. *Consider a mesh $\mathcal{T}_{h,p}$ consisting of a family of P^p-elements obtained using the master P^p-element on the reference domain K_t and the affine reference maps (3.21).*

(a) *Does this mesh conform to H^1? Show in detail.*

(b) *Show that Lagrange P^p-elements are affine-equivalent under the map (3.21).*

Exercise 3.9 *Consider a nodal finite element (K_q, P, Σ) on the reference square domain K_q with $P = \text{span}\{1, x_1, x_2, x_1^2 - x_2^2\}$ and $\Sigma = \{L_1, L_2, L_3, L_4\}$, where*

$$\begin{aligned} L_1(g) &= g(-1,0), \\ L_2(g) &= g(1,0), \\ L_3(g) &= g(0,-1), \\ L_4(g) &= g(0,1). \end{aligned}$$

1. *Check the unisolvency of this finite element.*

2. *Construct the corresponding nodal basis (if relevant).*

3. *Write the formula for the local element interpolant $\mathcal{I}_{K_q}(g)$, and apply it to the function $g(\boldsymbol{x}) = \cos(\pi(x_1 + x_2)) \in H^1(K_q)$. Present plots of both g and $\mathcal{I}_{K_q}(g)$.*

4. *Consider a finite element mesh $\mathcal{T}_{h,p}$ consisting of a family of such elements obtained using the reference maps (3.13).*

   (a) *Are these elements equivalent under the map (3.13)?*

   (b) *Does the mesh $\mathcal{T}_{h,p}$ conform to the space $H^1(\Omega_h)$? Show in detail.*

Exercise 3.10 *Consider a nodal finite element (K_q, P, Σ) on the reference square domain K_q with $P = \text{span}\{1, x_1, x_2, x_1^2 - x_2^2\}$ and $\Sigma = \{L_1, L_2, \ldots, L_4\}$, where*

$$\begin{aligned} L_1(g) &= \int_{-1}^{1} g(-1, x_2)\, dx_2, \\ L_2(g) &= \int_{-1}^{1} g(1, x_2)\, dx_2, \\ L_3(g) &= \int_{-1}^{1} g(x_1, -1)\, dx_1, \\ L_4(g) &= \int_{-1}^{1} g(x_1, 1)\, dx_1. \end{aligned}$$

1. *Check the unisolvency of this finite element.*

2. *Construct the corresponding nodal basis (if relevant).*

3. *Write the formula for the local element interpolant $\mathcal{I}_{K_q}(g)$, and apply it to the function $g(\boldsymbol{x}) = \cos(\pi(x_1 + x_2)) \in H^1(K_q)$. Present plots of both g and $\mathcal{I}_{K_q}(g)$.*

4. *Consider a finite element mesh $\mathcal{T}_{h,p}$ consisting of a family of such elements obtained using the reference maps (3.13).*

   (a) *Are these elements equivalent under the map (3.13)?*

   (b) *Does the mesh $\mathcal{T}_{h,p}$ conform to the space $H^1(\Omega_h)$? Show in detail.*

Exercise 3.11 *Consider a nodal finite element* (K_q, P, Σ) *on the reference square domain* K_q *with* $P = \text{span}\{1, x_1, x_2, x_1^2, x_1x_2, x_2^2, x_1x_2^2, x_1^2x_2\}$ *and* $\Sigma = \{L_1, L_2, \ldots, L_8\}$, *where*

$$\begin{aligned} L_1(g) &= g(-1,-1), \ L_2(g) = g(1,-1), \\ L_3(g) &= g(1,1), \quad L_4(g) = g(-1,1), \\ L_5(g) &= \int_{-1}^{1} g(-1,x_2)\,dx_2, \ L_6(g) = \int_{-1}^{1} g(1,x_2)\,dx_2, \\ L_7(g) &= \int_{-1}^{1} g(x_1,-1)\,dx_1, \ L_8(g) = \int_{-1}^{1} g(x_1,1)\,dx_1. \end{aligned}$$

1. *Check the unisolvency of this finite element.*
2. *Construct the corresponding nodal basis (if relevant).*
3. *Write the formula for the local element interpolant* $\mathcal{I}_{K_q}(g)$, *and apply it to the function* $g(\boldsymbol{x}) = \cos(\pi(x_1 + x_2)) \in H^1(K_q)$. *Present plots of both g and* $\mathcal{I}_{K_q}(g)$.
4. *Consider a finite element mesh* $\mathcal{T}_{h,p}$ *consisting of a family of such elements obtained using the reference maps (3.13).*
   - *(a) Are these elements equivalent under the map (3.13)?*
   - *(b) Does the mesh* $\mathcal{T}_{h,p}$ *conform to the space* $H^1(\Omega_h)$? *Show in detail.*

Exercise 3.12 *Consider a bounded one-dimensional domain* $\Omega = (a, b)$ *covered with a finite element mesh* $\mathcal{T}_{h,p}$ *consisting of M cubic Hermite elements* (K_i, P_i, Σ_i), $K_i = (x_{i-1}, x_i)$, $i = 1, 2, \ldots, M$. *The set of degrees of freedom* Σ_i *is defined as*

$$\begin{aligned} L_1^{(i)}(g) &= g(x_{i-1}), \\ L_2^{(i)}(g) &= g(x_i), \\ L_3^{(i)}(g) &= g'(x_{i-1}), \\ L_4^{(i)}(g) &= g'(x_i), \end{aligned}$$

for all $i = 1, 2, \ldots, M$ *and* $g \in P_i$.

1. *Find the minimum admissible polynomial degree* p_0 *for these elements.*
2. *Let* $P_i = P^{p_0}(K_i)$ *for all* $i = 1, 2, \ldots, M$. *Decide whether the elements are or are not unisolvent. Show in detail.*
3. *Construct a nodal basis* $\mathcal{B}_i$ *for every element* (K_i, P_i, Σ_i).
4. *Write the local element interpolants and the global interpolant.*
5. *Does the finite element mesh* $\mathcal{T}_{h,p}$ *conform to the space* $H^2(\Omega)$? *Show in detail. Hint: The* H^2*-conformity requirement in 1D is once-continuous differentiability.*
6. *Consider the space*

   $$V_{h,p} = \{v \in C^1(\Omega) \cap C(\overline{\Omega});\ v|_{K_i} \in P_i \text{ for all } i = 1, 2, \ldots, M\}.$$

   What is the dimension $N = dim(V_{h,p})$?
7. *Use the nodal basis functions on every element to design N suitable basis functions of the space* $V_{h,p}$. *Remember that every basis function has to be once continuously differentiable to lie in* $H^2(\Omega)$.

Partial Differential Equations and the Finite Element Method
by Pavel Šolín
Copyright © 2006 John Wiley & Sons, Inc.

CHAPTER 4

CONTINUOUS ELEMENTS FOR 2D PROBLEMS

After learning about the general concept of nodal finite elements in Chapter 3, the reader should know how to design general nodal finite elements of the form (K, P, Σ), and be able to perform the following operations:

- check the unisolvency of the element (K, P, Σ),
- construct the unique set of nodal shape functions $\theta_1, \theta_2, \ldots, \theta_{N_P}$ satisfying the delta property (3.4),
- use the set of degrees of freedom Σ and the nodal shape functions $\theta_1, \theta_2, \ldots, \theta_{N_P}$ to construct the local interpolant $\mathcal{I}_K$,
- construct the global interpolant $\mathcal{I}$ on a given finite element mesh and check whether or not the mesh conforms to a given space of functions,
- analyze the equivalence of nodal elements defined on different domains K and $\tilde{K}$.

In this chapter we apply these techniques to continuous finite elements for second-order PDEs in two spatial dimensions, extending the knowledge of one-dimensional continuous finite elements acquired in Chapter 2. The lowest-order Q^1/P^1-elements are introduced in Section 4.1. In Section 4.2 we discuss higher-order Gaussian quadrature in 2D. After that, the Q^1/P^1-elements are extended to higher-order Lagrange nodal elements in Section 4.3.

4.1 LOWEST-ORDER ELEMENTS

In this section, after introducing a suitable model problem and its weak formulation, we show in Paragraph 4.1.2 the sequence of geometrical and functional approximation needed to transform a PDE problem into a discrete finite element problem and we derive the approximate weak formulation of the model problem. The lowest-order basis functions of the finite element space $V_{h,p}$ are presented in Paragraph 4.1.3, and the weak formulation is transformed to the reference domains in Paragraph 4.1.4. Paragraph 4.1.5 is devoted to the constant coefficient case when precomputed template mass and stiffness integrals can be used. Paragraphs 4.1.6 and 4.1.7 discuss the data structures and implementation, and the section is closed with describing the interpolation on the lowest-order Q^1/P^1-meshes in Paragraph 4.1.6.

4.1.1 Model problem and its weak formulation

Consider a two-dimensional bounded domain Ω with a Lipschitz-continuous boundary $\partial\Omega$. Suppose that $\partial\Omega$ consists of two disjoint open parts Γ_D and Γ_N such that

$$\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N, \tag{4.1}$$

as illustrated in Figure 4.1.

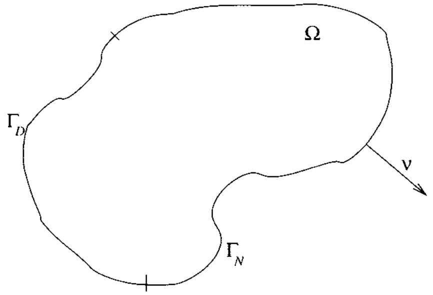


Figure 4.1 The domain Ω, its boundary $\partial\Omega$, and the unit outer normal vector ν to $\partial\Omega$.

Assume again the model equation (1.26),

$$-\nabla \cdot (a_1 \nabla u) + a_0 u = f \qquad \text{in } \Omega, \tag{4.2}$$

with a Dirichlet boundary condition

$$u(\boldsymbol{x}) = g_D(\boldsymbol{x}) \qquad \text{for all } \boldsymbol{x} \in \Gamma_D, \tag{4.3}$$

and a Neumann boundary condition

$$\frac{\partial u}{\partial \nu}(\boldsymbol{x}) = g_N(\boldsymbol{x}) \qquad \text{for all } \boldsymbol{x} \in \Gamma_N. \tag{4.4}$$

The existence of a unique solution is guaranteed if $\Gamma_D \neq \emptyset$ and

$$a_1(\boldsymbol{x}) \geq C_{min} > 0 \qquad \text{and} \qquad a_0(\boldsymbol{x}) \geq 0 \qquad \text{in } \Omega,$$

or if

$$a_1(\boldsymbol{x}) \geq C_{min} > 0 \quad \text{and} \quad a_0(\boldsymbol{x}) \geq \hat{C}_{min} > 0 \quad \text{in } \Omega,$$

in which case the Dirichlet part of the boundary can be empty.

In the following we assume that the coefficient functions a_1 and a_0 are constant; the extension to L^∞-functions is done analogously to Paragraph 2.2.1.

Weak formulation The weak formulation of this problem was discussed in detail in Paragraph 1.2.8. Hence consider some Dirichlet lift function $G \in H^1(\Omega)$ of the boundary data g_D, and look for the solution u in the form $u = U + G$. The new unknown U lies in the space (1.65),

$$V = \{v \in H^1(\Omega);\ v = 0 \text{ on } \Gamma_D\}. \tag{4.5}$$

The task is to find $U \in V$ such that

$$a(U, v) = l(v) \qquad \text{for all } v \in V, \tag{4.6}$$

where

$$\begin{aligned} a(U,v) &= \int_\Omega (a_1 \nabla U \cdot \nabla v + a_0 U v)\,\mathrm{d}\boldsymbol{x}, \qquad U, v \in V, \\ l(v) &= \int_\Omega (fv - a_1 \nabla G \cdot \nabla v - a_0 G v)\,\mathrm{d}\boldsymbol{x} + \int_{\Gamma_N} a_1 g_N v\,\mathrm{d}\boldsymbol{S}, \qquad v \in V. \end{aligned} \tag{4.7}$$

4.1.2 Approximations and variational crimes

Now let us go through the series of geometrical and functional approximation steps that turn the infinite-dimensional problem (4.6) into a finite-dimensional discrete problem of the form $\boldsymbol{SY} = \boldsymbol{F}$. At some points this requires a departure from the "mathematically clean" variational framework. Such operations are called variational crimes, and in practice it is not really possible to avoid them.

Step 1: Approximation of the domain Ω The domain Ω is approximated by a polygonal domain Ω_h, as shown in Figure 4.2.

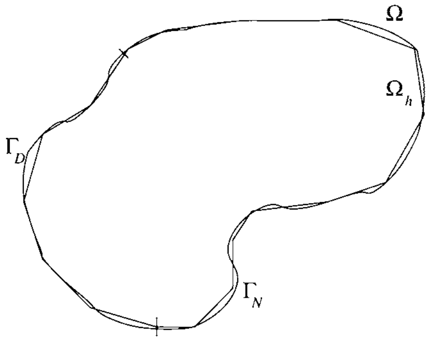


Figure 4.2 Polygonal approximation Ω_h of the domain Ω. Generally $\Omega_h \neq \Omega$ and even $\Omega_h \not\subset \Omega$.

If $\Omega_h \not\subset \Omega$, then the solution and other functions from the weak formulation (4.6) are not defined where they are to be approximated or evaluated. This is the first variational crime. If the boundary $\partial\Omega$ is piecewise-polynomial, then its approximation can be done exactly using curvilinear elements (see, e.g., [111]).

Step 2: Finite element mesh Assume the domain Ω_h be covered with a regular finite element mesh $\mathcal{T}_{h,p}$ (Definition 3.5) consisting of M nonoverlapping finite elements $K_1, K_2, \ldots, K_M$. Let all the elements be given the same polynomial degree $p = 1$. The discretization on irregular meshes is described, e.g., in [111]. Figure 4.3 shows examples of regular meshes on the domain Ω_h. The mesh is called hybrid when it contains both triangular and quadrilateral elements.

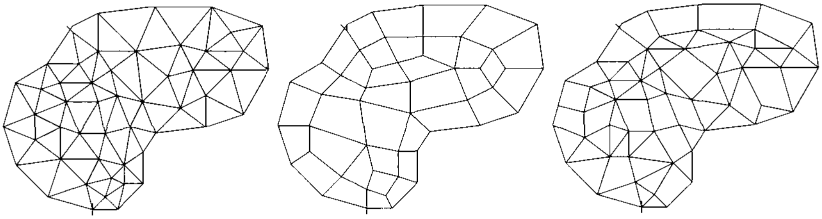


Figure 4.3 Regular triangular, quadrilateral and hybrid finite element meshes on Ω_h.

In order to facilitate the implementation, it is natural to require that the points $\overline{\Gamma}_D \cap \overline{\Gamma}_N$ coincide with some vertices of the mesh $\mathcal{T}_{h,p}$.

Step 3: Approximation of boundary conditions After replacing the original domain Ω by its polygonal approximation Ω_h, one loses the boundaries Γ_D and Γ_N, where the Dirichlet and Neumann boundary conditions were prescribed. The boundary conditions (4.3) and (4.4) have to be transferred in some suitable way to the polygonal parts $\Gamma_{D,h}$ and $\Gamma_{N,h}$ of the new boundary $\partial\Omega_h$. What one usually does is to define new boundary conditions by

$$\begin{aligned} u(\boldsymbol{x}) &= g_D(\boldsymbol{x}) \quad \text{for all } \boldsymbol{x} \in \Gamma_{D,h}, \\ \frac{\partial u}{\partial \nu}(\boldsymbol{x}) &= g_N(\boldsymbol{x}) \quad \text{for all } \boldsymbol{x} \in \Gamma_{N,h}. \end{aligned} \tag{4.8}$$

This is another variational crime, since the functions g_D and g_N are evaluated where they were not defined. Usually this goes through in the implementation, but it should be checked how much this approximation violates the underlying physical problem.

Step 4: Approximation of the space V According to the geometrical approximation $\Omega_h \approx \Omega$ from Step 1, the space $V(\Omega)$ is approximated by a piecewise- polynomial space $V_{h,p}(\Omega_h)$,

$$\begin{aligned} V_{h,p} = \{ & v \in C(\Omega_h);\ v_{h,p}|_{\Gamma_{D,h}} = 0; \\ & v|_{K_i} \in P^p(K_i) \text{ if } K_i \text{ is a triangle}, \\ & v|_{K_i} \in Q^p(K_i) \text{ if } K_i \text{ is a quadrilateral}\}. \end{aligned} \tag{4.9}$$

This also is a variational crime, since the Galerkin method does not admit a situation when $V_{h,p} \not\subset V$.

Step 5: Approximate weak formulation The discrete problem can be formulated after the Dirichlet lift $G \in H^1(\Omega)$ is "approximated" with a function $G_{h,p} \in H^1(\Omega_h)$. The approximate weak formulation of the model problem reads:

Find a function $U_{h,p} \in V_{h,p}$ such that the identity

$$\int_{\Omega_h} (a_1 \nabla U_{h,p} \cdot \nabla v + a_0 U_{h,p} v)\, \mathrm{d}\boldsymbol{x} \tag{4.10}$$
$$= \int_{\Omega_h} (fv - a_1 \nabla G_{h,p} \cdot \nabla v - a_0 G_{h,p} v)\, \mathrm{d}\boldsymbol{x} + \int_{\Gamma_{N,h}} a_1 g_N v \, \mathrm{d}\boldsymbol{S},$$

holds for all $v \in V_{h,p}$. As we mentioned before, the load function $f \in L^2(\Omega)$ as well as the coefficients $a_1, a_0 \in L^\infty(\Omega)$ are evaluated in the domain Ω_h where they may not be defined if $\Omega_h \not\subset \Omega$.

Step 6: The system of linear algebraic equations As usual, the unknown function $U_{h,p} \in V_{h,p}$ is expressed as a linear combination of some N basis functions $v_1, v_2, \dots, v_N \in V_{h,p}$ (a standard choice will be mentioned in Paragraph 4.1.3), with unknown coefficients $y_1, y_2, \dots, y_N$,

$$U_{h,p} = \sum_{j=1}^{N} y_j v_j. \tag{4.11}$$

Testing (4.17) by the basis functions v_i, $i = 1, 2, \dots, N$, one obtains a system of N linear algebraic equations,

$$\sum_{j=1}^{N} (y_j \int_{\Omega_h} a_1 \nabla v_j \cdot \nabla v_i + a_0 v_j v_i)\, \mathrm{d}\boldsymbol{x} \tag{4.12}$$
$$= \int_{\Omega_h} (f v_i - a_1 \nabla G_{h,p} \cdot \nabla v_i - a_0 G_{h,p} v_i)\, \mathrm{d}\boldsymbol{x} + \int_{\Gamma_{N,h}} a_1 g_N v_i \, \mathrm{d}\boldsymbol{S}$$

for all $i = 1, 2, \dots, N$. The system can be written in the matrix form (2.13),

$$\boldsymbol{S}\boldsymbol{Y} = \boldsymbol{F}, \tag{4.13}$$

where $\boldsymbol{S} \in \mathbb{R}^{N \times N}$ is the stiffness matrix, $\boldsymbol{Y} \in \mathbb{R}^N$ the vector of unknown coefficients and $\boldsymbol{F} \in \mathbb{R}^N$ the load vector.

In order to assemble the system (4.13), one needs to construct suitable basis functions $v_1, v_2, \dots, v_N$ of the space $V_{h,p}$. Let us do this in the next paragraph.

4.1.3 Basis of the space $V_{h,p}$

Assume a regular finite element mesh $\mathcal{T}_{h,p}$ consisting of M_q Q^1-elements and M_p P^1-elements, where $M_q + M_p = M \geq 1$. By $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N$ denote the N grid vertices that do not lie on the Dirichlet part $\Gamma_{D,h}$ of the boundary $\partial\Omega_h$ (we say that these vertices are unconstrained).

Proposition 4.1 *The dimension of the space $V_{h,p}$ is equal to N, where N is the number of unconstrained grid vertices.*

Proof: This follows easily from the definition (4.9) of the space $V_{h,p}$. ■

Because of the one-to-one relationship of the basis functions v_i and the unconstrained grid vertices $\boldsymbol{x}_i$, the lowest-order basis functions are called vertex functions. They have the form of "pyramids" shown in Figure 4.4, that naturally generalize the one-dimensional "hat functions" (2.25).

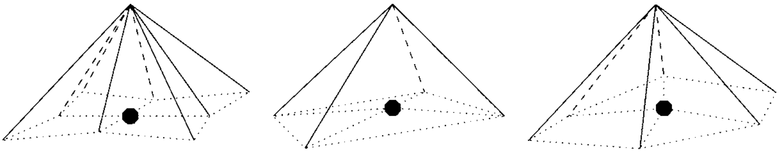


Figure 4.4 Vertex basis functions of the space $V_{h,p}$ on meshes consisting of Q^1- and P^1-elements.

These functions are defined as follows: Assume a vertex patch $S(i)$ consisting of all mesh triangles or quadrilaterals sharing the vertex $\boldsymbol{x}_i$,

$$S(i) = \bigcup_{k \in N(i)} \overline{K}_k, \tag{4.14}$$

where the index set $N(i)$ is defined as

$$N(i) = \{k;\ K_k \in \mathcal{T}_{h,p},\ \boldsymbol{x}_i \text{ is a vertex of } K_k\}. \tag{4.15}$$

The vertex function v_i is defined to be zero in $\Omega_h \setminus S(i)$, and in $S(i)$ it has the form

$$\begin{aligned} v_i(\boldsymbol{x})|_{K_k} &= (\varphi_q^{v_r} \circ \boldsymbol{x}_{K_k}^{-1})(\boldsymbol{x}) \quad \text{if } K_k \in S(i) \text{ is a quadrilateral,} \\ v_i(\boldsymbol{x})|_{K_k} &= (\varphi_t^{v_r} \circ \boldsymbol{x}_{K_k}^{-1})(\boldsymbol{x}) \quad \text{if } K_k \in S(i) \text{ is a triangle.} \end{aligned} \tag{4.16}$$

Here for every element $K_k \in S(i)$, $\varphi_q^{v_r}$ or $\varphi_t^{v_r}$ is the unique vertex nodal shape function on the reference domain K_q or K_t, respectively, such that $\varphi_q^{v_r}(\boldsymbol{x}_{K_k}^{-1}(\boldsymbol{x}_i)) = 1$ or $\varphi_t^{v_r}(\boldsymbol{x}_{K_k}^{-1}(\boldsymbol{x}_i)) = 1$.

Remark 4.1 *The reader does not have to worry about the presence of the inverse reference map in the relations (4.16), since the inverse map is not used explicitly in the computer code. All operations of the element-by-element loop will be performed on the reference domains, using suitable connectivity arrays. This will be discussed in Paragraph 4.1.7.*

Proposition 4.2 *For every unconstrained grid vertex $\boldsymbol{x}_i$, the corresponding function (4.16) is continuous in Ω_h. The functions $v_1, v_2, \ldots, v_N$ form a basis of the space $V_{h,p}$, and satisfy the delta property*

$$v_i(\boldsymbol{x}_j) = \delta_{ij}, \quad 1 \le i, j \le N.$$

Proof: The first part follows from the affinity of the functions v_i along edges in the patch $S(i)$. The rest is obvious from the construction. ■

4.1.4 Transformation of weak forms to the reference domain

The idea of the assembling algorithm is analogous to the one-dimensional case discussed in Chapter 2. Again, both the global stiffness matrix $\boldsymbol{S}$ and the right-hand side vector $\boldsymbol{F}$ will be filled in an element-by-element fashion. Therefore it is convenient to view identity (4.12) as a sum over all elements K_m, $m = 1, 2, \ldots, M$:

$$\sum_{j=1}^{N} y_j \sum_{m=1}^{M} \int_{K_m} (a_1 \nabla v_j \cdot \nabla v_i + a_0 v_j v_i)\, \mathrm{d}\boldsymbol{x} \tag{4.17}$$

$$= \sum_{m=1}^{M} \int_{K_m} (f v_i - a_1 \nabla G_{h,p} \cdot \nabla v_i - a_0 G_{h,p} v_i)\, \mathrm{d}\boldsymbol{x} + \sum_{m=1}^{M} \int_{\Gamma_{N,h} \cap \overline{K}_m} a_1 g_N v_i \mathrm{d}S,$$

to be satisfied for all basis functions v_i, $i = 1, 2, \ldots, N$.

First let us transform the element integrals from (4.17) to the appropriate reference domain, which is either K_q or K_t. Since the transformation on quadrilateral and triangular elements is analogous, it is sufficient to discuss, for example, the triangular case.

Transformation of function values A function $w(\boldsymbol{x}) \in C(K_m)$, $1 \leq m \leq M$, is transformed to the reference domain K_t in virtue of the affine reference map (3.21), $\boldsymbol{x}_{K_m}(\boldsymbol{\xi}) = (x_{K_m,1}(\boldsymbol{\xi}), x_{K_m,2}(\boldsymbol{\xi}))$, as follows:

$$\tilde{w}^{(m)}(\boldsymbol{\xi}) = (w \circ \boldsymbol{x}_{K_m})(\boldsymbol{\xi}) = w(x_{K_m,1}(\boldsymbol{\xi}), x_{K_m,2}(\boldsymbol{\xi})). \tag{4.18}$$

Transformation of partial derivatives This is a good exercise for the chain rule of differentiation. Assume that $w \in C^1(K_m)$. The partial derivatives of $\tilde{w}^{(m)}(\boldsymbol{\xi}) = (w \circ \boldsymbol{x}_{K_m})$ have the form

$$\begin{aligned}
\frac{\partial \tilde{w}^{(m)}}{\partial \xi_1}(\boldsymbol{\xi}) &= \frac{\partial w}{\partial x_1}|_{\boldsymbol{x}=\boldsymbol{x}_{K_m}(\boldsymbol{\xi})} \frac{\partial x_{K_m,1}}{\partial \xi_1}(\boldsymbol{\xi}) + \frac{\partial w}{\partial x_2}|_{\boldsymbol{x}=\boldsymbol{x}_{K_m}(\boldsymbol{\xi})} \frac{\partial x_{K_m,2}}{\partial \xi_1}(\boldsymbol{\xi}), \quad (4.19)\\
\frac{\partial \tilde{w}^{(m)}}{\partial \xi_2}(\boldsymbol{\xi}) &= \frac{\partial w}{\partial x_1}|_{\boldsymbol{x}=\boldsymbol{x}_{K_m}(\boldsymbol{\xi})} \frac{\partial x_{K_m,1}}{\partial \xi_2}(\boldsymbol{\xi}) + \frac{\partial w}{\partial x_2}|_{\boldsymbol{x}=\boldsymbol{x}_{K_m}(\boldsymbol{\xi})} \frac{\partial x_{K_m,2}}{\partial \xi_2}(\boldsymbol{\xi}).
\end{aligned}$$

This can be written as

$$\begin{pmatrix} \dfrac{\partial \tilde{w}^{(m)}}{\partial \xi_1} \\ \dfrac{\partial \tilde{w}^{(m)}}{\partial \xi_2} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial x_{K_m,1}}{\partial \xi_1} & \dfrac{\partial x_{K_m,2}}{\partial \xi_1} \\ \dfrac{\partial x_{K_m,1}}{\partial \xi_2} & \dfrac{\partial x_{K_m,2}}{\partial \xi_2} \end{pmatrix} \begin{pmatrix} \dfrac{\partial w}{\partial x_1} \\ \dfrac{\partial w}{\partial x_2} \end{pmatrix} = \left(\frac{D\boldsymbol{x}_{K_m}}{D\boldsymbol{\xi}} \right)^T \begin{pmatrix} \dfrac{\partial w}{\partial x_1} \\ \dfrac{\partial w}{\partial x_2} \end{pmatrix},$$

where $D\boldsymbol{x}_{K_m}/D\boldsymbol{\xi}$ stands for the Jacobi matrix of the map $\boldsymbol{x}_{K_m}$. Recall that nonsingular matrices satisfy

$$\left(\boldsymbol{A}^T\right)^{-1} = \left(\boldsymbol{A}^{-1}\right)^T = \boldsymbol{A}^{-T}.$$

Thus the gradient $\nabla w(\boldsymbol{x})$ at an arbitrary point $\boldsymbol{x} \in K_m$ is transformed to the point $\boldsymbol{\xi} = \boldsymbol{x}_{K_m}^{-1}(\boldsymbol{x}) \in K_t$ as follows,

$$\nabla w(\boldsymbol{x}) = \left(\frac{D\boldsymbol{x}_{K_m}}{D\boldsymbol{\xi}} \right)^{-T} \nabla \tilde{w}^{(m)}(\boldsymbol{\xi}). \tag{4.20}$$

According to (4.20), the stiffness term in (4.17) transforms from a mesh element K_m to the reference domain $\hat{K}$ as

$$\int_{K_m} \left[a_1(\boldsymbol{x}) \nabla v_j(\boldsymbol{x}) \cdot \nabla v_i(\boldsymbol{x}) + a_0(\boldsymbol{x}) v_j(\boldsymbol{x}) v_i(\boldsymbol{x}) \right] \, \mathrm{d}\boldsymbol{x} \tag{4.21}$$

$$= \int_{\hat{K}} J_{K_m} \left[\tilde{a}_1^{(m)}(\boldsymbol{\xi}) \left(\frac{D\boldsymbol{x}_{K_m}}{D\boldsymbol{\xi}} \right)^{-T} \nabla \tilde{v}_j^{(m)}(\boldsymbol{\xi}) \right] \cdot \left[\left(\frac{D\boldsymbol{x}_{K_m}}{D\boldsymbol{\xi}} \right)^{-T} \nabla \tilde{v}_i^{(m)}(\boldsymbol{\xi}) \right] \mathrm{d}\boldsymbol{\xi}$$

$$+ \int_{\hat{K}} \left[J_{K_m} \tilde{a}_0^{(m)}(\boldsymbol{\xi}) \tilde{v}_j^{(m)}(\boldsymbol{\xi}) \tilde{v}_i^{(m)}(\boldsymbol{\xi}) \right] \mathrm{d}\boldsymbol{\xi},$$

where

$$\tilde{v}_i^{(m)}(\boldsymbol{\xi}) = (v_i \circ \boldsymbol{x}_{K_m})(\boldsymbol{\xi}), \quad \tilde{v}_j^{(m)}(\boldsymbol{\xi}) = (v_j \circ \boldsymbol{x}_{K_m})(\boldsymbol{\xi}),$$

and

$$\tilde{a}_1^{(m)}(\boldsymbol{\xi}) = (a_1 \circ \boldsymbol{x}_{K_m})(\boldsymbol{\xi}), \quad \tilde{a}_0^{(m)}(\boldsymbol{\xi}) = (a_0 \circ \boldsymbol{x}_{K_m})(\boldsymbol{\xi}).$$

Since the reference map $\boldsymbol{x}_{K_m}$ in the triangular case is affine, the Jacobian

$$J_{K_m}(\boldsymbol{\xi}) = \det \left(\frac{D\boldsymbol{x}_{K_m}}{D\boldsymbol{\xi}} \right)$$

is constant, and without loss of generality, we can assume that it is positive: This is the case when $\boldsymbol{x}_{K_m}$ does not change the orientation of edges between the reference domain and the mesh element. In the quadrilateral case the Jacobian and both the Jacobi matrices on the right-hand side of (4.21) generally are not constant and have to be integrated numerically.

Remark 4.2 (Explicit inversion of 2×2 matrices) *The inverse of nonsingular 2×2 matrices can be done without the Gauss elimination procedure, using an explicit expansion formula*

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

In the case of the Jacobi matrix $D\boldsymbol{x}_{K_m}/D\boldsymbol{\xi}$ the denominator (which is the Jacobian J_{K_m}) cannot be zero since the map $\boldsymbol{x}_{K_m}$ is a bijection.

Denoting the constant entries of the inverse Jacobi matrix by

$$\left(\frac{\mathrm{D}\boldsymbol{x}_{K_m}}{\mathrm{D}\boldsymbol{\xi}} \right)^{-1} = \left\{ \frac{\partial \xi_r^{(m)}}{\partial x_n} \right\}_{r,n=1}^{d}, \tag{4.22}$$

where $d = 2$ is the spatial dimension, and assuming the constantness and positivity of J_{K_m}, one can rewrite (4.21) into

$$
\begin{aligned}
&\int_{K_m} \left[a_1(\boldsymbol{x})\nabla v_j(\boldsymbol{x}) \cdot \nabla v_i(\boldsymbol{x}) + a_0(\boldsymbol{x}) v_j(\boldsymbol{x}) \nabla v_i(\boldsymbol{x})\right] \, \mathrm{d}\boldsymbol{x} \qquad (4.23)\\
&= \sum_{n=1}^{d} \int_{\hat K} J_{K_m} \tilde a_1^{(m)}(\boldsymbol{\xi}) \left(\sum_{r=1}^{d} \frac{\partial \tilde v_j^{(m)}}{\partial \xi_r} \frac{\partial \xi_r^{(m)}}{\partial x_n} \right) \left(\sum_{s=1}^{d} \frac{\partial \tilde v_i^{(m)}}{\partial \xi_s} \frac{\partial \xi_s^{(m)}}{\partial x_n} \right) \mathrm{d}\boldsymbol{\xi}\\
&\quad + \int_{\hat K} J_{K_m} \tilde a_0^{(m)}(\boldsymbol{\xi}) \tilde v_j^{(m)}(\boldsymbol{\xi}) \tilde v_i^{(m)}(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi}.
\end{aligned}
$$

4.1.5 Simplified evaluation of stiffness integrals

Repeated numerical integration in (4.23) on every mesh element K_m, $m = 1, 2, \ldots, M$, can be avoided if the following two conditions are met:

1. The reference map $\boldsymbol{x}_{K_m}$ is affine. This is the case when either

   (a) the P^p-element K_m is triangular with straight edges

   or

   (b) the Q^p-element K_m is a rectangle.

2. All coefficients of the elliptic operator (1.1) are constant.

Then (4.23) can be simplified to

$$
\begin{aligned}
&\int_{K_m} \left[a_1 \nabla v_j(\boldsymbol{x}) \cdot \nabla v_i(\boldsymbol{x}) + a_0 v_j(\boldsymbol{x}) \nabla v_i(\boldsymbol{x})\right] \, \mathrm{d}\boldsymbol{x} \qquad (4.24)\\
&= J_{K_m} a_1 \sum_{n=1}^{d} \sum_{r=1}^{d} \frac{\partial \xi_r^{(m)}}{\partial x_n} \sum_{s=1}^{d} \frac{\partial \xi_s^{(m)}}{\partial x_n} \int_{\hat K} \frac{\partial \tilde v_j^{(m)}}{\partial \xi_r} \frac{\partial \tilde v_i^{(m)}}{\partial \xi_s} \, \mathrm{d}\boldsymbol{\xi}\\
&\quad + J_{K_m} a_0 \int_{\hat K} \tilde v_j^{(m)} \tilde v_i^{(m)} \, \mathrm{d}\boldsymbol{\xi},
\end{aligned}
$$

where $\hat K$ is either K_t or K_q. Hence for all $m = 1, 2, \ldots, M$, the stiffness terms (4.24) can be evaluated elementwise using the precomputed constants

$$
J_{K_m}, \quad \frac{\partial \xi_r^{(m)}}{\partial x_n}, \qquad 1 \le n, r \le 2, \tag{4.25}
$$

corresponding to the reference maps $\boldsymbol{x}_{K_m}$, and a few precomputed master element stiffness integrals of the form

$$
\int_{\hat K} \frac{\partial \tilde v_j^{(m)}}{\partial \xi_r} \frac{\partial \tilde v_i^{(m)}}{\partial \xi_s} \, \mathrm{d}\boldsymbol{\xi} = \int_{\hat K} \frac{\partial \varphi^{v_l}}{\partial \xi_r} \frac{\partial \varphi^{v_k}}{\partial \xi_s} \, \mathrm{d}\boldsymbol{\xi}. \tag{4.26}
$$

Here $\varphi^{v_k}, \varphi^{v_l}$ are the shape functions (3.20) defined on the reference domain $\hat K = K_t$, and $1 \le s, r \le d$. Appropriate shape functions are linked to the transformed basis functions $\tilde v_j^{(m)}$ and $\tilde v_i^{(m)}$ via connectivity arrays.

4.1.6 Connectivity arrays

Analogously to the one-dimensional case, the connectivity arrays lie at the heart of the element-by-element assembling algorithm. Assume that the hybrid Q^1/P^1 mesh $\mathcal{T}_{h,p}$ is represented via an element array

```
Element *Elem;
```

of the length M.

Element data structure The basic element data structure may be defined as follows:

```
struct {
  int nv;              //number of vertices
                       //(4 for quadrilaterals, 3 for triangles)
  int *vert;           //global vertex indices (length nv)
  int *vert_dir;       //vertex Dirichlet flags (length nv)
  int *vert_dof;       //vertex connectivity array (length nv)
  ...
} Element;
```

The variable `Elem[m].vert[j]`, $j = 1, 2, \ldots, nv$, contains the index of the vertex $x_{K_m}(v_j)$ of K_m (as it comes from the mesh generator, i.e., this is not the index of an unconstrained vertex). The ordering of vertices and edges of Q^1- and P^1-elements was shown in Figures 3.2 and 3.5 in Section 3.2. The flag `Elem[m].vert_dir[j]`, $j = 1, 2, \ldots, nv$, is zero if the vertex $x_{K_m}(v_j)$ of K_m is unconstrained, and one otherwise.

Construction of connectivity arrays Assume that for all elements $K_m \in \mathcal{T}_{h,p}$, $m = 1, 2, \ldots, M$, the number nv and the arrays vert and vert_dir have been defined. The first part requires reading a mesh file, and the latter linking Dirichlet boundary conditions to the constrained grid vertices.

The jth component of the connectivity array `Elem[m].vert_dof`, $j = 1, 2, \ldots, nv$, contains either

- the index of the vertex basis function of the space $V_{h,p}$ associated with the vertex $x_{K_m}(v_j)$ of K_m (if `Elem[m].vert_dir[j] == 0`),
- or a negative integer number -NBC (if `Elem[m].vert_dir[j] == 1`).

In the case of nonhomogeneous Dirichlet boundary conditions, the values of the Dirichlet lift G at the constrained vertices can be stored via an array of real numbers. For every constrained vertex, NBC may represent the corresponding index in this array. Using this construction, the implementation of nonhomogeneous Dirichlet boundary conditions is straightforward, and it does not need to be discussed here in more detail. The algorithm that creates the element connectivity arrays vert_dof for all Q^1- and P^1-elements in the mesh $\mathcal{T}_{h,p}$, looks as follows:

Algorithm 4.1 (Enumeration of vertex DOF)

```
By Nvert denote the total number of grid vertices in $\mathcal{T}_{h,p}$.
Allocate a temporary array int *DOFarray of the length Nvert.
//Initialize DOFarray with the numbers 1,2,...,Nvert:
for i=1,2,...,Nvert do DOFarray[i] := i;
```

```
//Deactivate constrained vertices of all elements:
for m=1,2,...,M do {     //global element loop
  for j=1,2,...,Elem[m].nv do {    //local vertex loop
    if (Elem[m].vert_dir[j] == 1 then {
      DOFarray[Elem[m].vert[j]] := -1;
    }
  }
}
//Re-enumerate vertices, leaving out the deactivated ones:
count := 1;
for i=1,2,...,Nvert do {
  if (DOFarray[i] > -1) then {
    DOFarray[i] := count;
    count := count+1;
  }
}
//Read the unconstrained vertex indices back into elements:
for m=1,2,...,M do {
  for j=1,2,...,Elem[m].nv do {
    if (Elem[m].vert_dir[j] == 0) then {
      Elem[m].vert_dof[j] := DOFarray[Elem[m].vert[j]];
    }
  }
}
N := count - 1;     //This is the dimension of the space $V_{h,p}$.
Deallocate the array DOFarray.
```

4.1.7 Assembling algorithm for Q^1/P^1-elements

Assume that the pair of simplifying conditions mentioned in Paragraph 4.1.5 hold and the stiffness term (4.23) reduces to (4.24), i.e., that the functions (4.25) are constant. In addition, assume that the problem (4.2) is equipped with homogeneous Dirichlet boundary data $\Gamma_{D,h} = \partial\Omega_h$ and $g_D = 0$ on $\Gamma_{D,h}$.

Preprocessing step (when (4.24) holds) In this case the global stiffness matrix S can be filled very efficiently based on (4.24). Begin with evaluating the constant Jacobi matrix of the reference map $\boldsymbol{x}_{K_m}$ on every element K_m, $m = 1, 2, \ldots, M$, using the formulae (3.21), (3.13). Store the constant absolute value of the determinant of the Jacobi matrix, for example, as

$$\texttt{Elem[m].jac} := |J_{K_m}|.$$

Invert the Jacobi matrix (as described in Remark 4.2), and store the constant inverse partial derivatives, for example, as

$$\texttt{Elem[m].inv\_j[r][n]} := \frac{\partial \xi_r^{(m)}}{\partial x_n}, \quad 1 \le n, r \le d.$$

Evaluate the master element stiffness integrals (4.26) for both the reference quadrilateral K_q and reference triangle K_t (whatever case is relevant). Store these constants, for example, in two separate global four-dimensional arrays

$$\texttt{MESI\_Q[k][l][r][s]} := \int_{K_q} \frac{\partial \varphi^{v_l}}{\partial \xi_r} \frac{\partial \varphi^{v_k}}{\partial \xi_s} \, \mathrm{d}\boldsymbol{\xi}, \qquad 1 \le k, l \le 4, 1 \le r, s \le d$$

with $\varphi^{v_k}, \varphi^{v_l}$ defined in (3.11), and

$$\texttt{MESI\_T[k][l][r][s]} := \int_{K_t} \frac{\partial \varphi^{v_l}}{\partial \xi_r} \frac{\partial \varphi^{v_k}}{\partial \xi_s} \, \mathrm{d}\boldsymbol{\xi}, \qquad 1 \le k, l \le 3, 1 \le r, s \le d,$$

where $\varphi^{v_k}, \varphi^{v_l}$ were defined in (3.20). Further, evaluate the master element mass integrals,

$$\texttt{MEMI\_Q[k][l]} := \int_{K_q} \varphi^{v_l} \varphi^{v_k} \, \mathrm{d}\boldsymbol{\xi}, \qquad 1 \le k, l \le 4,$$

and

$$\texttt{MEMI\_T[k][l]} := \int_{K_t} \varphi^{v_l} \varphi^{v_k} \, \mathrm{d}\boldsymbol{\xi}, \qquad 1 \le k, l \le 3.$$

Then for Q^1-elements the stiffness matrix contribution (4.24) attains the form

$$\begin{aligned} &\texttt{double SMC(Elem,k,l,m,MESI\_Q,MEMI\_Q)} := \texttt{Elem[m].jac} * \texttt{a1} \\ &* \sum_{n=1}^{d} \sum_{r=1}^{d} \texttt{Elem[m].inv\_j[r][n]} * \sum_{s=1}^{d} \texttt{Elem[m].inv\_j[s][n]} * \texttt{MESI\_Q[k][l][r][s]} \\ &+ \texttt{Elem[m].jac} * \texttt{a0} * \texttt{MEMI\_Q[k][l]}, \qquad 1 \le k, l \le 4, \end{aligned}$$

and analogously for P^1-elements one has

$$\begin{aligned} &\texttt{double SMC(Elem,k,l,m,MESI\_T,MEMI\_T)} := \texttt{Elem[m].jac} * \texttt{a1} \\ &* \sum_{n=1}^{d} \sum_{r=1}^{d} \texttt{Elem[m].inv\_j[r][n]} * \sum_{s=1}^{d} \texttt{Elem[m].inv\_j[s][n]} * \texttt{MESI\_T[k][l][r][s]} \\ &+ \texttt{Elem[m].jac} * \texttt{a0} * \texttt{MEMI\_T[k][l]}, \qquad 1 \le k, l \le 3. \end{aligned}$$

The idea of the element-by-element assembling algorithm is analogous to the one-dimensional case (Algorithm 2.5). With the connectivity arrays `vert_dof` available on all elements K_m, $k = 1, 2, \ldots, M$, and the constants precomputed above, it reads:

Algorithm 4.2 (Assembling algorithm)

```
N := $M_v$;
//Set the stiffness matrix $\boldsymbol{S}$ zero:
for i = 1,2,...,N do for j = 1,2,...,N do S[i][j] := 0;
//Set the right-hand side vector $\boldsymbol{F}$ zero:
for i = 1,2,...,N do F[i] := 0;
//Element loop:
for m = 1,2,...,M do {
  //Outer loop over shape functions:
  for i = 1,2,...,Elem[m].nv do {
    //Index of the test function $v_{m_1} \in V_{h,p}$ (row position in $\boldsymbol{S}$)
    m1 := Elem[m].vert_dof[i];
    //Inner loop over shape functions:
    //(Filling the $m_1$th row of $\boldsymbol{S}$)
    if (m1 > -1) then for j = 1,2,...,Elem[m].nv do {
      //Index of the basis function $v_{m_2} \in V_{h,p}$ (column position in $\boldsymbol{S}$)
      m2 := Elem[m].vert_dof[j];
      if (m2 > -1) then {
        if (Elem[m].nv == 4 then {
          S[m1,m2] := S[m1,m2] + SMC(Elem,i,j,m,MESI_Q,MEMI_Q);
        }
        else {
          S[m1,m2] := S[m1,m2] + SMC(Elem,i,j,m,MESI_T,MEMI_T);
        }
      }
    } //End of inner loop over shape functions
    //Contribution of the test function $v_{m_1}$ to the right-hand side $\boldsymbol{F}$:
    if (m1 > -1) then {
      F[m1] := F[m1] + Elem[m].jac*$\int_{\hat{K}} \tilde{f}^{(m)}(\boldsymbol{\xi})\varphi^{v_i}(\boldsymbol{\xi})\,d\boldsymbol{\xi}$;
    }
  } //End of outer loop over shape functions
} //End of element loop
//Notation $\tilde{f}^{(m)}(\boldsymbol{\xi}) = f(\boldsymbol{x}_{K_m}(\boldsymbol{\xi}))$ was used.
```

If the simplifying conditions formulated in Paragraph 4.1.5 do not apply, then the Jacobian, entries of the inverse Jacobi matrix, and other values are no longer constant in the elements. In such case, (4.24) has to be replaced with the more general relation (4.23), and instead of reading the precomputed entries from the MESI and MEMI arrays, the corresponding integrals have to be evaluated numerically.

4.1.8 Lagrange interpolation on Q^1/P^1-meshes

Assume a regular mesh $\mathcal{T}_{h,p}$ over a bounded domain Ω_h (Definition 3.5), consisting of Q^1- and/or P^1-elements (K_i, P_i, Σ_i), $i = 1, 2, \ldots, M$. For each quadrilateral element $Q^1(K_i)$, the polynomial space P_i and the set of degrees of freedom Σ_i have the form (3.16) and (3.17), and the unique nodal basis was defined in (3.18). For triangular elements $P^1(K_i)$, the space P_i, the set Σ_i, and the unique nodal basis were defined in (3.22), (3.23), and (3.24), respectively.

Proposition 4.3 *For any function $g \in C(\overline{\Omega}_h)$, the global Lagrange interpolant $\mathcal{I}(g)$ is continuous in $\overline{\Omega}_h$. Thus every regular mesh consisting of Q^1- and/or P^1-elements is conforming to the space $H^1(\Omega_h)$.*

Proof: The nodal basis functions (3.18) and (3.24) are affine along the edges of any quadrilateral and triangular element $K \in \mathcal{T}_{h,p}$, respectively. The definition of the degrees of freedom (3.17) and (3.23) implies that on every $K \in \mathcal{T}_{h,p}$ the local interpolant $\mathcal{I}_K(g)$ coincides with the interpolated function g at the vertices of K. Since the mesh $\mathcal{T}_{h,p}$ is regular, the global interpolant $\mathcal{I}(g)$ coincides with the interpolated function g at all mesh vertices and it is affine on the edges of all elements. Thus obviously it is continuous in $\overline{\Omega}_h$.■

The global interpolant is constructed according to Definition 3.6, elementwise, via the local interpolants (3.28). Given a function $g \in C(\overline{\Omega}_h)$ on an element $K_m \in \mathcal{T}_{h,p}$, the local interpolant on K_m is evaluated on the corresponding reference domain $\hat{K} = K_q$ or $\hat{K} = K_t$, using the set of vertex shape functions (3.18) or (3.24) on $\hat{K}$ and using the values of the function $g \circ \boldsymbol{x}_{K_m}$ at the vertices of $\hat{K}$. The result is transformed back to K_m. According to Proposition 4.3, one obtains a function which is continuous in Ω_h.

4.1.9 Exercises

Exercise 4.1 *Prove Proposition 4.1.*

Exercise 4.2 *Consider the problem*

$$-\Delta u = \left(\frac{3b}{2}x_2^2 - \frac{b^2}{2}x_2 - x_2^3\right)(6x_1 - 3a) + (6x_2 - 3b)\left(\frac{3a}{2}x_1^2 - \frac{a^2}{2}x_1 - x_1^3\right)$$

with homogeneous Dirichlet boundary conditions in a bounded domain $\Omega = (0, a) \times (0, b)$, *where* $0 < a, b \in \mathbb{R}$. *Let the domain* Ω *be covered with a Cartesian quadrilateral finite element mesh consisting of* $M = M_1 \times M_2$ *Lagrange* Q^1*-elements (the division is equidistant in both axial directions).*

1. *Write the weak formulation and approximate weak formulation of the problem.*

2. *Perform a unique enumeration of the interior grid points. Use an outer loop in the* x_1*-direction and an inner loop in the* x_2*-direction.*

3. *Write element connectivity arrays for general* M_1 *and* M_2.

4. *Print all master stiffness integrals of the form*

$$\int_{K_q} \frac{\partial \varphi^{v_k}}{\partial \xi_r} \frac{\partial \varphi^{v_l}}{\partial \xi_s} \, d\boldsymbol{x}, \quad 1 \le k, l \le 4, \quad 1 \le r, s \le 2,$$

that you will need for the assembling. (This is a little dull but it helps discover errors.)

5. *Write the reference map* $\boldsymbol{x}_{ij}$, $1 \le i \le M_1$, $1 \le j \le M_2$ *for the element* K_{ij} *on the position* (i, j) *in the mesh. Write its Jacobi matrix, Jacobian, and inverse Jacobi matrix.*

6. *Implement the element-by element assembling procedure (Algorithm 4.2).*

7. *Implement an algorithm for plotting the approximate solution* $u_{h,p}$ *in an element-by-element fashion. First construct the polynomial on the reference domain by means of the shape functions, the corresponding connectivity array and the coefficient vector. Then transform it to the physical element* K_{ij} *in virtue of the reference map* $\boldsymbol{x}_{ij}$.

8. *Present plots of the approximate solution* $u_{h,p}$ *for*

   (a) $a = 2,\ b = 1,\ M_1 = 10,\ M_2 = 5,$

   (b) $a = 2,\ b = 1,\ M_1 = 20,\ M_2 = 10,$

   (c) $a = 2,\ b = 1,\ M_1 = 40,\ M_2 = 20.$

   (d) $a = 2,\ b = 1,\ M_1 = 60,\ M_2 = 30.$

9. *The exact solution is*

$$u(x_1, x_2) = x_1 x_2 (a - x_1)(b - x_2) \left(\frac{a}{2} - x_1\right) \left(\frac{b}{2} - x_2\right).$$

Present the convergence curve of the above computations in the decimal-logarithmic scale. Use the $H^1(\Omega)$*-seminorm. Put the number of DOF on the horizontal axis.*

4.2 HIGHER-ORDER NUMERICAL QUADRATURE IN 2D

In this section we discuss higher-order Gaussian numerical quadrature rules on the reference domains K_q and K_t. For details regarding the theory and open problems in modern Gaussian numerical quadrature, we refer to [35, 46, 47, 48, 49, 70, 80, 104, 108] and [114]. For practical implementation, CD-ROM containing Gaussian quadrature data for various 2D and 3D reference domains and polynomials of the degree up to $p = 20$ is part of [111].

With the quadrature rules available on the reference domains K_t and K_q, the quadrature on arbitrary quadrilateral or triangular mesh elements is performed via the Substitution Theorem,

$$\int_K f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{\hat{K}} |J_K(\boldsymbol{\xi})|(f \circ \boldsymbol{x}_K)(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}.$$

Here either $\hat{K} = K_q$ or $\hat{K} = K_t$, and $J_K(\boldsymbol{\xi})$ is the Jacobian of the bijective reference map $\boldsymbol{x}_K : \hat{K} \to K$.

4.2.1 Gaussian quadrature on quads

Easiest to implement are quadrature formulae for Cartesian-product geometries, such as the reference square $K_q = (-1, 1)^2$.

Composite Gaussian quadrature Consider the formula

$$\int_{K_a} f(\xi)\mathrm{d}\xi \approx \sum_{i=1}^{M_a} w_{g_a,i} f(y_{g_a,i}),$$

where $y_{g_a,i}, w_{g_a,i}$ are Gaussian integration points and weights on the one-dimensional reference domain $K_a = (-1, 1)$ that integrate exactly all polynomials of the degree p and lower. It is easy to see that the product formula

$$\int_{K_a}\int_{K_a} g(\xi_1, \xi_2)\mathrm{d}\xi_1\mathrm{d}\xi_2 \approx \sum_{i=1}^{M_a}\sum_{j=1}^{M_a} w_{g_a,i}w_{g_a,j}g(y_{g_a,i}, y_{g_a,j}) \qquad (4.27)$$

has the order of accuracy p on K_q for functions of two variables (all bivariate polynomials up to the degree p in each variable are integrated exactly). An advantage of the composite quadrature is that it easily can be generalized to incomplete product polynomials (when the 1D polynomial spaces in the variables ξ_1, ξ_2 differ). In this way one can obtain quadrature rules of practically unlimited order of accuracy. More efficient formulae are available for spaces of complete polynomials (see [46]).

4.2.2 Gaussian quadrature on triangles

The triangular case is more difficult. First let us show a simple scheme based on the translation of the integration from the reference triangle K_t to the reference quadrilateral K_q.

Translation of quadrature from K_t to K_q This procedure can be viewed as "stretching" functions defined on K_t to be defined on K_q in such a way that their integrals remain unchanged. The following proposition defines the technique precisely.

Proposition 4.4 *Let $g(\boldsymbol{\xi})$ be a continuous bounded function defined on the reference triangle K_t. Then its integral over K_t is equal to the integral over K_q of an adjusted function given by the following formula,*

$$\int_{K_t} g(\boldsymbol{\xi})\, d\boldsymbol{\xi} = \int_{K_q} \frac{1-y_2}{2}\, g\left(-1 + \frac{1-y_2}{2}(y_1+1), y_2\right) d\boldsymbol{y}. \tag{4.28}$$

Proof: Consider the mapping

$$\boldsymbol{\xi}(\boldsymbol{y}) : \boldsymbol{y} \to \boldsymbol{\xi}(\boldsymbol{y}) = \begin{pmatrix} -1 + \dfrac{1-y_2}{2}(y_1+1) \\ y_2 \end{pmatrix}$$

that transforms K_q to K_t. Its Jacobian

$$\det\left(\frac{D\boldsymbol{\xi}}{D\boldsymbol{y}}\right) = \frac{1-y_2}{2}$$

is positive except for the upper edge $y_2 = 1$ where it vanishes. However, the mapping is one-to-one and the standard Substitution Theorem yields the result immediately. ■

Remark 4.3 *The transformation $\boldsymbol{\xi}(\boldsymbol{y})$ induced an additional affine factor $(1-y_2)/2$ in the integrand on the right-hand side of (4.28). The order of the quadrature rule in the y_2-variable should be increased accordingly.*

More difficult to implement, but certainly worth the effort, are the economical Gaussian quadrature schemes.

Economical Gaussian quadrature The fundamental equation for the construction of the integration points and weights for the reference triangle K_t reads

$$\int_{-1}^{1}\int_{-1}^{-\xi_1} f(\xi_1, \xi_2)\, d\xi_2\, d\xi_1 \approx \sum_{k=1}^{m} w_k f(\xi_{1,k}, \xi_{2,k}). \tag{4.29}$$

where m denotes the number of integration points. Each point is characterized by three unknowns: w_k, $\xi_{1,k}$, and $\xi_{2,k}$. After inserting a suitable polynomial basis into (4.29), one may obtain equations that are not independent. Systems with more unknowns than equations are obtained when the number n of terms in complete polynomials of the degree p is not divisible by three: for $p = 3$, for example, one has $n = (p+1)(p+2)/2 = 10$ independent polynomials, and thus at least four Gaussian points must be used (12 unknowns). Other standard difficulties are related to the nonuniqueness of solution to the nonlinear system, where weights can come out negative or points outside of the domain of integration. The design of optimal Gaussian quadrature formulae for higher polynomial degrees involves many open problems (see, e.g., [35, 46] and [47]).

Selected quadrature constants Tables 4.1–4.5 present the optimal Gaussian quadrature rules on the reference triangle K_t of the orders of accuracy $p = 1, 2, \ldots, 5$.

Table 4.1 Gaussian quadrature on K_t, order $p = 1$.

Point #	ξ_1-Coordinate	ξ_2-Coordinate	Weight
1.	-0.33333 33333 33333	-0.33333 33333 33333	2.00000 00000 00000

Table 4.2 Gaussian quadrature on K_t, order $p = 2$.

Point #	ξ_1-Coordinate	ξ_2-Coordinate	Weight
1.	-0.66666 66666 66667	-0.66666 66666 66667	0.66666 66666 66667
2.	-0.66666 66666 66667	0.33333 33333 33333	0.66666 66666 66667
3.	0.33333 33333 33333	-0.66666 66666 66667	0.66666 66666 66667

Table 4.3 Gaussian quadrature on K_t, order $p = 3$.

Point #	ξ_1-Coordinate	ξ_2-Coordinate	Weight
1.	-0.33333 33333 33333	-0.33333 33333 33333	-1.12500 00000 00000
2.	-0.60000 00000 00000	-0.60000 00000 00000	1.04166 66666 66667
3.	-0.60000 00000 00000	0.20000 00000 00000	1.04166 66666 66667
4.	0.20000 00000 00000	-0.60000 00000 00000	1.04166 66666 66667

Table 4.4 Gaussian quadrature on K_t, order $p = 4$.

Point #	ξ_1-Coordinate	ξ_2-Coordinate	Weight
1.	-0.10810 30181 68070	-0.10810 30181 68070	0.44676 31793 56022
2.	-0.10810 30181 68070	-0.78379 39636 63860	0.44676 31793 56022
3.	-0.78379 39636 63860	-0.10810 30181 68070	0.44676 31793 56022
4.	-0.81684 75729 80458	-0.81684 75729 80458	0.21990 34873 10644
5.	-0.81684 75729 80458	0.63369 51459 60918	0.21990 34873 10644
6.	0.63369 51459 60918	-0.81684 75729 80458	0.21990 34873 10644

Table 4.5 Gaussian quadrature on K_t, order $p = 5$.

Point #	ξ_1-Coordinate	ξ_2-Coordinate	Weight
1.	-0.33333 33333 33333	-0.33333 33333 33333	0.45000 00000 00000
2.	-0.05971 58717 89770	-0.05971 58717 89770	0.26478 83055 77012
3.	-0.05971 58717 89770	-0.88056 82564 20460	0.26478 83055 77012
4.	-0.88056 82564 20460	-0.05971 58717 89770	0.26478 83055 77012
5.	-0.79742 69853 53088	-0.79742 69853 53088	0.25187 83610 89654
6.	-0.79742 69853 53088	0.59485 39707 06174	0.25187 83610 89654
7.	0.59485 39707 06174	-0.79742 69853 53088	0.25187 83610 89654

4.3 HIGHER-ORDER NODAL ELEMENTS

In this section we extend the lowest-order Q^1- and P^1-elements to higher-order Lagrange elements. The quadrilateral case, based on the product Gauss–Lobatto points, is described in Paragraphs 4.3.1 and 4.3.2. The quality of the Lagrange interpolation is discussed in Paragraph 4.3.3. Higher-order triangular elements are constructed using the Fekete points in Paragraphs 4.3.4 and 4.3.5. The basis of the space $V_{h,p}$ for regular hybrid quadrilateral/triangular meshes is presented in Paragraph 4.3.6. Algorithmic aspects of the method, including concrete data structure and an extension of Algorithm 4.2 to higher-order Lagrange elements, are presented in Paragraphs 4.3.7–4.3.9. The interpolation on meshes consisting of higher-order Lagrange elements, along with the conformity to the space $H^1(\Omega_h)$, is discussed in Paragraph 4.3.10.

4.3.1 Product Gauss–Lobatto points

The favorable conditioning properties of the one-dimensional Lagrange shape functions based on the Gauss–Lobatto points in $\overline{K}_a$ (Figure 2.24) suggest that the Lagrange nodal element on the product geometry $K_q = K_a \times K_a$ should be designed using the Cartesian product of the Gauss–Lobatto points in both axial directions ξ_1 and ξ_2. Numerical experience confirms that indeed this is a good choice. For future reference let us define an orientation of the edges $e_1, e_2, \ldots, e_4$ as shown in Figure 4.5.

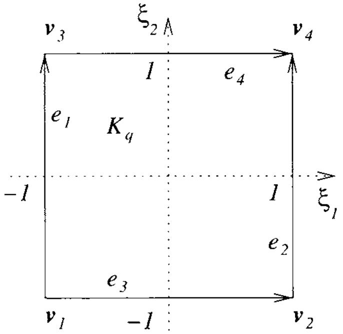


Figure 4.5 Orientation of edges on the reference quadrilateral K_q.

Quadrilateral elements admit two different directional orders of approximation $p, r \ge 1$. Let $y_i^{(p)} \in \overline{K}_a$ and $y_j^{(r)} \in \overline{K}_a$ be the one-dimensional Gauss–Lobatto points of the orders p and r, respectively (see Paragraph 2.4.5). For algorithmic purposes it is convenient to split the $(p+1)(r+1)$ product points in $\overline{K}_q$ into three groups as follows:

Four vertex nodes are defined as

$$\begin{aligned} \boldsymbol{v}^{v_1} &= \boldsymbol{v}_1 = (y_1^{(p)}, y_1^{(r)}) = (-1,-1), \\ \boldsymbol{v}^{v_2} &= \boldsymbol{v}_2 = (y_{p+1}^{(p)}, y_1^{(r)}) = (1,-1), \\ \boldsymbol{v}^{v_3} &= \boldsymbol{v}_3 = (y_1^{(p)}, y_{r+1}^{(r)}) = (-1,1), \\ \boldsymbol{v}^{v_4} &= \boldsymbol{v}_4 = (y_{p+1}^{(p)}, y_{r+1}^{(r)}) = (1,1). \end{aligned} \tag{4.30}$$

There are $r-1$ edge-interior nodes (edge nodes) on the edges e_1, e_2 and $p-1$ edge nodes on the edges e_3, e_4. For algorithmic purposes it is practical to sort them according to the orientation of the edges shown in Figure 4.5,

$$\begin{aligned} \boldsymbol{v}_1^{e_1} &= (y_1^{(p)}, y_2^{(r)}) = (-1, y_2^{(r)}), \\ \boldsymbol{v}_2^{e_1} &= (y_1^{(p)}, y_3^{(r)}) = (-1, y_3^{(r)}), \\ &\vdots \\ \boldsymbol{v}_{r-1}^{e_1} &= (y_1^{(p)}, y_r^{(r)}) = (-1, y_r^{(r)}), \end{aligned} \tag{4.31}$$

and so on.

The $(p-1)(r-1)$ element-interior nodes (bubble nodes) can be sorted in any unique way, for example as

$$\begin{aligned} \boldsymbol{v}_{1,1}^{b} &= (y_2^{(p)}, y_2^{(r)}), \\ \boldsymbol{v}_{1,2}^{b} &= (y_2^{(p)}, y_3^{(r)}), \\ &\vdots \\ \boldsymbol{v}_{p-1,r-1}^{b} &= (y_p^{(p)}, y_r^{(r)}). \end{aligned} \tag{4.32}$$

With this point set in hand, the Lagrange $Q^{p,r}$-element is constructed as follows:

4.3.2 Lagrange–Gauss–Lobatto $Q^{p,r}$-elements

It is natural to construct the master $Q^{p,r}$-element on the reference domain K_q first, and then to extend it to an arbitrary convex quadrilateral domain K.

$Q^{p,r}$-element on the reference domain K_q In the sense of Definition 3.1 the master element is a triad $(K_q, Q^{p,r}(K_q), \Sigma_q)$, where

$$Q^{p,r}(K_q) = \operatorname{span}\{\xi_1^k \xi_2^l;\ 1 \le k \le p;\ 1 \le l \le r;\ -1 \le \xi_1, \xi_2 \le 1\}, \tag{4.33}$$

and the set of degrees of freedom Σ_q contains linear forms associated with function values at the $(p+1)(r+1)$ nodal points in the usual sense. It is customary to write $Q^p = Q^{p,r}$ if $r = p$.

Nodal basis on K_q Let $\theta_1^{(p)}, \theta_2^{(p)}, \ldots, \theta_{p+1}^{(p)}$ be the set of the pth-order one-dimensional Lagrange nodal shape functions (2.57) on the reference domain K_a, satisfying the delta property (2.56),

$$\theta_k^{(p)}\left(y_l^{(p)}\right) = \delta_{kl}. \tag{4.34}$$

The nodal shape functions on the reference domain K_q are split into three groups according to the different types of nodes introduced above.

There are four vertex functions

$$\begin{aligned}\varphi_q^{v_1}(\boldsymbol{\xi}) &= \theta_1^{(p)}(\xi_1)\theta_1^{(r)}(\xi_2),\\ \varphi_q^{v_2}(\boldsymbol{\xi}) &= \theta_{p+1}^{(p)}(\xi_1)\theta_1^{(r)}(\xi_2),\\ \varphi_q^{v_3}(\boldsymbol{\xi}) &= \theta_1^{(p)}(\xi_1)\theta_{r+1}^{(r)}(\xi_2),\\ \varphi_q^{v_4}(\boldsymbol{\xi}) &= \theta_{p+1}^{(p)}(\xi_1)\theta_{r+1}^{(r)}(\xi_2).\end{aligned} \tag{4.35}$$

Further one has $r-1$ edge functions associated with the edges e_1 and e_2, and $p-1$ edge functions corresponding to the edges e_3 and e_4. On the edge e_1, for example, they have the form

$$\begin{aligned}\varphi_{1,q}^{e_1}(\boldsymbol{\xi}) &= \theta_1^{(p)}(\xi_1)\theta_2^{(r)}(\xi_2),\\ \varphi_{2,q}^{e_1}(\boldsymbol{\xi}) &= \theta_1^{(p)}(\xi_1)\theta_3^{(r)}(\xi_2),\\ &\vdots\\ \varphi_{r-1,q}^{e_1}(\boldsymbol{\xi}) &= \theta_1^{(p)}(\xi_1)\theta_r^{(r)}(\xi_2),\end{aligned} \tag{4.36}$$

and so on. Finally there are $(p-1)(r-1)$ bubble functions

$$\begin{aligned}\varphi_{1,1,q}^{b}(\boldsymbol{\xi}) &= \theta_2^{(p)}(\xi_1)\theta_2^{(r)}(\xi_1),\\ \varphi_{1,2,q}^{b}(\boldsymbol{\xi}) &= \theta_2^{(p)}(\xi_1)\theta_3^{(r)}(\xi_1),\\ &\vdots\\ \varphi_{p-1,r-1,q}^{b}(\boldsymbol{\xi}) &= \theta_p^{(p)}(\xi_1)\theta_r^{(r)}(\xi_1).\end{aligned} \tag{4.37}$$

The next two simple propositions state that indeed the above-defined shape functions form a basis of the finite element space, and that they satisfy the delta property (3.4).

Proposition 4.5 *The shape functions (4.35)–(4.37) form a basis in the space (4.33).*

Proof: The dimension of the space (4.33) is $(p+1)(r+1)$, which is equal to the number $4+2(p-1)+2(r-1)+(p-1)(r-1)$ of the shape functions (4.35)–(4.37). Their linear independence follows easily from (4.34). ■

Proposition 4.6 *The basis functions (4.35)–(4.37) satisfy the delta property (3.4) in the form*

$$\begin{aligned}\varphi_q^{v_i}(\boldsymbol{v}^{v_k}) &= \delta_{ik},\\ \varphi_{j,q}^{e_i}(\boldsymbol{v}_l^{e_k}) &= \delta_{ik}\delta_{jl},\\ \varphi_{i,j,q}^{b}(\boldsymbol{v}_{k,l}^{b}) &= \delta_{ik}\delta_{jl},\end{aligned} \tag{4.38}$$

and therefore they are the nodal basis of the space (4.33) in the sense of Definition 3.3.

Proof: This is left to the reader as a simple exercise. ■

Several useful geometrical properties of the nodal shape functions (4.35)–(4.37) are presented in Proposition 4.7, and the shape functions of the Q^2- and Q^3-elements are shown in Examples 4.1 and 4.2.

Proposition 4.7 *The Lagrange–Gauss–Lobatto shape functions (4.35)–(4.37) have the following properties:*

1. *The vertex shape function (4.35) corresponding to a vertex v^{v_i} of K_q vanishes at all remaining vertices and on the two opposite edges of K_q.*
2. *All edge shape functions (4.36) associated with an edge e_i vanish at all vertices of K_q and on all remaining edges.*
3. *All bubble shape functions (4.37) vanish on the whole boundary of K_q.*
4. *Each nodal shape function (4.35)–(4.37) is either zero or polynomial of degree exactly r when restricted to the edges e_1 and e_2, and either zero or polynomial of degree exactly p when restricted to the edges e_3 and e_4.*

Proof: This follows easily from (4.34), using the fact that every one-dimensional pth-degree polynomial is determined uniquely by its values at $p + 1$ distinct points. ■

■ EXAMPLE 4.1 (Lagrange–Gauss–Lobatto Q^2-element)

The nodal basis of the Q^2-element on the reference domain K_q is shown in Figures 4.6–4.8.

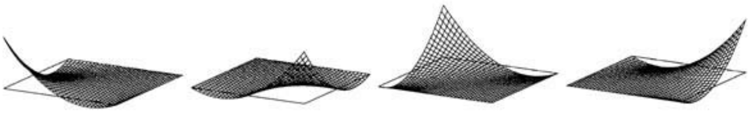


Figure 4.6 Nodal basis of the Q^2-element; the vertex functions φ_q^{v1}, φ_q^{v2}, φ_q^{v3} and φ_q^{v4}.

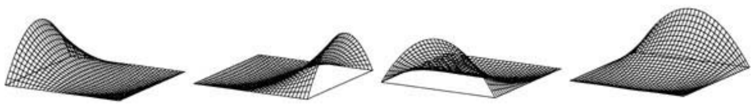


Figure 4.7 Nodal basis of the Q^2-element; the edge functions $\varphi_{1,q}^{e1}$, $\varphi_{1,q}^{e2}$, $\varphi_{1,q}^{e3}$ and $\varphi_{1,q}^{e4}$.

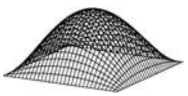


Figure 4.8 Nodal basis of the Q^2-element; the bubble function $\varphi_{1,1,q}^{b}$.

■ EXAMPLE 4.2 (Lagrange–Gauss–Lobatto Q^3-element)

The nodal basis of the Q^3-element on the reference domain K_q is shown in Figures 4.9–4.12.

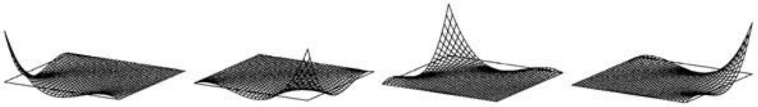


Figure 4.9 Nodal basis of the Q^3-element; the vertex functions φ_q^{v1}, φ_q^{v2}, φ_q^{v3} and φ_q^{v4}.

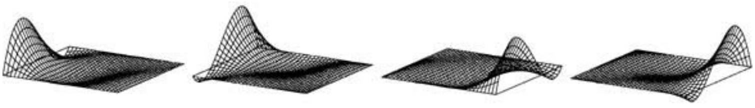


Figure 4.10 Nodal basis of the Q^3-element; the edge functions $\varphi_{1,q}^{e1}$, $\varphi_{2,q}^{e1}$, $\varphi_{1,q}^{e2}$ and $\varphi_{2,q}^{e2}$.

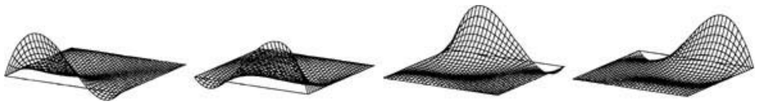


Figure 4.11 Nodal basis of the Q^3-element; the edge functions $\varphi_{1,q}^{e3}$, $\varphi_{2,q}^{e3}$, $\varphi_{1,q}^{e4}$ and $\varphi_{2,q}^{e4}$.

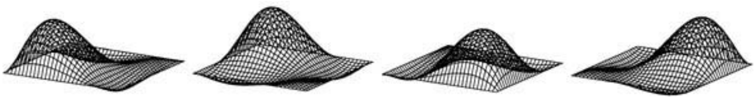


Figure 4.12 Nodal basis of the Q^3-element; the bubble functions $\varphi_{1,1,q}^{b}$, $\varphi_{1,2,q}^{b}$, $\varphi_{2,1,q}^{b}$ and $\varphi_{2,2,q}^{b}$.

The Gauss–Lobatto points in a convex quadrilateral $\overline{K} \subset \mathbb{R}^2$ Consider an arbitrary convex quadrilateral domain $K \subset \mathbb{R}^2$ with pairwise-distinct vertices $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_4$ and straight edges $s_1, s_2, \ldots, s_4$ (Figure 3.3 in Section 3.2). The corresponding isoparametric biaffine reference map $\boldsymbol{x}_K : K_q \rightarrow K$ was defined in (3.13). The vertex, edge, and interior (bubble) nodal points in $\overline{K}$ are defined as the images of the nodal points (4.30), (4.31), and (4.32) through the map $\boldsymbol{x}_K$. Let us list some of them for future reference:

There are four vertex nodes

$$\begin{aligned} \boldsymbol{x}^{v_1} &= \boldsymbol{x}_K(\boldsymbol{v}^{v_1}), \\ \boldsymbol{x}^{v_2} &= \boldsymbol{x}_K(\boldsymbol{v}^{v_2}), \\ \boldsymbol{x}^{v_3} &= \boldsymbol{x}_K(\boldsymbol{v}^{v_3}), \\ \boldsymbol{x}^{v_4} &= \boldsymbol{x}_K(\boldsymbol{v}^{v_4}), \end{aligned} \tag{4.39}$$

$r-1$ edge nodes on the edges s_1, s_2 and $p-1$ edge nodes on the edges s_3, s_4,

$$\begin{aligned} \boldsymbol{x}_1^{e_1} &= \boldsymbol{x}_K(\boldsymbol{v}_1^{e_1}), \\ \boldsymbol{x}_2^{e_1} &= \boldsymbol{x}_K(\boldsymbol{v}_2^{e_1}), \\ &\vdots \\ \boldsymbol{x}_{r-1}^{e_1} &= \boldsymbol{x}_K(\boldsymbol{v}_{r-1}^{e_1}) \end{aligned} \tag{4.40}$$

(similarly on the edges e_2, e_3, and e_4), and $(p-1)(r-1)$ interior (bubble) nodes

$$\begin{aligned} \boldsymbol{x}_{1,1}^{b} &= \boldsymbol{x}_K(\boldsymbol{v}_{1,1}^{b}), \\ \boldsymbol{x}_{1,2}^{b} &= \boldsymbol{x}_K(\boldsymbol{v}_{1,2}^{b}), \\ &\vdots \\ \boldsymbol{x}_{p-1,r-1}^{b} &= \boldsymbol{x}_K(\boldsymbol{v}_{p-1,r-1}^{b}), \end{aligned} \tag{4.41}$$

as illustrated in Figure 4.13.

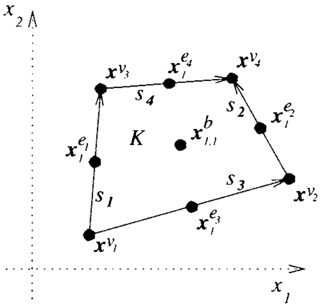


Figure 4.13 Gauss–Lobatto points in a quadrilateral $\overline{K} \subset \mathbb{R}^2$ with straight edges ($p = r = 2$).

$Q^{p,r}$*-element on *K In the sense of Definition 3.1, the domain K is equipped with the space

$$Q^{p,r}(K) = \{q \circ \boldsymbol{x}_K^{-1};\ g \in Q^{p,r}(K_q)\}. \tag{4.42}$$

The set of degrees of freedom Σ_K consists of $(p+1)(r+1)$ linear forms associated with function values at the same number of the Gauss–Lobatto nodal points (4.39), (4.40), and (4.41).

Nodal basis on K The nodal basis of the element $(K, Q^{p,r}(K), \Sigma_K)$ is defined as usual, i.e., by composing the shape functions on the reference domain K_q with the inverse map $\boldsymbol{x}_K^{-1}$. Let us stress again that the presence of the inverse map is formal, and $\boldsymbol{x}_K^{-1}$ does not need to be evaluated in the finite element code.

Proposition 4.8 *The set of* $(p+1)(r+1)$ *shape functions on* K,

$$\varphi_K(\boldsymbol{x}) \;=\; (\varphi_q \circ \boldsymbol{x}_K^{-1})(\boldsymbol{x}), \tag{4.43}$$

where φ_q *represents the nodal shape functions (4.35), (4.36) and (4.37) on the reference domain* K_q*, constitutes the unique nodal basis of the space (4.42). The finite element* $(K, Q^{p,r}(K), \Sigma_K)$ *is unisolvent.*

Proof: Left to the reader as an easy exercise. ■

The choice of optimal nodal points for triangular elements is much less trivial compared to the quadrilateral case, where the product Gauss–Lobatto points are known to have optimal interpolation properties. Therefore, before we present the higher-order triangular Lagrange P^p-elements in Paragraphs 4.3.4–4.3.6, let us devote Paragraph 4.3.3 to the analysis of the quality of the Lagrange interpolation in $d \geq 1$ spatial dimensions.

4.3.3 Lagrange interpolation and the Lebesgue constant

Assume a bounded convex domain $K \subset \mathbb{R}^d$, polynomial space $P(K)$ of the dimension N_P, and a set of N_P distinct points $\{z_i\}_{i=1}^{N_P} \subset \overline{K}$ that yield a unisolvent Lagrange nodal finite element $(K, P(K), \Sigma_K)$. Thus, given an arbitrary function $g \in C(\overline{K})$, there exists a unique polynomial $g_p \in P(K)$ such that $g_p(z_i) = g(z_i)$ for all $i = 1, 2, \ldots, N_P$. We use the notation

$$g_p(z) = (\mathcal{I}_{N_P} g)(z).$$

Let $G \in P(K)$ be the best approximation of the function g in the maximum norm,

$$\|g - G\|_{max} = \inf_{F \in P(K)} \|g - F\|_{max}.$$

It is not necessarily $G = \mathcal{I}_{N_P} g$, but since $G \in P(K)$, it holds $G = \mathcal{I}_{N_P} G$. Therefore we have

$$\begin{aligned} \|g - \mathcal{I}_{N_P} g\|_{max} &= \|g - G + \mathcal{I}_{N_P} G - \mathcal{I}_{N_P} g\|_{max} \\ &\leq \|g - G\|_{max} + \|\mathcal{I}_{N_P} G - \mathcal{I}_{N_P} g\|_{max} \\ &\leq \|g - G\|_{max} + \|\mathcal{I}_{N_P}\| \|G - g\|_{max} \\ &\leq (1 + \|\mathcal{I}_{N_P}\|) \|G - g\|_{max}, \end{aligned}$$

where

$$\|\mathcal{I}_{N_P}\| = \max_{\|f\|_{max}=1} \|\mathcal{I}_{N_P} f\|_{max}$$

is the standard operator maximum norm, and $\|\mathcal{I}_{N_P}\|$ is referred to as the Lebesgue constant.

The magnitude of the Lebesgue constant depends on the domain K, the polynomial space $P(K)$, and the interpolation points $\{z_i\}_{i=1}^{N_P}$. If the constant is small, then the interpolation operator $\mathcal{I}_{N_P}$ is good and vice versa. It is known that simple choices of interpolation points, such as equidistant, lead to disastrous exponential growth of $\|\mathcal{I}_{N_P}\|$ as the polynomial degree p is increased (this we saw already in Paragraph 2.7.4).

Thus, given a domain K and polynomial space $P(K)$, we face the problem to find optimal interpolation points that minimize the Lebesgue constant. Such points are called Lebesgue points. Unfortunately, nothing seems to be known about Lebesgue points in more than one spatial dimension. The best choices available today are the product Gauss–Lobatto points on quadrilaterals, and the Fekete points on triangles. Let us introduce the latter point set in the next paragraph.

4.3.4 The Fekete points

Let us first define the Fekete points and then discuss their properties and their application to the construction of higher-order Lagrange elements.

Definition 4.1 (Fekete points) *Let a bounded convex domain $K \subset \mathbb{R}^d$ be equipped with a polynomial space $P(K)$ of the dimension N_P. Given an arbitrary basis $\{\vartheta_i\}_{i=1}^{N_P}$ of the space $P(K)$, the* Fekete points *$\{\boldsymbol{y}_i\}_{i=1}^{N_P} \subset \overline{K}$ are a point set that maximizes the determinant*

$$\det \boldsymbol{L}(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{N_P}) = \max_{\{\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_{N_P}\} \subset \overline{K}} \det \boldsymbol{L}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_{N_P}), \tag{4.44}$$

where $\boldsymbol{L}$ is the generalized Vandermonde matrix (3.7) for the Lagrange degrees of freedom $L_i(g) = g(\boldsymbol{\xi}_i)$,

$$\boldsymbol{L}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_{N_P}) = \{L_i(\vartheta_j)\}_{i,j=1}^{N_P} = \{\vartheta_j(\boldsymbol{\xi}_i)\}_{i,j=1}^{N_P}. \tag{4.45}$$

Recall that the generalized Vandermonde matrix $\boldsymbol{L}$ is used to construct the unique nodal basis of nodal finite elements (see Theorem 3.1). It will be shown in Theorem 4.1 that the Fekete points are invariant under the choice of the basis $\{\vartheta_i\}_{i=1}^{N_P}$.

Construction Since no explicit formulae for the Fekete points are available, they have to be constructed by maximizing the determinant (4.44) numerically. This is a nonlinear optimization problem, and numerical methods may produce various solutions depending on the initial condition and other factors. The choice of the initial condition influences the result in a most significant way. Since the global optimality is unclear, the solutions are usually referred to as approximate Fekete points. A numerical algorithm for the construction of approximate Fekete points for triangles of polynomial degrees $p \leq 19$, based on a steepest ascent approach, was presented in [118].

Properties The key observation made in [15] and [16] (in the context of interpolation) was that in the one-dimensional case and in Cartesian product geometries, the Gauss–Lobatto and Fekete points are identical. The advantage of the Fekete points is that they can be defined for any geometry. Numerical experiments indicate that the Lagrange nodal shape functions on triangular elements built on the Fekete points have excellent conditioning properties (examples will be given later). However, there is no optimality proof, so it can be expected that even better point sets will appear in the future. Some known facts about the Fekete points are summarized below.

Theorem 4.1 *Let $p \geq 1$. The Fekete points have the following properties:*

1. *The Fekete points $\{y_i\}_{i=1}^{N_P} \subset \overline{K}_t$ are invariant under the choice of the basis $\{\vartheta_i\}_{i=1}^{N_P} \subset P^p(K_t)$.*

2. *In one-dimensional intervals and Cartesian product geometries the Fekete and Gauss–Lobatto points are the same.*

3. *On the edges of triangular domains the Fekete points coincide with the one-dimensional Gauss–Lobatto points.*

Proof: Assertion *1.* follows easily from the basic properties of determinants (see Paragraph A.1.9): The change of basis multiplies the determinant with a constant independent of the points. See [53] and [15] for *2.* Under the assumptions that the Vandermonde matrix is nonsingular, there exists a maximum number of points that can lie on the boundary. With a conjecture that the Fekete points in $\overline{K}_t$ attain this maximum number on the edges, *3.* was proved in [15]. ■

The Fekete points presented in Tables 4.6–4.8 and on the companion CD-ROM (for $1 \leq p \leq 19$) were drawn from [118] with permission of the authors.

Table 4.6 Fekete points in $\overline{K}_t$, $p = 1$.

Number of points	ξ_1-Coordinate	ξ_2-Coordinate
$n = 3$	1.000000000000	-1.000000000000
	-1.000000000000	1.000000000000
	-1.000000000000	-1.000000000000

Table 4.7 Fekete points in $\overline{K}_t$, $p = 2$.

Number of points	ξ_1-Coordinate	ξ_2-Coordinate
$n = 6$	0.000000000000	-1.000000000000
	-1.000000000000	-1.000000000000
	-1.000000000000	0.000000000000
	0.000000000000	0.000000000000
	-1.000000000000	1.000000000000
	1.000000000000	-1.000000000000

Table 4.8 Approximate Fekete points in $\overline{K}_t$, $p = 3$.

Number of points	ξ_1-Coordinate	ξ_2-Coordinate
$d = 10$	-0.333333333333	-0.333333333333
	-0.447213595500	-1.000000000000
	-1.000000000000	-1.000000000000
	-1.000000000000	-0.447213595500
	0.447213595500	-1.000000000000
	-0.447213595500	0.447213595500
	-1.000000000000	0.447213595500
	0.447213595500	-0.447213595500
	-1.000000000000	1.000000000000
	1.000000000000	-1.000000000000

The Fekete points are shown for $p = 1, 2, \ldots, 15$ in Figure 4.14.

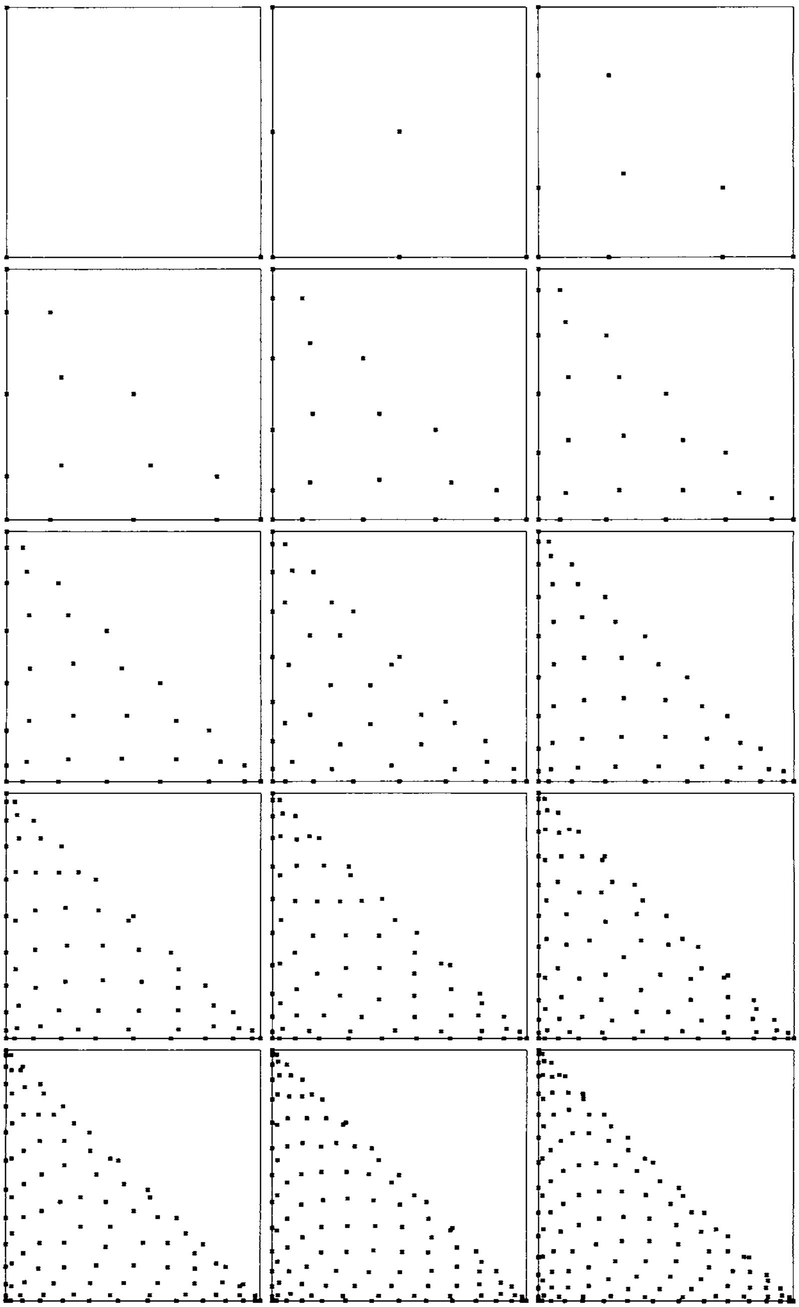


Figure 4.14 The Fekete points in $\overline{K_t}$, $p = 1, 2, \ldots, 15$.

Unique enumeration of the Fekete points For algorithmic purposes it is necessary to enumerate the Fekete points in K_t in a unique way. We assume that the edges of the reference triangle K_t are oriented as shown in Figure 4.15.

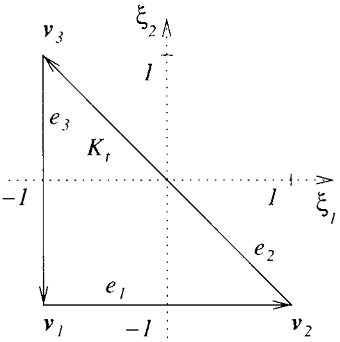


Figure 4.15 Orientation of edges on the reference triangle K_t.

Consider the one-dimensional pth-order Gauss–Lobatto points $y_i^{(p)} \in \overline{K_a}$ from Paragraph 2.4.5. By Theorem 4.1, the Fekete points exactly coincide with the Gauss–Lobatto points on edges of $\overline{K_t}$. The three vertex nodes are denoted by

$$\begin{aligned} \boldsymbol{v}^{v_1} &= \boldsymbol{v}_1 = (y_1^{(p)}, y_1^{(p)}) = (-1,-1), \\ \boldsymbol{v}^{v_2} &= \boldsymbol{v}_2 = (y_{p+1}^{(p)}, y_1^{(p)}) = (1,-1), \\ \boldsymbol{v}^{v_3} &= \boldsymbol{v}_3 = (y_1^{(p)}, y_{p+1}^{(p)}) = (-1,1). \end{aligned} \tag{4.46}$$

The $p-1$ edge nodes on each edge are sorted according to the orientation of the edge. For example, for the edge e_1 we have

$$\begin{aligned} \boldsymbol{v}_1^{e_1} &= (y_2^{(p)}, y_1^{(p)}) = (y_2^{(p)}, -1), \\ \boldsymbol{v}_2^{e_1} &= (y_3^{(p)}, y_1^{(p)}) = (y_3^{(p)}, -1), \\ &\vdots \\ \boldsymbol{v}_{p-1}^{e_1} &= (y_p^{(p)}, y_1^{(p)}) = (y_p^{(p)}, -1). \end{aligned} \tag{4.47}$$

Such enumeration of the edge nodes makes it possible to easily include both quadrilateral and triangular elements into hybrid quadrilateral/triangular meshes. The remaining $(p-1)(p-2)/2$ interior (bubble) nodes can be sorted in any unique way, and we denote them by $\boldsymbol{v}_1^b, \boldsymbol{v}_2^b, \dots, \boldsymbol{v}_{(p-1)(p-2)/2}^b$.

4.3.5 Lagrange–Fekete P^p-elements

The Lagrange P^p-element on the reference triangular domain K_t is equipped with the polynomial space $P^p(K_t)$, $\dim(P^p(K_t)) = N_P = (p+1)(p+2)/2$, and the set of the Lagrange degrees of freedom $\Sigma = \{L_1, L_2, \dots, L_{N_P}\}$ associated with the Fekete points $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \dots, \boldsymbol{\xi}_{N_P}$. The degrees of freedom are defined as the reader expects: $L_1(g) = g(\boldsymbol{\xi}_1)$, $L_2(g) = g(\boldsymbol{\xi}_2)$, ..., $L_{N_P}(g) = g(\boldsymbol{\xi}_{N_P})$ for all $g \in P^p(K_t)$. The unique Lagrange nodal basis satisfying the delta property (3.2) is obtained in the standard way by inverting the generalized Vandermonde matrix (4.45).

Enumeration of shape functions For algorithmic purposes we also need to split the Lagrange–Fekete shape functions into the vertex, edge and bubble functions. By $\varphi_t^{v_m}$ we denote the shape function associated with the mth vertex node $\boldsymbol{v}^{v_m}$, $m = 1,2,3$. The symbol $\varphi_{m,t}^{e_j}$ stands for the shape function corresponding to the mth edge node $\boldsymbol{v}_m^{e_j}$ (following the notation (4.47)), and $\varphi_{m,t}^{b}$ stands for the shape function associated with the mth bubble node $\boldsymbol{v}_m^b$, $m = 1,2,\dots,(p-1)(p-2)/2$. Proposition 4.9 describes the geometrical properties of the shape functions:

Proposition 4.9 *The Lagrange–Fekete shape functions have the following properties:*

1. *The vertex shape function $\varphi_t^{v_i}$, corresponding to the vertex node $\boldsymbol{v}^{v_i}$, vanishes at the two remaining vertices and on the opposite edge of K_t.*

2. *The edge shape function $\varphi_{j,t}^{e_i}$ associated with the edge e_i vanishes at all vertices and on all edges of K_t except for e_i.*

3. *All bubble shape functions vanish on the whole boundary of K_t.*

4. *Each Lagrange–Fekete shape function is either zero or a polynomial of the degree exactly p when restricted to the edges e_1, e_2, or e_3.*

Proof: Analogous to the proof of Proposition 4.7. ■

The next two examples show the Lagrange–Fekete shape functions for the Q^2- and Q^3-elements:

■ **EXAMPLE 4.3** **(P^2-element)**

The nodal basis of the P^2-element on the reference domain K_t is shown in Figures 4.16 and 4.17.

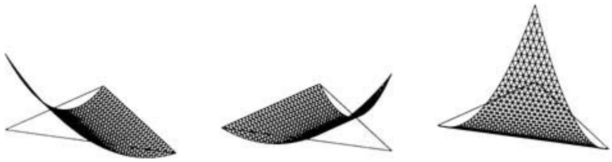


Figure 4.16 Nodal basis of the P^2-element; the vertex functions $\varphi_t^{v_1}$, $\varphi_t^{v_2}$, and $\varphi_t^{v_3}$.

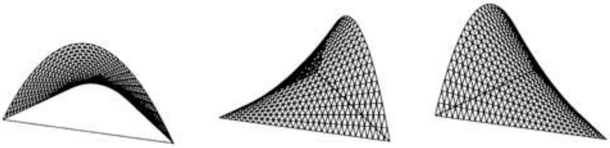


Figure 4.17 Nodal basis of the P^2-element; the edge functions $\varphi_{1,t}^{e_1}$, $\varphi_{1,t}^{e_2}$, and $\varphi_{1,t}^{e_3}$.

■ **EXAMPLE 4.4** **(P^3-element)**

The nodal basis of the P^3-element on the reference domain K_t is shown in Figures 4.18–4.21.

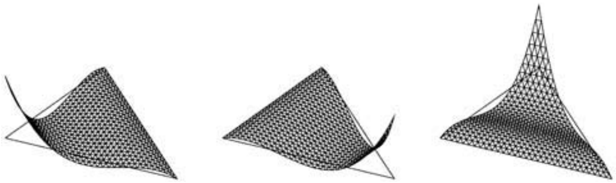


Figure 4.18 Nodal basis of the P^3-element; the vertex functions $\varphi_t^{v_1}$, $\varphi_t^{v_2}$, and $\varphi_t^{v_3}$.

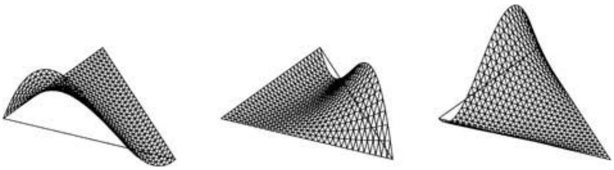


Figure 4.19 Nodal basis of the P^2-element; the edge functions $\varphi_{1,t}^{e_1}$, $\varphi_{1,t}^{e_2}$, and $\varphi_{1,t}^{e_3}$.

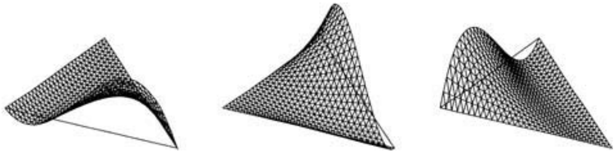


Figure 4.20 Nodal basis of the P^3-element; the edge functions $\varphi_{2,t}^{e_1}$, $\varphi_{2,t}^{e_2}$, and $\varphi_{2,t}^{e_3}$.

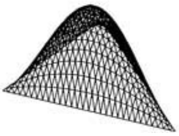


Figure 4.21 Nodal basis of the P^3-element; the bubble function $\varphi_{1,t}^{b}$.

4.3.6 Basis of the space $V_{h,p}$

Assume a regular hybrid mesh $\mathcal{T}_{h,p} = \{K_1, K_2, \ldots, K_M\}$ consisting of M_q Q^p-elements and M_p P^p-elements, $M_q + M_p = M \geq 1$. The requirement of a uniform polynomial degree p in the mesh is characteristic for nodal elements. The approximation could not be continuous with Lagrange elements of different polynomial degrees due to nonmatching nodal points on edges. This is illustrated in Figure 4.22.

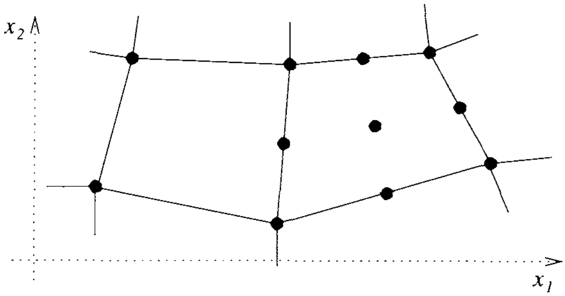


Figure 4.22 Mismatched nodal points on Q^1/Q^2-element interface.

It is our aim to use the shape functions defined Paragraphs 4.3.2 and 4.3.5 to construct the basis functions $v_1, v_2, \ldots, v_N$ of the space $V_{h,p}$. For this purpose, by x_i, $i = 1, 2, \ldots, M_v$ denote the unconstrained grid vertices, and by s_j, $j = 1, 2, \ldots, M_e$ the unconstrained mesh edges (by unconstrained we mean not lying on the Dirichlet boundary $\Gamma_{D,h}$).

Proposition 4.10 *The dimension of the finite element space $V_{h,p}$ is*

$$N = \dim(V_{h,p}) = M_v + (p-1)M_e + (p-1)^2 M_q + \frac{(p-1)(p-2)}{2} M_p.$$

Proof: Straightforward from the definition (4.9) of the space $V_{h,p}$. ■

There are M_v vertex functions associated with unconstrained grid vertices, $M_e(p-1)$ edge functions related to unconstrained mesh edges, and $M_q(p-1)^2 + M_p(p-1)(p-2)/2$ bubble functions associated with element interiors. These three types of basis functions are constructed as follows:

Vertex basis functions:

Assume the vertex element patch $S(i)$ corresponding to a grid vertex x_i, as illustrated in Figure 4.23.

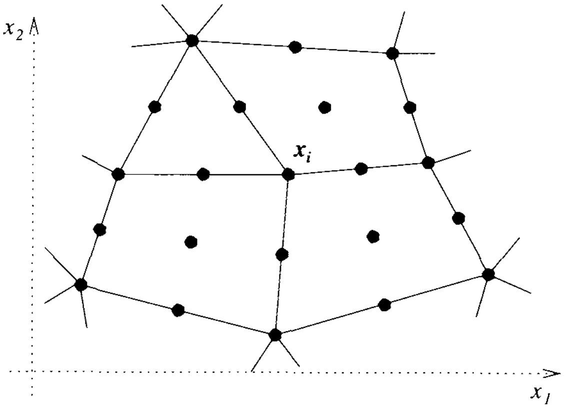


Figure 4.23 Element patch $S(i)$ corresponding to an unconstrained vertex x_i in a hybrid Q^2/P^2 mesh.

The vertex basis function v^{v_i} associated with $\boldsymbol{x}_i$ vanishes in $\Omega_h \setminus S(i)$, and in $S(i)$ it is defined analogously to (4.16),

$$\begin{aligned} v^{v_i}(\boldsymbol{x})|_{K_k} &= (\varphi_q^{v_r} \circ \boldsymbol{x}_{K_k}^{-1})(\boldsymbol{x}) \quad \text{if } K_k \in S(i) \text{ is a quadrilateral,} \\ v^{v_i}(\boldsymbol{x})|_{K_k} &= (\varphi_t^{v_r} \circ \boldsymbol{x}_{K_k}^{-1})(\boldsymbol{x}) \quad \text{if } K_k \in S(i) \text{ is a triangle.} \end{aligned} \tag{4.48}$$

Here $\varphi_q^{v_r}$ or $\varphi_t^{v_r}$ is the unique vertex nodal shape function of the polynomial degree p on K_q or K_t such that $\varphi_q^{v_r}(\boldsymbol{x}_{K_k}^{-1}(\boldsymbol{x}_i)) = 1$ or $\varphi_t^{v_r}(\boldsymbol{x}_{K_k}^{-1}(\boldsymbol{x}_i)) = 1$, respectively. The edge function $v^{v_i}(\boldsymbol{x})$ associated with a nodal point $\boldsymbol{x}_i$ vanishes at all remaining nodal points in the element patch $S(i)$.

Edge basis functions:

Assume an unconstrained mesh edge s_j with the endpoints $\boldsymbol{x}_{i_1}$ and $\boldsymbol{x}_{i_2}$. The global orientation of this edge can be defined, e.g., as the direction from the vertex with the lower index to the vertex with the greater index, i.e., $s_j = \boldsymbol{x}_{i_1}\boldsymbol{x}_{i_2}$ if $i_1 < i_2$ and $s_j = \boldsymbol{x}_{i_2}\boldsymbol{x}_{i_1}$ otherwise.

We define an edge element patch $S_e(j)$,

$$S_e(j) = \bigcup_{k \in N_e(j)} \overline{K}_k, \tag{4.49}$$

where

$$N_e(j) = \{k;\ K_k \in \mathcal{T}_{h,p},\ s_j \text{ is an edge of } K_k\}, \tag{4.50}$$

as shown in Figure 4.24.

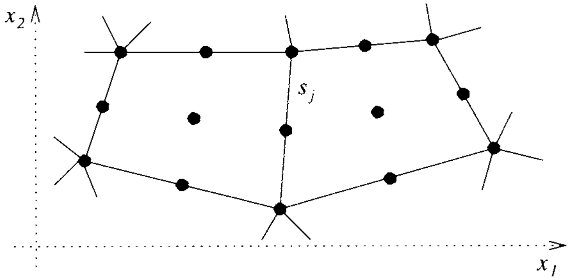


Figure 4.24 Element patch $S_e(j)$ corresponding to an unconstrained mesh edge s_j.

For each element $K_k \in S_e(i)$ by e_l denote the edge of the reference domain $\hat{K}$, such that $\boldsymbol{x}_{K_k}(e_l) = s_j$. Use the edge-interior nodal points on e_l and the reference map $\boldsymbol{x}_{K_k}$ to obtain coordinates of the edge-interior nodal points $\boldsymbol{x}_m^{s_j}$, $m = 1, 2, \dots, p-1$. These points are ordered on the edge s_j according to its global orientation (i.e., $\boldsymbol{x}_1^{s_j}$ is next to $\boldsymbol{x}_{i_1}$ and $\boldsymbol{x}_{p-1}^{s_j}$ is next to $\boldsymbol{x}_{i_2}$ if $s_j = \boldsymbol{x}_{i_1}\boldsymbol{x}_{i_2}$). There are $p-1$ edge basis functions $v_1^{s_j}, v_2^{s_j}, \dots, v_{p-1}^{s_j} \subset V_{h,p}$ associated with the points $\boldsymbol{x}_1^{s_j}, \boldsymbol{x}_2^{s_j}, \dots, \boldsymbol{x}_{p-1}^{s_j}$, respectively. Each edge function $v_m^{s_j}$, $1 \le m \le p-1$, is defined to be zero in $\Omega_h \setminus S_e(j)$, and in the patch $S_e(j)$ it satisfies

$$\begin{aligned} v_m^{s_j}(\boldsymbol{x})|_{K_k} &= (\varphi_{r,q}^{e_l} \circ \boldsymbol{x}_{K_k}^{-1})(\boldsymbol{x}) \quad \text{if } K_k \in S_e(j) \text{ is a quadrilateral,} \\ v_m^{s_j}(\boldsymbol{x})|_{K_k} &= (\varphi_{r,t}^{e_l} \circ \boldsymbol{x}_{K_k}^{-1})(\boldsymbol{x}) \quad \text{if } K_k \in S_e(j) \text{ is a triangle.} \end{aligned} \tag{4.51}$$

Here again $\varphi^{e_l}_{r,q}$ or $\varphi^{e_l}_{r,t}$ is the unique edge nodal shape function of the polynomial degree p on K_q or K_t, such that $\varphi^{e_l}_{r,q}(\boldsymbol{x}^{-1}_{K_k}(\boldsymbol{x}^{s_j}_m)) = 1$ or $\varphi^{e_l}_{r,t}(\boldsymbol{x}^{-1}_{K_k}(\boldsymbol{x}^{s_j}_m)) = 1$, respectively. The edge function $v^{s_j}_m(\boldsymbol{x})$ associated with a nodal point $\boldsymbol{x}^{s_j}_m$ vanishes at all remaining nodal points in the element patch $S(i)$.

<u>Bubble basis functions</u>:

Last to be defined are the bubble basis functions. There are $(p-1)^2$ bubble functions in each quadrilateral and $(p-1)(p-2)/2$ in each triangular element. The nodal points in the mesh element K_k are defined, as usual, to be the images of the interior nodal points in the corresponding reference domain $\hat{K}$ through the reference map $\boldsymbol{x}_{K_k} : \hat{K} \to K_k$.

Consider, for example, a triangular element $K_k \in \mathcal{T}_{h,p}$ and the interior nodal points $\boldsymbol{x}^{K_k}_1, \boldsymbol{x}^{K_k}_2, \ldots, \boldsymbol{x}^{K_k}_{(p-1)(p-2)/2}$. The bubble function $v^{K_k}_m$ associated with the nodal point $\boldsymbol{x}^{K_k}_m$ is defined to vanish in $\Omega_h \setminus K_k$, and in K_k we have

$$v^{K_k}_m(\boldsymbol{x}) = (\varphi^b_{r,t} \circ \boldsymbol{x}_{K_k})(\boldsymbol{x}). \tag{4.52}$$

Here $\varphi^b_{r,t} \in P(K_t)$ is the bubble shape function satisfying $\varphi^b_{r,t}(\boldsymbol{x}^{-1}_{K_k}(\boldsymbol{x}^{K_k}_m)) = 1$. The bubble function $v^{K_k}_m(\boldsymbol{x})$ associated with a nodal point $\boldsymbol{x}^{K_k}_m$ vanishes at all remaining nodal points in the element K_k, and thus also on its boundary ∂K_k.

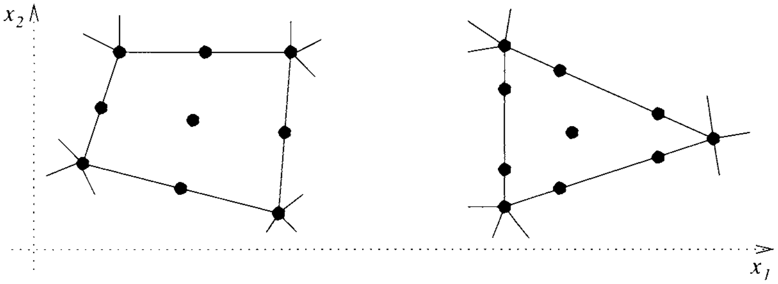


Figure 4.25 There is a single biquadratic bubble function on every Q^2-element, and a single cubic bubble function appears on P^3-elements.

Proposition 4.11 *The functions (4.48), (4.51), and (4.52) are continuous in Ω_h and constitute together a basis of the space $V_{h,p}$.*

Proof: This follows easily from the linear independence of basis functions associated with different nodal points in Ω_h. ■

4.3.7 Data structures

Before presenting the element-by-element assembling procedure in Paragraph 4.3.9, let us discuss the construction of the connectivity arrays. Again let the hybrid Q^p/P^p mesh $\mathcal{T}_{h,p}$ be represented via an element array `ElementP *Elem` of the length M.

Element data structure The `Element` data structure from Paragraph 4.1.6 can be extended to the higher-order case as follows:

```
struct {
  int nv;               //number of vertices
                        //(4 for quads, 3 for triangles)
  int *vert;            //global vertex indices (length nv)
  int *vert_dir;        //vertex Dirichlet flags (length nv)
  int *vert_dof;        //vertex connectivity array (length nv)
  int *edge_dir;        //edge Dirichlet flags (length nv)
  int **edge_dof;       //two-dimensional edge connectivity array
                        //(dimension nv*(MAXP-1))
  int *bubb_dof;        //bubble connectivity array
                        //(length (MAXP-1)*(MAXP-1) for quads,
                        //and (MAXP-1)*(MAXP-2)/2 for triangles)
  int *o;               //edge orientation flags (length nv)
  ...
} ElementP;
```

Here MAXP is the maximum allowed polynomial degree of the finite elements. The ElementP data structure can be optimized (the stored data are not independent) but we prefer this form for the sake of transparency. The optimization of data structures and algorithms will be described at the end of Paragraph 4.3.9. The vertex indices vert, vertex Dirichlet flags vert_dir and the vertex connectivity arrays vert_dof are used analogously to Paragraph 4.1.6. The meaning of the other variables is described below.

Edge Dirichlet flags The function of the edge Dirichlet flags Elem[m].edge_dir is analogous to the flags Elem[m].vert_dir: The variable Elem[m].edge_dir[j], $j = 1, 2, \ldots, nv$, is zero if the edge $\boldsymbol{x}_{K_m}(e_j)$ of K_m is unconstrained (i.e., not lying on the Dirichlet boundary $\Gamma_{D,h}$), and one otherwise. These flags are defined easily, using the fact that an edge is constrained if and only if both of its vertices are constrained (see Algorithm 4.3).

Edge orientation flags (for $p \geq 3$ only) When the number of edge-interior nodal points exceeds one (i.e., for $p \geq 3$), one has to take care about the orientation of the edges. Assume an element $K_m \in \mathcal{T}_{h,p}$, the appropriate reference domain $\hat{K} = K_q$ or $\hat{K} = K_t$, and the reference map $\boldsymbol{x}_{K_m} : \hat{K} \to K_m$. Let $s_j = \boldsymbol{x}_{i_1}\boldsymbol{x}_{i_2}$, $i_1 < i_2$, be an edge of K_m, and let e_k be the corresponding edge of $\hat{K}$, i.e., $s_j = \boldsymbol{x}_{K_m}(e_k)$. Since the orientations of s_j and e_k are independent, it is either

$$\text{(A)} \quad \boldsymbol{x}_{K_m}(e_k) = \boldsymbol{x}_{i_1}\boldsymbol{x}_{i_2} \quad \text{or} \quad \text{(B)} \quad \boldsymbol{x}_{K_m}(e_k) = \boldsymbol{x}_{i_2}\boldsymbol{x}_{i_1}.$$

The ElementP data structure contains the array Elem[m].o[] = ± 1 of the length nv for this purpose. In case (A) the orientations of s_j and e_k are compatible, i.e., the reference map $\boldsymbol{x}_{K_m}$ preserves the ordering of the edge-internal nodes,

$$\text{(A)} \quad \boldsymbol{x}_r^{s_j} = \boldsymbol{x}_{K_m}(\boldsymbol{v}_r^{e_k}) \qquad \text{for all } 1 \leq r \leq p-1,$$

and we define Elem[m].o[k] = 1. In the opposite case the ordering of the edge-internal nodes is reversed,

$$\text{(B)} \quad \boldsymbol{x}_r^{s_j} = \boldsymbol{x}_{K_m}(\boldsymbol{v}_{p-r}^{e_k}) \qquad \text{for all } 1 \leq r \leq p-1,$$

and we define Elem[m].o[k] = -1. This will be done in Algorithm 4.3.

Unique enumeration of edges Mesh generators always provide a list of vertices and a list of elements. This defines their unique enumeration as necessary for the definition of the vertex and bubble connectivities. Also a list of edges is needed for the definition of edge connectivities, but such a list usually is not provided by mesh generators. Therefore let us present a simple algorithm that enumerates unconstrained mesh edges. We begin with a data structure for the edges,

```
struct {
  int n1, n2;
  int e1, e2;
} TmpEdgeData;
```

Here `n1` $<$ `n2` are the indices of the vertices of the edge that define its orientation, and `e1` $<$ `e2` the indices of the adjacent elements. These entries will be defined for every unconstrained mesh edge in Algorithm 4.3. The list of the edges,

```
TmpEdgeData *EdgeList;
```

has the length $4M$. This is quite a crude upper bound, but `EdgeList` will be deallocated immediately after the element connectivity arrays are defined.

Algorithm 4.3 (Creating a temporary list of edges)

```
length := 0; //Current length of EdgeList
for m = 1,2,...,M do {
  if (Elem[m].nv == 4) then {    //$K_m$ is a quadrilateral
    //The first edge of $K_m$:  Defining the orientation flag:
    vA :=  Elem[m].vert[1];
    vB :=  Elem[m].vert[3];
    if (vA < vB) then Elem[m].o[1] := 1;
    else Elem[m].o[1] := -1;
    //The first edge of $K_m$:  Defining the Dirichlet flag:
    dirA := Elem[m].vert_dir[1];
    dirB := Elem[m].vert_dir[3];
    if (dirA*dirB == 1) then Elem[m].edge_dir[1] := 1;
    else Elem[m].edge_dir[1] := 0;
    //The first edge of $K_m$:  Adding to EdgeList
    //(if unconstrained and not visited before)
    if (Elem[m].edge_dir[1] == 0) then {
      CheckEdgeList(vA,vB,EdgeList,length,&found,&pos);
      if (found == 0) then { //The edge was not found in EdgeList
        length := length + 1;
        if (vA < vB) then {
          EdgeList[length].n1 := vA;
          EdgeList[length].n2 := vB;
        }
        else {
          EdgeList[length].n1 := vB;
          EdgeList[length].n2 := vA;
        }
        EdgeList[length].e1 := m;
        EdgeList[length].e2 := -1;
      }
      else { //The edge was found in EdgeList on the position pos
        EdgeList[pos].e2 := m;
      }
    }
    ...  //The same for the remaining three edges:
    ...  //2nd edge:  vA = Elem[m].vert[2], vB = Elem[m].vert[4]
    ...  //3rd edge:  vA = Elem[m].vert[1], vB = Elem[m].vert[2]
    ...  //4th edge:  vA = Elem[m].vert[3], vB = Elem[m].vert[4]
  }
  else {   //$K_m$ is a triangle
    //The first edge of $K_m$:  Defining the orientation flag:
    vA :=  Elem[m].vert[1];
    vB :=  Elem[m].vert[2];
    if (vA < vB) then Elem[m].o[1] := 1;
    else Elem[m].o[1] := -1;
    //The first edge of $K_m$:  Defining the Dirichlet flag:
    dirA := Elem[m].vert_dir[1];
    dirB := Elem[m].vert_dir[2];
    if (dirA*dirB == 1) then Elem[m].edge_dir[1] := 1;
    else Elem[m].edge_dir[1] := 0;
    //The first edge of $K_m$:  Adding to EdgeList
    //(if unconstrained and not visited before)
    if (Elem[m].edge_dir[1] == 0) then {
      CheckEdgeList(vA,vB,EdgeList,length,&found,&pos);
      if (found == 0) then { //The edge was not found in EdgeList
        length := length + 1;
        if (vA < vB) then {
          EdgeList[length].n1 := vA;
          EdgeList[length].n2 := vB;
        }
        else {
          EdgeList[length].n1 := vB;
          EdgeList[length].n2 := vA;
        }
        EdgeList[length].e1 := m;
        EdgeList[length].e2 := -1;
      }
      else { //The edge was found in EdgeList on the position pos
        EdgeList[pos].e2 := m;
      }
    }
    ...  //The same for the remaining two edges:
    ...  //2nd edge:  vA = Elem[m].vert[2], vB = Elem[m].vert[3]
    ...  //3rd edge:  vA = Elem[m].vert[3], vB = Elem[m].vert[1]
  }
}
$M_e$ := length - 1; //The number of unconstrained mesh edges
```

Here, the function CheckEdgeList(vA,vB,EdgeList,length,&found,&pos) parses the EdgeList and tests if either $\{$vA,vB$\}$ or $\{$vB,vA$\}$ are present. If found, it returns found := 1 and the corresponding position pos, otherwise it returns found := 0.

4.3.8 Connectivity arrays

Now the edge and bubble connectivity arrays edge_dof and bubb_dof can be defined. The jth component of the array Elem[m].edge_dof[i], $1 \leq i \leq nv$, $1 \leq j \leq p-1$, contains either

- the index of the edge basis function of the space $V_{h,p}$ associated with the jth internal node $\boldsymbol{x}_{K_m}(\boldsymbol{v}_j^{e_i})$ on the ith edge of K_m (if Elem[m].edge_dir[i] == 0)
- or a negative integer number -NBC (if Elem[m].edge_dir[i] == 1).

In the case of nonhomogeneous boundary conditions, the values of the Dirichlet lift G at the edge-internal nodes of constrained edges can be stored via an array of real numbers. The index NBC can be used to indicate a position in this array, where the value of the Dirichlet lift G at the corresponding edge-internal node of the element K_m is stored (analogously to the treatment of constrained vertices in Paragraph 4.1.6). With this construction, the implementation of nonhomogeneous Dirichlet boundary conditions is straightforward.

The algorithm for the edge connectivities is based on the temporary array EdgeList and proceeds in an edge-by-edge fashion. As before, let M_v be the number of unconstrained grid vertices (vertex DOF) and M_e the number of unconstrained mesh edges. The algorithm will add $p - 1$ edge-internal DOF to every unconstrained edge.

Algorithm 4.4 (Enumeration of edge DOF)

```
//Loop over unconstrained edges:
for e = 1,2,...,Me do {
  //Lower-index element adjacent to the edge EdgeList[e]:
  e1 = EdgeList[e].e1;
  if (Elem[e1].nv == 4) then {   //$K_{e_1}$ is a quadrilateral
    //Locate the edge EdgeList[e] in the element Elem[e1]:
    a1 := Elem[e1].vert[1]; a2 := Elem[e1].vert[2];
    a3 := Elem[e1].vert[3]; a4 := Elem[e1].vert[4];
    b1 := EdgeList[e].n1; b2 := EdgeList[e].n2;
    if ((b1==a1 and b2==a3) or (b1==a3 and b2==a1)) then ee:=1;
    if ((b1==a2 and b2==a4) or (b1==a4 and b2==a2)) then ee:=2;
    if ((b1==a1 and b2==a2) or (b1==a2 and b2==a1)) then ee:=3;
    if ((b1==a3 and b2==a4) or (b1==a4 and b2==a3)) then ee:=4;
    //Enumerate the edge-internal DOF on the ee-th edge of Elem[e1]:
    if (Elem[e1].o[ee] == 1) then for j = 1,2,...,p-1 do {
      //(the local and global orientations are compatible)
      Elem[e1].edge_dof[ee][j] := Mv + (p-1)*(e-1) + j;
      //Here:  Mv is the number of vertex DOF, and (p-1)*(e-1) is the
      //number of edge-internal DOF assigned to previously visited edges.
    }
    else {
      //(incompatible orientations -- the ordering of local DOF is reversed)
      Elem[e1].edge_dof[ee][p-j] := Mv + (p-1)*(e-1) + j;
    }
  }
  else {   //$K_{e_1}$ is a triangle
    //Locate the edge EdgeList[e] in the element Elem[e1]:
    a1 := Elem[e1].vert[1]; a2 := Elem[e1].vert[2];
    a3 := Elem[e1].vert[3];
    b1 := EdgeList[e].n1; b2 := EdgeList[e].n2;
    if ((b1==a1 and b2==a2) or (b1==a2 and b2==a1)) then ee:=1;
    if ((b1==a2 and b2==a3) or (b1==a3 and b2==a2)) then ee:=2;
    if ((b1==a3 and b2==a1) or (b1==a1 and b2==a3)) then ee:=3;
    //Enumerate the edge-internal DOF on the ee-th edge of Elem[e1]:
    if (Elem[e1].o[ee] == 1) then for j = 1,2,...,p-1 do {
      //(the local and global orientations are compatible)
      Elem[e1].edge_dof[ee][j] := Mv + (p-1)*(e-1) + j;
      //Here:  Mv is the number of vertex DOF, and (p-1)*(e-1) is the
      //number of edge-internal DOF assigned to previously visited edges.
    }
    else {
      //(incompatible orientations -- the ordering of local DOF is reversed)
      Elem[e1].edge_dof[ee][p-j] := Mv + (p-1)*(e-1) + j;
    }
  }
  //Higher-index element adjacent to the edge EdgeList[e]:
  e2 = EdgeList[e].e2;
  if (e2 > -1) then {
    //Perform now the same operations as for Elem[e1] above.
    ...
  }
}
Deallocate EdgeList
```

The distribution of the remaining $M_q(p-1)^2 + M_p(p-1)(p-2)/2$ bubble DOF to element interiors is simpler:

Algorithm 4.5 (Enumeration of bubble DOF)

```
bubb_dof_count := Mv + (p-1)*Me;  //Number of previously assigned DOF
for m = 1,2,...,M do {
  if (Elem[m].nv == 4) then {   //$K_m$ is a quadrilateral
    for i = 1,2,...,(p-1)*(p-1) do {
      bubb_dof_count := bubb_dof_count + 1;
      Elem[m].bubb_dof[i] := bubb_dof_count;
    }
  }
  else {   //$K_m$ is a triangle
    for i = 1,2,...,(p-1)*(p-2)/2 do {
      bubb_dof_count := bubb_dof_count + 1;
      Elem[m].bubb_dof[i] := bubb_dof_count;
    }
  }
}
```

The connectivity arrays Elem[m].vert_dof, Elem[m].edge_dof and Elem[m].bubb _dof on all elements $K_m \in \mathcal{T}_{h,p}$ are now ready. The connectivity algorithms can be written without storing the edge orientation flags Elem[m].o explicitly. The reader can remove them after getting more familiar with the algorithm.

■ **EXAMPLE 4.5 (Connectivity arrays)**

Consider a mesh consisting of four quadratic Lagrange elements as shown in Figure 4.23. Let the reference maps be chosen in such a way that the lower-left vertex of the reference domain always is linked to the lower-left corner of the physical element. If we consider, for example, a problem with homogeneous Dirichlet boundary conditions, then the dimension of the space $V_{h,p}$ equals 8, and the basis functions are enumerated as shown in Figure 4.26.

4.3.9 Assembling algorithm for Q^p/P^p-elements

The extension of Algorithm 4.2 to the Q^p/P^p-meshes is not complicated. Let us consider the same setting as in Paragraph 4.3.9, i.e., the model problem (4.2) with homogeneous Dirichlet boundary conditions. Moreover we assume that the simplifying conditions on the data formulated in Paragraph 4.1.5 are met. The following constants stay unchanged on all elements K_m, $1 \leq m \leq M$: The Jacobian

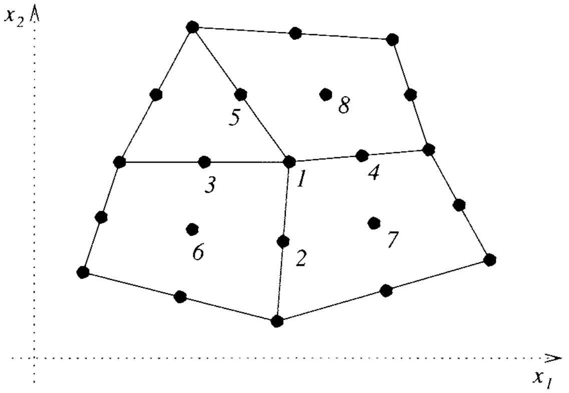


Figure 4.26 Enumeration of basis functions for a simple mesh consisting of four second-order nodal elements.

$$\texttt{Elem[m].jac} := |J_{K_m}|,$$

and the entries of the inverse Jacobi matrix

$$\texttt{Elem[m].inv\_j[r][n]} := \frac{\partial \xi_r^{(m)}}{\partial x_n},$$

for all $1 \le n, r \le d$.

The four-dimensional array MESI_Q is extended to cover all combinations of shape functions on the reference domain K_q,

$$\texttt{MESI\_Q[k][l][r][s]} := \int_{\hat{K}} \frac{\partial \varphi_l}{\partial \xi_r} \frac{\partial \varphi_k}{\partial \xi_s} \, \mathrm{d}\boldsymbol{\xi}, \quad 1 \le k, l \le (p+1)^2, \ 1 \le r, s \le d,$$

where $\varphi_1, \varphi_2, \ldots, \varphi_{(p+1)^2}$ are the four vertex functions (4.35) associated with the nodes (4.30), followed by the $4(p-1)$ edge functions (4.36) related to the nodes (4.31) for each edge $e_1, e_2, \ldots, e_4$, and by the $(p-1)^2$ bubble functions (4.37) corresponding to the interior nodes (4.32). All these shape functions were uniqely enumerated. The array MESI_T is extended to

$$\texttt{MESI\_T[k][l][r][s]} := \int_{\hat{K}} \frac{\partial \varphi_l}{\partial \xi_r} \frac{\partial \varphi_k}{\partial \xi_s} \, \mathrm{d}\boldsymbol{\xi}, \quad 1 \le k, l \le (p+1)(p+2)/2, \ 1 \le r, s \le d,$$

where $\varphi_1, \varphi_2, \ldots, \varphi_{(p+1)(p+2)/2}$ stand for the three vertex functions associated with the nodes (4.46), followed by the $3(p-1)$ edge functions related to the nodes (4.47) for each edge e_1, e_2, e_3, and by the $(p-1)(p-2)/2$ bubble functions corresponding to the interior nodes. In the same way the master element mass integrals MEMI are extended to cover all combinations of the shape functions. The functions `double SMC(Elem,k,l,m,MESI_Q,MEMI_Q)` and `double SMC(Elem,k,l,m,MESI_T,MEMI_T)`, that calculate the stiffness matrix contribution (4.24), stay unchanged.

The assembling algorithm is analogous to Algorithm 4.2, only now it covers all combinations of the shape functions on the reference domain.

Algorithm 4.6 (Assembling algorithm for higher-order Lagrange elements)

```
N := $M_v + (p-1)M_e + (p-1)^2 M_q + (p-1)(p-2)/2M_p$;
//Set the stiffness matrix $\boldsymbol{S}$ zero:
for i = 1,2,...,N do for j = 1,2,...,N do S[i][j] := 0;
//Set the right-hand side vector $\boldsymbol{F}$ zero:
for i = 1,2,...,N do F[i] := 0;
//Element loop:
for m = 1,2,...,M do {
  //Loop over vertex test functions:
  for i = 1,2,...,Elem[m].nv do {
    //Index of the vertex test function $v_{m_1} \in V_{h,p}$
    //(row position in $\boldsymbol{S}$)
    m1 := Elem[m].vert_dof[i];
    //Loop over all vertex, edge and bubble basis functions:
    //(Filling the $m_1$th row of $\boldsymbol{S}$)
    //1.  loop over vertex basis functions:
    if (m1 > -1) then for j = 1,2,...,Elem[m].nv do {
      //Index of the vertex basis function $v_{m_2} \in V_{h,p}$
      //(column position in $\boldsymbol{S}$)
      m2 := Elem[m].vert_dof[j];
      if (m2 > -1) then {
        if (Elem[m].nv == 4 then {
          S[m1,m2] := S[m1,m2] + SMC(Elem,i,j,m,MESI_Q,MEMI_Q);
        }
        else {
          S[m1,m2] := S[m1,m2] + SMC(Elem,i,j,m,MESI_T,MEMI_T);
        }
      }
    } //End of loop over vertex basis functions
    //2.  loop over edge basis functions:
    if (m1 > -1) then for j = 1,2,...,Elem[m].nv do {
      for k = 1,2,...,p-1 do {
        //Index of the edge basis function $v_{m_2} \in V_{h,p}$
        //(column in $\boldsymbol{S}$)
        m2 := Elem[m].edge_dof[j][k];
        if (m2 > -1) then {
          if (Elem[m].nv == 4 then {
            S[m1,m2] := S[m1,m2] + SMC(Elem,i,4+j,m,MESI_Q,MEMI_Q);
          }
          else {
            S[m1,m2] := S[m1,m2] + SMC(Elem,i,3+j,m,MESI_T,MEMI_T);
          }
        }
      }
    } //End of loop over edge basis functions
    //3.  loop over bubble basis functions:
    if (Elem[m].nv == 4 then {
      if (m1 > -1) then for k = 1,2,...,(p-1)*(p-1) do {
        //Index of the bubble basis function $v_{m_2} \in V_{h,p}$
        //(column in $\boldsymbol{S}$)
        m2 := Elem[m].bubb_dof[k];
        S[m1,m2] := S[m1,m2] + SMC(Elem,i,4+4*(p-1)+j,m,MESI_Q,MEMI_Q);
      }
    }
    else {
      if (m1 > -1) then for k = 1,2,...,(p-1)*(p-2)/2 do {
        //Index of the bubble basis function $v_{m_2} \in V_{h,p}$
        //(column in $\boldsymbol{S}$)
        m2 := Elem[m].bubb_dof[k];
        S[m1,m2] := S[m1,m2] + SMC(Elem,i,4+4*(p-1)+j,m,MESI_T,MEMI_T);
      }
    } //End of loop over bubble basis functions
    //Now the $m_1$th row of the stiffness matrix $\boldsymbol{S}$ is filled.
    //Contribution of the vertex test function $v_{m_1}$ to the right-hand side $\boldsymbol{F}$:
    if (m1 > -1) then {
      F[m1] := F[m1] + Elem[m].jac*$\int_{\hat{K}} \tilde{f}^{(m)}(\boldsymbol{\xi})\varphi^{v_i}(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}$;
    }
  } //End of loop over vertex test functions
  //Now the $M_v$ rows in the linear algebraic system $\boldsymbol{SY} = \boldsymbol{F}$
  //corresponding to all vertex basis functions of the space $V_{h,p}$
  //are filled.
  //Next fill the rows of $\boldsymbol{SY} = \boldsymbol{F}$ corresponding to all edge test functions:
  //Loop over edge test functions:
  for i = 1,2,...,Elem[m].nv do {
    for l = 1,2,...,p-1 do {
      //Index of the edge test function $v_{m_1} \in V_{h,p}$
      //(row in $\boldsymbol{S}$)
      m1 := Elem[m].edge_dof[i][l];
      //Loop over all vertex, edge and bubble basis functions:
      ...
      //Contribution of the edge test function $v_{m_1}$ to the right-hand side $\boldsymbol{F}$:
      if (m1 > -1) then {
        F[m1] := F[m1] + Elem[m].jac*$\int_{\hat{K}} \tilde{f}^{(m)}(\boldsymbol{\xi})\varphi_l^{e_i}(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}$;
      }
    }
  } //End of loop over edge test functions
  //At last fill the rows of $\boldsymbol{SY} = \boldsymbol{F}$ corresponding to all bubble test functions:
  if (Elem[m].nv == 4 then {
    for k = 1,2,...,(p-1)*(p-1) do {
      //Index of the bubble test function $v_{m_1} \in V_{h,p}$
      //(row in $\boldsymbol{S}$)
      m1 := Elem[m].bubb_dof[k];
      //Loop over all vertex, edge and bubble basis functions:
      ...
      //Contribution of the bubble test function $v_{m_1}$ to right-hand side $\boldsymbol{F}$:
      F[m1] := F[m1] + Elem[m].jac*$\int_{\hat{K}} \tilde{f}^{(m)}(\boldsymbol{\xi})\varphi_k^b(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}$;
    }
  }
  else {
    for k = 1,2,...,(p-1)*(p-1)/2 do {
      //Index of the bubble test function $v_{m_1} \in V_{h,p}$
      //(row in $\boldsymbol{S}$)
      m1 := Elem[m].bubb_dof[k];
      //Loop over all vertex, edge and bubble basis functions:
      ...
      //Contribution of the bubble test function $v_{m_1}$ to right-hand side $\boldsymbol{F}$:
      F[m1] := F[m1] + Elem[m].jac*$\int_{\hat{K}} \tilde{f}^{(m)}(\boldsymbol{\xi})\varphi_k^b(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}$;
    }
  } //End of loop over bubble test functions
} //End of element loop
```

If the simplifying conditions formulated in Paragraph 4.1.5 do not apply, then the Jacobian, the entries of the inverse Jacobi matrix, and other values are no longer constant in the elements. In such case, (4.24) has to be replaced with the more general relation (4.23), and instead of reading the precomputed entries from the MESI and MEMI arrays, the corresponding integrals have to be evaluated numerically.

Optimization of Algorithm 4.6 Significant part of Algorithm 4.6 (the application of a given test function to all vertex, edge and bubble basis functions) was repeated with minor changes four times. The algorithm was presented in this full form for the sake of transparency, but in practice the repeated part can be handled via a subroutine whose input parameters identify the given test function. The corresponding reformulation of Algorithm 4.6 is straightforward.

Moreover, all indices in the connectivity arrays `Elem[m].edge_dof[j], j = 1, 2, ..., Elem[m].nv` are determined uniquely by the first index `Elem[m].edge_dof[j][1]` and the orientation flag `Elem[m].o[j]`. Therefore they do not need be stored explicitly. Analogously, it is sufficient to store just the first index of the bubble connectivity array, `Elem[m].bubb_dof[1]`, instead of the whole array `Elem[m].bubb_dof`. It can be recommended that the reader performs these optimization steps after a first version of the code is working.

4.3.10 Lagrange interpolation on Q^p/P^p-meshes

The global interpolant of a function $g \in C(\overline{\Omega}_h)$ on a regular mesh $\mathcal{T}_{h,p}$ consisting of Q^p- and/or P^p-elements is obtained analogously to the Q^1/P^1-case from Paragraph 4.1.8.

Proposition 4.12 *The global Lagrange interpolant $\mathcal{I}(g)$ is continuous in $\overline{\Omega}_h$ for every function $g \in C(\overline{\Omega}_h)$. Thus every regular mesh consisting of Q^p- and/or P^p-elements is conforming to the space $H^1(\Omega_h)$.*

Proof: This is left to the reader as an easy exercise. ■

As usual, the global interpolant is evaluated elementwise on the reference domains, using the sets of Gauss–Lobatto and Fekete points, and the Lagrange–Gauss–Lobatto and Lagrange–Fekete nodal shape functions. With $p = 1$, one obtains the lowest-order case discussed in Paragraph 4.1.8.

4.3.11 Exercises

Exercise 4.3 *Prove Proposition 4.6.*

Exercise 4.4 *Prove Proposition 4.7.*

Exercise 4.5 *Prove Proposition 4.8.*

Exercise 4.6 *Prove Proposition 4.10.*

Exercise 4.7 *In Algorithm 4.6, replace the repeated application of a given test function to all vertex, edge and bubble basis functions with a suitable subroutine.*

Exercise 4.8 *Extend your code from Exercise 4.2 to Q^2 elements using Algorithm 4.6.*

1. *Present plots of the approximate solution for the parameters*

   (a) $a = 2,\ b = 1,\ M_1 = 4,\ M_2 = 2,$

   (b) $a = 2,\ b = 1,\ M_1 = 10,\ M_2 = 5,$

   (c) $a = 2,\ b = 1,\ M_1 = 20,\ M_2 = 10,$

   (d) $a = 2,\ b = 1,\ M_1 = 40,\ M_2 = 20.$

2. *Present the convergence curve of the above computations in the $H^1(\Omega)$-seminorm. Compare it with the convergence curve from Exercise 4.2.*

Exercise 4.9 *Prove Proposition 4.12.*

CHAPTER 5

TRANSIENT PROBLEMS AND ODE SOLVERS

The nature changes at every time instant, and the numerical simulation of evolutionary processes plays an important role in applied sciences and engineering. Transient problems can be very complicated and the spectrum of numerical methods for their solution is accordingly wide.

At the introductory level it is natural to begin with the Method of lines (MOL), which has a prominent position due to its ability to add temporal evolution to all numerical methods for stationary PDEs without altering the spatial discretization. This is demonstrated in Section 5.1, where we exploit the finite element technology developed in Chapters 2–4. With the MOL this is the "easy part", and the "real work" is done by solving the arising system of ordinary differential equations (ODEs). Therefore, the largest part of this chapter is devoted to modern ODE solvers.

Section 5.2 introduces the general concept of one-step methods, which are the best candidates to be used for MOL in combination with adaptive finite element methods. The discussion continues with the properties and implementation aspects of explicit and implicit Euler methods and higher-order Runge–Kutta (RK) schemes. Section 5.3 introduces the reader to stability analysis of ODEs and ODE solvers. Basic understanding of stability helps the reader to use the ODE solvers adequately and efficiently. Section 5.4 presents the nowadays most popular implicit higher-order methods, including the Gauss and Radau implicit Runge–Kutta (IRK) schemes. Presented are both the classical and simplified Newton's methods for the solution of nonlinear algebraic systems arising in implicit ODE solvers.

5.1 METHOD OF LINES

In this section the reader will need some basic facts about second-order parabolic problems from Chapter 1. Recall the general second-order parabolic equation (1.76),

$$\frac{\partial u}{\partial t} + Lu = f, \tag{5.1}$$

where L is a second-order elliptic operator of the form (1.1), $\Omega \subset \mathbb{R}^2$ is a bounded domain with Lipschitz-continuous boundary, $T > 0$, $Q_T = \Omega \times (0, T)$ is the corresponding space-time cylinder, and $f \in C(Q_T)$. The classical regularity assumptions are weakened after the problem is stated in the weak sense. For example, in the special case

$$Lu = -\frac{k}{\varrho c}\Delta u, \qquad f = \frac{q}{\varrho c},$$

equation (5.1) describes the temporal evolution of the temperature u induced by heat sources of the density q in a domain Ω filled with an isotropic material. Here k is the thermal conductivity, ϱ the material density, and c the specific heat of the material.

5.1.1 Model problem

Assume that $\partial\Omega$ consists of two disjoint open pieces Γ_D and Γ_N such that

$$\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$$

(see Figure 4.1). Equation (5.1) is equipped with the Dirichlet boundary conditions

$$u(\boldsymbol{x}, t) = g_D(\boldsymbol{x}) \qquad \text{for all } (\boldsymbol{x}, t) \in \Gamma_D \times (0, T), \tag{5.2}$$

Neumann boundary conditions

$$\frac{\partial u}{\partial \boldsymbol{\nu}}(\boldsymbol{x}, t) = g_N(\boldsymbol{x}) \qquad \text{for all } (\boldsymbol{x}, t) \in \Gamma_N \times (0, T), \tag{5.3}$$

and, moreover, with an initial condition

$$u(\boldsymbol{x}, 0) = u_0(\boldsymbol{x}) \qquad \text{for all } \boldsymbol{x} \in \Omega. \tag{5.4}$$

For simplicity. let the functions g_D and g_N be time-independent in the following.

5.1.2 Weak formulation

We learned in Chapter 1 how to formulate problem (5.1)–(5.4) in the weak sense: The nonhomogeneous boundary data g_D is represented by a suitable Dirichlet lift $G \in H^1(\Omega)$, such that $G = g_D$ on Γ_D in the sense of traces. The solution u is written as a sum

$$u(\boldsymbol{x}, t) = G(\boldsymbol{x}) + U(\boldsymbol{x}, t),$$

where for all $t \in (0, T)$

$$U(t) \in V = \{v \in H^1(\Omega);\ v|_{\Gamma_D} = 0\}.$$

Using the notation $(U(t))(\cdot) = U(\cdot, t)$, the weak formulation reads:

Given $f \in L^2(Q_T)$ and $U_0 = u_0 - G \in V$, find $U \in L^2(0,T;V) \cap C^0([0,T];L^2(\Omega))$ such that

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega (U(t))(\boldsymbol{x})v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + a(U(t),v) &= l(v) \quad \text{for all } v \in V, && (5.5)\\ U(0) &= U_0, && (5.6) \end{aligned}$$

in the sense of distributions. Both the bilinear form $a(\cdot,\cdot)$ and the linear form $l(\cdot)$ were defined in Chapter 4.

5.1.3 The ODE system

The basic idea of the Method of lines is to keep the temporal variable t continuous while the spatial part of the problem is discretized analogously to time-independent problems. This technique is called semidiscretization in space. The outline of the procedure is as follows: Perform all spatial approximation steps described in Paragraph 4.1.2 and design the piecewise-polynomial space $V_{h,p} \subset V$ according to the finite element mesh $\mathcal{T}_{h,p}$. Construct a suitable basis

$$\{v_1, v_2, \ldots, v_N\} \subset V_{h,p}.$$

Express the sought function $U_{h,p}$ as a linear combination of the basis functions v_j, $j = 1, 2, \ldots, N$, with time-dependent coefficients $y_j(t)$,

$$U_{h,p}(\boldsymbol{x},t) = \sum_{j=1}^{N} y_j(t)v_j(\boldsymbol{x}) \tag{5.7}$$

[compare to (4.11)].

The variational formulation (5.5) is approximated using the sequence of approximations listed in Paragraph 4.1.2: The domain Ω is replaced with a simpler domain Ω_h suitable for meshing, boundary conditions are moved from $\partial\Omega$ to the new boundary $\partial\Omega_h$, coefficients and data are extended to Ω_h if $\Omega_h \not\subset \Omega$, the space V is replaced with a piecewise-polynomial space $V_{h,p}$ built on the finite element mesh, exact integration is replaced with the Gaussian quadrature, etc. After inserting the construction (5.7) into the approximate variational formulation, one obtains

$$\sum_{j=1}^{N} \dot{y}_j(t) \underbrace{\int_{\Omega_h} v_j(\boldsymbol{x})v_i(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}}_{m_{ij}} + \sum_{j=1}^{N} y_j(t) \underbrace{a(v_j, v_i)}_{s_{ij}} = l(v_i), \tag{5.8}$$

$i = 1, 2, \ldots, N$. Written in matrix form, (5.8) reads

$$\boldsymbol{M}\dot{\boldsymbol{Y}}(t) + \boldsymbol{S}\boldsymbol{Y}(t) = \boldsymbol{F}(t). \tag{5.9}$$

Here, $\boldsymbol{M}$ is the mass matrix,

$$m_{ij} = (v_j, v_i)_{L^2} = \int_{\Omega_h} v_j(\boldsymbol{x}) v_i(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}, \quad 1 \le i, j \le N,$$

$\boldsymbol{S}$ is the stiffness matrix,

$$s_{ij} = a(v_j, v_i), \qquad 1 \le i, j \le N,$$

$\boldsymbol{F}$ is the right-hand side vector,

$$f_j = l(v_j), \qquad 1 \le j \le N,$$

and $\boldsymbol{Y}(t)$ is the vector of unknown time-dependent coefficients $y_j(t), j = 1, 2, \ldots, N$.

Now (5.9) no longer depends on the spatial variable $\boldsymbol{x}$, and thus (5.9) is a system of linear ODEs. Defining

$$\boldsymbol{\Phi}(\boldsymbol{Y}(t), t) = \boldsymbol{M}^{-1}[\boldsymbol{F}(t) - \boldsymbol{S}\boldsymbol{Y}(t)], \tag{5.10}$$

one obtains a standard initial value problem

$$\begin{aligned} \dot{\boldsymbol{Y}}(t) &= \boldsymbol{\Phi}(\boldsymbol{Y}(t), t), && (5.11) \\ \boldsymbol{Y}(0) &= \boldsymbol{Y}^0. && (5.12) \end{aligned}$$

We assume that the right-hand side function $\boldsymbol{\Phi}(\boldsymbol{Y}, t)$ is continuous and locally Lipschitz in $\boldsymbol{Y}$ (these are the assumptions of the existence and uniqueness theorem for ODEs [25]).

5.1.4 Construction of the initial vector

In the finite element context, the initial coefficient vector $\boldsymbol{Y}^0 = (y_{1,0}, y_{2,0}, \ldots, y_{N,0})^T$ is determined uniquely by any interpolant $U_{h,p,0} \in V_{h,p}$ of the initial condition $U(0) = u_0 - G \in V$ via the expansion

$$U_{h,p,0} = \sum_{i=1}^{N} y_{i,0} v_i.$$

Here $\{v_1, v_2, \ldots, v_N\}$ is the finite element basis of the space $V_{h,p}$. Since the interpolation is done in a Hilbert space setting, there are at least three basic interpolation options with different quality and cost:

1. Best interpolant minimizing the norm $\|(u_0 - G) - U_{h,p,0}\|_V$, is obtained via the global orthogonal projection of $u_0 - G$ onto the space $V_{h,p}$. In this case, one has to solve a system of $N = \dim(V_{h,p})$ linear algebraic equations of the form (2.82),

$$\left((u_0 - G) - \sum_{j=1}^{N} y_{j,0} v_j, v_i \right)_V = 0 \quad \text{for all } i = 1, 2, \ldots, N. \tag{5.13}$$

2. Projection-based interpolant that combines the Lagrange interpolation of vertex values with the orthogonal projection on the edges and in the element interiors. The

one-dimensional version of this technique is simple (see Section 2.82), but in 2D it involves the nontrivial space $H_{00}^{1/2}$, which exceeds the scope of this text (see, e.g., [111] for details). This technique only involves the orthogonal projection locally, and therefore it is faster but less accurate than the full orthogonal projection.

3. Lagrange nodal interpolant. This is the fastest but at the same time the least accurate technique. One proceeds as described in Paragraph 4.1.8 for Q^1/P^1-meshes and in Paragraph 4.3.10 for meshes consisting of higher-order Q^p/P^p-elements.

Evaluation of the vector $\dot{\boldsymbol{Y}}(t)$ In most computations the mass matrix $\boldsymbol{M}$ is not inverted explicitly since $\boldsymbol{M}^{-1}$ is a large dense matrix. Instead, one usually resolves $\dot{\boldsymbol{Y}}(t)$ from a system of linear equations

$$\boldsymbol{M}\dot{\boldsymbol{Y}}(t) = \boldsymbol{B} \tag{5.14}$$

with the right-hand side

$$\boldsymbol{B} = \boldsymbol{F}(t) - \boldsymbol{S}\boldsymbol{Y}(t).$$

Iterative matrix solvers perform efficiently on the system (5.14) since the mass matrix $\boldsymbol{M}$ is well-conditioned (usually much better than the stiffness matrix $\boldsymbol{S}$). It is worth mentioning that certain spectral element methods yield a diagonal mass matrix $\boldsymbol{M}$ (see, e.g., [69]).

5.1.5 Autonomous systems and phase flow

The notions of autonomous system and phase flow will be used frequently in this chapter. By the symbol

$$\mathcal{Y}(\boldsymbol{X}^0, t, t_0) \tag{5.15}$$

we denote the solution $\boldsymbol{Y}(t)$ to (5.11) at the time $t \in \mathbb{R}$, starting from the initial vector $\boldsymbol{X}^0 \in \mathbb{R}^N$ and initial time $t_0 \in \mathbb{R}$. Without loss of generality, we can assume that $t_0 = 0$. In the special case of autonomous systems,

$$\dot{\boldsymbol{Y}}(t) = \boldsymbol{\Phi}(\boldsymbol{Y}) \tag{5.16}$$

the time only enters relatively via time differences, and therefore one can leave out the initial time t_0 from (5.15). Then the symbol

$$\mathcal{Y}(\boldsymbol{X}, \Delta t) \tag{5.17}$$

is used to denote the solution to (5.16) starting at $\boldsymbol{X} \in \mathbb{R}^N$ after the time-increment Δt. Autonomous systems occur frequently in practice (for example, if coefficients and data to a parabolic PDE do not depend on time explicitly) and they are the basis for the stability analysis of numerical methods for ODEs (to be discussed in Section 5.3).

Under the assumption that the solution $Y(t)$ exists for all $t \in \mathbb{R}$, the $\mathbb{R}^N \to \mathbb{R}^N$ transformations

$$\mathcal{F}^{\Delta t}\boldsymbol{X} = \mathcal{Y}(\boldsymbol{X}, \Delta t) \qquad \text{for all } \boldsymbol{X} \in \mathbb{R}^N,\ \Delta t \in \mathbb{R}.$$

form a one-parameter Abelian (commutative) group. This group [and sometimes also the function $\mathcal{Y}(\boldsymbol{X}, \Delta t)$ itself] is called phase flow of equation (5.16). The corresponding binary operation '$\star$' is the continuation,

$$(\mathcal{F}^{\Delta s} \star \mathcal{F}^{\Delta t})\boldsymbol{X} = \mathcal{F}^{\Delta s}\mathcal{F}^{\Delta t}\boldsymbol{X} = \mathcal{Y}(\mathcal{Y}(\boldsymbol{X}, \Delta t), \Delta s) \qquad \text{for all } \boldsymbol{X} \in \mathbb{R}^N,\ \Delta t, \Delta s \in \mathbb{R}.$$

It is easy to check the commutativity of this operation,

$$\begin{aligned}(\mathcal{F}^{\Delta s} \star \mathcal{F}^{\Delta t})\boldsymbol{X} &= \mathcal{F}^{\Delta s}\mathcal{F}^{\Delta t}\boldsymbol{X} \\ &= \mathcal{F}^{\Delta t}\mathcal{F}^{\Delta s}\boldsymbol{X} \\ &= (\mathcal{F}^{\Delta t} \star \mathcal{F}^{\Delta s})\boldsymbol{X} \qquad \text{for all } \boldsymbol{X} \in \mathbb{R}^N,\ \Delta t, \Delta s \in \mathbb{R}.\end{aligned}$$

The identity element of the phase flow is the identity transformation

$$\mathcal{F}^0\boldsymbol{X} = \mathcal{Y}(\boldsymbol{X}, 0) = \boldsymbol{X} \qquad \text{for all } \boldsymbol{X} \in \mathbb{R}^N,$$

and the inverse element to $\mathcal{F}^{\Delta t}$ is defined as the reader expects,

$$\mathcal{F}^{-\Delta t}\boldsymbol{X} = \mathcal{Y}(\boldsymbol{X}, -\Delta t) \qquad \text{for all } \boldsymbol{X} \in \mathbb{R}^N,\ \Delta t \in \mathbb{R}.$$

The verification of the associativity law,

$$(\mathcal{F}^{\Delta r} \star \mathcal{F}^{\Delta s}) \star \mathcal{F}^{\Delta t} = \mathcal{F}^{\Delta r} \star (\mathcal{F}^{\Delta s} \star \mathcal{F}^{\Delta t}),$$

is left to the reader as a simple exercise.

5.2 SELECTED TIME INTEGRATION SCHEMES

There exist many excellent papers and books on the numerical solution of ODEs, and numerous sophisticated ODE packages can be downloaded from the Internet. However, one should not think that all important problems in the theory and numerics of ODEs have been solved. On the contrary: Significant progress has been made recently in the development of new methods and in understanding of the existing ones, and the numerical solution of ODEs continues being a very active research area.

The initial-value ODE problems resulting from the MOL exhibit specific features that have to be considered when selecting an appropriate ODE solver. Often, stiffness makes the application of explicit schemes prohibitive and requires implicit methods. The ODE solver should be of a higher order of accuracy: Higher-order schemes are preferable even for lower-order spatial discretizations because of their efficiency. Third, the increasing popularity of self-adaptive finite element schemes prefers one-step ODE solvers. Summing up, higher-order implicit one-step methods are one of the nowadays' most popular choices.

In Paragraph 5.2.1 we introduce the general concept of one-step methods and define their consistency and convergence. Paragraph 5.2.2 begins with the explicit and implicit Euler methods, and it describes their application to the initial-value ODE system (5.11), (5.12) with emphasis on the case with the linear right-hand side (5.10). The concept of stiffness is discussed in Paragraph 5.2.3, and a discussion of modern explicit one-step Runge–Kutta (RK) methods for nonstiff problems is given in Paragraph 5.2.4. A feasible algorithm for automatic adaptivity based on embedded RK methods is described in Paragraph 5.2.5. General (implicit) RK methods are discussed in Paragraph 5.2.6.

5.2.1 One-step methods, consistency and convergence

The general one-step method for equation (5.11) calculates an approximation $\boldsymbol{X}^{t+\Delta t}$ of the solution at the time $t + \Delta t$ using the approximation $\boldsymbol{X}$ at the time t and a time step Δt. This can be expressed using the notation

$$\boldsymbol{X}^{t+\Delta t} = \mathcal{E}(\boldsymbol{X}, t, \Delta t). \tag{5.18}$$

Analogously to the continuous case (5.17), in autonomous systems of the form (5.16) one can drop the temporal variable t and define

$$\boldsymbol{X}^{\Delta t} = \mathcal{E}(\boldsymbol{X}, \Delta t). \tag{5.19}$$

The function $\mathcal{E}$ sometimes is referred to as the discrete phase flow of the autonomous system.

Consider a finite time interval $(0, T)$, and introduce its partition $0 = t_0 < t_1 < t_2 < \ldots < t_K = T$, where t_k is the kth temporal level and $\Delta t_k = t_{k+1} - t_k$ the kth time step, $k = 0, 1, \ldots, K-1$. Then the one-step method (5.18) starting at the initial condition $\boldsymbol{Y}^0 = \boldsymbol{Y}(0)$ creates an approximation of the exact solution $\boldsymbol{Y}(t) = \mathcal{Y}(\boldsymbol{Y}^0, t, 0)$ of the problem (5.11), (5.12) in the form of a sequence of discrete states $\boldsymbol{Y}^1, \boldsymbol{Y}^2, \ldots, \boldsymbol{Y}^K$ at the times $t_1, t_2, \ldots, t_K$:

$$\begin{aligned} \boldsymbol{Y}^1 &= \mathcal{E}(\boldsymbol{Y}^0, 0, \Delta t_0), \\ \boldsymbol{Y}^2 &= \mathcal{E}(\boldsymbol{Y}^1, t_1, \Delta t_1), \\ &\vdots \\ \boldsymbol{Y}^K &= \mathcal{E}(\boldsymbol{Y}^{K-1}, t_{K-1}, \Delta t_{K-1}). \end{aligned} \tag{5.20}$$

The consistency error of the one-step method (5.18) is defined naturally as the difference between the approximation and the exact solution to (5.11) after one time step, when starting from the same state $\boldsymbol{X}$. The following definition expresses this difference using the functions $\mathcal{Y}$ and $\mathcal{E}$.

Definition 5.1 (Consistency error) *The* consistency error *of the one-step method (5.18) at $\boldsymbol{X} \in \mathbb{R}^N$ and $t > 0$ for sufficiently small $\Delta t > 0$ is defined as*

$$\epsilon(\boldsymbol{X}, t, \Delta t) = \mathcal{Y}(\boldsymbol{X}, t, \Delta t) - \mathcal{E}(\boldsymbol{X}, t, \Delta t).$$

In order to distinguish between the lowest- and higher-order time integration schemes, it is natural to define the order of consistency.

Definition 5.2 (Order of consistency) *The* order of consistency *of the one-step method (5.18) equals p if*

$$\epsilon(X, t, \Delta t) = O(\Delta t^{p+1}) \tag{5.21}$$

holds for sufficiently small Δt locally uniformly *for all $\boldsymbol{X}$ and t. The method is said to be* consistent *if its order of consistency p is at least one.*

The following result is frequently used in the numerical analysis of ODEs:

Lemma 5.1 (Consistency of one-step methods) *Assume that the function $\mathcal{E}$ is continuously differentiable in the variable Δt for all sufficiently small $0 < \Delta t \leq \Delta t^*$. Then the one-step method (5.18) is consistent if and only if there exists an increment function $\psi(\boldsymbol{X}, t, \Delta t)$ continuous in t for all $\boldsymbol{X}$, such that $\psi(\boldsymbol{X}, t, 0) = \boldsymbol{\Phi}(\boldsymbol{X}, t)$ and*

$$\mathcal{E}(\boldsymbol{X}, t, \Delta t) = \boldsymbol{X} + \Delta t \psi(\boldsymbol{X}, t, \Delta t).$$

Proof: Based on the Taylor expansion of the functions $\mathcal{Y}$ and $\mathcal{E}$ (see, e.g., [25]). ■

Naturally, one wants to analyze whether and when the approximate solution approaches the exact one as the time step converges to zero. For this purpose we define the discretization error and convergence of one-step methods:

Definition 5.3 (Discretization error) *Assume that the system (5.11), (5.12) has an exact solution $\boldsymbol{Y}(t)$ in the interval $(0, T)$. Consider a partition $0 = t_0 < t_1 < t_2 < \ldots < t_K = T$ and let $\boldsymbol{Y}^0, \boldsymbol{Y}^1, \ldots, \boldsymbol{Y}^K$ be the approximate solution obtained by means of the one-step method (5.18). The* discretization error *is defined as*

$$\epsilon_{dt} = \max_{k=1,2,\ldots,K} \|\boldsymbol{Y}(t_k) - \boldsymbol{Y}^k\|.$$

The symbol dt stands for the diameter of the time partition,

$$dt = \max_{k=1,2,\ldots,K} (t_k - t_{k-1}).$$

The notion of convergence of the general one-step method (5.18) is defined as follows:

Definition 5.4 (Convergence) *The one-step method (5.18) is said to be* convergent *with the order $p \geq 1$ if there exists a constant $dt^* > 0$ such that*

$$\epsilon_{dt} = O(dt^p)$$

for all temporal partitions of the interval $(), T)$ whose diameter $dt \leq dt^*$.

The following theorem is the basic convergence result for one-step methods:

Theorem 5.1 (Convergence of one-step methods) *Let $\mathcal{E}(\boldsymbol{X}, t, \Delta t)$ be a one-step method whose increment function $\psi(\boldsymbol{X}, t, \Delta t)$ is locally Lipschitz-continuous in the variable $\boldsymbol{X}$. Assume that along a trajectory*

$$\boldsymbol{Y}(t) \in C^1([0, T], \mathbb{R}^N)$$

the consistency error satisfies

$$\boldsymbol{Y}(t + \Delta t) - \mathcal{E}(\boldsymbol{Y}(t), t, \Delta t) = O(\Delta t^{p+1}).$$

Then the one-step method is convergent to $\boldsymbol{Y}(t)$ with the order p.

Proof: The proof is based on Lemma 5.1. See, e.g., [25]. ■

The simplest concrete examples of the general one-step method (5.18) are the explicit and implicit Euler methods.

5.2.2 Explicit and implicit Euler methods

Euler methods are the oldest and least sophisticated ODE solvers. The explicit Euler method is popular because of its very simple implementation and minimum overhead cost, but it also is known to be unstable unless the time step is extremely small. The implicit Euler method is more stable, but for nonlinear ODEs it requires the solution of a system of nonlinear algebraic equations in every time step. In the case of linear ODEs the application of both the explicit and implicit Euler schemes is equally simple.

Explicit Euler scheme

The explicit Euler method is obtained by approximating the temporal derivative in (5.11) by the forward time difference,

$$\dot{\boldsymbol{Y}}(t_k) \approx \frac{\boldsymbol{Y}^{k+1} - \boldsymbol{Y}^k}{\Delta t_k},$$

and leaving the right-hand side of (5.11) on the kth temporal level. In this way one obtains

$$\boldsymbol{Y}^0 = \boldsymbol{Y}(t_0), \tag{5.22}$$

$$\boldsymbol{Y}^{k+1} = \boldsymbol{Y}^k + \Delta t_k \boldsymbol{\Phi}(\boldsymbol{Y}^k, t_k), \tag{5.23}$$

which is a one-step method of the class (5.18),

$$\mathcal{E}(\boldsymbol{X}, t, \Delta t) = \boldsymbol{X} + \Delta t \boldsymbol{\Phi}(\boldsymbol{X}, t).$$

Since $\boldsymbol{\Phi}$ is continuous, this method is evidently consistent with the order $p = 1$ in the sense of Definition 5.2. If the right-hand side $\boldsymbol{\Phi}(\boldsymbol{Y}, t)$ is locally Lipschitz-continuous in the variable $\boldsymbol{Y}$, the increment function

$$\psi(\boldsymbol{X}, t, \Delta t) = \boldsymbol{\Phi}(\boldsymbol{X}, t)$$

satisfies the assumptions of Theorem 5.1, and therefore the one-step method is convergent with the order $p = 1$.

It follows from (5.10) that on each time level one obtains a system of linear algebraic equations of the form

$$\boldsymbol{M}\boldsymbol{Y}^{k+1} = \boldsymbol{B}^k, \tag{5.24}$$

where

$$\boldsymbol{B}^k = \boldsymbol{M}\boldsymbol{Y}^k + \Delta t_k(\boldsymbol{F}(t_k) - \boldsymbol{S}\boldsymbol{Y}^k).$$

The presence of the mass matrix $\boldsymbol{M}$ on the left-hand side of (5.24) is not very pleasant for an explicit method, since the time step is very small and the system (5.24) has to be solved many times. Therefore, in practice $\boldsymbol{M}$ sometimes is truncated to its diagonal,

$$\boldsymbol{M} \approx \mathrm{diag}(m_{11}, m_{22}, \dots, m_{NN}). \tag{5.25}$$

This operation is called mass lumping.

Limitations The truncation (5.25) produces a higher-order temporal error term that often can be neglected with a low-order FEM discretization in space. Generally it is not practical to combine low-order ODE schemes with higher-order FEM. As we said above, the explicit Euler method is known to be unstable unless the time step Δt is very small. For parabolic problems, the theory says that Δt must be proportional to the square of the volume of the smallest element in the mesh, i.e.,

$$\Delta t = O(\Delta h^2). \tag{5.26}$$

This criterion makes the explicit Euler method extremely time-consuming and almost impossible to combine with spatial adaptivity, where $\Delta h \to 0$. The situation is less severe in the case of hyperbolic problems, where the criterion (5.26) is replaced with the less constraining CFL condition (see, e.g., [52, 77] and [78]),

$$\Delta t = O(\Delta h).$$

The stability of one-step methods will be discussed in more detail in Section 5.3.

Implicit Euler scheme

The implicit Euler method is obtained by approximating the temporal derivative in (5.11) by the backward time difference,

$$\dot{\boldsymbol{Y}}(t_{k+1}) \approx \frac{\boldsymbol{Y}^{k+1} - \boldsymbol{Y}^k}{\Delta t_k},$$

and assuming the right-hand side of (5.11) on the $(k+1)$th time level. The ODE problem (5.11), (5.12) yields a discrete system

$$\begin{array}{rcl} \boldsymbol{Y}^0 &=& \boldsymbol{Y}(t_0), \\ \boldsymbol{Y}^{k+1} &=& \boldsymbol{Y}^k + \Delta t_k \boldsymbol{\Phi}(\boldsymbol{Y}^{k+1}, t_k + \Delta t_k), \end{array} \tag{5.27}$$

In general the function $\boldsymbol{\Phi}$ is nonlinear and requires a special treatment (such as, e.g., some sort of fixed point or Newton's method). However, the linearity of the model problem (5.1) yields

$$\boldsymbol{M}\frac{\boldsymbol{Y}^{k+1} - \boldsymbol{Y}^k}{\Delta t_k} = \boldsymbol{F}(t_{k+1}) - \boldsymbol{S}\boldsymbol{Y}^{k+1},$$

and as a result, the system one has to solve on each time level is linear,

$$\boldsymbol{S}_k \boldsymbol{Y}^{k+1} = \boldsymbol{B}^{k+1}. \tag{5.28}$$

Here

$$\boldsymbol{S}_k = \boldsymbol{M} + \Delta t_k \boldsymbol{S} \tag{5.29}$$

and

$$\boldsymbol{B}^{k+1} = \Delta t_k \boldsymbol{F}(t_{k+1}) + \boldsymbol{M}\boldsymbol{Y}^k.$$

Stability and accuracy It is well known that the implicit Euler scheme is absolutely stable, i.e., it works with any size of the time step (this will be discussed in more detail in Section 5.3). One should not forget that this method only is first-order accurate. In most cases the performance of iterative matrix solvers deteriorates when the time step Δt_k grows too large, since usually

$$\kappa(\boldsymbol{M}) \ll \kappa(\boldsymbol{S}),$$

and thus the matrix (5.29) becomes ill-conditioned.

Remark 5.1 *Without the truncation (5.25) of the mass matrix $\boldsymbol{M}$ the implementation cost of both the explicit and implicit Euler schemes is the same. Thus for linear problems it certainly is a good idea to use the implicit scheme.*

5.2.3 Stiffness

It is customary to say that stiffness is a property of ODEs that complicates their numerical solution. In reality the stiffness is a more complex phenomenon that involves at least three basic ingredients: the solved equation or system, the numerical method, and the time step. It is known that stiffness is associated with the behavior of perturbations to a given solution. To illustrate this, let $\boldsymbol{Y}(t)$ be an exact solution of equation (5.11),

$$\dot{\boldsymbol{Y}}(t) = \boldsymbol{\Phi}(\boldsymbol{Y}(t), t), \tag{5.30}$$

and let $\epsilon \boldsymbol{Z}(t)$, where ϵ is a very small real number, be a perturbation of $\boldsymbol{Y}(t)$. When replacing $\boldsymbol{Y}(t)$ with the perturbed solution $\boldsymbol{Y}(t) + \epsilon \boldsymbol{Z}(t)$,

$$\dot{\boldsymbol{Y}}(t) + \epsilon \dot{\boldsymbol{Z}}(t) = \boldsymbol{\Phi}(\boldsymbol{Y}(t) + \epsilon \boldsymbol{Z}(t), t),$$

and neglecting the quadratic and higher-order terms in the Taylor expansion, one obtains

$$\dot{\boldsymbol{Y}}(t) + \epsilon \dot{\boldsymbol{Z}}(t) = \boldsymbol{\Phi}(\boldsymbol{Y}(t), t) + \epsilon \boldsymbol{J}(t) \boldsymbol{Z}(t). \tag{5.31}$$

Here

$$\boldsymbol{J}(t) = \frac{\mathrm{D}\boldsymbol{\Phi}}{\mathrm{D}\boldsymbol{Y}}(\boldsymbol{Y}(t), t) \tag{5.32}$$

is the Jacobi matrix of the right-hand side $\boldsymbol{\Phi}$. Subtracting (5.30) from (5.31), one obtains an equation governing the evolution of the perturbation,

$$\dot{\boldsymbol{Z}}(t) = \boldsymbol{J}(t) \boldsymbol{Z}(t),$$

Now, in a time interval where neither the solution $\boldsymbol{Y}(t)$ nor the Jacobi matrix $\boldsymbol{J}(t)$ change significantly, the growth of the components of the perturbation $\boldsymbol{Z}(t)$ is determined by the eigenvalues of $\boldsymbol{J}(t)$. In general, the existence of one or more eigenvalues whose real part is negative and large in magnitude is a sign that stiffness almost certainly is present. This is demonstrated on a simple linear ODE system:

Example of a stiff problem Let us solve a system of two linear ODEs,

$$\begin{aligned} \dot{y}_1(t) &= -y_1(t), \\ \dot{y}_2(t) &= -100y_2(t), \end{aligned} \tag{5.33}$$

equipped with the initial condition

$$\begin{aligned} y_1(0) &= 1, \\ y_2(0) &= 1. \end{aligned} \tag{5.34}$$

Equations (5.33) are autonomous and they can be written in the matrix form

$$\dot{\boldsymbol{Y}}(t) = \boldsymbol{A}\boldsymbol{Y}(t),$$

where

$$\boldsymbol{A} = \begin{pmatrix} -1 & 0 \\ 0 & -100 \end{pmatrix}.$$

Hence the Jacobi matrix (5.32) is directly $\boldsymbol{A}$. The eigenvalues of $\boldsymbol{A}$, $\lambda_1 = -1$ and $\lambda_2 = -100$, determine the form of the exact solution,

$$\begin{aligned} y_1(t) &= e^{-t}, \\ y_2(t) &= e^{-100t}, \end{aligned} \tag{5.35}$$

which is depicted in Figure 5.1.

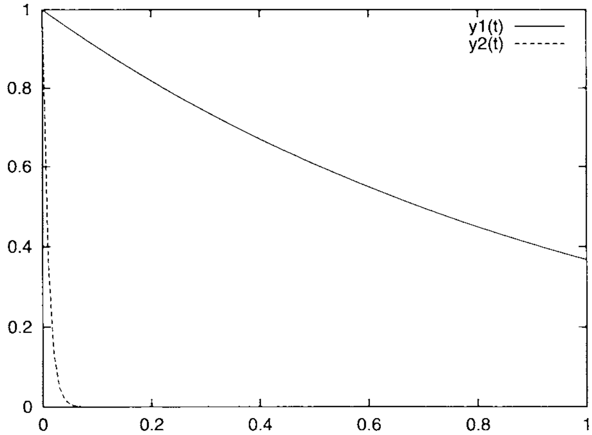


Figure 5.1 Two different temporal scales in the solution of the stiff problem (5.33), (5.34).

We see that the solution components $y_1(t)$ and $y_2(t)$ vanish at different temporal scales. It is well known that explicit methods applied to stiff problems like (5.33), (5.34) are unstable unless the time step is absurdly small. The best known one-step schemes for stiff problems higher-order implicit RK methods, will be discussed in Section 5.4.

5.2.4 Explicit higher-order RK schemes

Fortunately not all ODEs are stiff, and explicit methods are useful for numerous types of ODE problems. The RK methods are sophisticated one-step methods that generalize the explicit Euler scheme to higher orders of accuracy. The first method of this kind was introduced as a generalization of the Taylor's method in 1895 by Carle David Tolmé Runge [101],

$$\begin{aligned} z_1 &= \Phi(Y^k, t_k), \\ z_2 &= \Phi\left(Y^k + \frac{\Delta t_k}{2} z_1, t_k + \frac{\Delta t_k}{2}\right), \\ Y^{k+1} &= Y^k + \Delta t_k z_2. \end{aligned} \tag{5.36}$$

Figure 5.2 Carle David Tolmé Runge (1856–1927).

C.D.T. Runge contributed significantly to the fields of differential geometry, interpolation, and numerical solution of algebraic and ordinary differential equations. He also was active in experimental physics, where he investigated the wavelengths of the spectral lines of elements. Nowadays his explicit second-order method (5.36) is widely used in the slightly more general form

$$\begin{aligned} z_1 &= \Phi(Y^k, t_k), \\ z_2 &= \Phi\left(Y^k + \frac{\Delta t_k}{2\omega_2} z_1, t_k + \frac{\Delta t_k}{2\omega_2}\right), \\ Y^{k+1} &= Y^k + \Delta t_k[(1 - \omega_2) z_1 + \omega_2 z_2], \end{aligned} \tag{5.37}$$

where possible choices of the parameters are $\omega_2 = 1/2$, $\omega_2 = 1/3$ or $\omega_2 = 1$.

As shown in 1901 by W. Kutta [74], (5.36) can be extended to more general nested evaluations of the right-hand side, resulting into the s-stage explicit RK method

$$z_i = \Phi\left(Y^k + \Delta t_k \sum_{j=1}^{i-1} a_{ij} z_j, t_k + c_i \Delta t_k\right), \quad c_1 = 0, \tag{5.38}$$

$$Y^{k+1} = Y^k + \Delta t_k \sum_{i=1}^{s} b_i z_i. \tag{5.39}$$

The parameters a_{ij} and c_i, satisfying

$$c_i = \sum_{j=1}^{i-1} a_{ij},$$

are determined from the Taylor expansion of the function $\dot{Y}(t)$ in (5.11) in such a way that the order of the truncation error is maximized (see, e.g., [106]). The values of these parameters are sufficiently well tabulated.

Butcher's arrays Both the explicit and implicit RK methods can be written economically in terms of the Butcher's arrays [25],

$$\begin{array}{c|cccc}
c_1 & a_{11} & a_{12} & \dots & a_{1s} \\
c_2 & a_{21} & a_{22} & \dots & a_{2s} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
c_s & a_{s1} & a_{s2} & \dots & a_{ss} \\
\hline
 & b_1 & b_2 & \dots & b_s
\end{array}$$

The RK method given by such array is referred to as the $(\boldsymbol{b}, \boldsymbol{c}, \mathcal{A})$ RK method. For example, the array

$$\begin{array}{c|c}
0 & 0 \\
\hline
 & 1
\end{array}$$

corresponds to the explicit Euler method (one-stage explicit RK method)

$$\begin{aligned}
\boldsymbol{z}_1 &= \boldsymbol{\Phi}(\boldsymbol{Y}^k, t_k), \\
\boldsymbol{Y}^{k+1} &= \boldsymbol{Y}^k + \Delta t_k \boldsymbol{z}_1.
\end{aligned} \tag{5.40}$$

The original second-order Runge's method (5.36) can be written as

$$\begin{array}{c|cc}
0 & & \\
1/2 & 1/2 & \\
\hline
 & 0 & 1
\end{array}$$

The famous Kutta's fourth-order "classical RK method" (1901) has the form

$$\begin{array}{c|cccc}
0 & & & & \\
1/2 & 1/2 & & & \\
1/2 & 0 & 1/2 & & \\
1 & 0 & 0 & 1 & \\
\hline
 & 1/6 & 1/3 & 1/3 & 1/6
\end{array}$$

To give one more example, the so-called Kutta's 3/8 formula is given by

$$\begin{array}{c|cccc}
0 & & & & \\
1/3 & 1/3 & & & \\
2/3 & -1/3 & 1 & & \\
1 & 1 & -1 & 1 & \\
\hline
 & 1/8 & 3/8 & 3/8 & 1/8
\end{array}$$

A Runge–Kutta method is explicit if the diagonal and the upper triangular submatrix of $\mathcal{A}$ are zero. The sufficient and necessary condition for the consistency of RK methods is formulated in the following lemma.

Lemma 5.2 (Consistency of explicit RK methods) *An explicit* $(\boldsymbol{b}, \boldsymbol{c}, \mathcal{A})$ *RK method (5.38), (5.39) is consistent for all continuous right-hand sides* $\boldsymbol{\Phi}$ *if and only if*

$$\sum_{i=1}^{s} b_i = 1. \tag{5.41}$$

This condition will be extended to general implicit RK methods in Paragraph 5.4.1.

Proof: The method (5.38), (5.39) can be written in the incremental form

$$\mathcal{E}(\boldsymbol{X}, t, \Delta t) = \boldsymbol{X} + \Delta t \psi(\boldsymbol{X}, t, \Delta t)$$

with

$$\psi(\boldsymbol{X}, t, \Delta t) = \boldsymbol{X} + \Delta t \sum_{i=1}^{s} b_i \boldsymbol{z}_i.$$

For $\Delta t = 0$ we have $\boldsymbol{z}_i(\boldsymbol{X}, t, 0) = \boldsymbol{\Phi}(\boldsymbol{X}, t)$ and therefore

$$\psi(\boldsymbol{X}, t, 0) = \boldsymbol{X} + \Delta t \left(\sum_{i=1}^{s} b_i \right) \boldsymbol{\Phi}(\boldsymbol{X}, t).$$

By Lemma 5.1 consistency is equivalent to $\psi(\boldsymbol{X}, t, \Delta t) = \boldsymbol{\Phi}(\boldsymbol{X}, t)$. This is the case if and only if (5.41) holds. ■

In reality, the coefficients b_i and c_i are the weights and points of a quadrature formula in the interval $(0, 1)$ (to be discussed in more detail later). Now let us have a look at the maximum order of consistency of an s-stage explicit RK method.

Lemma 5.3 (Order of consistency of explicit RK methods) *Let an explicit s-stage* $(\boldsymbol{b}, \boldsymbol{c}, \mathcal{A})$ *RK method have the order of consistency p for all infinitely smooth right-hand sides* $\boldsymbol{\Phi}$*. Then necessarily*

$$p \leq s.$$

Proof: Applying the method to a scalar initial value problem

$$\begin{aligned} \dot{y}(t) &= y(t), \\ y(0) &= 1, \end{aligned} \tag{5.42}$$

we find that

$$\mathcal{E}(1, 0, \Delta t) = e^{\Delta t} = 1 + \Delta t + \frac{\Delta t^2}{2!} + \ldots + \frac{\Delta t^p}{p!} + O(\Delta t^{p+1}).$$

Thus necessarily $z_i(1, 0, \Delta t)$ is a polynomial of the degree less than or equal to $i - 1$. Hence $\mathcal{E}(1, 0, \Delta t)$ is a polynomial in Δt of the degree at most s, and for the consistency error $\epsilon(1, 0, \Delta t)$ to be $O(\Delta t^{p+1})$ it must be $p \le s$. ■

Explicit RK methods are very popular because of their very simple implementation. More precisely, their implementation is very simple in combination with the truncation (5.25) of the mass matrix $\boldsymbol{M}$. The loss of accuracy due to this operation is less significant than in the case of the explicit Euler method.

Most popular are RK methods of the orders $p = 1$ (Euler methods), $p = 2$, and $p = 4$, since for higher p it is $p < s$ (see Table 5.1), and the ratio of the cost and performance becomes less optimal.

Table 5.1 Minimum number of stages for a pth-order RK method.

p	1	2	3	4	5	6	7	8	≥ 9
s_{min}	1	2	3	4	6	7	9	11	$\ge p+3$

5.2.5 Embedded RK methods and adaptivity

An ODE solver designed to perform well on a wide range of problems should be adaptive and control at least the local error. In this context, worth mentioning are the embedded RK methods. These are simple adaptive schemes that estimate the local error via the difference

$$\boldsymbol{E}^{k+1} \approx \boldsymbol{Y}^{k+1} - \widehat{\boldsymbol{Y}}^{k+1}, \tag{5.43}$$

where the values $\boldsymbol{Y}^{k+1}$ and $\widehat{\boldsymbol{Y}}^{k+1}$ are obtained by performing the same time step twice with RK methods of the order $m + 1$ and m, respectively. This is an idea as old as error estimation itself, and moreover, it seems to more than double the amount of work per time step. However, the Fehlberg's trick [51] makes it possible to keep the amount of work proportional to the $(m + 1)$th-order method.

Fehlberg's trick The basic idea of the Fehlberg's trick is to let the first stage of the new time step be the same as the last stage of the current step, i.e.,

$$\begin{aligned}
\boldsymbol{\Phi}(\boldsymbol{Y}^k + \Delta t_k \sum_{j=1}^{s-1} a_{sj} z_j^k, t_k + c_s \Delta t_k) &= z_s^k \\
&= z_1^{k+1} \\
&= \boldsymbol{\Phi}(\boldsymbol{Y}^{k+1}, t_k + \Delta t_k) \\
&= \boldsymbol{\Phi}(\boldsymbol{Y}^k + \Delta t_k \sum_{j=1}^{s} b_j z_j^k, t_k + \Delta t_k).
\end{aligned}$$

This holds for all right-hand sides if the coefficients satisfy

$$\begin{aligned}
c_s &= 1, \\
b_s &= 0, \\
a_{sj} &= b_j \quad \text{for all } j = 1, 2, \ldots, s.
\end{aligned}$$

The remaining coefficients are determined routinely (see, e.g., [25]). The total number of evaluations of the right-hand side Φ in n steps of the algorithm is only $n \cdot (s-1)+1$ instead of $n \cdot s$. Therefore we speak about an effectively $(s-1)$-stage method, of the type RK$p(p-1)$.

Among numerous known embedded RK$p(p-1)$ methods, the most mature were constructed by J. R. Dormand and P. J. Prince (see [42] and [43]). The coefficients of their effectively 6-stage RK5(4) method are given in Table 5.2.

Table 5.2 Coefficients of the Dormand–Prince RK5(4) method.

0							
$\frac{1}{5}$	$\frac{1}{5}$						
$\frac{3}{10}$	$\frac{3}{40}$	$\frac{9}{40}$					
$\frac{4}{5}$	$\frac{44}{45}$	$-\frac{56}{15}$	$\frac{32}{9}$				
$\frac{8}{9}$	$\frac{19372}{6561}$	$-\frac{25360}{2187}$	$\frac{64448}{6561}$	$-\frac{212}{729}$			
1	$\frac{9017}{3168}$	$-\frac{355}{33}$	$\frac{46732}{5247}$	$\frac{49}{176}$	$-\frac{5103}{18656}$		
1	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	
	$\frac{35}{384}$	0	$\frac{500}{1113}$	$\frac{125}{192}$	$-\frac{2187}{6784}$	$\frac{11}{84}$	0
	$\frac{5179}{57600}$	0	$\frac{7571}{16695}$	$\frac{393}{640}$	$-\frac{92097}{339200}$	$\frac{187}{2100}$	$\frac{1}{40}$

Various adaptive algorithms can be built upon embedded RK$p(p-1)$ methods, using either the rather primitive error estimate (5.43), or some more sophisticated estimate that typically involves the stages z_i (see, e.g., [25]).

The following basic adaptive algorithm reduces the time step Δt to $DTRED * \Delta t$ if (5.43) exceeds a given tolerance TOL, and it increases Δt to $DTINC * \Delta t$ if (5.43) is less than $ERRMIN * TOL$. If the ODE system is rooted in a parabolic PDE, the initial time step may be defined as $\Delta t_0 := (\Delta h)^2$ (where Δh is the volume of the smallest element in the mesh $\mathcal{T}_{h,p}$). Otherwise, some other appropriate value of Δt_0 may be chosen.

Algorithm 5.1 (Adaptive RK5(4) method)

```
Read the local error tolerance parameter $TOL$;
Read the final time $T_{final}$;
Set the time step reduction parameter $DTRED$ (for example) to 1/2;
Set the time step increase parameter $DTINC$ (for example) to $3/2$;
Set the parameter $ERRMIN$ (for example) to $0.05$;
Define the initial time step as $\Delta t := \Delta t_0$;
Define the initial condition $\boldsymbol{Y}^0 := \boldsymbol{Y}(0)$;
Set $t := 0$ and $k := 0$;
Do {
  Estimate $\boldsymbol{Y}(t_{k+1})$ by $\boldsymbol{Y}^{k+1}$ using $\boldsymbol{Y}^k$, $\Delta t$ and the RK5 method;
  Estimate $\boldsymbol{Y}(t_{k+1})$ by $\widehat{\boldsymbol{Y}}^{k+1}$ using $\boldsymbol{Y}^k$, $\Delta t$ and the embedded RK4 method;
  Calculate a local error estimate $\boldsymbol{E}^{k+1}$ via (5.43);
  If ($ERRMIN * TOL \le \|\boldsymbol{E}^{k+1}\| \le TOL$) then {
    $k := k + 1$;
  }
  else {
    if ($\|\boldsymbol{E}^{k+1}\| > TOL$) then $\Delta t := DTRED * \Delta t$;
    else {
      $t := t + \Delta t$;
      $k := k + 1$;
      $\Delta t := DTINC * \Delta t$;
    }
  }
} while ($t < T_{final}$);
```

The application of this algorithm to a concrete problem may show the need for adjustment of the parameters Δt_0, $DTRED$, $ERRMIN$ and $DTINC$. Let us remark that embedded RK methods also exists in the implicit version (see [25] and the references therein).

5.2.6 General (implicit) RK schemes

Implicit RK methods were introduced in 1964 by J. C. Butcher by allowing the coefficient matrix $\mathcal{A}$ in (5.38), (5.39) to be a full matrix. This generalization yields an s-stage RK method

$$z_i = \Phi(\boldsymbol{Y}^k + \Delta t_k \sum_{j=1}^{s} a_{ij} z_j, t_k + c_i \Delta t_k), \tag{5.44}$$

$$\boldsymbol{Y}^{k+1} = \boldsymbol{Y}^k + \Delta t_k \sum_{i=1}^{s} b_i z_i. \tag{5.45}$$

The summation in (5.44) runs over all $i = 1, 2, \ldots, s$, and therefore identical z_is appear on both sides of the equation whenever $a_{ii} \neq 0$, and unknown higher-index z_js, $i < j$, are present if $a_{ij} \neq 0$. In these cases the RK method (5.44), (5.45) is implicit. In turn, the explicit RK methods (5.38), (5.39) are obtained if $a_{ij} = 0$ for all $j \ge i$, $i = 1, 2, \ldots, s$. The currently best known Gauss and Radau higher-order IRK methods are introduced in Paragraph 5.4.2, after we discuss in more detail the role of higher-order numerical quadrature rules in Paragraph 5.4.1. For now, let us mention a few simpler IRK methods and illustrate their application to problem (5.11), (5.12) with the right-hand side (5.10).

The Butcher's array

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$$

represents the implicit Euler method (5.27),

$$\begin{aligned} z_1 &= \Phi(\boldsymbol{Y}^k + \Delta t_k z_1, t_k + \Delta t_k), \\ \boldsymbol{Y}^{k+1} &= \boldsymbol{Y}^k + \Delta t_k z_1. \end{aligned}$$

Another one-stage method

$$\begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array}$$

corresponds to the "implicit midpoint rule". Notice that $p = 2s$ here, which by Lemma 5.3 would not be possible with an explicit RK method. A third array,

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}$$

defines another second-order IRK based on the "implicit trapezoidal rule"

$$\begin{aligned} z_1 &= \Phi(Y^k, t_k), \\ z_2 &= \Phi\left(Y^k + \frac{\Delta t_k}{2} z_1 + \frac{\Delta t_k}{2} z_2, t_k + \Delta t_k\right), \\ Y^{k+1} &= Y^k + \frac{\Delta t_k}{2} z_1 + \frac{\Delta t_k}{2} z_2. \end{aligned} \tag{5.46}$$

With a general nonlinear right-hand side Φ, each iteration of an s-stage IRK method requires the solution of a system of nonlinear algebraic equations (see Paragraph 5.4.3). The natural questions of existence and uniqueness of solution to this system are highly nontrivial, but the answer to both of them is positive, for sufficiently small values of Δt (see, e.g., [25]).

In our particular case, the linearity of the model problem (5.1) translates into (5.10),

$$\Phi(Y(t), t) = M^{-1}[F(t) - SY(t)].$$

In turn the algebraic system to be solved is linear, as it was in the case of the implicit Euler method in Paragraph 5.2.6. For example, the above second-order IRK (5.46) yields

$$\begin{aligned} z_1 &= M^{-1}(F^k - SY^k), \\ z_2 &= M^{-1}\left[F^{k+1} - S\left(Y^k + \frac{\Delta t_k}{2} z_1 + \frac{\Delta t_k}{2} z_2\right)\right], \end{aligned} \tag{5.47}$$

and thus z_1 and z_2 are obtained by solving two linear algebraic systems

$$\begin{aligned} M z_1 &= F^k - SY^k, \\ \left(M + \frac{\Delta t_k}{2} S\right) z_2 &= F^{k+1} - S\left(Y^k + \frac{\Delta t_k}{2} z_1\right). \end{aligned} \tag{5.48}$$

More about higher-order RK schemes will be said in Section 5.4, after introducing basic stability concepts of ODEs and one-step methods in Section 5.3.

5.3 INTRODUCTION TO STABILITY

The stability domains of the functions $\mathcal{Y}$ and $\mathcal{E}$ are fairly independent, and the performance of the ODE solver is determined by their intersection. Let us begin with defining the classical concept of Ljapunov stability:

Definition 5.5 (Stability, asymptotic stability) *Let $(\boldsymbol{Y}^0, t_0)$ be such that the solution $\boldsymbol{Y}(t) = \mathcal{Y}(\boldsymbol{Y}^0, t, t_0)$ of the ODE (5.11) exists for all $t \geq t_0$. The solution $\boldsymbol{Y}(t)$ is said to be* stable *at $(\boldsymbol{Y}^0, t_0)$ (in the forward direction) if for every $\epsilon > 0$ there exists a $\delta > 0$ such that*

$$\boldsymbol{Y}(t) \in B_\epsilon(\mathcal{Y}(\boldsymbol{Y}_*^0, t, t_0))$$

for all $t \geq t_0$ and all perturbed initial states $\boldsymbol{Y}_*^0 \in B_\delta(\boldsymbol{Y}^0)$. In addition, if there exists a $\delta_0 > 0$ such that

$$\lim_{t\to\infty} \|\mathcal{Y}(\boldsymbol{Y}^0, t, t_0) - \mathcal{Y}(\boldsymbol{Y}_*^0, t, t_0)\| = 0$$

for all perturbed initial states $\boldsymbol{Y}_*^0 \in B_{\delta_0}(\boldsymbol{Y}^0)$, the solution $\boldsymbol{Y}(t)$ is said to be asymptotically stable *at $(\boldsymbol{Y}^0, t_0)$. In such case we say that sufficiently small perturbations of the initial state are "damped out". Solution $\boldsymbol{Y}(t)$ is* unstable *if it is not stable. An ODE is called* stable *if it has a stable solution for all initial conditions $(\boldsymbol{Y}^0, t_0)$.*

In other words, the function $\mathcal{Y}$ of an ODE is said to be stable (in the sense of Ljapunov) if small changes of the initial state cannot cause excessive changes in the temporal evolution. An analogous definition can be formulated for the stability in the backward direction. It can be shown that a solution $\boldsymbol{Y}(t)$ that is asymptotically stable in the forward direction is unstable in the backward direction. The notion of stability is invariant under the choice of the norm $\|\cdot\|$ in $\mathbb{R}^N$, since all norms in finite-dimensional spaces are equivalent (Definition A.34, Theorem A.5).

5.3.1 Autonomization of RK methods

The stability of numerical methods for ODEs is analysed in the context of linear autonomous systems of the form

$$\begin{aligned} \dot{\boldsymbol{Y}}(t) &= \boldsymbol{A}\boldsymbol{Y}(t), \\ \boldsymbol{Y}(0) &= \boldsymbol{Y}^0, \end{aligned} \tag{5.49}$$

where $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ is a constant real (or complex) matrix. But, does the study of the autonomous system (5.49) have some relation to the original nonautonomous system (5.11), (5.12),

$$\begin{aligned} \dot{\boldsymbol{Y}}(t) &= \boldsymbol{\Phi}(\boldsymbol{Y}(t), t), \\ \boldsymbol{Y}(0) &= \boldsymbol{Y}^0? \end{aligned} \tag{5.50}$$

The answer is yes. It is well known that the system (5.11), (5.12) can be autonomized by defining a new (augmented) state variable

$$\boldsymbol{Z}(t) := \begin{pmatrix} \boldsymbol{Y}(t) \\ t \end{pmatrix}.$$

The ODE system is changed accordingly to

$$\begin{aligned} \dot{\boldsymbol{Z}}(t) &= \begin{pmatrix} \boldsymbol{\Phi}(\boldsymbol{Y}(t), t) \\ 1 \end{pmatrix}, \\ \boldsymbol{Z}(0) &= \begin{pmatrix} \boldsymbol{Y}^0 \\ 0 \end{pmatrix}. \end{aligned} \tag{5.51}$$

Let $\mathcal{Y}(\boldsymbol{Y}^0, t, 0)$ describe the evolution of the original system (5.50) and the function $\mathcal{Z}(\boldsymbol{Z}^0, t, 0)$ the evolution of the autonomized system (5.51). Then $\mathcal{Y}$ and $\mathcal{Z}$ are equivalent if the condition

$$\begin{pmatrix} \mathcal{Y}(\boldsymbol{Y}(t), t, t+\Delta t) \\ t+\Delta t \end{pmatrix} = \mathcal{Z}\left(\begin{pmatrix} \boldsymbol{Y}(t) \\ t \end{pmatrix}, t, t+\Delta t\right) \tag{5.52}$$

is satisfied. It seems not to be widely known that virtually all RK methods that are used in practice account for this equivalence by producing identical results when applied to the systems (5.50) and (5.51). Such RK methods are said to be invariant under autonomization. However, the fulfillment of (5.52) is not automatic:

Lemma 5.4 *A general $(\boldsymbol{b}, \boldsymbol{c}, \mathcal{A})$ RK method (5.44) is invariant under autonomization if and only if it is consistent and*

$$c_i = \sum_{j=1}^{s} a_{ij} \quad \textit{for all } i = 1, 2, \dots, s. \tag{5.53}$$

Proof: See, e.g., [25]. ■

It is customary to use the notation $(\boldsymbol{b}, \mathcal{A})$ for RK methods with the property (5.53). The formalism of Butcher's arrays reveals easily that all the explicit and implicit RK methods presented until now (including both the explicit and implicit Euler methods), were invariant under autonomization. Without loss of generality, we restrict ourselves to RK methods invariant under autonomization also in the rest of this chapter.

5.3.2 Stability of linear autonomous systems

The invariance of RK methods under autonomization justifies the study of their performance on the linear autonomous system (5.49). It is well known that in this case the exact solution has the form

$$\boldsymbol{Y}(t) = \mathcal{Y}(\boldsymbol{Y}^0, t) = \exp(\boldsymbol{A}t)\boldsymbol{Y}^0, \tag{5.54}$$

where the matrix exponential $\exp(\boldsymbol{A}t)$ is defined via the absolutely convergent series

$$\exp(\boldsymbol{A}t) = \sum_{n=1}^{\infty} \frac{(t\boldsymbol{A})^n}{n!}. \tag{5.55}$$

For every $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ this series converges locally uniformly in $\mathbb{R}$, i.e., uniformly in all finite intervals $(-T, T)$, $T \in \mathbb{R}$ (for a proof see, e.g., [25]).

Because of (5.54) and (5.55), the complex exponential function $\exp(z)$, $z \in \mathbb{C}$, is called the stability function of the linear autonomous system (5.49). We will see in Paragraphs 5.3.3 and 5.3.4 that explicit and implicit one-step methods of the order p are based on its pth-degree polynomial or rational approximation of the form

$$\exp(z) = R(z) + O(z^{p+1}), \tag{5.56}$$

respectively. The following theorem characterizes the stability of the matrix exponential $\exp(\boldsymbol{A}t)$ in terms of the eigenvalues of $\boldsymbol{A}$. Recall Definition A.18 of the spectrum $\sigma(\boldsymbol{A})$.

Theorem 5.2 *The linear autonomous ODE system (5.49) is stable if and only if the following two conditions are met:*

1. *$Re(\lambda) \leq 0$ for all $\lambda \in \sigma(\boldsymbol{A})$,*
2. *All eigenvalues $\lambda \in \sigma(\boldsymbol{A})$ such that $Re(\lambda) = 0$ have index exactly one. (The index of an eigenvalue is the size of the associated Jordan blocks in the Jordan canonical form of the matrix $\boldsymbol{A}$).*

The system is asymptotically stable if and only if $Re(\lambda) < 0$ for all $\lambda \in \sigma(\boldsymbol{A})$.

Proof: See, e.g., [25]. ■

The reader can see that neither the stability nor the asymptotical stability of the solution $\boldsymbol{Y}(t)$ to linear autonomous systems depend on the initial condition $\boldsymbol{Y}^0$.

5.3.3 Stability functions and stability domains

In practice we need to know to what extent the discrete phase flow $\mathcal{E}$ of a given one-step scheme,

$$\boldsymbol{Y}^{k+1} = \mathcal{E}(\boldsymbol{Y}^k, \Delta t_k), \tag{5.57}$$

preserves the stability of the continuous phase flow $\mathcal{Y}$ to the original autonomous ODE problem. For this we need to introduce the notion of stability domains for both functions $\mathcal{Y}$ and $\mathcal{E}$. Typically, these two stability domains are different, and the numerical method is stable in their intersection.

For simplicity let us begin with a scalar version of the linear autonomous problem (5.49) of the form

$$\begin{aligned} \dot{y}(t) &= \lambda y(t), \qquad t \in (0, \infty), \\ y(0) &= y^0, \end{aligned} \tag{5.58}$$

where $0 \neq \lambda \in \mathbb{C}$ is a constant. By Theorem 5.2 the function

$$\mathcal{Y}(x, t) = \exp(t\lambda)x \tag{5.59}$$

is stable if $Re(\lambda) \leq 0$. This motivates the following definition:

Definition 5.6 (Stability domain of $\mathcal{Y}$) *Let the continuous phase flow have the form (5.59). Then the* stability domain *of $\mathcal{Y}$ is the set*

$$\mathcal{S}_{exp} = \{z \in \mathbb{C};\ Re(z) \leq 0\}. \tag{5.60}$$

Now let us look at the stability domains of the explicit and implicit Euler methods:

Explicit Euler method The approximation of $y(t) = \mathcal{Y}(y^0, t)$ with a constant time step Δt has the form

$$\begin{aligned} y^1 &= (1 + \Delta t\lambda)y^0, \\ y^2 &= (1 + \Delta t\lambda)y^1 = (1 + \Delta t\lambda)^2 y^0, \\ y^3 &= (1 + \Delta t\lambda)y^2 = (1 + \Delta t\lambda)^3 y^0, \\ &\vdots \end{aligned} \tag{5.61}$$

Hence the discrete phase flow $\mathcal{E}$ can be written as

$$\mathcal{E}(x, \Delta t) = R(\Delta t \lambda)x, \tag{5.62}$$

where the affine polynomial $R(z)$,

$$R(z) = 1 + z = \exp(z) + O(z^2), \tag{5.63}$$

is said to be the stability function of the explicit Euler method. The function $R(z)$ is a consistent approximation of $\exp(z)$ in the sense of the following definition.

Definition 5.7 *We say that a rational approximation $R(z)$ of the complex exponential* $\exp(z)$ *has* consistency order p *if*

$$\exp(z) = R(z) + O(z^{p+1}).$$

The function $R(z)$ is said to be consistent *if $p \geq 1$.*

As the reader may expect. the consistency order of the stability function R is tightly related to the consistency order of the function $\mathcal{E}$ (in the sense of Definition 5.2). This will be formulated precisely in Lemma 5.6. The stability requirement

$$\lim_{n \to \infty} y^n = 0,$$

applied to the method (5.61), yields the stability condition

$$|1 + \Delta t \lambda| < 1.$$

This is equivalent to the well known time step restriction for the explicit Euler method,

$$\Delta t < \frac{-2\mathrm{Re}(\lambda)}{|\lambda|^2}. \tag{5.64}$$

In the real case ($0 \neq \lambda \in \mathbb{R}$) condition (5.64) reduces to

$$\Delta t < \frac{2}{|\lambda|}.$$

Definition 5.8 (Stability domain of $\mathcal{E}$) *Let the discrete phase flow $\mathcal{E}$ have the form (5.62). Then its* stability domain *is the set*

$$\mathcal{S}_R = \{z \in \mathbb{C};\ |R(z)| < 1\}.$$

In the case of the explicit Euler method, $R(z)$ is given by (5.63) and therefore the stability domain of $\mathcal{E}$ is the open complex circle with the center at $-1 + 0i$ and radius 1,

$$\mathcal{S}_R = \{z \in \mathbb{C};\ |1 + z| < 1\},$$

as illustrated in Figure 5.3.

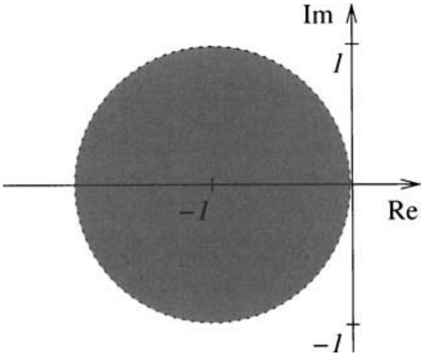


Figure 5.3 The stability domain $\mathcal{S}_R$ of the discrete phase flow $\mathcal{E}$ of the explicit Euler method.

We see that $\mathcal{S}_R \subset \mathcal{S}_{exp}$, and this is why the time step restriction (5.64) exists. The relation of the stability domains $\mathcal{S}_R$ and $\mathcal{S}_{exp}$ will be different in the case of the implicit Euler method:

Implicit Euler method The method (5.27), applied to the linear scalar equation (5.58), yields

$$y^{k+1} = y^k + \lambda \Delta t y^{k+1}.$$

From here, the value of y^{k+1} is calculated via the relation

$$y^{k+1} = (1 - \Delta t \lambda)^{-1} y^k.$$

In the vector-valued linear autonomous case (5.49), this operation corresponds to the solution of a system of linear equations (to be discussed in more detail in Paragraph 5.3.4). With a constant time step Δt, the method approximates $y(t) = \mathcal{Y}(y^0, t)$ with a series of discrete values

$$\begin{array}{rcl} y^1 & = & (1 - \Delta t\lambda)^{-1} y^0, \\ y^2 & = & (1 - \Delta t\lambda)^{-1} y^1 = (1 + \Delta t\lambda)^{-2} y^0, \\ y^3 & = & (1 - \Delta t\lambda)^{-1} y^2 = (1 + \Delta t\lambda)^{-3} y^0, \\ & \vdots & \end{array} \tag{5.65}$$

Thus the discrete phase flow $\mathcal{E}$ attains a form similar to (5.62),

$$\mathcal{E}(x, \Delta t) = R(\Delta t \lambda) x,$$

but now the stability function $R(z)$ is rational,

$$R(z) = \frac{1}{1 - z} = 1 + z + z^2 + z^3 + \ldots, \quad \exp(z) = R(z) + O(z^2). \tag{5.66}$$

We see that the function $R(z)$ is a consistent approximation of $\exp(z)$. The stability domain of the function $\mathcal{E}$ is

$$\mathcal{S}_R = \mathbb{C} \setminus \{z \in \mathbb{C};\ |R(z)| < 1\} = \mathbb{C} \setminus \{z \in \mathbb{C};\ |1 - z| < 1\},$$

which is the complement of the closed complex circle with the center at $1 + 0i$ and radius 1. In particular,

$$\mathcal{S}_{exp} \subset \mathcal{S}_R,$$

and therefore the implicit Euler method is stable for all values of $\lambda \in \mathbb{C}$ and all time steps Δt (we say that it is absolutely stable).

5.3.4 Stability functions for general RK methods

The stability functions to general higher-order RK methods are obtained by extending the results from the previous paragraph. Let us give one additional example prior to introducing the general result in Theorem 5.3.

Applying the second-order "implicit trapezoidal rule" IRK method (5.46) with constant time step Δt to the vector-valued linear autonomous system (5.49), one obtains

$$\boldsymbol{Y}^{k+1} = \mathcal{E}(\boldsymbol{Y}^k, \Delta t) = \boldsymbol{Y}^k + \frac{\Delta t}{2}\left(\boldsymbol{A}\boldsymbol{Y}^{k+1} + \boldsymbol{A}\boldsymbol{Y}^k\right), \tag{5.67}$$

which is equivalent to

$$\left(\boldsymbol{I} - \frac{\Delta t}{2}\boldsymbol{A}\right)\boldsymbol{Y}^{k+1} = \boldsymbol{I} + \frac{\Delta t}{2}\boldsymbol{A}. \tag{5.68}$$

Hence the invertibility of the matrix $\boldsymbol{I} - \Delta t\boldsymbol{A}/2$ has to be checked. Regarding this, the following lemma is helpful.

Proposition 5.1 *Let $\boldsymbol{A} \in \mathbb{R}^{N\times N}$ be a constant matrix and $R(z) = P(z)/Q(z)$ a rational function, where P and Q are mutually prime polynomials. Then the definition of $R(\boldsymbol{A}) = P(\boldsymbol{A})Q^{-1}(\boldsymbol{A})$ makes sense only if the matrix $Q(\boldsymbol{A})$ is invertible, and this is the case if and only if no eigenvalue of $\boldsymbol{A}$ is a root of the function $Q(z)$.*

Proof: Depending on the reader's background, this can be shown simply using the machinery of functional calculus (see, e.g., [100]), or the statement can be proved discretely using the Jordan canonical form of the matrix $\boldsymbol{A}$ (see, e.g., [25] and [60]). ■

Returning to (5.68): Evidently the only pole of the function

$$R(z) = \frac{1 + z/2}{1 - z/2} \tag{5.69}$$

is $z^* = 2$. By Theorem 5.2 all eigenvalues of the matrix $\boldsymbol{A}$ in a stable linear autonomous system (5.11) are nonpositive. Using the fact that

$$\lambda \in \sigma(\boldsymbol{A}) \Leftrightarrow \lambda t \in \sigma(\boldsymbol{A}t),$$

by Proposition 5.1 we can write

$$\boldsymbol{Y}^{k+1} = \frac{\boldsymbol{I} + \Delta t\boldsymbol{A}/2}{\boldsymbol{I} - \Delta t\boldsymbol{A}/2}\boldsymbol{Y}^k,$$

(where the fraction is understood in the sense of multiplication by the inverse matrix to the denominator). In the following we encounter numerous rational approximations $R(z)$ of the complex exponential $\exp(z)$, whose poles will always have positive real parts for the same reason as here.

The stability function $R(z)$ approximates $\exp(z)$ with the consistency order $p = 2$,

$$R(z) = \frac{1 + z/2}{1 - z/2} = 1 + z + \frac{z^2}{2} + \frac{z^3}{4} + O(z^4) = \exp(z) + O(z^3).$$

Since $|R(z)| < 1$ for all $z \in \mathbb{C}$ such that $Re(z) < 0$, the inclusion $\mathcal{S}_{exp} \subset \mathcal{S}_R$ holds. Therefore the "implicit trapezoidal rule" IRK method (5.67) is absolutely stable.

Theorem 5.3 *The discrete phase flow $\mathcal{E}$ of a general s-stage RK method $(\boldsymbol{b}, \mathcal{A})$ has the form*

$$\mathcal{E}(\boldsymbol{X}, \Delta t) = R(\Delta t \boldsymbol{A})\boldsymbol{X},$$

where the stability function $R(z)$ is rational,

$$R(z) = 1 + z\boldsymbol{b}^T(\boldsymbol{I} - z\mathcal{A})^{-1}\mathbf{1} \tag{5.70}$$

(here by $\mathbf{1}$ we mean the vector $(1, 1, \ldots, 1)^T \in \mathbb{R}^s$). Moreover, $R(z)$ can be written uniquely in the form

$$R(z) = \frac{P(z)}{Q(z)} \tag{5.71}$$

with mutually prime polynomials P and Q such that $\deg P \leq s$, $\deg Q \leq s$, and $P(0) = Q(0) = 1$.

The expression $z\mathcal{A}$ is interpreted as the tensor product of two matrices, $\boldsymbol{A}t \otimes \mathcal{A}$, when $\boldsymbol{A}t$ is substituted for z in (5.71).

Proof: It is sufficient to consider the scalar linear autonomous ODE $\dot{y}(t) = \lambda y(t)$, $y(0) = 1$. The RK method $(\boldsymbol{b}, \mathcal{A})$ with the time step Δt yields the linear system,

$$\mathcal{E}(1, \Delta t) = R(\Delta t \lambda) = 1 + \Delta t \sum_{j=1}^{s} b_j \lambda g_j,$$

where

$$g_i = 1 + \Delta t \sum_{j=1}^{s} a_{ij} \lambda g_j, \quad i = 1, 2, \ldots, s.$$

Putting $z = \Delta t \lambda$ and $\boldsymbol{g} = (g_1, g_2, \ldots, g_s)^T \in \mathbb{R}^s$, this system yields

$$R(z) = 1 + z\boldsymbol{b}^T\boldsymbol{g}, \quad \boldsymbol{g} = \mathbf{1} + z\mathcal{A}\boldsymbol{g},$$

and we obtain (5.70).

To the second part of the assertion: The system can be solver by Cramer's rule, in which case we find that

$$g_i = \frac{P_i}{\det(\boldsymbol{I} - z\mathcal{A})}, \quad i = 1, 2, \ldots, s,$$

where P_i are polynomials of $\deg P_i \le s - 1$. Since $\tilde{Q}(z) = \det(\boldsymbol{I} - z\mathcal{A})$ is a polynomial of $\deg \tilde{Q} \le s$ satisfying $\tilde{Q}(0) = 1$, it follows that the rational function $R(z)$,

$$R(z) = \frac{\tilde{Q}(z) + z\sum_{j=1}^{s} b_j P_j}{\tilde{Q}(z)}.$$

assumes the form (5.71) once all common divisors in the numerator and denominator have been removed. ■

An immediate consequence of Theorem 5.3 is Lemma 5.5.

Lemma 5.5 *The stability function $R(z)$ of explicit higher-order RK methods $(\boldsymbol{b}, \mathcal{A})$ is polynomial.*

Proof: Left to the reader as an exercise. ■

A simple example documenting this fact is the fourth-order "classical" RK method mentioned Paragraph 5.2.4, whose stability function $R(z)$ (consistent with the order $p = 4$) has the form

$$R(z) = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24} = \exp(z) + O(z^5). \tag{5.72}$$

The verification is left to the reader as an exercise.

Finally we can formulate the intuitively clear relation between the order of consistency of the rational stability function $R(z)$ and the consistency order of the function $\mathcal{E}$:

Lemma 5.6 *Consider a linear autonomous system (5.49) with the continuous phase flow $\mathcal{Y}(\boldsymbol{X}, t)$. Let the discrete phase flow $\mathcal{E}(\boldsymbol{X}, \Delta t)$ be defined as*

$$\mathcal{E}(\boldsymbol{X}, \Delta t) = R(\boldsymbol{A}t)\boldsymbol{X},$$

where $R(z)$ is a rational stability function that approximates the complex exponential $\exp(z)$ *with the consistency order $p \ge 1$. Then the consistency order of the function $\mathcal{E}$ is p,*

$$\mathcal{Y}(\boldsymbol{X}, \Delta t) - \mathcal{E}(\boldsymbol{X}, \Delta t) = O((\Delta t\boldsymbol{A})^{p+1}).$$

Proof: The result follows immediately from (5.55) and Definition 5.2. ■

5.3.5 Maximum consistency order of IRK methods

While by Lemma 5.3 the consistency order p of explicit RK methods can never exceed the number of stages s, already the simplest "implicit midpoint rule" IRK in Paragraph 5.2.6 exhibited twice better consistency order $p = 2s$. This is true for implicit RK methods in general. The reason is that the rational stability function $R(z) = P(z)/Q(z)$, where $\deg P = \deg Q = s$, can approximate the complex exponential $\exp(z)$ with the consistency order up to $p = 2s$:

Lemma 5.7 *Let $R(z)$ be a rational approximation of the complex exponential such that*

$$\exp(z) = R(z) + O(z^{p+1}). \tag{5.73}$$

Then any representation $R(z) = P(z)/Q(z)$ satisfies

$$p \leq \deg P + \deg Q.$$

Proof: By contradiction suppose that (5.73) holds with $\deg P \leq k$, $\deg Q \leq j$, such that $k + j \leq p$. Hence

$$\frac{P(z)}{Q(z)} - \exp(z) = O(z^{k+j+2}) \quad \text{as } z \to 0,$$

and by multiplication with $Q(z)$,

$$P(z) - Q(z)\exp(z) = O(z^{k+j+2}). \tag{5.74}$$

Relation (5.74) already implies that necessarily $P = Q = 0$ – this is the desired contradiction that will be shown by induction.

Consider first $k = 0$, in which case P is a constant. Multiplying (5.74) by $\exp(-z)$ we obtain that

$$P(z)\exp(-z) - Q(z) = O(z^{j+2}). \tag{5.75}$$

It is $\deg Q \leq j$, and therefore differentiating (5.75) $j + 1$ times, we find that

$$(-1)^{j+1}P = O(z), \tag{5.76}$$

which means that $P = 0$, and from (5.75) necessarily $Q = 0$. Now the induction step: Assume that the statement holds for $k - 1 \geq 0$. Differentiating (5.74) one obtains

$$P'(z) - (Q'(z) + Q(z))\exp(z) = O(z^{k+j+1}).$$

Since $\deg P' \leq k - 1$ and $\deg(Q' + Q) \leq j$, by the induction hypothesis we can conclude that $P' = 0$. This again yields $P = Q = 0$. ■

Several examples of s-stage implicit RK methods with the maximum consistency order will be presented in Paragraph 5.4.2. Such methods are constructed elegantly by embedding higher-order numerical quadrature rules into a general collocation framework.

5.3.6 A-stability and L-stability

Being familiar with the rational stability functions $R(z)$ and the stability domains of both the continuous and discrete phase flows $\mathcal{Y}$ and $\mathcal{E}$, we can introduce the concepts of A- and L-stability.

***A*-stability** A one-step method for the autonomous system (5.49) is absolutely stable only if the inclusion

$$\mathcal{S}_{exp} \subset \mathcal{S}_R \tag{5.77}$$

holds. [This was the case, e.g., with the implicit Euler method (5.27) and with the "implicit trapezoidal rule" IRK (5.46)]. Generally, one-step methods with the property (5.77) are called A-stable (G. Dahlquist, 1963). Lemma 5.8 gives an important characterization of A-stable methods:

Lemma 5.8 *One-step method $\mathcal{E}(\boldsymbol{X}, \Delta t) = R(\Delta t \boldsymbol{A})\boldsymbol{X}$, whose stability function $R(z)$ is polynomial, cannot be A-stable.*

Proof: Every nontrivial polynomial $R(z)$ satisfies

$$\lim_{z \longrightarrow \infty} |R(z)| = \infty,$$

and therefore its stability domain

$$\mathcal{S}_R = \{z \in \mathbb{C};\ |R(z)| < 1\}$$

is compact. Thus $\mathcal{S}_R$ never can contain the whole negative complex half-plane $\mathcal{S}_{exp}$. ■

Consequently, every A-stable one-step method necessarily is implicit. Moreover, every explicit one-step method comes with some stability restriction on the time step.

***L*-stability** Until now we discussed the stability of the recursive procedure (5.20),

$$\begin{aligned} \boldsymbol{Y}^1 &= \mathcal{E}(\boldsymbol{Y}^0, \Delta t) = R(\Delta t \boldsymbol{A})\boldsymbol{Y}^0, \\ \boldsymbol{Y}^2 &= \mathcal{E}(\boldsymbol{Y}^1, \Delta t) = R^2(\Delta t \boldsymbol{A})\boldsymbol{Y}^0, \\ \boldsymbol{Y}^3 &= \mathcal{E}(\boldsymbol{Y}^2, \Delta t) = R^3(\Delta t \boldsymbol{A})\boldsymbol{Y}^0, \\ &\vdots \end{aligned}$$

with a fixed time step Δt. However, one also is interested in the behavior of implicit methods as the size of the time step Δt grows. We are asking whether and when the condition

$$\lim_{\Delta t \longrightarrow \infty} \mathcal{E}(\boldsymbol{X}, \Delta t) = 0$$

holds. This motivates the definition of L-stability:

Definition 5.9 *Let the linear autonomous problem (5.49) be asymptotically stable. Then the one-step method $\mathcal{E}(\boldsymbol{X}, \Delta t) = R(\Delta t \boldsymbol{A})\boldsymbol{X}$ is said to be* L-stable *if and only if it is A-stable and*

$$\lim_{z \longrightarrow \infty} R(z) = 0. \tag{5.78}$$

This terminology was introduced by B.L. Ehle in 1969. Again, explicit one-step methods are not considered here. The following example shows that not all A-stable methods are L-stable:

■ EXAMPLE 5.1 (A- and L-stable methods)

1. We know from Paragraph 5.3.3 that the implicit Euler method is A-stable. Its rational stability function (5.66),

$$R(z) = \frac{1}{1-z},$$

satisfies (5.78). Therefore this method also is L-stable.

2. Another A-stable method that we know is the second-order "implicit trapezoidal rule" IRK method with the rational stability function (5.69),

$$R(z) = \frac{1+z/2}{1-z/2}.$$

Since

$$\lim_{z\longrightarrow\infty} R(z) \neq 0,$$

this method is not L-stable.

3. Next consider the second-order IRK method based on the "implicit midpoint rule"

$$\begin{array}{c|c} 1/2 & 1/2 \\ \hline & 1 \end{array}$$

This method is not L-stable (the proof is left to the reader as an exercise).

4. Last consider the 2-stage third-order Radau method

$$\begin{array}{c|cc} 1/3 & 5/12 & -1/12 \\ 1 & 3/4 & 1/4 \\ \hline & 3/4 & 1/4 \end{array}$$

This method is A-stable and also L-stable, which again is left to the reader as an exercise. More about Gauss and Radau methods will be said in Section 5.4.

Condition (5.78), which is necessary for the L-stability of the last IRK method in Example 5.1, can be verified literally at a glance, using the following theorem:

Theorem 5.4 *Suppose that for a general RK method $(\boldsymbol{b}, \mathcal{A})$ the matrix $\mathcal{A}$ is invertible, and the row vector $\boldsymbol{b}^T$ is identical to some row of the matrix $\mathcal{A}$. Then*

$$\lim_{z\longrightarrow\infty} R(z) = 0.$$

Proof: Since $\mathcal{A}$ is invertible, Theorem 5.3 implies that

$$\lim_{z\longrightarrow\infty} R(z) = 1 - \boldsymbol{b}^T\mathcal{A}^{-1}\mathbf{1}.$$

Let the jth row of $\mathcal{A}$ be identical with $\boldsymbol{b}^T$. Then

$$\boldsymbol{b} = \boldsymbol{e}_j^T \mathcal{A},$$

where $\boldsymbol{e}_j$ denotes the jth unit vector. The conclusion

$$\lim_{z \to \infty} R(z) = 1 - \boldsymbol{e}_j^T \mathcal{A}\mathcal{A}^{-1}\mathbf{1} = 0$$

follows immediately. ■

At this point we believe to have given a sufficient introduction to the stability of ODEs and one-step methods. For a deeper study of this topic we refer the reader to the books [25] and [60]. In the last section of this chapter let us discuss the most sophisticated one-step ODE solvers: the higher-order implicit Runge–Kutta methods.

5.4 HIGHER-ORDER IRK METHODS

It was discovered in early 1970s that general (implicit) RK methods could be generated by combining classical collocation methods with higher-order numerical quadrature rules. Several RK methods derived via the traditional Taylor expansion techniques turned out to actually be collocation methods. A classical book on higher-order IRK methods is [60].

5.4.1 Collocation methods

Let us return to the (nonautonomous) ODE system (5.11), (5.12) resulting from the MOL,

$$\begin{aligned} \dot{\boldsymbol{Y}}(t) &= \boldsymbol{\Phi}(\boldsymbol{Y}(t), t), && (5.79)\\ \boldsymbol{Y}(0) &= \boldsymbol{Y}^0. && (5.80) \end{aligned}$$

The collocation constructs the approximate solution $\boldsymbol{X}(t) \approx \boldsymbol{Y}(t)$ as a continuous (vector-valued) function which is a polynomial of degree s in every interval $[t_k, t_k + \Delta t_k]$, $k = 0, 1, \ldots, K-1$,

$$\boldsymbol{X}(t_k + \tau) = \mathcal{E}(\boldsymbol{Y}^k, t_k, \tau), \quad \tau \in [0, \Delta t_k].$$

In the interval $[t_k, t_k + \Delta t_k]$ the function $\boldsymbol{X}$ not only must fulfill the initial condition

$$\boldsymbol{X}(t_k) = \mathcal{E}(\boldsymbol{Y}^k, t_k, 0) = \boldsymbol{Y}^k, \qquad (5.81)$$

but also it has to satisfy ("collocate") equation (5.79) at additional s internal points $t_k + c_1\Delta t_k, t_k + c_2\Delta t_k, \ldots, t_k + c_s\Delta t_k$ of the interval $[t_k, t_k + \Delta t_k]$,

$$\begin{aligned} \dot{\boldsymbol{X}}(t_k + c_1\Delta t_k) &= \boldsymbol{\Phi}(\boldsymbol{X}(t_k + c_1\Delta t_k), t_k + c_1\Delta t_k), && (5.82)\\ \dot{\boldsymbol{X}}(t_k + c_2\Delta t_k) &= \boldsymbol{\Phi}(\boldsymbol{X}(t_k + c_2\Delta t_k), t_k + c_2\Delta t_k), \\ &\vdots \\ \dot{\boldsymbol{X}}(t_k + c_s\Delta t_k) &= \boldsymbol{\Phi}(\boldsymbol{X}(t_k + c_s\Delta t_k), t_k + c_s\Delta t_k), \end{aligned}$$

where $0 \le c_1 < c_2 < \ldots, < c_s \le 1$ are suitable constants. These s parameters fully determine the method (its consistency, convergence, stability, and all other aspects). With the approximate solution $\boldsymbol{X}(t)$ in hand, the approximate solution on the new time level is defined as

$$\boldsymbol{Y}^{k+1} = \boldsymbol{X}(t_k + \Delta t_k) = \mathcal{E}(\boldsymbol{Y}^k, t_k, \Delta t_k). \tag{5.83}$$

It is known that $\boldsymbol{X}(t)$ exists and is unique for sufficiently small time steps and a sufficiently regular right-hand side $\boldsymbol{\Phi}$. However, the proof is by no means trivial, and we refer the reader to [25] and [60]. In the following let us discuss the selection of the parameters c_i, $i = 1, 2, \ldots, s$.

The collocation procedure Consider a set of collocation points $0 \le c_1 < c_2 < \ldots, < c_s \le 1$, along with the standard Lagrange nodal basis $\theta_1, \theta_2, \ldots, \theta_s$ of the polynomial space

$$P = P^{s-1}(0,1),$$

satisfying the condition

$$\theta_i(c_j) = \delta_{ij}.$$

For brevity, by z_i denote the derivative of $\boldsymbol{X}(t)$ at the collocation point c_i,

$$\boldsymbol{z}_i = \dot{\boldsymbol{X}}(t_k + c_i \Delta t_k) \qquad \text{for all } 1 \le i \le s.$$

Exploiting the Lagrange interpolation polynomial (A.75), the derivative $\dot{\boldsymbol{X}}(t)$ in the interval $[t_k, t_k + \Delta t_k]$ can be written as

$$\dot{\boldsymbol{X}}(t_k + \xi \Delta t_k) = \sum_{j=1}^{s} \boldsymbol{z}_j \theta_j(\xi), \quad \xi \in [0,1]. \tag{5.84}$$

Integrating (5.84) and using the initial condition (5.81), we find that

$$\boldsymbol{X}(t_k + c_i \Delta t_k) = \boldsymbol{Y}^k + \Delta t_k \int_0^{c_i} \dot{\boldsymbol{X}}(t_k + \xi \Delta t_k)\,\mathrm{d}\xi = \boldsymbol{Y}^k + \Delta t_k \sum_{j=1}^{s} a_{ij} \boldsymbol{z}_j, \tag{5.85}$$

where

$$a_{ij} = \int_0^{c_i} \theta_j(\xi)\,\mathrm{d}\xi, \quad i, j = 1, 2, \ldots, s. \tag{5.86}$$

Substituting these values into the collocation condition (5.82), one obtains

$$\boldsymbol{z}_i = \boldsymbol{\Phi}\left(\boldsymbol{Y}^k + \Delta t_k \sum_{j=1}^{s} a_{ij} \boldsymbol{z}_j, t_k + c_i \Delta t_k\right), \quad i = 1, 2, \ldots, s.$$

By (5.83) and (5.85) the approximation at the $(k+1)$th time level has the form

$$\boldsymbol{Y}^{k+1} = \boldsymbol{X}(t_k + \Delta t_k) = \boldsymbol{Y}^k + \Delta t_k \int_0^1 \dot{\boldsymbol{X}}(t_k + \xi \Delta t_k)\,\mathrm{d}\xi = \boldsymbol{Y}^k + \Delta t_k \sum_{j=1}^{s} b_j \boldsymbol{z}_j, \tag{5.87}$$

where

$$b_j = \int_0^1 \theta_j(\xi)\,\mathrm{d}\xi, \quad j = 1, 2, \dots, s. \tag{5.88}$$

Let us see that the points $\boldsymbol{c}$ and weights $\boldsymbol{b}$ represent a quadrature rule that is exact for polynomials of the degree $s-1$: Every such polynomial φ can be written in terms of the Lagrange basis,

$$\varphi(\xi) = \sum_{j=1}^{s} \varphi(c_j)\theta_j(\xi),$$

and for its integral one obtains

$$\int_0^1 \varphi(\xi)\,\mathrm{d}\xi = \int_0^1 \sum_{j=1}^{s} \varphi(c_j)\theta_j(\xi)\,\mathrm{d}\xi = \sum_{j=1}^{s} \varphi(c_j) \int_0^1 \theta_j(\xi)\,\mathrm{d}\xi = \sum_{j=1}^{s} \varphi(c_j) b_j.$$

Finally let us define

$$\mathcal{A} = \{a_{ij}\}_{i,j=1}^{s}, \quad \boldsymbol{b} = (b_1, b_2, \dots, b_s)^T, \quad \boldsymbol{c} = (c_1, c_2, \dots, c_s)^T.$$

The relation (5.87) represents the implicit RK method $(\boldsymbol{b}, \boldsymbol{c}, \mathcal{A})$ defined in (5.44). The consistency condition for explicit RK methods (5.41),

$$\sum_{j=1}^{s} b_j = 1,$$

extends naturally to implicit RK methods via (5.88),

$$\sum_{j=1}^{s} b_j = \sum_{j=1}^{s} \int_0^1 \theta_j(\xi)\,\mathrm{d}\xi = \int_0^1 \sum_{j=1}^{s} \theta_j(\xi) \cdot 1\,\mathrm{d}\xi = \int_0^1 1\,\mathrm{d}\xi = 1. \tag{5.89}$$

Moreover, we have the following lemma:

Lemma 5.9 *The coefficients of an implicit RK method $(\boldsymbol{b}, \boldsymbol{c}, \mathcal{A})$ defined by collocation satisfy the conditions*

$$\sum_{j=1}^{s} b_j c_j^{q-1} = \frac{1}{q}, \quad q = 1, 2, \dots, s, \tag{5.90}$$

and

$$\sum_{j=1}^{s} a_{ij} c_j^{q-1} = \frac{1}{q} c_i^q, \quad q = 1, 2, \dots, s \tag{5.91}$$

(with the convention $0^0 = 1$). In particular, the method is consistent and invariant under autonomization.

Proof: It follows from (5.88) that

$$\sum_{j=1}^{s} b_j c_j^{q-1} = \sum_{j=1}^{s} \int_0^1 c_j^{q-1} \theta_j(\xi)\,\mathrm{d}\xi = \int_0^1 \sum_{j=1}^{s} c_j^{q-1} \theta_j(\xi)\,\mathrm{d}\xi.$$

Looking at the integrand in more detail, we discover that at each collocation point c_r it achieves the value c_r^{q-1}, $r = 1, 2, \ldots, s$. Thus necessarily

$$\sum_{j=1}^{s} c_j^{q-1}\theta_j(\xi) = \xi^{q-1},$$

and (5.90) follows,

$$\sum_{j=1}^{s} b_j c_j^{q-1} = \int_0^1 \xi^{q-1}\,\mathrm{d}\xi = \frac{1}{q}.$$

Using the same technique for (5.86), one easily obtains (5.91). The consistency follows from (5.89) which is a special case of (5.90), and the invariance under autonomization follows immediately from Lemma 5.4. ■

The formula (5.90) states that the quadrature rule

$$\sum_{j=1}^{s} b_j\varphi(c_j) \approx \int_0^1 \varphi(\xi)\,\mathrm{d}\xi$$

is exact for all polynomials $\varphi \in P^{s-1}(0, 1)$.

The following result, which relates the consistency order of an implicit RK method constructed via collocation to the order of accuracy of the underlying quadrature rule, was first proved under slightly simplified assumptions by J. C. Butcher in 1964. A different idea of the proof was presented by S.P. Norsett and G. Wanner [89] in 1979.

Theorem 5.5 *An implicit RK method $(\boldsymbol{b}, \boldsymbol{c}, \mathcal{A})$ generated by collocation has for a p-times continuously differentiable right-hand side $\boldsymbol{\Phi}$ the consistency order p if and only if the quadrature formula defined by the nodes* **c** *and weights* **b** *has the order of accuracy p.*

Proof: See, e.g., [59]. ■

5.4.2 Gauss and Radau IRK methods

Theorem 5.5 suggests an efficient strategy for the design of s-stage implicit RK schemes of the consistency order $1 \leq p \leq 2s$:

1. Choose a quadrature rule $(\boldsymbol{c}, \boldsymbol{b})$ that is exact for polynomials of order $p - 1$.

2. Use (5.86) to define the Butcher's array $(\boldsymbol{b}, \boldsymbol{c}, \mathcal{A})$.

Gauss IRK methods From Section 2.3 we know that every Gaussian quadrature rule $(\boldsymbol{c}, \boldsymbol{b})$ with s quadrature points is exact for polynomials of the degree up to $2s - 1$.

Lemma 5.10 *Every s-stage Gauss IRK method has the consistency order $p = 2s$ for all $2s$-times continuously differentiable right-hand sides $\boldsymbol{\Phi}$.*

Proof: Immediate consequence of Theorem 5.5. ■

Thus the Gauss IRK methods attain the maximum consistency order $p = 2s$ of s-stage IRK methods, derived in Paragraph 5.3.5. In contrast to this result, the maximum order of s-stage explicit RK methods is an open problem (see Table 5.1). The Gauss IRK method for $s = 1$ is the "implicit midpoint rule" that we are familiar with from Paragraph 5.2.6

and from Example 5.1. Another Gauss IRK method with the consistency order $p = 4$, corresponding to the stage count $s = 2$, is

$$\begin{array}{c|cc}
1/2 - \sqrt{3}/6 & 1/4 & 1/4 - \sqrt{3}/6 \\
1/2 + \sqrt{3}/6 & 1/4 + \sqrt{3}/6 & 1/4 \\
\hline
 & 1/2 & 1/2
\end{array}$$

Lemma 5.11 *All Gauss IRK methods are A-stable. Moreover, their stability domain $\mathcal{S}_R$ exactly coincides with the negative complex half plane $\mathcal{S}_{exp}$ (5.60).*

Proof: See, e.g., [60]. ■

Let us mention that $\mathcal{S}_R = \mathcal{S}_{exp}$ means that Gauss IRK methods preserve isometry. Moreover it is known that these methods are reversible. Both these properties are positive for the performance of the methods, as the reader may expect. These and more interesting aspects of IRK methods are thoroughly discussed in [60].

One of the few drawbacks of Gauss IRK methods is that generally they are not L-stable. This is a consequence of the fact that the Gaussian quadrature points do not lie at interval endpoints, and therefore the approximate solution obtained via collocation has jumps in the temporal derivative at all times t_k, $k = 1, 2, \ldots$.

Radau IRK methods The above-mentioned lack of L-stability is eliminated via collocation methods based on Radau quadrature rules, that place collocation points at the interval endpoints (see, e.g., [111] for details on this numerical quadrature and for a CD-ROM with Radau quadrature data).

Lemma 5.12 *Every s-stage Radau IRK method has the consistency order $p = 2s - 1$ for all $(2s - 1)$-times continuously differentiable right-hand sides Φ. All Radau IRK methods are A-stable and also L-stable.*

Proof: The consistency order follows from the fact that a Lobatto-Radau quadrature rule with s nodes has the order of accuracy $p = 2s - 1$, and from Theorem 5.5. For the rest see, e.g., [60]. ■

The reader already encountered the 1-stage Radau method (implicit Euler scheme) and the 2-stage third-order Radau method in Example 5.1. Let us present the 3-stage fifth-order Radau method,

$$\begin{array}{c|ccc}
(4 - \sqrt{6})/10 & (88 - 7\sqrt{6})/360 & (296 - 169\sqrt{6})/1800 & (-2 + 3\sqrt{6})/225 \\
(4 + \sqrt{6})/10 & (296 + 169\sqrt{6})/1800 & (88 + 7\sqrt{6})/360 & (-2 - 3\sqrt{6})/225 \\
1 & (16 - \sqrt{6})/36 & (16 + \sqrt{6})/36 & 1/9 \\
\hline
 & (16 - \sqrt{6})/36 & (16 + \sqrt{6})/36 & 1/9
\end{array}$$

The L-stability of this method follows from Theorem 5.4 immediately. For an implementation of this method, enhanced with a step size control based on an embedded third-order method, see code RADAU5 in [60].

5.4.3 Solution of nonlinear systems

Although we mostly deal with linear problems in this introductory text, let us devote one paragraph to the approximate solution to system of nonlinear algebraic equations arising from the application of higher-order IRK methods. We describe both the quadratically convergent classical Newton's method, and a simplified Newton's method that converges linearly, but without the need to reconstruct the Jacobi matrix of the right-hand side in every iteration.

One step of the IRK method Recall that the implicit s-stage RK method (5.44), (5.45) consists of two operations: the calculation of the stage derivatives z_i via a system of generally nonlinear algebraic equations,

$$z_i = \Phi(Y^k + \Delta t_k \sum_{j=1}^{s} a_{ij} z_j, t_k + c_i \Delta t_k), \quad i = 1, 2, \ldots, s, \tag{5.92}$$

and the evaluation of the solution on the new time level,

$$Y^{k+1} = Y^k + \Delta t_k \sum_{i=1}^{s} b_i z_i. \tag{5.93}$$

In order to ease the operation with the Jacobi matrix of the right-hand side, it is advantageous to introduce a set of new vectors,

$$g_i = \Delta t_k \sum_{j=1}^{s} a_{ij} z_j, \quad i = 1, 2, \ldots, s. \tag{5.94}$$

Substituting (5.94) into (5.92), one obtains

$$z_i = \Phi(Y^k + g_i, t_k + c_i \Delta t_k), \quad i = 1, 2, \ldots, s, \tag{5.95}$$

and by (5.94) this further yields

$$g_i = \Delta t_k \sum_{j=1}^{s} a_{ij} \Phi(Y^k + g_j, t_k + c_j \Delta t_k), \quad i = 1, 2, \ldots, s. \tag{5.96}$$

Let us postpone the solution of this nonlinear system for a moment, and assume that the vectors $g_1, g_2, \ldots, g_s$ are known. In order to accomplish the step of the IRK method by (5.93), one needs to distinguish two situations depending on the coefficient matrix $\mathcal{A}$.

$\mathcal{A}^{-1}$ exists:

In this case the evaluation of Y^{k+1} is easier. Consider the matrix $Z = (z_1, z_2, \ldots, z_s)$ of the type $N \times s$, that has the stage derivatives z_i in its columns, and the matrix $H = (g_1, g_2, \ldots, g_s)$. For later use, by z_i^* and g_i^*, $i = 1, 2, \ldots, N$, denote the rows of the matrices Z and H, respectively. The relation (5.94) between the vectors z_i and g_i can be expressed as

$$H^T = \Delta t_k \mathcal{A} Z^T,$$

and in particular it is

$$Z^T = \frac{1}{\Delta t_k}\mathcal{A}^{-1}H^T. \tag{5.97}$$

Thus the original stage derivatives z_i can be recovered from (5.97) via a single matrix multiplication and (5.93) can be used directly,

$$Y^{k+1} = Y^k + \sum_{j=1}^{s} d_j g_j, \quad d^T = b^T \mathcal{A}^{-1}. \tag{5.98}$$

This result greatly reduces if b^T is identical to some row of $\mathcal{A}$ (see Theorem 5.4). This was the case, e.g., with the third-order Radau method from Example 5.1 as well as with the fifth-order Radau method presented in Paragraph 5.4.2. For example, when this is the last row, then

$$b^T = e_s^T \mathcal{A}, \quad e_s = (0, 0, \dots, 0, 1)^T \in \mathbb{R}^s.$$

Hence,

$$d = e_s,$$

and (5.98) simplifies to

$$Y^{k+1} = Y^k + g_s.$$

<u>$\mathcal{A}$ is not invertible:</u>

In this case generally one cannot access the vectors z_i. Substituting (5.95) into (5.93), one obtains

$$Y^{k+1} = Y^k + \Delta t_k \sum_{i=1}^{s} b_i \Phi(Y^k + g_i, t_k + c_i \Delta t_k). \tag{5.99}$$

Thus the vectors g_i can be used instead, but at the price of s additional evaluations of the right-hand side Φ.

The classical Newton's iteration Let us now turn our attention to the solution of the nonlinear algebraic system (5.96) for the vectors g_i, $i = 1, 2, \dots, s$. For the sake of clarity, let us define the vector

$$G = (g_1, g_2, \dots, g_s)^T \in \mathbb{R}^{Ns},$$

and write the system (5.96) in a compact form

$$\Psi(G) = G - \Delta t_k \begin{pmatrix} \sum_{j=1}^{s} a_{1j}\Phi(Y^k + g_j, t_k + c_j\Delta t_k) \\ \sum_{j=1}^{s} a_{2j}\Phi(Y^k + g_j, t_k + c_j\Delta t_k) \\ \vdots \\ \sum_{j=1}^{s} a_{sj}\Phi(Y^k + g_j, t_k + c_j\Delta t_k) \end{pmatrix} = 0.$$

Since the solution components g_i are expected to be small, it makes sense to use zero vector as the initial guess. Then the classical Newton's method assumes the form

$$\begin{aligned} G^0 &= 0, \\ \frac{\mathrm{D}\Psi}{\mathrm{D}G}(G^n)\Delta G^n &= -\Psi(G^n), \qquad (5.100) \\ G^{n+1} &= G^n + \Delta G^n. \qquad (5.101) \end{aligned}$$

Thus in each step one has to solve a system of Ns linear algebraic equations with the Jacobi matrix

$$\frac{\mathrm{D}\Psi}{\mathrm{D}G}(G) = \begin{pmatrix} B_{11} & B_{12} & \dots & B_{1s} \\ B_{21} & B_{22} & \dots & B_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ B_{s1} & B_{s2} & \dots & B_{ss} \end{pmatrix}. \qquad (5.102)$$

Each block B_{ij} has the size $N \times N$. The diagonal blocks B_{ii} are defined as

$$B_{ii} = I - \Delta t_k a_{ii} \frac{\mathrm{D}\Phi}{\mathrm{D}Y}(Y^k + g_i, t_k + c_i \Delta t_k),$$

and the nondiagonal blocks B_{ij}, $i \neq j$, have the form

$$B_{ij} = -\Delta t_k a_{ij} \frac{\mathrm{D}\Phi}{\mathrm{D}Y}(Y^k + g_j, t_k + c_j \Delta t_k).$$

The standard convergence theorem for the Newton's method implies quadratic convergence for sufficiently small Δt_k. But one has to realize that this convergence rate comes at a high price: At each step of the loop (5.100)–(5.101) the complete Jacobi matrix (5.102) of the size $Ns \times Ns$ has to be reconstructed. Therefore in practice one may consider a simpler iterative process:

Simplified Newton's method It is known that the iterative process (5.100)–(5.101) stays convergent for sufficiently small Δt_k at a reduced linear rate if the matrix $\mathrm{D}\Psi/\mathrm{D}G(G^n)$ in (5.100) is replaced with $\mathrm{D}\Psi/\mathrm{D}G(G^0)$. With

$$J_k = \frac{\mathrm{D}\Phi}{\mathrm{D}Y}(Y^k, t_k)$$

we have

$$\frac{\mathrm{D}\Psi}{\mathrm{D}G}(G^0) = \begin{pmatrix} I - \Delta t_k a_{11} J_k & -\Delta t_k a_{12} J_k & \dots & -\Delta t_k a_{1s} J_k \\ -\Delta t_k a_{21} J_k & I - \Delta t_k a_{22} J_k & \dots & -\Delta t_k a_{2s} J_k \\ \vdots & \vdots & \ddots & \vdots \\ -\Delta t_k a_{s1} J_k & -\Delta t_k a_{s2} J_k & \dots & I - \Delta t_k a_{ss} J_k \end{pmatrix}.$$

Using the tensor product of the matrices $\mathcal{A}$ and J_k, this can be written as

$$\frac{\mathrm{D}\Psi}{\mathrm{D}G}(G^0) = I - \Delta t_k \mathcal{A} \otimes J_k.$$

The simplified Newton's method assumes the form

$$\begin{aligned} J_k &= \frac{\mathrm{D}\Phi}{\mathrm{D}Y}(Y^k, t_k), \\ G^0 &= 0, \\ (I - \Delta t_k \mathcal{A} \otimes J_k)\Delta G^n &= -\Psi(G^n), \qquad (5.103) \\ G^{n+1} &= G^n + \Delta G^n. \qquad (5.104) \end{aligned}$$

This iterative procedure is economical, since a single LU-decomposition of the matrix $I - \Delta t_k \mathcal{A} \otimes J_k$ only is needed at each time step, i.e., an order of $2(Ns)^3/3$ operations.

Termination criterion for the simplified Newton's method Suppose that the time step Δt_k is sufficiently small so that the iteration (5.103)–(5.104) converges linearly, i.e., that there exists some contraction factor $0 \le \omega < 1$ so that

$$|\Delta G^{n+1}| \le \omega |\Delta G^n|, \quad n = 1, 2, \dots$$

Let G be the unknown exact solution. From relevant estimates for linearly convergent fixed point iterations (see, e.g., [99]) it is known that

$$|G - G^{n+1}| \le \frac{\omega}{1-\omega} |\Delta G^n|.$$

In practice the unknown contraction coefficient ω is replaced with the known quotient

$$\omega_n = \frac{|\Delta G^n|}{|\Delta G^{n-1}|}, \quad n = 1, 2, \dots$$

It is our aim to stop the iterative process as soon as $|G - G^{n+1}| \le TOL$, where $0 < TOL$ is a suitable small parameter. Thus the stopping criterion has the form

$$\frac{\omega_n}{1-\omega_n} |\Delta G^n| \le TOL,$$

i.e.,

$$\frac{|\Delta G^n|^2}{||\Delta G^{n-1}| - |\Delta G^n||} \le TOL.$$

5.5 EXERCISES

Exercise 5.1 *Use Definition 5.5 to prove that a solution of (5.11) that is asymptotically stable in the forward direction, is necessarily unstable in the backward direction.*

Exercise 5.2 *Consider the ODE (5.11) with the right-hand side (5.10). Extend the procedure of construction of the linear system (5.48) to the general $(b, c, \mathcal{A})$ RK method.*

Exercise 5.3 *Use Theorem 5.3 to prove that the stability function $R(z) = 1 + zb^T(I - z\mathcal{A})^{-1}\mathbf{1}$ of every explicit higher-order RK method $(b, \mathcal{A})$ is polynomial. Hint: Exploit the characteristic structure of the Butcher's matrix $\mathcal{A}$ for the inversion of $I - z\mathcal{A}$.*

Exercise 5.4 *Use the formula $R(z) = 1 + zb^T(I - z\mathcal{A})^{-1}\mathbf{1}$ to verify that the fourth-order "classical" RK method introduced in Paragraph 5.2.4 has the stability function (5.72).*

Exercise 5.5 *Verify that the second-order IRK method based on the "implicit midpoint rule" from Example 5.1 is not L-stable.*

Exercise 5.6 *Show (without using Theorem 5.4) that the third-order Radau IRK method from Example 5.1 is L-stable. Hint: Apply the method to the linear autonomous system (5.49), write its stability function $R(z)$, and verify the conditions for A- and L-stability.*

Exercise 5.7 *Prove the second formula (5.91) in Lemma 5.9 using the same technique as for the first formula (5.90).*

Exercise 5.8 *Consider the 2-stage Gauss IRK method from Paragraph 5.4.2.*

1. *Write the method in the full form (5.44).*
2. *What is the consistency order of this IRK method and why?*
3. *Is the method A-stable?*
4. *Show that the method is not L-stable. Hint: Apply it to a linear autonomous problem, write explicitly the stability function $R(z)$ and use Definition 5.9.*
5. *Write an algorithm that applies this method to the ODE system (5.11), (5.12) with the right-hand side (5.10) (resulting from the semidiscretization of the linear parabolic model problem by the MOL). Define carefully all systems of linear algebraic equations that are to be solved.*

Exercise 5.9 *Perform Exercise 5.8 with the Gaussian quadrature rule defined in Table 2.3 (with values transformed to the interval $(0, 1)$).*

Exercise 5.10 *With the experience gained in Exercises 5.8 and 5.9, try to write an algorithm for a general s-stage Gauss IRK method for an ODE system (5.11), (5.12) with the right-hand side (5.10).*

Exercise 5.11 *Write the stability function $R(z)$ of the fifth-order Radau IRK method from Paragraph 5.4.2. Verify the conditions for A- and L-stability.*

Exercise 5.12 *Assume a second-order elliptic operator L of the form (1.1).*

1. *Use the classification of PDEs acquired in Section 1.1 to show that the time-dependent extension $\partial/\partial t + L$ of the operator L in (5.1) indeed is parabolic.*
2. *Decide if the operator L can be extended to a time-dependent second-order elliptic operator. If yes, give an example.*
3. *Decide if a (time-independent) second-order parabolic operator can be extended to a time-dependent second-order parabolic operator. If yes, give an example.*
4. *Decide if a (time-independent) second-order hyperbolic operator can be extended to a time-dependent second-order parabolic operator. If yes, give an example.*

Exercise 5.13 *Consider the domain* $\Omega = (0, a) \times (0, b) \subset \mathbb{R}^2$*, and extend your code from Exercise 4.2 to solve the heat transfer equation*

$$\begin{aligned}
\frac{\partial u}{\partial t} - \Delta u &= \sin(\pi t)\left(\frac{3b}{2}x_2^2 - \frac{b^2}{2}x_2 - x_2^3\right)(6x_1 - 3a) \\
&\quad + \sin(\pi t)(6x_2 - 3b)\left(\frac{3a}{2}x_1^2 - \frac{a^2}{2}x_1 - x_1^3\right) \\
&\quad + \pi\cos(\pi t)x_1x_2(a - x_1)(b - x_2)\left(\frac{a}{2} - x_1\right)\left(\frac{b}{2} - x_2\right),
\end{aligned}$$

equipped with a zero initial condition

$$u(\boldsymbol{x}, 0) = 0 \quad \textit{in } \Omega,$$

and homogeneous Dirichlet boundary conditions on the boundary $\partial\Omega$*. Let* $a = 3$, $b = 2$, $M_1 = 60$, $M_2 = 40$.

1. *Implement the implicit Euler method (5.28) for the ODE system (5.11), (5.12). For time steps* $\Delta t = 0.01, 0.05, 0.1, 0.5, 5/6$ *present plots of the approximate solution* $u_{h,p}$ *at the time* $T = 5/2$*. Write the computational times.*

2. *Implement the explicit Euler method (5.24) as well as its version with the diagonal truncation (5.25) of the mass matrix* $\boldsymbol{M}$*. In the latter case, do not forget to simplify the procedure of solution of the system (5.24) accordingly. Use the criterion (5.26) to propose a suitable initial size of the time step* Δt*.*

   (a) *In both cases try to increase* Δt *until the time integration becomes unstable. What are the critical values of the constant in the relation* $\Delta t = C(\Delta h)^2$*?*

   (b) *In both cases present a plot of the approximate solution* $u_{h,p}$ *at the time* $T = 5/2$ *(± the size of one time step). Write the computational times in both cases.*

3. *Implement the adaptive RK*$5(4)$ *method given by Table 5.2.*

   (a) *Run the program for the values of* $TOL = 0.0001, 0.001, 0.01, 0.1$*. Plot the solution at* $T = 5/2$ *(again, ± the size of one time step). Write the initial value* $\Delta t_0 := (\Delta h)^2$*, and in all four cases the total number of time steps, the computational time and the size of the time step* Δt *at the end of the computation.*

   (b) *Investigate the sensitivity of Algorithm 5.1 on the initial time step. Hint: Run the computation with* $TOL = 0.001$ *and* $\Delta t := 0.01\Delta t_0$, $\Delta t := 0.1\Delta t_0$, $\Delta t := 10\Delta t_0$, $\Delta t := 100\Delta t_0$*. In all four cases write the total number of time steps, the computational time and the size of the time step* Δt *at the end of the computation.*

4. *As conclusion, compare all four methods from the point of view of accuracy, efficiency and stability.*

5. *The exact solution is*

$$u(x, t) = \sin(\pi t)x_1x_2(a - x_1)(b - x_2)\left(\frac{a}{2} - x_1\right)\left(\frac{b}{2} - x_2\right).$$

CHAPTER 6

BEAM AND PLATE BENDING PROBLEMS

In Chapters 2–4 we considered second-order PDEs whose weak formulations took place in the Sobolev space $H^1(\Omega)$. This space required the piecewise-polynomial finite element approximations to be globally continuous (i.e., continuous across element interfaces). Now we are going to study fourth-order problems with the weak formulations in $H^2(\Omega)$. Finite element approximations conforming to $H^2(\Omega)$ are required to be once continuously differentiable. Since the fourth-order PDEs are encountered in practice less frequently compared to second-order problems, we devote more attention to their physical background and derivation.

In Section 6.1 we derive the Euler–Bernoulli model for the bending of elastic beams, discuss various types of boundary conditions, derive the weak formulation of the problem, and prove the existence and uniqueness of the weak solution. In Section 6.2 we discretize the weak formulation by the lowest-order Hermite elements. Higher-order approximations with both nodal and hierarchic Hermite elements are discussed in Section 6.3. Two-dimensional Hermite elements (which do not conform to $H^2(\Omega)$ but are useful for many other applications) are presented in Section 6.4. Section 6.5 describes the Reissner–Mindlin and Kirchhoff plate bending models. The finite element discretization of the Kirchhoff thin plate model via the H^2-conforming lowest- and higher-order nodal Argyris elements is discussed in Section 6.6.

6.1 BENDING OF ELASTIC BEAMS

There are two basic one-dimensional models for the bending of elastic beams: The Euler–Bernoulli model consisting of one fourth-order PDE, and the Timoshenko model based on a pair of coupled second-order equations.

The Timoshenko model is simpler to solve in the sense that standard H^1-conforming elements can be used for its discretization, and it is known to better capture the purely three-dimensional behavior of the structure (such as large deformations). On the other hand, the higher-order elements used to discretize the Euler–Bernoulli case yield significantly better convergence rates. In this text we focus on the Euler–Bernoulli model in order to show the application of the Hermite and Argyris elements. The Timoshenko model is discussed quite frequently in monographs and textbooks (see, e.g., [95] and the reference therein).

6.1.1 Euler–Bernoulli model

This paragraph requires the knowledge of some elementary topics in continuum mechanics that can be found, e.g., in [20, 95] or [124]. The one-dimensional Hooke's law has the form

$$\sigma = E\epsilon, \tag{6.1}$$

where σ is the stress induced by the strain ϵ, and E is the modulus of elasticity of the material.

Consider a prismatic beam of a homogeneous isotropic Hookean material with a rectangular cross section, whose longitudinal axis coincides with the x-axis of the given Cartesian system of coordinates. The position of the centroids of the end-faces is fixed, and a pair of bending moments acting on the ends of the beams is illustrated in Figure 6.1. (The downward-pointing z-axis is used to make the signs of both the force f and deflection u relative to the direction of gravity.)

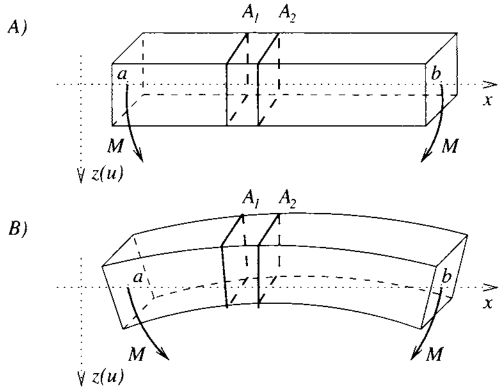


Figure 6.1 Bending of a prismatic beam; (A) initial configuration, (B) deformed state under moment M acting on the ends.

The basic assumption of the simple beam theory is that the (normal) deflection is relatively small compared to the length of the beam, so that every pair of adjacent cross-sections A_1 and A_2, which are perpendicular to the axis of the beam in the original configuration, remain planar and perpendicular to the beam axis during the deformation.

The deflection of the beam can be described as the vertical displacement of the centroidal surface that corresponds to $u = 0$ in the original configuration. In the situation shown in Figure 6.1 the deflection curve must be a circular arc, since because of the homogeneity of the material every cross section is subjected to the same stress and strain. By R denote the radius of the deformed beam axis. A small portion of the axis of the beam before and after deformation is shown in Figure 6.2.

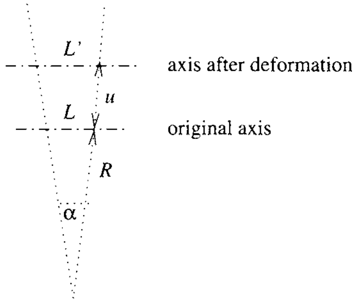


Figure 6.2 Strain induced by the deflection of the beam. Here L and L' stand for the distance of the midpoints of the cross-sections A_1 and A_2 before and after deformation.

For a large radius R and small angle α we can write

$$L = R \sin\alpha, \quad L' = (R + u)\sin\alpha,$$

and thus the axial strain ϵ, which is defined as the ratio of the length increment and the original length, has the form

$$\epsilon = \frac{L' - L}{L} = \frac{u}{R}.$$

As a response to the strain ϵ, there is a stress σ, which according to the above assumptions is one-dimensional in the direction of the x-axis. Hooke's law (6.1) yields

$$\sigma = E\epsilon = E\frac{u}{R}. \tag{6.2}$$

It follows from here that the centroidal plane $u = 0$ remains unstressed during the bending, i.e., that material particles on it are not strained in the axial direction. The plane is therefore called the neutral surface of the beam.

The resultant moment of the bending stress σ on every beam cross-section A must be equal to the external moment M,

$$M = \int_A u\sigma \mathrm{d}A.$$

Substituting (6.2) into this equation, we obtain

$$M = \frac{E}{R}\int_{A'} u^2 \, \mathrm{d}A = \frac{EI}{R}, \tag{6.3}$$

where the area moment of inertia I of the beam is defined as

$$I = \int_A y^2 \sigma \, \mathrm{d}A. \tag{6.4}$$

Let us point out that I is a geometrical property of the beam, while E is a material property.

By f denote the transversal load and by F_s the corresponding shear force (which is perpendicular to the beam axis). The curvature of a circular arc is given by

$$\frac{\mathrm{d}^2 u}{\mathrm{d}x^2} = \frac{1}{R}. \tag{6.5}$$

The standard relations

$$f = \frac{\mathrm{d}F_s}{\mathrm{d}x} \quad \text{and} \quad F_s = \frac{\mathrm{d}M}{\mathrm{d}x}$$

yield

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2} M = f. \tag{6.6}$$

Substituting (6.3) with (6.5) into (6.6) and denoting $b(x) = E(x)I(x)$, we obtain the Euler–Bernoulli beam model

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(b(x)\frac{\mathrm{d}^2 u}{\mathrm{d}x^2}(x)\right) = f(x) \qquad \text{for all } x \in \Omega, \tag{6.7}$$

where $\Omega = (a, b)$ is an open bounded one-dimensional interval representing the beam. Equation (6.7) requires b to be twice-continuously differentiable, u four times continuously differentiable, and f continuous in Ω. These quite strong regularity requirements will be reduced after the problem is formulated in the weak sense in Pararaph 6.1.3.

6.1.2 Boundary conditions

Equation (6.7) is a fourth-order problem, and therefore four suitable boundary conditions are needed to guarantee the existence and uniqueness of solution (this will be discussed in more detail in Paragraph 6.1.4). Analogously to second-order problems, the boundary conditions can be split into essential and natural, depending on whether or not they influence the form of the space V in the weak formulation. Most frequently one prescribes the following quantities:

Essential boundary conditions:

- deflection

$$u_a = u(a) \quad \text{and/or} \quad u_b = u(b),$$

- slope

$$du_a = u'(a) \quad \text{and/or} \quad du_b = u'(b).$$

Natural boundary conditions:

- moment

$$M_a = \left(EI \frac{\mathrm{d}^2 u}{\mathrm{d}x^2} \right)(a) \quad \text{and/or} \quad M_b = \left(EI \frac{\mathrm{d}^2 u}{\mathrm{d}x^2} \right)(b), \tag{6.8}$$

- shear force

$$F_a = \left[\frac{\mathrm{d}}{\mathrm{d}x} \left(EI \frac{\mathrm{d}^2 u}{\mathrm{d}x^2} \right) \right](a) \quad \text{and/or} \quad F_b = \left[\frac{\mathrm{d}}{\mathrm{d}x} \left(EI \frac{\mathrm{d}^2 u}{\mathrm{d}x^2} \right) \right](b). \tag{6.9}$$

Some combinations that lead to a unique solution are shown in Figures 6.3–6.5. The transversal force is indicated by the arrows.

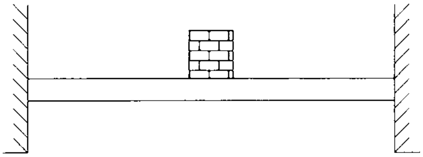


Figure 6.3 Clamped beam: Prescribed is the deflection and slope at both ends.

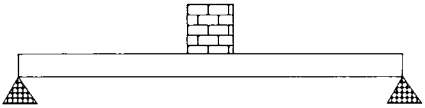


Figure 6.4 Simply supported beam: Prescribed is the deflection and moment at both ends.

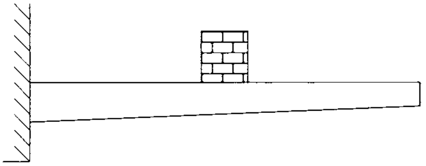


Figure 6.5 Cantilever beam: Prescribed is the deflection and slope at one end, and moment and shear force at the other end.

6.1.3 Weak formulation

Let us formulate equation (6.7) in the weak sense, employing the symbols

$$\nabla u = \frac{\mathrm{d}u}{\mathrm{d}x} \quad \text{and} \quad \Delta u = \frac{\mathrm{d}^2 u}{\mathrm{d}x^2}$$

for brevity. The standard first step consists of multiplying the equation with a sufficiently regular test function v, and integrating over the domain Ω,

$$\int_\Omega \Delta(b\Delta u)v \, \mathrm{d}x = \int_\Omega fv \, \mathrm{d}x.$$

Green's theorem applied to the term with the highest derivatives yields

$$-\int_\Omega \nabla(b\Delta u) \cdot \nabla v \, \mathrm{d}x + [\nabla(b\Delta u)v]_a^b = \int_\Omega fv \, \mathrm{d}x,$$

where $|g|_a^b = g(b) - g(a)$. Using Green's theorem once more, we obtain

$$\int_\Omega b\Delta u \Delta v \, \mathrm{d}x - [b\Delta u \nabla v]_a^b + [\nabla(b\Delta u)v]_a^b = \int_\Omega fv \, \mathrm{d}x. \tag{6.10}$$

The integrals in (6.10) exist if $b \in L^\infty(\Omega)$, $u, v \in H^2(\Omega)$ and $f \in L^2(\Omega)$ (in fact f can be chosen from a larger space $H^{-2}(\Omega)$, which is the dual to $H^2(\Omega)$).

Now essential boundary conditions are implemented by further constraining the space $H^2(\Omega)$. For example, the choice of the clamped boundary conditions from Figure 6.3,

$$u(a) = u(b) = \nabla u(a) = \nabla u(b) = 0, \tag{6.11}$$

leads to $u, v \in V$, where the space $V \subset H^2(\Omega)$ is defined by

$$V = H_0^2(\Omega) = \{v \in H^2(\Omega);\ v(a) = v(b) = \nabla v(a) = \nabla v(b) = 0\}. \tag{6.12}$$

Since both u and v vanish at the endpoints together with their first derivatives, also the square brackets in (6.10) disappear and one obtains

$$\int_\Omega b\Delta u \Delta v \, \mathrm{d}x = \int_\Omega fv \, \mathrm{d}x. \tag{6.13}$$

Finally, if some terms in the square brackets are present after the essential boundary conditions were applied (as it might be the case, e.g., with the cantilever beam from Figure 6.5), natural boundary conditions are incorporated by properly substituting into these terms from (6.8) and/or (6.9).

6.1.4 Existence and uniqueness of solution

Let us show the existence and uniqueness of the weak solution first for the case of the clamped boundary conditions (6.11), i.e., for the following weak formulation:

For given $b \in L^\infty(\Omega)$ and $f \in L^2(\Omega)$ find $u \in V$ such that

$$a(u,v) = l(v) \quad \text{for all } v \in V, \tag{6.14}$$

where the linear forms $a : V \times V \to \mathbb{R}$ and $l \in V'$ are given by

$$a(u,v) = \int_\Omega b\Delta u \Delta v \,\mathrm{d}x,$$
$$l(v) = \int_\Omega fv \,\mathrm{d}x.$$

In order to obtain a unique solution, we have to add the assumption of strict positivity of $b(x)$:

$$0 < C_{EI} \leq b(x) \quad \text{for all } x \in \Omega. \tag{6.15}$$

This requirement is intuitively clear, and it holds unless the elasticity modulus E or the area moment of inertia I vanish in Ω. Condition (6.15) will play a role in the proof of the following lemma:

Lemma 6.1 *Under the assumption (6.15), the problem (6.14) has a unique solution $u \in V$.*

Proof: We need to verify that the assumptions of the Lax–Milgram lemma (Theorem 1.5) are satisfied. Let us begin with the case $0 < b(x) = \text{const}$, where the main idea is free of technical details. Suppose that b is removed from the equation by redefining $f := f/b$. We use Hölder's inequality (Theorem A.10) to see that the form $a(\cdot,\cdot)$ is bounded,

$$\begin{aligned}
|a(u,v)| &= \left| \int_\Omega \Delta u \Delta v \,\mathrm{d}x \right| \leq \left(\int_\Omega (\Delta u)^2 \,\mathrm{d}x \right)^{\frac{1}{2}} \left(\int_\Omega (\Delta v)^2 \,\mathrm{d}x \right)^{\frac{1}{2}} \\
&\leq \left(\int_\Omega u^2 + |\nabla u|^2 + (\Delta u)^2 \,\mathrm{d}x \right)^{\frac{1}{2}} \left(\int_\Omega v^2 + |\nabla v|^2 + (\Delta v)^2 \,\mathrm{d}x \right)^{\frac{1}{2}} \\
&= \|u\|_V^2 \|v\|_V^2 \quad \text{for all } u, v \in V,
\end{aligned}$$

where $\|\cdot\|_V$ is the H^1-norm $\|\cdot\|_{1,2}$ (see Definition A.57). Since

$$v \in H_0^1(\Omega) \quad \text{and} \quad \nabla v \in H_0^1(\Omega),$$

the Poincaré–Friedrichs inequality (Theorem A.26) with $k = 1$ can be applied to both v and ∇v. Hence, there exist positive constants C_0, C_1 such that

$$\int_\Omega v^2 + |\nabla v|^2 \,\mathrm{d}x \leq C_0 \int_\Omega |\nabla v|^2 \,\mathrm{d}x \quad \text{for all } v \in V, \tag{6.16}$$

and

$$\int_\Omega |\nabla v|^2 + (\Delta v)^2 \,\mathrm{d}x \leq C_1 \int_\Omega (\Delta v)^2 \,\mathrm{d}x \quad \text{for all } v \in V. \tag{6.17}$$

From (6.17) it follows that

$$a(v,v) = \int_\Omega (\Delta v)^2 \,\mathrm{d}x \geq \frac{1}{C_1} \int_\Omega |\nabla v|^2 + (\Delta v)^2 \,\mathrm{d}x \quad \text{for all } v \in V.$$

Relation (6.16) yields

$$\frac{1}{C_1}\int_\Omega |\nabla v|^2 + (\Delta v)^2 \,\mathrm{d}x \geq \frac{1}{C_1 C_0}\int_\Omega v^2 + |\nabla v|^2 \,\mathrm{d}x + \frac{1}{C_1}\int_\Omega (\Delta v)^2 \,\mathrm{d}x \qquad \text{for all } v \in V.$$

Finally we obtain

$$a(v,v) \geq \frac{1}{C_1 \max(C_0, 1)}\|v\|_V^2 \quad \text{for all } v \in V.$$

Thus with $0 < b(x) = \text{const}$, the form $a(\cdot,\cdot)$ is continuous and V-elliptic, and according to the Lax–Milgram lemma problem (6.14) has a unique solution $u \in V$. In the case of nonconstant strictly positive $b \in L^\infty(\Omega)$, the idea of the proof is the same except for a slightly more technical manipulation with the inequalities, which is left to the reader as an exercise. ■

It should be stressed that, analogously to the second-order elliptic case, some combinations of the boundary conditions are prohibited since they do not lead to a unique solution. This would be the case, e.g., if the deflection u was not prescribed at either end. The form $a(\cdot,\cdot)$ remains V-elliptic when the beam is only clamped at one end, since the Poincaré–Friedrichs' inequality (Theorem A.26) still holds (see Remark A.8).

6.2 LOWEST-ORDER HERMITE ELEMENTS IN 1D

For the first exposition of the Hermite elements let us consider problem (6.7) with the clamped boundary conditions (6.11) from Paragraph 6.1.4. The weak formulation for this case was derived in (6.14).

6.2.1 Model problem

Consider a subdivision $a = x_0 < x_1 < \ldots < x_M = b$ of the domain Ω. For $i = 1, 2, \ldots, M$ denote $K_i = (x_{i-1}, x_i)$. Recall from Paragraph A.4.2 that H^1-functions are continuous in one spatial dimension,

$$w \in H^1(\Omega) \quad \Rightarrow \quad w \in C(\Omega). \tag{6.18}$$

Applying (6.18) to the derivative of $w = v' \in H^1(\Omega)$, one obtains

$$v \in H^2(\Omega) \quad \Rightarrow \quad v \in C^1(\Omega). \tag{6.19}$$

Any approximate solution to problem (6.14) has to be once continuously differentiable (globally smooth) in Ω. It follows from here that the approximation has to be at least piecewise quadratic. However, the space of smooth, piecewise-quadratic functions is not frequently used for reasons to be explained in Paragraph 6.2.2. It is standard to employ cubic and higher-degree polynomials.

With a general polynomial distribution $3 \leq p_i = p_i(K_i)$, $i = 1, 2, \ldots, M$, the space $V_{h,p}$ has the form

$$\begin{aligned} V_{h,p} &= \{v \in C^1(\Omega);\ v(a) = v(b) = v'(a) = v'(b) = 0; \\ &\quad v|_{K_i} \in P^{p_i}(K_i)\} \subset V, \end{aligned} \tag{6.20}$$

and the approximate weak formulation reads:

Approximate weak formulation Find a function $u_{h,p} \in V_{h,p}$ such that

$$\int_\Omega b\Delta u_{h,p}\Delta v \,\mathrm{d}x = \int_\Omega fv \,\mathrm{d}x \quad \text{for all } v \in V_{h,p}. \tag{6.21}$$

Since the bilinear form $a(\cdot,\cdot)$ is continuous and V-elliptic on $V_{h,p}$, the Lax–Milgram lemma implies that the discrete problem (6.21) has a unique solution $u_{h,p} \in V_{h,p}$.

Basis of $V_{h,p}$ and the linear algebraic system Assume a basis $\{v_1, v_2, \ldots, v_N\}$ of the space $V_{h,p}$. Express

$$u_{h,p} = \sum_{j=1}^{N} y_j v_j, \tag{6.22}$$

where $y_1, y_2, \ldots, y_N$ are unknown coefficients in the usual sense. Using (6.22) and employing $v := v_1, v_2, \ldots, v_N$, identity (6.21) comes over to a system of linear algebraic equations of the form

$$\sum_{j=1}^{N} y_j \int_\Omega b\Delta v_j \Delta v_i \,\mathrm{d}x = \int_\Omega f v_i \,\mathrm{d}x, \quad i = 1, 2, \ldots, N, \tag{6.23}$$

which can be written in the compact form

$$\boldsymbol{S}\boldsymbol{Y} = \boldsymbol{F}. \tag{6.24}$$

We saw in the proof of Lemma 6.1 that problem (6.14) is V-elliptic. Therefore the bilinear form $a(\cdot,\cdot)$ defines an energetic inner product on $V \times V$, and the standard orthogonality property of the type (2.14) holds,

$$a(u - u_{h,p}, v) = 0 \quad \text{for all } v \in V_{h,p}. \tag{6.25}$$

This in turn means that the approximate solution $u_{h,p}$ is independent of the choice of the basis $\{v_1, v_2, \ldots, v_N\}$ of the space $V_{h,p}$ (see Remark 2.2).

However, as we know from Paragraph 2.5.3, the choice of the basis in $V_{h,p}$ influences the condition number of the stiffness matrix $\boldsymbol{S}$, and in turn the performance of the iterative matrix solvers for the linear system $\boldsymbol{S}\boldsymbol{Y} = \boldsymbol{F}$ dramatically. This is why one has to design the basis functions $v_1, v_2, \ldots, v_N$ very carefully. Before introducing general higher-order Hermite elements in Section 6.3, let us review the standard cubic case in Paragraph 6.2.2.

6.2.2 Cubic Hermite elements

As mentioned earlier, the smallest space $V_{h,p}$ consisting of piecewise-quadratic polynomials is not used very frequently in practice. The reason is that the support of a smooth, piecewise quadratic basis function has to extend over at least three elements. But more importantly, it is not possible to find a set of degrees of freedom Σ_i such that $(K_i, P^2(K_i), \Sigma_i)$ would constitute a unisolvent nodal finite element conforming to the space $H^2(\Omega)$.

The lowest-order element $K_i = (x_{i-1}, x_i)$ conforming to $H^2(\Omega)$ is the Hermite element with degrees of freedom associated with both the function values $u(x_{i-1}), u(x_i)$ and derivatives $u'(x_{i-1}), u'(x_i)$ at the endpoints. This makes it four degrees of freedom per element, i.e., the local polynomial space on the interval K_i has to be $P^3(K_i)$.

Cubic Hermite element on the reference domain K_a As always, let us first define the element on a reference domain, which in this case is the interval $K_a = (-1, 1)$. The cubic Hermite element is a triad $(K_a, P^3(K_a), \Sigma_a)$, where the set of degrees of freedom Σ_a consists of the linear forms $L_i : P^3(K_a) \to \mathbb{R}$,

$$\begin{aligned} L_1(g) &= g(-1), \\ L_2(g) &= g(1), \\ L_3(g) &= g'(-1), \\ L_4(g) &= g'(1). \end{aligned} \tag{6.26}$$

Let us check the unisolvency of this finite element and construct the unique nodal basis of the space $P^3(K_a)$. We choose an arbitrary basis of the space $P^3(K_a)$, say,

$$\{g_1, g_2, g_3, g_4\} = \{1, \xi, \xi^2, \xi^3\}.$$

The generalized Vandermonde matrix $\boldsymbol{L} = \{L_i(g_j)\}_{i,j=1}^4$ has the form

$$\boldsymbol{L} = \begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & -2 & 3 \\ 0 & 1 & 2 & 3 \end{pmatrix}$$

(see Theorem 3.1). Since $\boldsymbol{L}$ is nonsingular, the element is unisolvent. The inverse matrix

$$\boldsymbol{L}^{-1} = \begin{pmatrix} 1/2 & 1/2 & 1/4 & -1/4 \\ -3/4 & 3/4 & -1/4 & -1/4 \\ 0 & 0 & -1/4 & 1/4 \\ 1/4 & -1/4 & 1/4 & 1/4 \end{pmatrix}$$

contains in its columns the coefficients defining the nodal basis $\mathcal{B} = \{\theta_1, \theta_2, \theta_3, \theta_4\}$ in terms of the original basis functions $\{1, \xi, \xi^2, \xi^3\}$,

$$\begin{aligned} \theta_1(\xi) &= \omega_0(\xi) = \frac{1}{2} - \frac{3}{4}\xi + \frac{1}{4}\xi^3, \\ \theta_2(\xi) &= \omega_1(\xi) = \frac{1}{2} + \frac{3}{4}\xi - \frac{1}{4}\xi^3, \\ \theta_3(\xi) &= \omega_2(\xi) = \frac{1}{4} - \frac{1}{4}\xi - \frac{1}{4}\xi^2 + \frac{1}{4}\xi^3, \\ \theta_4(\xi) &= \omega_3(\xi) = -\frac{1}{4} - \frac{1}{4}\xi + \frac{1}{4}\xi^2 + \frac{1}{4}\xi^3 \end{aligned} \tag{6.27}$$

(the symbols $\omega_0, \omega_1, \ldots, \omega_3$ are introduced for later reference). It is easy to verify that these functions satisfy the delta property (3.4) in the form

$$L_j(\theta_k) = \delta_{jk} \qquad \text{for all } 1 \le j, k \le 4. \tag{6.28}$$

These four nodal shape functions are depicted in Figures 6.6 and 6.7.

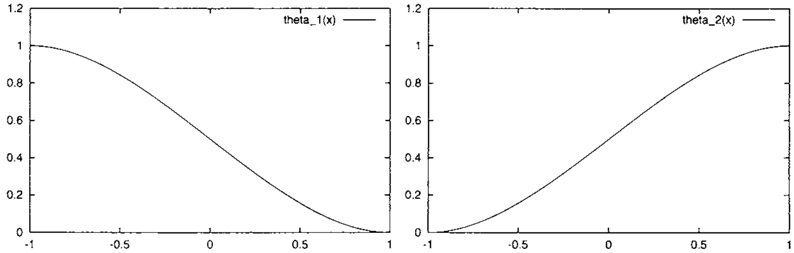


Figure 6.6 Cubic shape functions θ_1 and θ_2, representing function values at the endpoints of K_a.

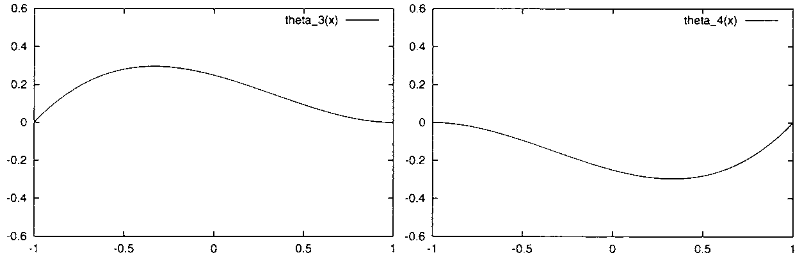


Figure 6.7 Cubic shape functions θ_3 and θ_4, representing derivatives at the endpoints of K_a.

The notion of vertex and bubble shape functions for the Hermite elements differs from what we had for the Lagrange elements:

Definition 6.1 *Given a one-dimensional Hermite element (K, P, Σ), a shape function $\theta \subset P$ is said to be* bubble function *if it vanishes, together with its first derivative θ', at both endpoints of the interval K. Shape functions which are not bubble functions are said to be* vertex functions.

All shape functions depicted in Figures 6.6 and 6.7 are vertex functions.

Cubic Hermite element on a general interval $K_i \subset \mathbb{R}$ The cubic Hermite element in the interval $K_i = (x_{i-1}, x_i)$ is defined as $(K_i, P^3(K_i), \Sigma_i)$, where the set of degrees of freedom $\Sigma_i = \{L_1^{(i)}, L_2^{(i)}, \ldots, L_4^{(i)}\}$ comprises the linear forms $L_1^{(i)}(g) = g(x_{i-1})$, $L_2^{(i)}(g) = g(x_i)$, $L_3^{(i)}(g) = g'(x_{i-1})$, $L_4^{(i)}(g) = g'(x_i)$ defined on $P^3(K_i)$.

In Example 3.5 we saw that finite elements with derivatives are not affine-equivalent, and this applies to the Hermite elements as well. Hence it is natural to split the nodal shape functions into two groups related to Lagrange and Hermite degrees of freedom, respectively. (The Lagrange DOF are associated with the function values and the Hermite with the derivatives.) The Lagrange shape functions θ_1, θ_2 are affine equivalent in the standard way, and the affine reference map $x_{K_i} : K_a \to K_i$ from (2.37) can be used to define the nodal shape functions $\theta_1^{(i)}$ and $\theta_2^{(i)}$ on K_i,

$$\begin{aligned} \theta_1^{(i)} &= \omega_0^{(i)} = \theta_1 \circ x_{K_i}^{-1}, \\ \theta_2^{(i)} &= \omega_1^{(i)} = \theta_2 \circ x_{K_i}^{-1}. \end{aligned} \tag{6.29}$$

The Hermite shape functions $\theta_3^{(i)}$ and $\theta_4^{(i)}$ on K_i have the form

$$\begin{aligned} \theta_3^{(i)} &= \omega_2^{(i)} = J_{K_i}\theta_3 \circ x_{K_i}^{-1}, \\ \theta_4^{(i)} &= \omega_3^{(i)} = J_{K_i}\theta_4 \circ x_{K_i}^{-1}, \end{aligned} \tag{6.30}$$

where the additional multiplication with the constant Jacobian J_{K_i} is needed in order to preserve the values of the derivatives at the endpoints,

6.3 HIGHER-ORDER HERMITE ELEMENTS IN 1D

This section is devoted to the design and properties of higher-order Hermite elements. The cubic Hermite elements are extended to arbitrarily high polynomial degrees in both the nodal and hierarchic fashion in Paragraphs 6.3.1 and 6.3.2. The conditioning properties of the nodal and hierarchic higher-order shape functions are compared in Paragraph 6.3.3. The basis of the space $V_{h,p}$ is constructed in Paragraph 6.3.4 and the integrals from the weak formulation (6.21) are transformed to the reference domain K_a in Paragraph 6.3.5. Algorithmic aspects, such as the construction of the connectivity arrays and the assembling algorithm, are discussed in Paragraphs 6.3.6 and 6.3.7. The three basic ways of interpolation (i.e., the best, projection-based and nodal interpolants) are discussed in the context of higher-order Hermite elements in Paragraph 6.3.8.

6.3.1 Nodal higher-order elements

The cubic Hermite element on the reference domain $K_a = (-1, 1)$ can be extended to a nodal Hermite element of the order $p > 3$ by adding $p - 3$ new degrees of freedom. Since the two Lagrange and two Hermite degrees of freedom at the interval endpoints already guarantee the conformity to the space $H^2(\Omega)$ (this will be discussed in more detail later), it is natural to add Lagrange degrees of freedom. Hence, choose some $p - 3$ additional nodal points $y_2, y_3, \ldots, y_{p-2}$ in the interval K_a, so that

$$-1 = y_1 < y_2 < \ldots < y_{p-2} = 1$$

The Lagrange shape function associated with the nodal point y_k, $2 \le k \le p-1$, vanishes at both endpoints of K_a together with its first derivative, and therefore it is a bubble function according to Definition 6.1.

Fourth-order Hermite element on the reference domain K_a The properties of the nodal Hermite elements strongly depend on the choice of the nodal points $y_2, y_3, \ldots, y_{p-2}$. The situation is simple in the fourth-order case, where the choice $y_2 = 0$ is dictated by the symmetry requirement. Thus we obtain an element $(K_a, P^4(K_a), \Sigma_a)$, where the set of degrees of freedom Σ_a comprises the linear forms $L_i : P^4(K_a) \to \mathbb{R}$,

$$\begin{aligned} L_1(g) &= g(-1), \\ L_2(g) &= g(0), \\ L_3(g) &= g(1), \\ L_4(g) &= g'(-1), \\ L_5(g) &= g'(1). \end{aligned} \tag{6.31}$$

The pair of Hermite degrees of freedom is placed at the end of the list for algorithmic reasons. Following the standard procedure, we obtain a unique set of nodal shape functions in the form

$$\begin{aligned} \theta_1(\xi) &= -\frac{3}{4}\xi + \xi^2 + \frac{1}{4}\xi^3 - \frac{1}{2}\xi^4, \\ \theta_2(\xi) &= 1 - 2\xi^2 + \xi^4, \\ \theta_3(\xi) &= \frac{3}{4}\xi + \xi^2 - \frac{1}{4}\xi^3 - \frac{1}{2}\xi^4, \\ \theta_4(\xi) &= -\frac{1}{4}\xi + \frac{1}{4}\xi^2 + \frac{1}{4}\xi^3 - \frac{1}{4}\xi^4, \\ \theta_5(\xi) &= -\frac{1}{4}\xi - \frac{1}{4}\xi^2 + \frac{1}{4}\xi^3 + \frac{1}{4}\xi^4. \end{aligned} \tag{6.32}$$

These functions are depicted in Figures 6.8–6.10.

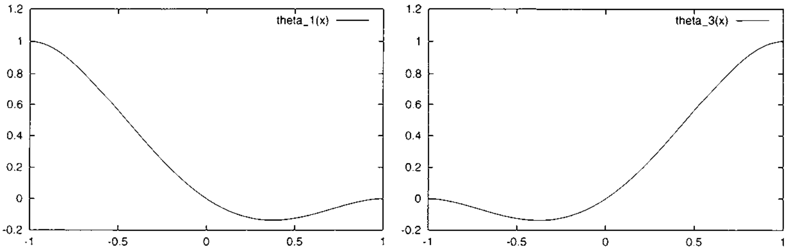


Figure 6.8 Fourth-order vertex functions θ_1 and θ_3 representing function values at the endpoints.

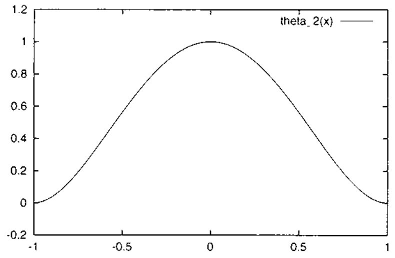


Figure 6.9 Fourth-order bubble function θ_2 representing the function value at the midpoint.

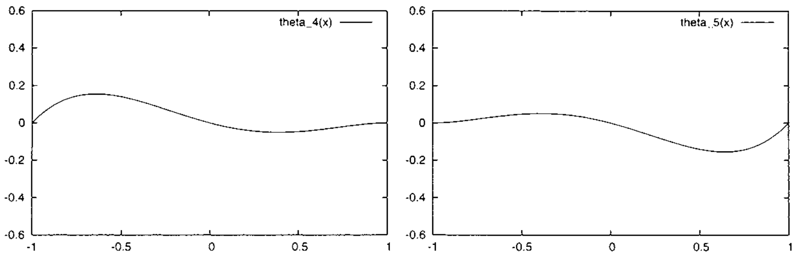


Figure 6.10 Fourth-order vertex functions θ_4 and θ_5 representing derivatives at the endpoints.

The fourth-order Hermite element on a general interval K_i as well as its nodal basis are constructed analogously to the cubic case: The Lagrange shape functions θ_1, θ_2, and θ_3 are transformed to K_i via the inverse reference map $x_{K_i}^{-1}$ analogously to (6.29), and the correction by the Jacobian J_{K_i} is applied to the Hermite shape functions θ_4 and θ_5 similarly to (6.30). The inverse of the reference map $x_{K_i}^{-1}$ is not required by the element-by-element assembling procedure.

For higher polynomial degrees $p \geq 5$ the $p-1$ Lagrange nodal points can be identified, for example, with the $p-1$ Gauss–Lobatto points $-1 = y_1 < y_2 < y_3 < \ldots < y_{p-1} = 1$ of the order $p-2$. Later in Paragraph 6.3.3 we show that this choice is advantageous from the point of view of the condition number of the resulting stiffness matrix. Next let us turn our attention to the hierarchic Hermite elements.

6.3.2 Hierarchic higher-order elements

The basic idea of hierarchic elements, explained at the beginning of Paragraph 2.4.6, applies to Hermite elements as well. The lowest-order basis $\mathcal{B}_3$ comprises the four cubic Hermite vertex functions (6.27),

$$\mathcal{B}_3 = \{\omega_0, \omega_1, \ldots, \omega_3\}.$$

For every $p \geq 3$, the basis $\mathcal{B}_{p+1}$ is defined by

$$\mathcal{B}_{p+1} = \mathcal{B}_p \cup \{\omega_{p+1}\}, \tag{6.33}$$

where the polynomial ω_{p+1} of the degree $p+1$ is a suitable bubble function (i.e., $\omega_{p+1} \in P^p(K_a), \omega_{p+1}(\pm 1) = \omega'_{p+1}(\pm 1) = 0$). Since the choice of a hierarchic basis is not unique, one has certain freedom to optimize the conditioning properties of the higher-order shape functions.

Recall that the excellent conditioning properties of the Lobatto hierarchic shape functions (2.63) were due to the orthonormality of the corresponding higher-order bubble functions in the H_0^1-product,

$$(u, v)_{H_0^1(K_a)} = \int_{-1}^{1} \nabla u \nabla v \, \mathrm{d}\xi \qquad \text{for all } u, v \in H_0^1(K_a). \tag{6.34}$$

In Paragraph 2.5.3 this orthonormality was achieved by using the integrated Legendre polynomials. Analogously, the weak formulation (6.14) involves the H_0^2-product,

$$(u,v)_{H_0^2(K_a)} = \int_{-1}^{1} \Delta u \Delta v \, \mathrm{d}\xi \qquad \text{for all } u,v \in H_0^2(K_a). \tag{6.35}$$

Higher-order bubble functions orthonormal in this inner product will possess optimality analogous to the Lobatto shape functions (2.63). Hence, after integrating the Lobatto bubble functions $l_2, l_3, \ldots$,

$$\bar{l}_k(\xi) = \int_{-1}^{\xi} l_{k-2}(\zeta)\, \mathrm{d}\zeta, \quad 5 \le k,$$

we see that

$$\begin{aligned} \bar{l}_k(-1) &= \int_{-1}^{-1} l_{k-2}(\zeta)\, \mathrm{d}\zeta = 0, \\ \bar{l}'_k(-1) &= l_{k-2}(-1) = 0, \\ \bar{l}'_k(1) &= l_{k-2}(1) = 0, \end{aligned}$$

which is exactly what we want. However, at the same time we see that

$$\bar{l}_k(1) = \int_{-1}^{1} l_{k-2}(\zeta)\, \mathrm{d}\zeta = 0$$

only holds for all odd $k \ge 5$. In this case we can define the bubble functions directly,

$$\omega_k(\xi) = \bar{l}_{k-2}(\xi) \qquad \text{for all} k \ge 5,$$

but extra work needs to be done if k is an even number. For all $k > 5$ even, the explicit formulae of the bubble functions $\omega_k(\xi)$ are obtained from the orthogonality and (anti)symmetry requirements. This leads to a nonlinear system of algebraic equations, that can be solved with some effort (see [110] for details).

The formulae of $\omega_4, \omega_5, \ldots, \omega_{10}$ are shown below for reference.

$$\begin{aligned} \omega_4(\xi) &= \sqrt{\frac{5}{128}} \left(1-\xi^2\right)^2, \\ \omega_5(\xi) &= \sqrt{\frac{7}{128}} \left(1-\xi^2\right)^2 \xi, \\ \omega_6(\xi) &= \frac{1}{6}\sqrt{\frac{9}{128}} \left(1-\xi^2\right)^2 \left(-7\xi^2+1\right), \\ \omega_7(\xi) &= \frac{1}{2}\sqrt{\frac{11}{128}} \left(1-\xi^2\right)^2 \left(3\xi^2-1\right)\xi, \\ \omega_8(\xi) &= \frac{1}{16}\sqrt{\frac{13}{128}} \left(1-\xi^2\right)^2 \left(33\xi^4-18\xi^2+1\right), \\ \omega_9(\xi) &= \frac{1}{48}\sqrt{\frac{15}{128}} \left(1-\xi^2\right)^2 \left(143\xi^4-110\xi^2+15\right)\xi, \\ \omega_{10}(\xi) &= \frac{1}{32}\sqrt{\frac{17}{128}} \left(1-\xi^2\right)^2 \left(143\xi^6-143\xi^4+33\xi^2-1\right). \end{aligned} \tag{6.36}$$

Bubble functions constructed in this way are orthonormal in the inner product (6.35) not only among themselves, but also to the four vertex functions $\omega_0, \omega_1, \ldots, \omega_3$. Therefore the master element stiffness matrix $\boldsymbol{S}_{K_a}$ for the biharmonic operator in the reference interval K_a has the form

$$\boldsymbol{S}_{K_a} = \left(\begin{array}{ccccccc} 3/2 & -3/2 & 3/2 & 3/2 & 0 & \ldots & 0 \\ -3/2 & 3/2 & -3/2 & -3/2 & 0 & \ldots & 0 \\ 3/2 & -3/2 & 2 & 1 & 0 & \ldots & 0 \\ 3/2 & -3/2 & 1 & 2 & 0 & \ldots & 0 \\ 0 & 0 & 0 & 0 & 1 & & 0 \\ \vdots & \vdots & \vdots & \vdots & & & \vdots \\ 0 & 0 & 0 & 0 & 0 & \ldots & 1 \end{array} \right). \tag{6.37}$$

The bubble functions $\omega_4, \omega_5, \ldots, \omega_{11}$ are shown in Figures 6.11–6.14.

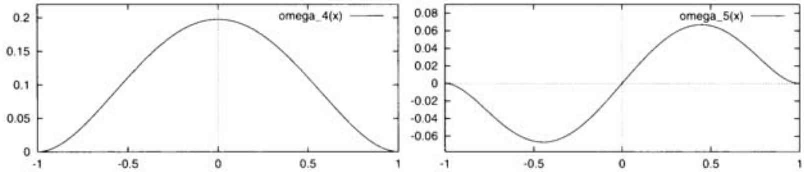


Figure 6.11 H_0^2-orthonormal hierarchic shape functions ω_4, ω_5.

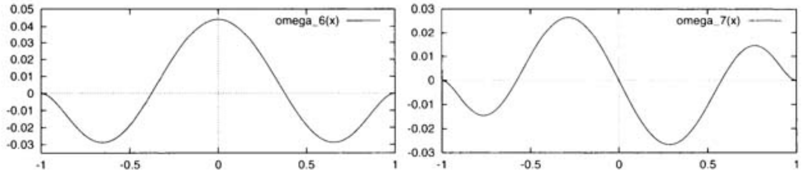


Figure 6.12 H_0^2-orthonormal hierarchic shape functions ω_6, ω_7.

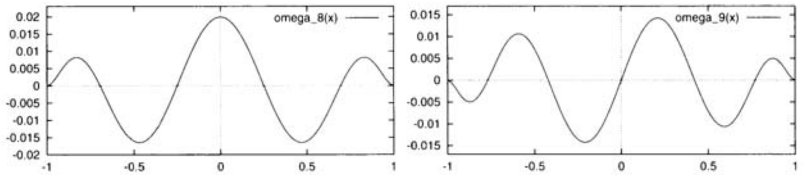


Figure 6.13 H_0^2-orthonormal hierarchic shape functions ω_8, ω_9.

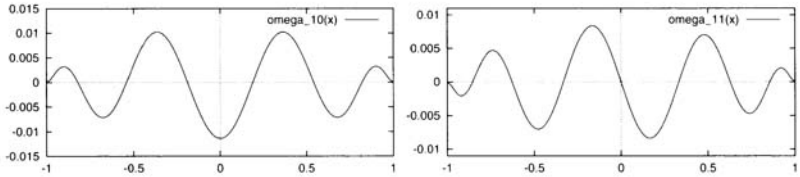


Figure 6.14 H_0^2-orthonormal hierarchic shape functions ω_{10}, ω_{11}.

6.3.3 Conditioning of shape functions

The conditioning properties of higher-order shape functions are essential for the performance of the iterative matrix solvers on the discrete problem. Therefore let us follow up with the discussion from Paragraph 2.5.3 and study the conditioning properties of the nodal and hierarchic shape functions for higher-order Hermite elements.

Conditioning in the H_0^2-product For simplicity let us consider the biharmonic problem (6.7) with $b \equiv 1$ and the boundary conditions (6.11) on the reference domain K_a. A one-element mesh $\mathcal{T}_{h,p} = \{K_a\}$ will be used for its discretization. The stiffness matrix S_0 is obtained by leaving out from the master element stiffness matrix the four rows and four columns corresponding to the vertex functions. Thus in the hierarchic case the master element stiffness matrix (6.37) reduces to the $(p-3) \times (p-3)$ identity matrix.

Figure 6.15 compares the condition numbers of the stiffness matrix S_0 obtained using four different sets of higher-order shape functions: the nodal shape functions defined on equidistant, Chebyshev and Gauss–Lobatto points, and the hierarchic shape functions. The horizontal axis represents the polynomial degree of the element.

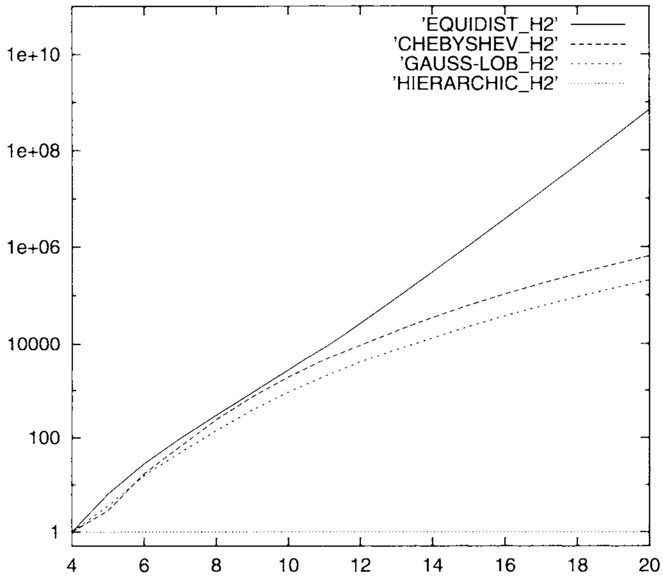


Figure 6.15 Conditioning of various sets of bubble functions in the H_0^2-product on the reference domain K_a. The horizontal axis represents the polynomial degree p of the element.

While the nodal shape functions on the equidistant points are uniformly worst and the hierarchic shape functions optimal (both as expected), it is interesting to see that the Gauss–Lobatto points are a better choice than the Chebyshev points for $p \geq 5$. These two point sets performed similarly in the discretization of the Laplace operator.

Conditioning in the H_0^1-product Despite the Laplace operator is not present in the Euler–Bernoulli beam model explicitly, it may be involved in more general fourth-order problems. This is the case, for example, with the equation $\Delta(b\Delta u) - \nabla(c\nabla u) = f$. This equation, when equipped with the boundary conditions (6.11), has the weak form

$$\int_\Omega b\Delta u \Delta v \,\mathrm{d}x + \int_\Omega c\nabla u \nabla v \,\mathrm{d}x = \int_\Omega f v \,\mathrm{d}x.$$

Hence, in this case the condition number of the resulting stiffness matrix also depends on the conditioning of the shape functions in the H_0^1-product (6.34). For reference, the corresponding comparison is shown in Figure 6.16.

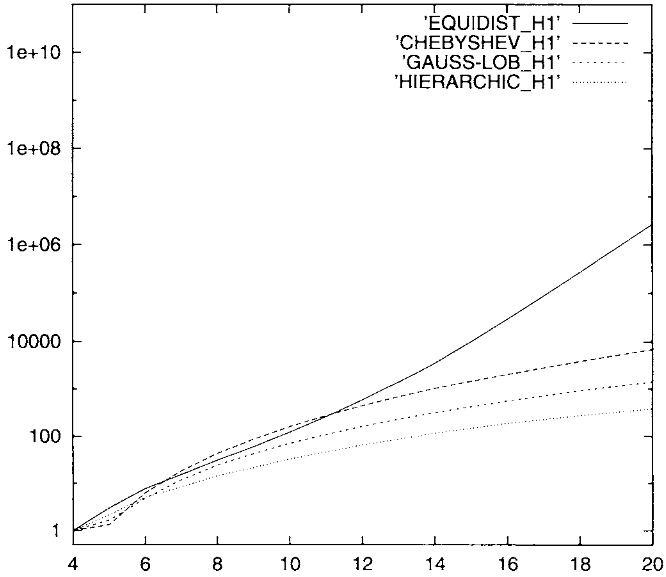


Figure 6.16 Conditioning of the higher-order shape functions in the H_0^1-product (6.34). The hierarchic shape functions (6.36) are not optimal anymore, but they still are better than the other three choices for $p \geq 7$.

It follows from Figure 6.16 that (a) surprisingly, the equidistant nodal points perform better than the Chebyshev points for $7 \leq p \leq 11$, and (b) for every $p \geq 7$ the hierarchic shape functions give the best result.

6.3.4 Basis of the space $V_{h,p}$

With suitable shape functions on the reference domain K_a in hand, the basis functions $v_1, v_2, \ldots, v_N$ of the space $V_{h,p} \subset H_0^2(\Omega)$ can be designed. We shall work work with the hierarchic shape functions $\omega_0, \omega_1, \ldots$ in what follows.

Assume a bounded domain $\Omega = (a, b) \subset \mathbb{R}$ and a finite element mesh $\mathcal{T}_{h,p}$ consisting of $M \geq 1$ Hermite elements K_i on subintervals $K_i = (x_{i-1}, x_i)$, equipped with arbitrary polynomial degrees $3 \le p_i = p(K_i)$. The space $V_{h,p}$ was defined in (6.20),

$$\begin{aligned} V_{h,p} \quad = \quad & \{v \in C^1(\Omega);\ v(a) = v(b) = v'(a) = v'(b) = 0; \\ & v|_{K_i} \in P^{p_i}(K_i)\}. \end{aligned} \tag{6.38}$$

It is easy to calculate the dimension of this space,

$$N = \dim(V_{h,p}) = -M - 2 + \sum_{i=1}^{M} p_i. \tag{6.39}$$

In view of Definition 6.1, the basis functions are split into vertex and bubble functions.

Vertex basis functions: The vertex functions are associated with the internal grid points x_i, $i = 1, 2, \ldots, M_1$, and they always extend over two adjacent elements K_i and K_{i+1}. A first set of $M - 1$ vertex functions $v_i^{(v,0)}$ represent the function values,

$$v_i^{(v,0)} = \begin{cases} \omega_1 \circ x_{K_i}^{-1}, & x \in K_i, \\ \\ \omega_0 \circ x_{K_{i+1}}^{-1}, & x \in K_{i+1}. \end{cases} \tag{6.40}$$

The other $M - 1$ vertex functions $v_i^{(v,1)}$ represent the derivatives,

$$v_i^{(v,1)} = \begin{cases} J_{K_i}\omega_3 \circ x_{K_i}^{-1}, & x \in K_i, \\ \\ J_{K_{i+1}}\omega_2 \circ x_{K_{i+1}}^{-1}, & x \in K_{i+1}. \end{cases} \tag{6.41}$$

The delta property (6.28) of the cubic Hermite shape functions $\omega_0, \omega_1, \ldots, \omega_3$ translates into

$$v_i^{(v,0)}(x_j) = \delta_{ij}, \qquad \left(v_i^{(v,0)}\right)'(x_j) = 0,$$

and

$$v_i^{(v,1)}(x_j) = 0, \qquad \left(v_i^{(v,1)}\right)'(x_j) = \delta_{ij},$$

for all $1 \le i \le M - 1$ and $0 \le j \le M$.

Bubble basis functions: On every element K_i we define $p_i - 3$ bubble functions

$$v_{i,k}^b = \omega_k \circ x_{K_i}^{-1},\ x \in K_i,\ k = 4, 5, \ldots, p_i. \tag{6.42}$$

In the nodal higher-order case the bubble functions are defined by

$$v_{i,k}^b = \theta_{k-2} \circ x_{K_i}^{-1},\ x \in K_i,\ k = 4, 5, \ldots, p_i.$$

where θ_k are the bubble shape functions associated with the Lagrange degrees of freedom at the $p_i - 3$ nodal points $-1 < y_2 < y_3 < \ldots < y_{p_i - 2} < 1$.

Lemma 6.2 *Functions (6.40)–(6.42) belong to $C^1(\Omega)$, and form a basis of the space (6.38).*

Proof: This is clear from their construction. ■

6.3.5 Transformation of weak forms to the reference domain

For the element-by-element assembling algorithm, the approximate weak formulation (6.23) needs to be written as a sum over all elements,

$$\sum_{m=1}^{M}\sum_{j=1}^{N} y_j \int_{K_m} b\Delta v_j \Delta v_i \,\mathrm{d}x = \sum_{m=1}^{M}\int_{K_m} f v_i \,\mathrm{d}x \qquad \text{for all } i = 1,2,\ldots,N. \tag{6.43}$$

Every integral in the sum is transformed to the reference domain K_a via the Substitution Theorem. Consider a mesh element K_m, $1 \le m \le M$, and the standard one-dimensional affine reference map $x_{K_m} : K_a \to K_m$ defined in (2.37). Assume that the Jacobian $J_{K_m} > 0$ for every K_m. With the notation from Paragraph 2.4.3,

$$\tilde{v}_i^{(m)}(\xi) = (v_i \circ x_{K_m})(\xi),$$

one has

$$v_i'(x) = \frac{1}{J_{K_m}}[\tilde{v}_i^{(m)}]'(\xi), \qquad x = x_{K_m}(\xi),$$

and further

$$\Delta v_i(x) = \frac{1}{J_{K_m}^2}\Delta\tilde{v}_i^{(m)}(\xi), \qquad x = x_{K_m}(\xi).$$

This means that the biharmonic stiffness integrals for the model problem (6.14) attain the form

$$\int_{K_m} b(x)\Delta v_j(x)\Delta v_i(x)\,\mathrm{d}x = \int_{K_a} \frac{1}{J_{K_m}^3}\tilde{b}^{(m)}(\xi)\Delta\tilde{v}_j^{(m)}(\xi)\Delta\tilde{v}_i^{(m)}(\xi)\,\mathrm{d}\xi, \tag{6.44}$$

where $\tilde{b}^{(m)} = b \circ x_{K_m}$. The right-hand side integrals from (6.43) transform to the reference domain K_a simply as

$$\int_{K_m} f v_i \,\mathrm{d}x = \int_{K_a} J_{K_m}\tilde{f}^{(m)}\tilde{v}_i^{(m)}\,\mathrm{d}\xi,$$

where $\tilde{f}^{(m)} = f \circ x_{K_m}$.

6.3.6 Connectivity arrays

The extension of the data structures and algorithms from Paragraphs 2.4.8 and 2.4.9 to Hermite elements in one spatial dimension is straightforward. In what follows let us assume the model problem (6.7) with the homogeneous Dirichlet conditions (6.11) for both the solution u and its first derivative $u'(x)$. The extension to nonhomogeneous Dirichlet boundary conditions is done analogously to Paragraph 2.6.

Element data structure Consider a finite element mesh $\mathcal{T}_{h,p} = \{K_1, K_2, \ldots, K_M\}$ consisting of Hermite elements of arbitrary polynomial degrees $3 \le p_i = p(K_i)$, $i = 1,2,\ldots,M$. Choose a reasonable upper bound MAXP and define:

```
struct {
  int p;              //polynomial degree of element
  int vert_dir[4];    //vertex Dir. flags for both u and du/dx
  int vert_dof[4];    //vertex connectivity arrays
  int *bubb_dof;      //bubble connectivity array
                      //(length MAXP-3)
  ...
} Element;
```

The Dirichlet flag `Elem[m].vert_dir[j]`, $j = 1, 2$, has the following meaning: It is zero if the left vertex of $K_m = (x_{m-1}, x_m)$ is unconstrained by a Dirichlet boundary condition for the solution u, and it equals to one otherwise. The flag `Elem[m].vert_dir[2]` is related to the right vertex of K_m in the same way, and the Dirichlet flags `Elem[m].vert_dir[j]`, $j = 3, 4$, have an analogous meaning for the first derivative $u'(x)$.

Unique enumeration of basis functions According to (6.39), the dimension of the space $V_{h,p}$ is

$$N = \dim(V_{h,p}) = -M - 2 + \sum_{i=1}^{M} p_i.$$

The N basis functions have to be enumerated uniquely so that the connectivity links can be defined. We use the following scheme: First enumerate all vertex functions representing the solution values at the internal grid points,

$$v_i = v_i^{(v,0)} \qquad \text{for all } i = 1, 2, \ldots, M - 1.$$

Then add all vertex functions representing the derivatives at the internal grid points,

$$v_{M-1+i} = v_i^{(v,1)} \qquad \text{for all } i = 1, 2, \ldots, M - 1.$$

At the end of the list put the bubble functions, using an outer element loop $m = 1, 2, \ldots, M$ and an inner loop $p = 4, 5, \ldots, p_m$. This will be implemented in Algorithm 6.1 below.

Element connectivity arrays Analogously to Paragraph 2.4.8, the values of the Dirichlet lifts for both the solution u and the first derivative $u'(x)$, which are only nonzero in the case of nonhomogeneous Dirichlet boundary conditions, are stored in a global array `double DIR_BC_ARRAY[4]` $= \{G(a), G(b), G_{der}(a), G_{der}(b)\}$. The variable `Elem[m].vert_dof [1]` contains either

- a positive index i of a vertex basis function $v_i = v_i^{(v,0)}$ associated with the left vertex of the element K_m (if the vertex is not constrained by a Dirichlet boundary condition for the solution u, i.e., if `Elem[m].vert_dir [1] == 0`),
- or `-1`, so that $G(a) =$ `DIR_BC_ARRAY[-Elem[m].vert_dof[1]]` (if `Elem[m].vert_dir[1] == 1`).

Similarly one defines `Elem[m].vert_dof[2]` for the right vertex of the element K_m. If `Elem [m].vert_dir[2] == 1`, then `Elem[m].vert_dof[2] == -2`. The variable `Elem[m].vert_dof[3]` contains either

- the index $M - 1 + i$ of a vertex basis function $v_{M-i+i} = v_i^{(v,1)}$ corresponding to the left vertex of the element K_m (if the vertex is not constrained by a Dirichlet boundary condition for the first derivative u', i.e., if Elem[m].vert_dir[3] == 0),
- or -3, so that $G_{der}(a) =$ DIR_BC_ARRAY[-Elem[m].vert_dof[3]] (if Elem[m].vert_dir[3] == 1).

The value Elem[m].vert_dof[4] for the right vertex of the element K_m is defined analogously. If Elem [m].vert_dir[4] == 1, then Elem[m].vert_dof[4] == -4. The bubble functions are always unconstrained, and therefore the value Elem[m].bubb_dof[j], j = 1,2,...,Elem[m].p-3, contains the index of the bubble basis functions of the polynomial degree $j + 3$ associated with the element K_m.

The following connectivity algorithm is similar to Algorithm 2.4, except that the Hermite shape functions ω_2 and ω_3 are treated differently from the rest of the shape functions.

Algorithm 6.1 (Connectivity algorithm for Hermite elements)

```
count := 1;
index := 1;  //For Lagrange vertex functions $\omega_0, \omega_1$
//Block A: Enumerate Lagrange vertex functions ($v_i^{v,0}$):
//Visiting the element $K_1$:
if (Elem[1].vert_dir[index] == 1) then {
  Elem[1].vert_dof[index] := -index;
}
else {
  Elem[1].vert_dof[index] := count;
  count := count + 1;
}
Elem[1].vert_dof[index+1] := count;
//Visiting interior elements $K_2, K_3, \ldots, K_{M-1}$:
for m = 2,3,...,M-1 do {
  Elem[m].vert_dof[index] := count;
  count := count + 1;
  Elem[m].vert_dof[index+1] := count;
}
//Visiting the element $K_M$:
Elem[M].vert_dof[index] := count;
count := count + 1;
if (Elem[M].vert_dir[index+1] == 1) then {
  Elem[M].vert_dof[index+1] := -index-1;
}
else {
  Elem[M].vert_dof[index+1] := count;
  count := count + 1;
}
//Block B: Enumerate Hermite vertex functions ($v_i^{v,1}$):
//(run block A once more with  index := 3)
//Block C: Enumerate all (Lagrange) bubble functions:
for m = 1,2,...,M do {
  for j = 1,2,...,Elem[m].p-3 do {
    Elem[m].bubb_dof[j] := count;
    count := count + 1;
  }
}
```

The function of Algorithm 6.1 can be illustrated on a simple example:

■ EXAMPLE 6.1

Consider the model problem (6.7) with homogeneous Dirichlet boundary conditions (6.11), and a mesh $\mathcal{T}_{h,p}$ consisting of three elements K_1, K_2 and K_3 of the polynomial degrees $p_1 = 4$, $p_2 = 6$ and $p_3 = 5$. According to (6.39), the dimension of the space $V_{h,p}$ is $N = 10$. The input data for the connectivity Algorithm 6.1 is

```
Elem[1].p = 4;
Elem[1].vert_dir = {1,0,1,0};
Elem[2].p = 6;
Elem[2].vert_dir = {0,0,0,0};
Elem[3].p = 5;
Elem[3].vert_dir = {0,1,0,1};
```

The resulting element connectivity arrays have the form

```
Elem[1].vert_dof = {-1,1,-3,3};
Elem[1].bubb_dof = {5};
Elem[2].vert_dof = {1,2,3,4};
Elem[2].bubb_dof = {6,7,8};
Elem[3].vert_dof = {2,-2,4,-4};
Elem[3].bubb_dof = {9,10};
```

6.3.7 Assembling algorithm

The complexity of the assembling algorithm depends on whether or not the function b in (6.7) is constant. If it is constant, then the global stiffness matrix S can be assembled using the set (6.37) of few precomputed master element stiffness integrals. Otherwise explicit numerical integration needs to be done on every mesh element K_m, $m = 1, 2, \dots, M$. For simplicity, assume that $0 < b = \text{const}$. Then b can be removed from the equation by replacing f with f/b. The master element stiffness matrix (6.37) can be represented via a two-dimensional array MESI,

$$\texttt{MESI[i][j]} := \int_{-1}^{1} \Delta\omega_{i-1}(x)\Delta\omega_{j-1}(x)\,\mathrm{d}x \qquad \text{for all } 1 \le i, j \le \texttt{MAXP} + 1.$$

Further we define the function

```
double Jac(double jac, int index) {
  if(index < 3) return 1.;
  else return jac;
}
```

that is used to distinguish between the Lagrange and Hermite shape functions in the assembling algorithm:

Algorithm 6.2 (Assembling algorithm for Hermite elements in 1D)

```
//Calculate the dimension of the space $V_{h,p}$:
N := -2 - M;
for m = 1,2,...,M do N := N + Elem[m].p;
//Calculate the value of Elem[m].jac for all elements $K_m$, $m = 1,2,\dots,M$:
for m = 1,2,...,M do Elem[m].jac := $(x_m - x_{m-1})/2$;
//Set the stiffness matrix $S$ zero:
for i = 1,2,...,N do for j = 1,2,...,N do S[i][j] := 0;
```

```
//Set the right-hand side vector $\boldsymbol{F}$ zero:
for i = 1,2,...,N do F[i] := 0;
//Element loop:
for m = 1,2,...,M do {
  //Loop over vertex test functions:
  for i = 1,2,...,4 do {
    //If > -1, this is index of a  test function $v_{m_1} \in V_{h,p}$,
    //i.e.,  row position in $\boldsymbol{S}$:
    m1 := Elem[m].vert_dof[i];
    //Loop over vertex basis functions:
    if (m1 > -1) then for j = 1,2,...,4 do {
      //If > -1, this is index of a  basis function $v_{m_2} \in V_{h,p}$,
      //i.e.,  column position in $\boldsymbol{S}$:
      m2 := Elem[m].vert_dof[j];
      if (m2 > -1) then {
        //Multiply each Hermite shape function with an extra Jacobian:
        jac := Elem[m].jac;
        jactb := Jac(jac,i)*Jac(jac,j);
        S[m1][m2] := S[m1][m2] + jactb*MESI[i][j]/jac$^3$;
      }
    } //End of inner loop over vertex functions
    //Loop over bubble basis functions:
    for j = 1,2,...,Elem[m].p-3 do {
      m2 := Elem[m].bubb_dof[j];
      if (m2 > -1) then {
        jac := Elem[m].jac;
        S[m1][m2] := S[m1][m2] + Jac(jac,i)*MESI[i][j+4]/jac$^3$;
      }
    } //End of inner loop over bubble functions
    //Contribution of the vertex test function $v_{m_1}$ to the right-hand side $\boldsymbol{F}$:
    if (m1 > -1) then {
      jac := Elem[m].jac;
      F[m1] := F[m1] + Jac(jac,i)*$\int_{K_a} |J_{K_m}| \tilde{f}^{(m)}(\xi)\omega_{i-1}(\xi)\,\mathrm{d}\xi$;
    }
  } //End of outer loop over vertex functions
  //Loop over bubble test functions:
  for i = 1,2,...,Elem[m].p-3 do {
    m1 := Elem[m].bubb_dof[i];
    //Loop over vertex basis functions:
    if (m1 > -1) then for j = 1,2,...,4 do {
      m2 := Elem[m].vert_dof[j];
      if (m2 > -1) then {
        jac := Elem[m].jac;
        S[m1][m2] := S[m1][m2] + Jac(jac,j)*MESI[i+4][j]/jac$^3$;
      }
    } //End of inner loop over vertex functions
    //Loop over bubble basis functions:
    if (m1 > -1) then for j = 1,2,...,Elem[m].p-3 do {
      m2 := Elem[m].bubb_dof[j];
      if (m2 > -1) then S[m1][m2] := S[m1][m2] + MESI[i+4][j+4]/Elem[m].jac$^3$;
    } //End of inner loop over bubble functions
    //Contribution of the bubble test function $v_{m_1}$ to the right-hand side $\boldsymbol{F}$:
    if (m1 > -1) then F[m1] := F[m1] + $\int_{K_a} |J_{K_m}| \tilde{f}^{(m)}(\xi)\omega_{i-1}(\xi)\,\mathrm{d}\xi$;
  } //End of outer loop over bubble functions
} //End of element loop
```

6.3.8 Interpolation on Hermite elements

Similarly to the H^1-conforming case discussed in Paragraphs 2.7.2–2.7.4, also on the Hermite elements we have at least three basic interpolation options with different quality and cost:

1. the best interpolant exploiting a global orthogonal projection (best quality but highest cost),
2. the projection-based interpolant combining the nodal interpolation of vertex values and derivatives with local orthogonal projections in element interiors (slightly less accurate than the best interpolant, but much more efficient, especially with orthonormal higher-order bubble functions),
3. the traditional explicit nodal interpolant (fastest but worst quality).

Best interpolant Consider a bounded domain $\Omega = (a, b) \subset \mathbb{R}$ and a finite element mesh $\mathcal{T}_{h,p}$ consisting of $M \geq 1$ Hermite elements $K_i = (x_{i-1}, x_i)$ equipped with arbitrary polynomial degrees $3 \leq p_i = p(K_i)$. The best interpolant of a function $g \in V = H_0^2(\Omega)$ in the finite-dimensional subspace $V_{h,p} \subset V$ is obtained as follows: The function $g_{h,p}$ is expressed as

$$g_{h,p} = \sum_{j=1}^{N} y_j v_j, \tag{6.45}$$

where $\{v_1, v_2, \ldots, v_N\}$ is a basis of $V_{h,p}$. The unknown coefficients $y_1, y_2, \ldots, y_N$ are determined from the orthogonality condition

$$(g - g_{h,p}) \perp V_{h,p},$$

that is equivalent to a system of linear algebraic equations,

$$\sum_{j=1}^{N} y_j (v_j, v_i)_V = (g, v_i)_V \qquad \text{for all } i = 1, 2, \ldots, N. \tag{6.46}$$

Since the Poincaré–Friedrichs' inequality (Theorem A.26) holds in the space V, one can use either the full H^2-product,

$$(u, v)_V = \int_\Omega uv + \nabla u \nabla v + \Delta u \Delta v \, \mathrm{d}x$$

or, equivalently, the simpler H_0^2-product

$$(u, v)_V = \int_\Omega \Delta u \Delta v \, \mathrm{d}x.$$

Projection-based interpolant The interpolant $g_{h,p}$ is sought in two steps, as a sum of the vertex and bubble interpolants

$$g_{h,p} = g_{h,p}^v + g_{h,p}^b. \tag{6.47}$$

Vertex interpolant: The elementwise cubic function $g^v_{h,p} \in C^1(\Omega)$ satisfies

$$g^v_{h,p}(x_i) = g(x_i),\ \left(g^v_{h,p}\right)'(x_i) = g'(x_i) \quad \text{for all } i = 0, 1, \ldots, M. \tag{6.48}$$

Using the basis functions (6.40) and (6.41), on every $K_i = (x_{i-1}, x_i)$ we obtain

$$g^v_{h,p}|_{K_i} = v^{(v,0)}_{i-1} g(x_{i-1}) + v^{(v,0)}_i g(x_i) + J_{K_i} v^{(v,1)}_{i-1} g'(x_{i-1}) + J_{K_i} v^{(v,1)}_i g'(x_i). \tag{6.49}$$

Bubble interpolant: Since the residual $g - g^v_{h,p}$ vanishes at all grid points x_i together with its first derivative, on every element K_i, $i = 1, 2, \ldots, M$ it belongs to the space (A.92),

$$H^2_0(K_i) = \{v \in H^2(K_i);\ v(x_{i-1}) = v(x_i) = v'(x_{i-1}) = v'(x_i) = 0\}.$$

On every element K_i with $p_i \geq 4$ consider the polynomial subspace

$$P^{p_i}_{00}(K_i) = \{v \in P^{p_i}_0(K_i);\ v'(x_{i-1}) = v'(x_i) = 0\} \subset H^2_0(K_i)$$

of the dimension $p_i - 3$. Let us stay with the H^2_0-product

$$(u, v)_{H^2_0(K_i)} = \int_{K_i} \Delta u \Delta v \, \mathrm{d}x \tag{6.50}$$

for simplicity. The unique bubble interpolant $g^b_{h,p}$ on the element K_i is determined from the orthogonality condition

$$(g - g_{h,p}) \perp P^{p_i}_{00}(K_i).$$

Using the bubble functions (6.42) that generate the space $P^{p_i}_{00}(K_i)$, this is equivalent to

$$\left(g - g^v_{h,p} - g^b_{h,p}, v^b_{i,k}\right)_{H^2_0(K_i)} = 0 \quad \text{for all } k = 4, 5, \ldots, p_i. \tag{6.51}$$

Expressing

$$g^b_{h,p}|_{K_i} = \sum_{m=4}^{p_i} \alpha^{(i)}_m v^b_{i,m},$$

and inserting this linear combination into (6.51), one obtains a system of $p_i - 3$ linear algebraic equations,

$$\int_{K_i} \Delta\left(g - g^v_{h,p} - \sum_{m=4}^{p_i} \alpha^{(i)}_m v^b_{i,m}\right)(x)\, \Delta\left(v^b_{i,k}\right)(x)\, \mathrm{d}x = 0,\ \ k = 4, 5, \ldots, p_i, \tag{6.52}$$

for the unknown coefficients $\alpha^{(i)}_m$. Transformed to the reference domain K_a, with the orthogonal hierarchic shape functions (6.36) this linear system simplifies substantially to

$$\sum_{m=4}^{p_i} \alpha_m^{(i)} \underbrace{\int_{K_a} \Delta\omega_m(\xi)\Delta\omega_k(\xi)\,\mathrm{d}\xi}_{\delta_{mk}} = \int_{K_i} \Delta(\tilde{g} - \tilde{g}_{h,p}^v)(\xi)\Delta\omega_k(\xi)\,\mathrm{d}\xi, \tag{6.53}$$

which means that

$$\alpha_k^{(i)} = \int_{K_i} \Delta(\tilde{g} - \tilde{g}_{h,p}^v)(\xi)\Delta\omega_k(\xi)\,\mathrm{d}\xi, \tag{6.54}$$

where $k = 4, 5, \ldots, p_i$. Hence no system of linear algebraic equations needs to be solved. Here $\tilde{g}(\xi) = g(x_{K_i}(\xi))$ and

$$\begin{aligned} \tilde{g}_{h,p}^v(\xi) &= (g_{h,p}^v(x_{K_i}(\xi)) \\ &= \omega_0(\xi)g(x_{i-1}) + \omega_1(\xi)g(x_i) + J_{K_i}\omega_2(\xi)g'(x_{i-1}) + J_{K_i}\omega_3(\xi)g'(x_i). \end{aligned}$$

After obtaining the coefficients $\alpha_k^{(i)}$, $k = 4, 5, \ldots, p_i$, for every element K_i, $i = 1, 2, \ldots, M$, the construction of the projection-based interpolant $g_{h,p} = g_{h,p}^v + g_{h,p}^b$ is accomplished.

Lemma 6.3 (Local optimality of the projection-based interpolant) *Let $\Omega = (a, b) \subset \mathbb{R}$ be covered with a finite element mesh $\mathcal{T}_{h,p}$ consisting of M Hermite elements $K_i = (x_{i-1}, x_i)$ equipped with the polynomial degrees $3 \le p_i = p(K_i)$. Let $g \in H^2(\Omega) \cap C^0(\overline{\Omega})$, $g_{h,p} \in V_{h,p}$ its projection-based interpolant (6.47) and $\tilde{g}_{h,p} \in V_{h,p}$ an arbitrary other interpolant satisfying $\tilde{g}_{h,p}(x_j) = g(x_j)$, $\tilde{g}'_{h,p}(x_j) = g'(x_j)$ for all $j = 0, 1, \ldots, M$. Then*

$$|g - g_{h,p}|_{2,2,K_i} \le |g - \tilde{g}_{h,p}|_{2,2,K_i} \quad \textit{for all } i = 1, 2, \ldots, M, \tag{6.55}$$

and therefore also

$$|g - g_{h,p}|_{2,2,\Omega} \le |g - \tilde{g}_{h,p}|_{2,2,\Omega}. \tag{6.56}$$

If the bubble interpolant $g_{h,p}^b$ is calculated using the full H^2-product $(\cdot, \cdot)_{2,2}$ instead of (6.50), the inequalities (6.55) and (6.56) hold with the full H^2-norm $\|\cdot\|_{2,2}$.

Proof: The proof is analogous to the proof of Lemma 2.5 and it is left to the reader as an exercise. ■

Nodal interpolant The Hermite elements are automatically endowed with the standard nodal interpolant (3.28),

$$\mathcal{I}_K(g) = \sum_{i=1}^{N_P} L_i(g)\theta_i.$$

Here θ_i are the nodal basis functions of the space $P^{p_i}(K_i)$ meeting the delta property (3.4), g is an arbitrary function from some space V such that $P^{p_i}(K_i) \subset V(K_i)$, and it is assumed that all linear forms L_i, $i = 1, 2, \ldots, N_P$ are defined for g.

Lemma 6.4 (Conformity to $H^2(\Omega)$) *The finite element mesh $\mathcal{T}_{h,p}$ consisting of arbitrary-order Hermite elements, introduced in Paragraph 6.3.4, is conforming to the space $H^2(\Omega)$.*

Proof: It is sufficient to verify that the global interpolant $\mathcal{I}(g)$ of an arbitrary function $g \in H^2(\Omega) \cap C(\overline{\Omega})$ is smooth at all grid vertices. The continuity of $(\mathcal{I}(g))(x)$ at all x_i, $i = 1, 2, \ldots, M$, follows from the delta property (3.4) and the continuity of the basis functions (6.40). Analogously the continuity of the derivative $(\mathcal{I}(g))'(x)$ follows from (3.4) and the smoothness of the basis functions (6.40). ■

Since the nodal interpolant $\mathcal{I}_{K_i}(g)$ matches the values and first derivatives of the interpolated function g at all grid vertices, by Lemma 6.3 its quality cannot be better than the quality of the projection-based interpolant.

6.4 HERMITE ELEMENTS IN 2D

In contrast to the one-dimensional case, Hermite elements do not conform to the Sobolev space H^2 in 2D and 3D. Nevertheless, they find important application in nonconforming approximations to fourth-order problems, computational geometry, surface reconstruction, and elsewhere. We focus on triangular elements, since the construction of Hermite quadrilaterals can be done easily using the product geometry of the reference domain K_q.

6.4.1 Lowest-order elements

The cubic Hermite element $(K_t, P^3(K_t), \Sigma)$ on the triangular reference domain K_t is equipped with the set Σ consisting of three degrees of freedom per vertex (one for the function value and two for directional derivatives) and one complementary interior degree of freedom that is added for the sake of unisolvency. This degree of freedom usually is associated with the function value at the center of gravity of K_t.

There are two equivalent choices for the directional derivatives: either the directions of the coordinate axes (i.e., the partial derivatives $\partial/\partial x_1$ and $\partial/\partial x_2$) or the directions of the edges of K_t. This is illustrated in Figure 6.17.

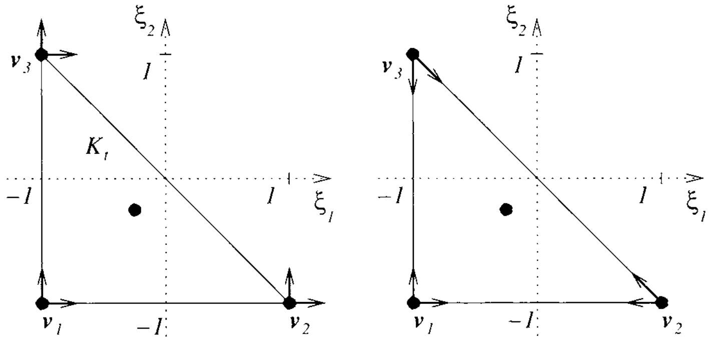


Figure 6.17 Two equivalent types of cubic Hermite elements.

Let us discuss, e.g., the former case (left part of Figure 6.17). This element has four Lagrange degrees of freedom

$$\begin{aligned} L_1(g) &= g(-1, -1), \\ L_2(g) &= g(1, -1), \end{aligned} \tag{6.57}$$

$$\begin{aligned} L_3(g) &= g(-1,1), \\ L_4(g) &= g(-1/3,-1/3), \end{aligned}$$

and six Hermite degrees of freedom

$$\begin{aligned} L_5(g) &= \frac{\partial g}{\partial \xi_1}(-1,-1), \\ L_6(g) &= \frac{\partial g}{\partial \xi_1}(1,-1), \\ L_7(g) &= \frac{\partial g}{\partial \xi_1}(-1,1), \\ L_8(g) &= \frac{\partial g}{\partial \xi_2}(-1,-1), \\ L_9(g) &= \frac{\partial g}{\partial \xi_2}(1,-1), \\ L_{10}(g) &= \frac{\partial g}{\partial \xi_2}(-1,1). \end{aligned} \tag{6.58}$$

The corresponding nodal basis, satisfying the delta property (3.4), is constructed using the standard procedure which was described in Paragraph 3.1.2:

Consider, for example, the monomial basis of the polynomial space $P^3(K_t)$,

$$\mathcal{B} = \{g_1, g_2, \ldots, g_{10}\} = \{1, \xi_1, \xi_2, \xi_1^2, \xi_1\xi_2, \xi_2^2, \xi_1^3, \xi_1^2\xi_2, \xi_1\xi_2^2, \xi_2^3\}.$$

Inverting the Vandermonde matrix $\boldsymbol{L} = \{L_i(g_j)\}_{i,j=1}^{10}$ and reading the coefficients for the functions $g_1, g_2, \ldots, g_{10}$ from its columns, one arrives at the nodal shape functions

$$\begin{aligned} \varphi_t^{v_1}(\boldsymbol{\xi}) &= \frac{7}{8}\xi_1 + \frac{7}{8}\xi_2 + \frac{13}{8}\xi_1^2 + \frac{13}{4}\xi_1\xi_2 + \frac{13}{8}\xi_2^2 + \frac{1}{4}\xi_1^3 + \frac{13}{8}\xi_1^2\xi_2 + \frac{13}{8}\xi_1\xi_2^2 \\ &\quad + \frac{1}{4}\xi_2^3, \\ \varphi_t^{v_2}(\boldsymbol{\xi}) &= \frac{1}{2} + \frac{13}{8}\xi_1 + \frac{7}{8}\xi_2 + \frac{7}{8}\xi_1^2 + \frac{7}{4}\xi_1\xi_2 + \frac{7}{8}\xi_2^2 - \frac{1}{4}\xi_1^3 + \frac{7}{8}\xi_1^2\xi_2 + \frac{7}{8}\xi_1\xi_2^2, \\ \varphi_t^{v_3}(\boldsymbol{\xi}) &= \frac{1}{2} + \frac{7}{8}\xi_1 + \frac{13}{8}\xi_2 + \frac{7}{8}\xi_1^2 + \frac{7}{4}\xi_1\xi_2 + \frac{7}{8}\xi_2^2 + \frac{7}{8}\xi_1^2\xi_2 + \frac{7}{8}\xi_1\xi_2^2 - \frac{1}{4}\xi_2^3, \\ \varphi_t^{b}(\boldsymbol{\xi}) &= -\frac{27}{8}\xi_1 - \frac{27}{8}\xi_2 - \frac{27}{8}\xi_1^2 - \frac{27}{4}\xi_1\xi_2 - \frac{27}{8}\xi_2^2 - \frac{27}{8}\xi_1^2\xi_2 - \frac{27}{8}\xi_1\xi_2^2, \\ \varphi_t^{v_1,1}(\boldsymbol{\xi}) &= \frac{1}{4}\xi_1 + \frac{1}{4}\xi_2 + \frac{1}{2}\xi_1^2 + \xi_1\xi_2 + \frac{1}{2}\xi_2^2 + \frac{1}{4}\xi_1^3 + \frac{3}{4}\xi_1^2\xi_2 + \frac{1}{2}\xi_1\xi_2^2, \\ \varphi_t^{v_2,1}(\boldsymbol{\xi}) &= -\frac{1}{4} - \frac{3}{4}\xi_1 - \frac{1}{2}\xi_2 - \frac{1}{4}\xi_1^2 - \xi_1\xi_2 - \frac{1}{2}\xi_2^2 + \frac{1}{4}\xi_1^3 - \frac{1}{2}\xi_1^2\xi_2 - \frac{1}{2}\xi_1\xi_2^2, \\ \varphi_t^{v_3,1}(\boldsymbol{\xi}) &= \frac{1}{4} + \frac{1}{2}\xi_1 + \frac{3}{4}\xi_2 + \frac{1}{4}\xi_1^2 + \xi_1\xi_2 + \frac{1}{2}\xi_2^2 + \frac{1}{4}\xi_1^2\xi_2 + \frac{1}{2}\xi_1\xi_2^2, \\ \varphi_t^{v_1,2}(\boldsymbol{\xi}) &= \frac{1}{4}\xi_1 + \frac{1}{4}\xi_2 + \frac{1}{2}\xi_1^2 + \xi_1\xi_2 + \frac{1}{2}\xi_2^2 + \frac{1}{2}\xi_1^2\xi_2 + \frac{3}{4}\xi_1\xi_2^2 + \frac{1}{4}\xi_2^3, \\ \varphi_t^{v_2,2}(\boldsymbol{\xi}) &= \frac{1}{4} + \frac{3}{4}\xi_1 + \frac{1}{2}\xi_2 + \frac{1}{2}\xi_1^2 + \xi_1\xi_2 + \frac{1}{4}\xi_2^2 + \frac{1}{2}\xi_1^2\xi_2 + \frac{1}{4}\xi_1\xi_2^2, \\ \varphi_t^{v_3,2}(\boldsymbol{\xi}) &= -\frac{1}{4} - \frac{1}{2}\xi_1 - \frac{3}{4}\xi_2 - \frac{1}{2}\xi_1^2 - \xi_1\xi_2 - \frac{1}{4}\xi_2^2 - \frac{1}{2}\xi_1^2\xi_2 - \frac{1}{2}\xi_1\xi_2^2 + \frac{1}{4}\xi_2^3. \end{aligned} \tag{6.59}$$

The first four shape functions $\varphi_t^{v_1}, \varphi_t^{v_2}, \varphi_t^{v_3}$ and φ_t^b in (6.59) correspond to the Lagrange DOF L_1, L_2, L_3 and L_4, respectively, while the rest correspond to the Hermite DOF. The nodal basis is depicted in Figures 6.18–6.21.

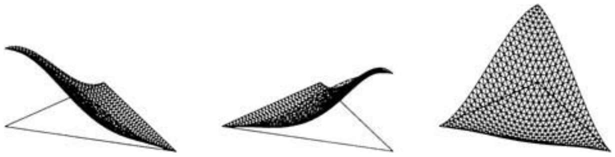


Figure 6.18 Nodal basis of the cubic Hermite element; the vertex functions $\varphi_t^{v_1}$, $\varphi_t^{v_2}$, and $\varphi_t^{v_3}$.

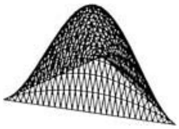


Figure 6.19 Nodal basis of the cubic Hermite element; the bubble function φ_t^b.

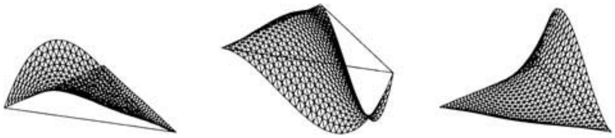


Figure 6.20 Nodal basis of the cubic Hermite element; the vertex functions $\varphi_t^{v_1,1}$, $\varphi_t^{v_2,1}$, and $\varphi_t^{v_3,1}$.

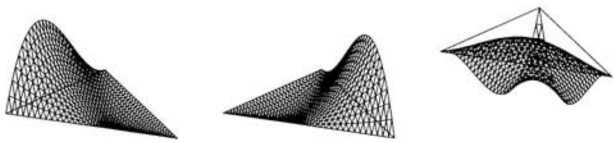


Figure 6.21 Nodal basis of the cubic Hermite element; the vertex functions $\varphi_t^{v_1,2}$, $\varphi_t^{v_2,2}$, and $\varphi_t^{v_3,2}$.

6.4.2 Higher-order Hermite–Fekete elements

Next let us consider a polynomial degree $p \geq 4$. Recall that the dimension of the space $P^p(K_t)$ is $N_P = (p+1)(p+2)/2$. The vertex degrees of freedom L_1, L_2, L_3 and $L_5, L_6, \ldots, L_{10}$ from the cubic case guarantee the continuity and smoothness of the approximation at the vertices.

Hence, $(p+1)(p+2)/2 - 9$ new Lagrange degrees of freedom need to be defined in such a way that the finite element is unisolvent and the approximation continuous along element interfaces.

A pth-degree polynomial restricted to an edge of K_t is determined via $p+1$ parameters, of which four are the vertex degrees of freedom at the endpoints. Thus $p-3$ additional Lagrange degrees of freedom need to be placed into the interior of each edge of K_t. The one-dimensional interior Gauss–Lobatto points of the order $p-2$ are a suitable choice for this purpose (although better point sets may exist). In addition to that,

$$\frac{(p+1)(p+2)}{2} - 9 - 3(p-3) = \frac{(p-1)(p-2)}{2}$$

interior Lagrange degrees of freedom remain to be chosen. It is natural to associate them with the $(p-1)(p-2)/2$ interior Fekete points of the order p. For $p=3$ this set of degrees of freedom exactly coincides with the cubic case described in Paragraph 6.4.1.

The distributions of the degrees of freedom for a fourth- and fifth-order Hermite–Fekete elements are illustrated in Figure 6.22.

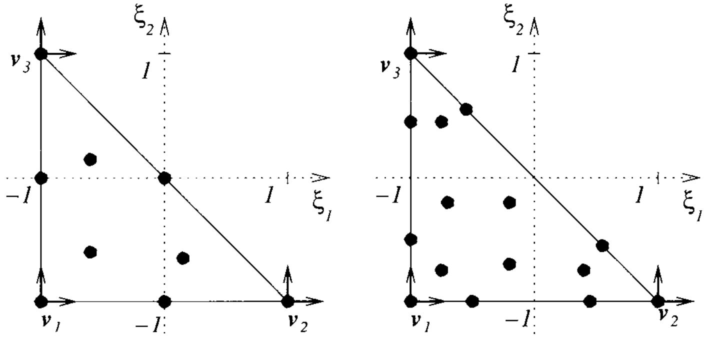


Figure 6.22 Fourth- and fifth-order Hermite–Fekete elements on K_t.

Lemma 6.5 *The Hermite–Fekete element $(K_t, P^p(K_t), \Sigma)$, where $p \geq 3$ and Σ consists of the $(p+1)(p+2)/2$ above-defined degrees of freedom, is unisolvent.*

Proof: Let $P = P^p(K_t)$ and $N_P = \dim(P) = (p+1)(p+2)/2$. It is clear from above that $\text{card}(\Sigma) = 9 + 3(p-3) + (p-1)(p-2)/2 = N_P$. Let $g \in P$ be arbitrary. It is sufficient to show that if $L(g) = 0$ for all $L \in \Sigma$, then necessarily $g \equiv 0$. First observe the values of g on an edge e: the two derivatives at the endpoints, two function values at the endpoints and the $p-3$ values at the interior Gauss–Lobatto (Fekete) points together constitute $p+1$ parameters that determine a unique one-dimensional polynomial on e. Since all these values are zero, necessarily $g \equiv 0$ on e and in turn on the whole boundary of K_t. It follows from the unisolvency of the Lagrange–Fekete elements that $g \equiv 0$ also in the element interior. ∎

The nodal shape functions for a general polynomial degree p are constructed analogously to the cubic case, by choosing a suitable basis $\mathcal{B} = \{g_1, g_2, \ldots, g_{N_P}\}$ of $P^p(K_t)$ and inverting the corresponding generalized Vandermonde matrix $\{L_i(g_j)\}_{i,j=1}^{N_P}$.

6.4.3 Design of basis functions

Assume a bounded polygonal domain $\Omega_h \subset \mathbb{R}^2$ covered with a finite element mesh $\mathcal{T}_{h,p}$ consisting of $M \geq 1$ Hermite elements of a uniform polynomial degree $p \geq 3$. Then the finite element space $V_{h,p}$ (that for simplicity is not constrained with any essential boundary conditions) has the form

$$\begin{aligned} V_{h,p} \;\; = \;\; & \{v \in C(\overline{\Omega}_h);\ v|_{K_i} \in P^p(K_i); \qquad (6.60) \\ & \frac{\partial v}{\partial x_1} \text{ and } \frac{\partial v}{\partial x_2} \text{ are continuous at every grid vertex}\}. \end{aligned}$$

Proposition 6.1 *The dimension of the space $V_{h,p}$ is*

$$N = dim(V_{h,p}) = 3M_v + (p-3)M_e + \frac{(p-1)(p-2)}{2}M, \qquad (6.61)$$

where M_v is the number of grid vertices and M_e the number of mesh edges.

Proof: There are one Lagrange and two Hermite degrees of freedom associated with each grid vertex, $p-3$ Lagrange degrees of freedom on each edge, and $(p-1)(p-2)/2$ Lagrange degrees of freedom in the interior of each element. Each of these degrees of freedom is represented by one basis function in the basis of $V_{h,p}$. ■

The basis of the space $V_{h,p}$ consists of two types of basis functions:

- Lagrange basis functions associated with the Lagrange degrees of freedom,
- Hermite basis functions representing the partial derivatives at grid vertices.

Lagrange vertex, edge and bubble basis functions

For a polynomial degree $p \geq 3$ there are three Lagrange vertex shape functions associated with the function values at the vertices of K_t, $p-3$ Lagrange edge functions per edge of K_t corresponding to the edge-interior Gauss–Lobatto points, and $(p-1)(p-2)/2$ Lagrange bubble shape functions associated with the Fekete nodal points in the interior of K_t. The fact that the partial derivatives $\partial/\partial\xi_1$ and $\partial/\partial\xi_2$ of the Lagrange shape functions vanish at the vertices of K_t implies that the corresponding Lagrange vertex, edge and bubble basis functions of the space $V_{h,p}$ can be designed in the same fashion as the vertex, edge and bubble basis functions on Lagrange elements.

Hermite vertex basis functions

The design of the Hermite vertex basis functions of the space $V_{h,p}$, which are associated with the partial derivatives of the approximation at grid vertices, is worth discussing in more detail. It is clear that the standard affine reference map $\boldsymbol{x}_K : K_t \to K$ does not preserve the degrees of freedom,

$$\frac{\partial}{\partial x_k}[\varphi \circ \boldsymbol{x}_K^{-1}](\boldsymbol{x}_K(\boldsymbol{\xi})) \neq \frac{\partial}{\partial \xi_k}\varphi(\boldsymbol{\xi})$$

(where φ stands for a Hermite shape function defined in the reference domain K_t). Consider a grid vertex $\boldsymbol{x}_i$ along with the vertex patch (4.14) of all elements adjacent to $\boldsymbol{x}_i$,

$$S(i) = \bigcup_{k \in N(i)} \overline{K}_k,$$

where

$$N(i) = \{k;\ K_k \in \mathcal{T}_{h,p},\ \boldsymbol{x}_i \text{ is a vertex of } K_k\}.$$

There is a pair of Hermite vertex basis functions in $V_{h,p}$ that represent $\partial/\partial x_1$ and $\partial/\partial x_2$ at $\boldsymbol{x}_i$; let us denote them by $v_i^{(1)}$ and $v_i^{(2)}$. These functions are continuous in the whole domain Ω_h and vanish in $\Omega_h \setminus S(i)$. They also vanish at all nodal points in $S(i)$, and their first partial derivatives vanish at all boundary vertices of the patch $S(i)$. At the vertex $\boldsymbol{x}_i$ these functions satisfy

$$\frac{\partial}{\partial x_j} v_i^{(k)}(\boldsymbol{x}_i) = \delta_{jk}, \quad j, k = 1, 2.$$

Assume the restriction of $v_i^{(1)}$ to an element $K \in S(i)$. Let $\varphi_t^{v_m,1}$ and $\varphi_t^{v_m,2}$ be the pair of Hermite vertex functions associated with the corresponding vertex v_m of K_t, in such a way that $\boldsymbol{x}_K(v_m) = \boldsymbol{x}_i$. The trick is to find a linear combination

$$\varphi^{(1)} = \alpha_1 \varphi_t^{v_m,1} + \alpha_2 \varphi_t^{v_m,2}$$

such that

$$v_i^{(1)} = \varphi^{(1)} \circ \boldsymbol{x}_K^{-1}$$

and

$$\frac{\partial}{\partial x_j} v_i^{(1)}(\boldsymbol{x}_i) = \delta_{1j}.$$

The rule (4.20) for the transformation of gradients yields

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \nabla v_i^{(1)}(\boldsymbol{x}_i) = \left(\frac{D\boldsymbol{x}_K}{D\boldsymbol{\xi}}\right)^{-T} \nabla \varphi^{(1)}(v_m)$$

$$= \left(\frac{D\boldsymbol{x}_K}{D\boldsymbol{\xi}}\right)^{-T} \begin{pmatrix} \alpha_1 \underbrace{\frac{\partial \varphi_t^{v_m,1}}{\partial \xi_1}(v_m)}_{1} + \alpha_2 \underbrace{\frac{\partial \varphi_t^{v_m,2}}{\partial \xi_1}(v_m)}_{0} \\ \alpha_1 \underbrace{\frac{\partial \varphi_t^{v_m,1}}{\partial \xi_2}(v_m)}_{0} + \alpha_2 \underbrace{\frac{\partial \varphi_t^{v_m,2}}{\partial \xi_2}(v_m)}_{1} \end{pmatrix} = \left(\frac{D\boldsymbol{x}_K}{D\boldsymbol{\xi}}\right)^{-T} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}.$$

Hence

$$\alpha_1 = \frac{\partial \boldsymbol{x}_{K,1}}{\partial \xi_1}, \quad \alpha_2 = \frac{\partial \boldsymbol{x}_{K,1}}{\partial \xi_2}.$$

The other Hermite vertex basis function $v_i^{(2)}$ corresponding to the grid vertex $\boldsymbol{x}_i$ is constructed analogously in the form

$$v_i^{(2)} = \varphi^{(2)} \circ \boldsymbol{x}_K^{-1}, \quad \varphi^{(2)} = \beta_1 \varphi_t^{v_m,1} + \beta_2 \varphi_t^{v_m,2},$$

so that

$$\frac{\partial}{\partial x_j} v_i^{(2)}(\boldsymbol{x}_i) = \delta_{2j}.$$

Performing analogous calculation as above, we find that the coefficients β_1 and β_2 have to be

$$\beta_1 = \frac{\partial \boldsymbol{x}_{K,2}}{\partial \xi_1}, \quad \beta_2 = \frac{\partial \boldsymbol{x}_{K,2}}{\partial \xi_2}.$$

Thus finally we find that in $K \subset S(i)$ the Hermite vertex basis functions $v_i^{(1)}$ and $v_i^{(2)}$ are defined as

$$\begin{pmatrix} v_i^{(1)} \\ v_i^{(2)} \end{pmatrix} = \left[\left(\frac{\mathrm{D}\boldsymbol{x}_K}{\mathrm{D}\boldsymbol{\xi}} \right) \begin{pmatrix} \varphi_t^{v_m,1} \\ \varphi_t^{v_m,2} \end{pmatrix} \right] \circ \boldsymbol{x}_K^{-1}.$$

6.4.4 Global nodal interpolant and conformity

Let $\Omega_h \subset \mathbb{R}^2$ be a polygonal domain covered with a finite element mesh $\mathcal{T}_{h,p}$ consisting of M triangular (and/or quadrilateral) Hermite elements of a uniform polynomial degree $p \geq 3$. Let $g \in H^2(\Omega_h)$ be a function for which all degrees of freedom on all elements in the mesh are defined. Then the global nodal interpolant $\mathcal{I}(g)$ is constructed routinely via the elementwise local nodal interpolants (3.28),

$$\mathcal{I}_K(g) = \sum_{i=1}^{N_P} L_i(g)\theta_i,$$

where $K \in \mathcal{T}_{h,p}$, $N_P = (p+1)(p+2)/2$ and θ_i are the corresponding nodal shape functions on the element K (see Section 3.3.1).

The continuity of $\mathcal{I}(g)$ at all element vertices is guaranteed by the fact that there is a Lagrange degree of freedom associated with every mesh vertex. On every edge e in the mesh there are two Hermite and two Lagrange degrees of freedom associated with the endpoints of e, and additional $p - 3$ Lagrange degrees of freedom associated with the $p - 3$ Gauss–Lobatto points in the interior of e. These $p + 1$ parameters determine a unique pth-degree polynomial on the edge e. Therefore $\mathcal{I}(g) \in C(\overline{\Omega}_h)$.

In the next section we derive and analyze partial differential equations that describe the bending of elastic plates.

6.5 BENDING OF ELASTIC PLATES

Plates are three-dimensional solids whose thickness is very small compared to their other dimensions. The bending of such structures, and indeed an extension to shells, were the first subjects to which the finite element method was applied in the early 1960s. Usually the complete three-dimensional numerical treatment of such problems is not practical, since the discrete problems are both very large and ill-conditioned.

Therefore, several classical assumptions were introduced long time ago in order to simplify the solution of plate problems. There are two basic plate models, of which the historically older Kirchhoff (thin) plate model, based on the early works of Sophie Germain [24, 119] and [96] in 1811, was completed and formalized by G. Kirchhoff [71] in 1851. The thin plate assumptions were relaxed by E. Reissner [97] in 1945 and in a slightly different way by R. D. Mindlin [81] in 1951. The Reissner–Mindlin plate model extends the field of application to shear-deformable thick plates, and therefore sometimes it is referred to as thick or shear-deformable plate model.

The Reissner–Mindlin plate model is a second-order problem that naturally corresponds to the Timoshenko beam model, while the Kirchhoff model is a fourth-order problem that generalizes the Euler–Bernoulli beam model. Although at the first glance the numerical treatment of the Reissner–Mindlin plate seems to be easier, it turns out that it conceals all basic difficulties plaguing the Kirchhoff model, and in reality its numerical solution is in some sense even more difficult. On the other hand, the variational formulation of the fourth-order Kirchhoff model takes place in the space $H^2(\Omega)$ which is much smaller than the space $H^1(\Omega)$. In order to conform to $H^2(\Omega)$, the approximations have to be once continuously differentiable (see Paragraph A.4.3). Since the assembling procedures for the H^2-conforming elements are nontrivial, nowadays it is popular to resort to mixed methods that lead to the standard H^1-conforming elements (see, e.g., [18, 95] and [124]).

Instead of reviewing existing results on mixed methods, we find it more useful to focus on the application of the less frequently encountered C^1-elements in this text. These elements are natural for the variational setting in the space $H^2(\Omega)$. It was demonstrated in Paragraphs 6.3.7 and 6.4.3 that the key to a transparent element-by-element assembling procedure are the correct transformation relations for the weak forms and shape functions from the mesh element to the reference domain and vice versa. These topics are addressed in Section 6.6.

Prior to introducing the finite elements, in Paragraphs 6.5.1–6.5.4 we present the derivation of the thick and thin plate models, list various types of boundary conditions, construct the variational formulation and discuss the existence and uniqueness of the weak solution. It turns out that the approximation of a smooth boundary via a nonsmooth curve (a common technique for second-order problems) may change the physics of the fourth-order problem completely. This phenomenon, known as the Babuška's paradox of thin plates, is mentioned in Paragraph 6.5.5.

6.5.1 Reissner–Mindlin (thick) plate model

Although the shear-deformable plate model is historically younger than the Kirchhoff model, the natural order of their presentation is opposite. Consider a plate of a constant thickness $t > 0$ whose middle plane coincides with the (x_1x_2)-plane and whose projection to the (x_1x_2)-plane occupies a bounded domain $\Omega \subset \mathbb{R}^2$ with Lipschitz boundary. Thus the three-dimensional body of the plate is $\Omega \times (-t/2, t/2)$. We assume that the plate is subject to external forces which are normal to its middle plane $x_3 = 0$. The Reissner–Mindlin model is based on the following four postulates:

- (P1) Planar cross-sections normal to the middle plane remain planes during the deformation, and segments lying on normals to the middle surface are deformed linearly.

- (P2) The displacement in the x_3-direction does not depend on the x_3-coordinate.

- (P3) The displacement of points lying on the middle plane occurs in the x_3-direction only.

- (P4) The normal stress $\sigma_{33} = 0$.

The postulate (P1) is the most important assumption of the theory of plates and shells. It follows from (P1)–(P3) that the displacement $\boldsymbol{u}(\boldsymbol{x}) = (u_1, u_2, u_3)^T(\boldsymbol{x})$ has the form

$$\begin{aligned} u_1(\boldsymbol{x}) &= -x_3\phi_1(x_1, x_2), \\ u_2(\boldsymbol{x}) &= -x_3\phi_2(x_1, x_2), \\ u_3(\boldsymbol{x}) &= w(x_1, x_2), \end{aligned} \tag{6.62}$$

where w is the transversal displacement or (normal) deflection, and $\phi = (\phi_1, \phi_2)$ is the rotation of the transverse normal vector. Let us introduce the equations governing the quantities w, ϕ_1 and ϕ_2:

The bending strains associated with the displacement field (6.62) have the form

$$\begin{aligned} \epsilon_1 &= \frac{\partial u_1}{\partial x_1} = -x_3\frac{\partial \phi_1}{\partial x_1}, \\ \epsilon_2 &= \frac{\partial u_2}{\partial x_2} = -x_3\frac{\partial \phi_2}{\partial x_2}, \\ \gamma_{12} &= \frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1} = -x_3\left(\frac{\partial \phi_1}{\partial x_2} + \frac{\partial \phi_2}{\partial x_1}\right), \\ \gamma_{13} &= \frac{\partial w}{\partial x_1} + \frac{\partial u_1}{\partial x_3} = \frac{\partial w}{\partial x_1} - \phi_1, \\ \gamma_{23} &= \frac{\partial w}{\partial x_2} + \frac{\partial u_2}{\partial x_3} = \frac{\partial w}{\partial x_2} - \phi_2, \\ \epsilon_3 &= 0. \end{aligned} \tag{6.63}$$

Here ϵ_1, ϵ_2, and γ_{12} are the in-plane strain components, and

$$(\gamma_{13}, \gamma_{23}) = \nabla w - (\theta_1, \theta_2) \tag{6.64}$$

are the strain components corresponding to the transverse shear. [This relation models the shear-deformability in the Reissner–Mindlin model. In the Kirchhoff model, the left-hand side of (6.64) is zero]. In the linear isotropic case the shear force resultants have the form

$$\begin{aligned} Q_1 &= \int_{-t/2}^{t/2} \tau_{13}\,\mathrm{d}x_3 = \kappa Gt\left(\frac{\partial w}{\partial x_1} - \phi_1\right), \\ Q_2 &= \int_{-t/2}^{t/2} \tau_{23}\,\mathrm{d}x_3 = \kappa Gt\left(\frac{\partial w}{\partial x_2} - \phi_2\right), \end{aligned} \tag{6.65}$$

where $\{\tau_{ij}\}_{i,j=1}^{3}$ is the stress tensor and G the shear elasticity modulus. Directional shear rigidities G_{13} and G_{23} may be used instead of G to model linear orthotropic elasticity (see, e.g., [95]). The constant κ is a shear correction coefficient introduced to account for the fact that the shear stresses are not constant across the section. A value of $\kappa = 5/6$ is exact for a rectangular, homogeneous section and it corresponds to a parabolic shear stress distribution.

Using appropriate constitutive relations, all momentum components can be related to displacement derivatives. In the linear elastic case the bending moments M_{11} and M_{22}, and the twisting moment $M_{12} = M_{21}$ are defined as

$$\begin{aligned} M_{11} &= \int_{-t/2}^{t/2} \tau_{11} x_3 \, \mathrm{d}x_3 = -D \left(\frac{\partial \phi_1}{\partial x_1} + \gamma \frac{\partial \phi_2}{\partial x_2} \right), \qquad (6.66) \\ M_{22} &= \int_{-t/2}^{t/2} \tau_{22} x_3 \, \mathrm{d}x_3 = -D \left(\nu \frac{\partial \phi_1}{\partial x_1} + \frac{\partial \phi_2}{\partial x_2} \right), \\ M_{12} = M_{21} &= \int_{-t/2}^{t/2} \tau_{12} x_3 \, \mathrm{d}x_3 = -\frac{D(1-\gamma)}{2} \left(\frac{\partial \phi_1}{\partial x_2} + \frac{\partial \phi_2}{\partial x_1} \right). \end{aligned}$$

Here D is the bending stiffness, defined by

$$D = \frac{Et^3}{12(1-\gamma^2)}, \qquad (6.67)$$

where E is the direct (in-plane) elasticity modulus and γ the Poisson's ratio. In the case of a linear orthotropic material, the elasticity modulus has two directional components E_1 and E_2, and the Poisson's ratio γ is replaced with the directional values γ_{12}, γ_{21}. Accordingly, the bending stiffness D has the components

$$\begin{aligned} D_{11} &= \frac{E_1 t^3}{12(1-\gamma_{12}\gamma_{21})}, \\ D_{12} &= \frac{\gamma_{12} E_2 t^3}{12(1-\gamma_{12}\gamma_{21})}, \\ D_{22} &= \frac{E_2 t^3}{12(1-\gamma_{12}\gamma_{21})}, \\ D_{66} &= \frac{1}{12} G_{12} t^3, \end{aligned}$$

and the moments attain the form

$$\begin{aligned} M_1 &= -\left(D_{11} \frac{\partial \phi_1}{\partial x_1} + D_{12} \frac{\partial \phi_2}{\partial x_2} \right), \\ M_2 &= -\left(D_{12} \frac{\partial \phi_1}{\partial x_1} + D_{22} \frac{\partial \phi_2}{\partial x_2} \right), \\ M_{12} &= -2 D_{66} \left(\frac{\partial \phi_1}{\partial x_2} + \frac{\partial \phi_2}{\partial x_1} \right). \end{aligned}$$

See, e.g., [95] for details. The Reissner–Mindlin model consists of three equilibrium equations that relate the transversal force f to the shear resultants, and the shear resultants to the momentum components. To begin with, the equilibrium equation

$$\int_{-t/2}^{t/2} \left(\frac{\partial \tau_{13}}{\partial x_1} + \frac{\partial \tau_{23}}{\partial x_2} + \frac{\partial \tau_{33}}{\partial x_3} \right) \mathrm{d}x_3 = 0,$$

written in the form

$$\frac{\partial}{\partial x_1}\int_{-t/2}^{t/2}\tau_{13}\,\mathrm{d}x_3+\frac{\partial}{\partial x_2}\int_{-t/2}^{t/2}\tau_{23}\,\mathrm{d}x_3+\underbrace{\tau_{33}|_{t/2}-\tau_{33}|_{-t/2}}_{f}=0,$$

(where the transverse loading f arises from the resultant of the normal traction on the top and bottom surfaces) yields via (6.65) the first thick plate equation,

$$\frac{\partial Q_1}{\partial x_1}+\frac{\partial Q_2}{\partial x_2}+f=0. \tag{6.68}$$

The momentum equilibrium conditions

$$\int_{-t/2}^{t/2} z\left(\frac{\partial\tau_{11}}{\partial x_1}+\frac{\partial\tau_{12}}{\partial x_2}+\frac{\partial\tau_{13}}{\partial x_3}\right)\mathrm{d}x_3=0,$$
$$\int_{-t/2}^{t/2} z\left(\frac{\partial\tau_{21}}{\partial x_1}+\frac{\partial\tau_{22}}{\partial x_2}+\frac{\partial\tau_{23}}{\partial x_3}\right)\mathrm{d}x_3=0,$$

assumed together with relations (6.65) and (6.66), complete the thick plate model by the relations

$$\begin{aligned}\frac{\partial M_{11}}{\partial x_1}+\frac{\partial M_{12}}{\partial x_2}-Q_1&=0,\\ \frac{\partial M_{12}}{\partial x_1}+\frac{\partial M_{22}}{\partial x_2}-Q_2&=0.\end{aligned} \tag{6.69}$$

Some of the above-defined quantities are depicted in Figure 6.23.

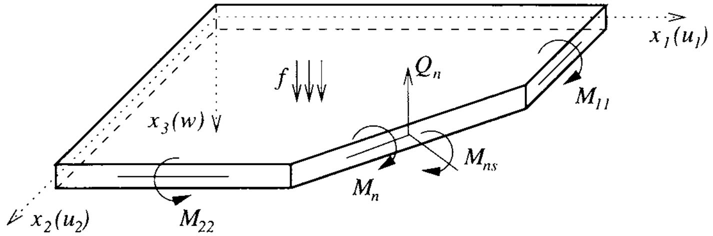


Figure 6.23 The transversal force, shear resultant, and bending and twisting moments.

Further details on the derivation of the thick plate model can be found, e.g., in [124]. In the next paragraph, equations (6.68), (6.69) will be used to deduce the Kirchhoff thin plate model.

6.5.2 Kirchhoff (thin) plate model

In addition to the hypotheses (P1)–(P4) of the Reissner–Mindlin plate model, the Kirchhoff model imposes the normal (Love's, Kirchhoff's) hypothesis, which is analogous to the basic assumption of the Euler–Bernoulli beam theory (Paragraph 6.1.1):

- (P5) Vectors which are normal to the middle surface $x_3 = 0$ remain normal to the (deformed) middle surface during the deformation.

This assumption neglects the shear deformations γ_{13} and γ_{23} in (6.64), and thus the rotations ϕ_1 and ϕ_2 are related to the partial derivatives of the normal deflection w via the relation

$$\phi(x_1, x_2) = \nabla w(x_1, x_2). \tag{6.70}$$

The situation is depicted in Figure 6.24.

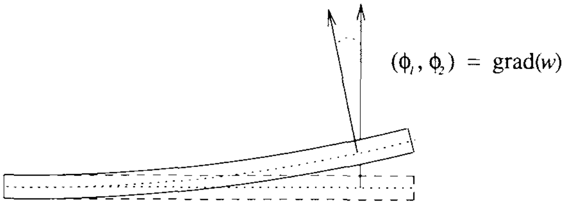


Figure 6.24 The hypothesis (P5) relates the rotations ϕ_1, ϕ_2 to the deflection w via its gradient.

With (6.70) the displacements (6.62) take the form

$$\begin{aligned} u_1(\boldsymbol{x}) &= -x_3\frac{\partial w}{\partial x_1}(x_1, x_2), \\ u_2(\boldsymbol{x}) &= -x_3\frac{\partial w}{\partial x_2}(x_1, x_2), \\ u_3(\boldsymbol{x}) &= w(x_1, x_2), \end{aligned} \tag{6.71}$$

and (6.66) yields the momentum components

$$\begin{aligned} M_{11} &= -D\left(\frac{\partial^2 w}{\partial x_1^2} + \nu\frac{\partial^2 w}{\partial x_2^2}\right), \\ M_{22} &= -D\left(\nu\frac{\partial^2 w}{\partial x_1^2} + \frac{\partial^2 w}{\partial x_2^2}\right), \\ M_{12} = M_{21} &= -D(1-\gamma)\frac{\partial^2 w}{\partial x_1 x_2}. \end{aligned} \tag{6.72}$$

Substituting into (6.69), we obtain

$$\begin{aligned} D\frac{\partial}{\partial x_1}\left(\frac{\partial^2 w}{\partial x_1^2} + \nu\frac{\partial^2 w}{\partial x_2^2}\right) + D(1-\gamma)\frac{\partial}{\partial x_2}\left(\frac{\partial^2 w}{\partial x_1 x_2}\right) + Q_1 &= 0, \\ D(1-\gamma)\frac{\partial}{\partial x_1}\left(\frac{\partial^2 w}{\partial x_1 x_2}\right) + D\frac{\partial}{\partial x_2}\left(\nu\frac{\partial^2 w}{\partial x_1^2} + \frac{\partial^2 w}{\partial x_2^2}\right) + Q_2 &= 0. \end{aligned} \tag{6.73}$$

Substituting the shear resultants Q_1, Q_2 from (6.73) into (6.68), we finally obtain the Kirchhoff thin plate model in its well-known form,

$$D\left(\frac{\partial^4 w}{\partial x_1^4} + 2\frac{\partial^4 w}{\partial x_1^2 x_2^2} + \frac{\partial^4 w}{\partial x_2^4}\right) = D\Delta^2 w = f. \tag{6.74}$$

In the following we introduce several types of boundary conditions for equation (6.74), derive its weak formulation, and prove the existence and uniqueness of the weak solution.

6.5.3 Boundary conditions

The boundary conditions for plates are typically prescribed in local $(\boldsymbol{\nu}, \boldsymbol{s})$ coordinates, where $\boldsymbol{\nu} = (\nu_1, \nu_2)$ and $\boldsymbol{s} = (s_1, s_2)$ are the unit normal and tangential vectors to the boundary $\partial\Omega$, respectively. The shear force resultants Q_i and the moments M_{ij} in the Reissner–Mindlin model were defined by (6.65) and (6.66),

$$\begin{aligned} Q_1 &= \kappa G t \left(\frac{\partial w}{\partial x_1} - \phi_1 \right), \\ Q_2 &= \kappa G t \left(\frac{\partial w}{\partial x_2} - \phi_2 \right), \\ M_{11} &= -D \left(\frac{\partial \phi_1}{\partial x_1} + \gamma \frac{\partial \phi_2}{\partial x_2} \right), \\ M_{22} &= -D \left(\gamma \frac{\partial \phi_1}{\partial x_1} + \frac{\partial \phi_2}{\partial x_2} \right), \\ M_{12} = M_{21} &= -\frac{D(1-\gamma)}{2} \left(\frac{\partial \phi_1}{\partial x_2} + \frac{\partial \phi_2}{\partial x_1} \right). \end{aligned} \tag{6.75}$$

In the Kirchhoff model we have relations (6.69) and (6.72),

$$\begin{aligned} Q_1 &= \frac{\partial M_{11}}{\partial x_1} + \frac{\partial M_{12}}{\partial x_2}, \\ Q_2 &= \frac{\partial M_{12}}{\partial x_1} + \frac{\partial M_{22}}{\partial x_2}, \\ M_{11} &= -D \left(\frac{\partial^2 w}{\partial x_1^2} + \gamma \frac{\partial^2 w}{\partial x_2^2} \right), \\ M_{22} &= -D \left(\gamma \frac{\partial^2 w}{\partial x_1^2} + \frac{\partial^2 w}{\partial x_2^2} \right), \\ M_{12} = M_{21} &= -D(1-\gamma) \frac{\partial^2 w}{\partial x_1 x_2}. \end{aligned} \tag{6.76}$$

These quantities are transformed into the local coordinates as follows:

$$\begin{aligned} M_i &= \sum_{j=1}^{2} M_{ij} \nu_j, \quad j = 1, 2, \\ M_\nu &= \sum_{i,j=1}^{2} M_{ij} \nu_i \nu_j = M_1 \nu_1 + M_2 \nu_2, \\ Q_\nu &= Q_1 \nu_1 + Q_2 \nu_2, \\ \phi_\nu &= \phi_1 \nu_1 + \phi_2 \nu_2, \\ \phi_s &= \phi_2 \nu_1 - \phi_1 \nu_2, \\ M_{\nu s} &= (M_{22} - M_{11}) \nu_1 \nu_2 + M_{12} (\nu_1^2 - \nu_2^2) = M_2 \nu_1 - M_1 \nu_2. \end{aligned} \tag{6.77}$$

The transformation rule for the partial derivatives is

$$\frac{\partial}{\partial \boldsymbol{\nu}} = \nu_1 \frac{\partial}{\partial x_1} + \nu_2 \frac{\partial}{\partial x_2}, \quad \frac{\partial}{\partial \boldsymbol{s}} = \nu_1 \frac{\partial}{\partial x_2} - \nu_2 \frac{\partial}{\partial x_1}.$$

Clamped boundary Clamped boundary conditions are used when the transversal deflection w and both the rotations ϕ_ν and ϕ_s are given. This is the case, for example, when a portion of the plate is built into a solid wall. In the Reissner–Mindlin model one prescribes

$$\begin{aligned} w &= w^*, \\ \phi_\nu &= \phi_\nu^*, \\ \phi_s &= \phi_s^*. \end{aligned} \tag{6.78}$$

In the Kirchhoff model the tangential rotation ϕ_s is determined uniquely by the values of w on the boundary,

$$\phi_s = \frac{\partial w}{\partial s},$$

and moreover, the normal rotation ϕ_ν is related to the normal derivative of w via (6.70),

$$\phi_\nu = \phi_1\nu_1 + \phi_2\nu_2 = \frac{\partial w}{\partial x_1}\nu_1 + \frac{\partial w}{\partial x_2}\nu_2 = \frac{\partial w}{\partial \boldsymbol{\nu}}.$$

Therefore for thin plates one prescribes the deflection w and its normal derivative,

$$\begin{aligned} w &= w^*, \\ \frac{\partial w}{\partial \boldsymbol{\nu}} &= \left(\frac{\partial w}{\partial \boldsymbol{\nu}}\right)^* = \phi_\nu^*. \end{aligned} \tag{6.79}$$

Traction boundary The Reissner–Mindlin model admits the prescription of the moments $M_\nu, M_{\nu s}$ and the stress resultant Q_ν,

$$\begin{aligned} M_\nu &= M_\nu^*, \\ M_{\nu s} &= M_{\nu s}^*, \\ Q_\nu &= Q_\nu^*. \end{aligned} \tag{6.80}$$

An important special case of the traction boundary is free boundary with

$$\begin{aligned} M_\nu &= 0, \\ M_{\nu s} &= 0, \\ Q_\nu &= 0. \end{aligned}$$

It is easy to see that these three quantities are linearly dependent in the Kirchhoff model, and therefore only two conditions are prescribed. Usually these are

$$\begin{aligned} M_\nu &= M_\nu^*, \\ Q_\nu + \frac{\partial M_{\nu s}}{\partial s} &= Q_\nu^* + \frac{\partial M_{\nu s}^*}{\partial s} = V^*. \end{aligned} \tag{6.81}$$

The latter quantity is usually called effective shearing force of the plate.

Simply supported boundary These boundary condition combine both the fixed and traction boundary conditions in order to model situations when the plate lies on a solid support (with unknown values of the rotations or $\partial w/\partial \boldsymbol{\nu}$ on the boundary), etc. We distinguish between hard- and soft-supported boundary.

Hard-supported boundary: The Reissner–Mindlin model admits the prescription of w, ϕ_ν, and M_ν,

$$\begin{aligned} w &= w^*, \\ \phi_\nu &= \phi_\nu^*, \\ M_\nu &= M_\nu^*. \end{aligned} \tag{6.82}$$

In the Kirchhoff model by (6.70) we have $\phi_\nu = \partial w/\partial \boldsymbol{\nu}$, which means that the prescription of ϕ_ν on the boundary would lead to a clamped case (with M_ν not prescribed). But we may give M_ν without constraining ϕ_ν, which leads to the boundary conditions

$$\begin{aligned} w &= w^*, \\ M_\nu &= M_\nu^*. \end{aligned} \tag{6.83}$$

Soft-supported boundary: In the thick plate model one prescribes w together with two traction boundary conditions for M_ν and $M_{\nu s}$,

$$\begin{aligned} w &= w^*, \\ M_\nu &= M_\nu^*, \\ M_{\nu s} &= M_{\nu s}^*. \end{aligned} \tag{6.84}$$

In the Kirchhoff model one only prescribes two conditions, usually w and the normal moment M_ν,

$$\begin{aligned} w &= w^*, \\ M_\nu &= M_\nu^* \end{aligned} \tag{6.85}$$

(i.e., the hard- and soft-support boundary conditions are the same for thin plates).

6.5.4 Weak formulation and unique solvability

In this paragraph we consider the Kirchhoff thin plate model. Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with Lipschitz-continuous boundary that is split into three open (not necessarily connected) disjoint parts Γ_{cl}, Γ_{ss}, and Γ_{tr}, as shown in Figure 6.25. The boundary part Γ_{tr} can be empty, and also at most one of the remaining two parts Γ_{cl} and Γ_{ss} can be empty as long as Γ_{ss} is not contained in a single line. These conditions are among the assumptions of the existence and uniqueness Theorem 6.1.

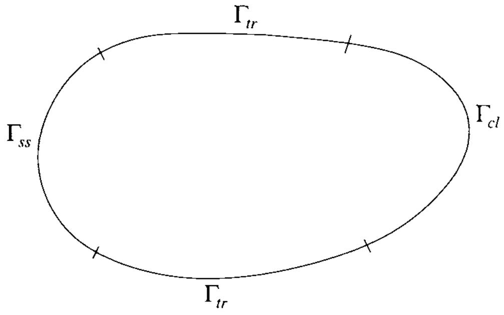


Figure 6.25 The boundary is split into three parts Γ_{cl}, Γ_{ss}, and Γ_{tr}, representing the clamped, simply supported, and traction boundary conditions.

We consider the (essential) clamped and simply-supported boundary conditions

$$\begin{aligned} w(\boldsymbol{x}) &= w^* \quad \text{on } \Gamma_{ss}, \\ w(\boldsymbol{x}) &= w^* \quad \text{on } \Gamma_{cl}, \\ \frac{\partial w}{\partial \boldsymbol{\nu}}(\boldsymbol{x}) &= \left(\frac{\partial w}{\partial \boldsymbol{\nu}}\right)^*(\boldsymbol{x}) = \phi_\nu^*(\boldsymbol{x}) \quad \text{on } \Gamma_{cl} \end{aligned} \tag{6.86}$$

and the (natural) traction boundary conditions

$$\begin{aligned} Q_\nu + \frac{\partial M_{\nu s}}{\partial s} &= Q_\nu^* + \frac{\partial M_{\nu s}^*}{\partial s} \text{ and } M_\nu = M_\nu^* \quad \text{on } \Gamma_{tr}, \\ M_\nu &= M_\nu^* \quad \text{on } \Gamma_{ss}. \end{aligned} \tag{6.87}$$

The space where the weak formulation takes place is the corresponding subspace of the Sobolev space $H^2(\Omega)$,

$$V(\Omega) = \{w \in H^2(\Omega);\ w|_{\Gamma_{ss}} = 0;\ w|_{\Gamma_{cl}} = (\partial w/\partial \boldsymbol{\nu})|_{\Gamma_{cl}} = 0\}.$$

As usual we choose some sufficiently regular Dirichlet lift $G(\boldsymbol{x})$ representing the essential boundary conditions (6.86), i.e.,

$$\begin{aligned} G(\boldsymbol{x}) &= w^* \quad \text{on } \Gamma_{ss}, \\ G(\boldsymbol{x}) &= w^* \quad \text{on } \Gamma_{cl}, \\ \frac{\partial G}{\partial \boldsymbol{\nu}}(\boldsymbol{x}) &= \phi_\nu^*(\boldsymbol{x}) \quad \text{on } \Gamma_{cl}. \end{aligned}$$

The solution $w(\boldsymbol{x})$ is sought in the form

$$w(\boldsymbol{x}) = W(\boldsymbol{x}) + G(\boldsymbol{x}), \tag{6.88}$$

where the unknown function $W \in V(\Omega)$ satisfies the homogeneous boundary conditions

$$\begin{aligned} W(\boldsymbol{x}) &= 0 \quad \text{on } \Gamma_{ss}, \\ W(\boldsymbol{x}) &= 0 \quad \text{on } \Gamma_{cl}, \\ \frac{\partial W}{\partial \boldsymbol{\nu}}(\boldsymbol{x}) &= 0 \quad \text{on } \Gamma_{cl}. \end{aligned} \tag{6.89}$$

It is advantageous to develop the weak formulation from the equation

$$-\left(\frac{\partial^2 M_{11}}{\partial x_1^2} + 2\frac{\partial^2 M_{12}}{\partial x_1 \partial x_2} + \frac{\partial^2 M_{22}}{\partial x_2^2}\right) = f, \tag{6.90}$$

which is equivalent to (6.74) through (6.68) and (6.76). We multiply (6.90) with a test function $\psi(\boldsymbol{x}) \in V(\Omega)$, whose minimum regularity will be determined later from the final weak forms, and integrate over Ω,

$$-\int_\Omega \frac{\partial}{\partial x_1}\underbrace{\left(\frac{\partial M_{11}}{\partial x_1} + \frac{\partial M_{12}}{\partial x_2}\right)}_{Q_1}\psi + \frac{\partial}{\partial x_2}\underbrace{\left(\frac{\partial M_{12}}{\partial x_1} + \frac{\partial M_{22}}{\partial x_2}\right)}_{Q_2}\psi \, \mathrm{d}\boldsymbol{x} = \int_\Omega f\psi \, \mathrm{d}\boldsymbol{x}.$$

Green's theorem yields

$$\int_\Omega \left(\frac{\partial M_{11}}{\partial x_1} + \frac{\partial M_{12}}{\partial x_2}\right) \frac{\partial \psi}{\partial x_1}\nu_1 + \left(\frac{\partial M_{12}}{\partial x_1} + \frac{\partial M_{22}}{\partial x_2}\right) \frac{\partial \psi}{\partial x_2}\nu_2 \, \mathrm{d}\boldsymbol{x}$$
$$- \int_{\Gamma_{tr}} Q_\nu \psi \, \mathrm{d}S = \int_\Omega f\psi \, \mathrm{d}\boldsymbol{x}.$$

Applying Green's theorem to the first integral once more, we obtain

$$- \int_\Omega M_{11}\frac{\partial^2 \psi}{\partial x_1^2} + 2M_{12}\frac{\partial^2 \psi}{\partial x_1 \partial x_2} + M_{22}\frac{\partial^2 \psi}{\partial x_2^2} \, \mathrm{d}\boldsymbol{x} \tag{6.91}$$
$$+ \int_{\partial\Omega} \underbrace{(M_{11}\nu_1 + M_{12}\nu_2)}_{M_1} \frac{\partial \psi}{\partial x_1} + \underbrace{(M_{12}\nu_1 + M_{22}\nu_2)}_{M_2} \frac{\partial \psi}{\partial x_2} \, \mathrm{d}S$$
$$- \int_{\Gamma_{tr}} Q_\nu \psi \, \mathrm{d}S = \int_\Omega f\psi \, \mathrm{d}\boldsymbol{x}.$$

The transformation relations (6.77) yield

$$\int_{\partial\Omega} M_1 \frac{\partial \psi}{\partial x_1} + M_2 \frac{\partial \psi}{\partial x_2} \, \mathrm{d}S = \int_{\partial\Omega} M_\nu \frac{\partial \psi}{\partial \boldsymbol{\nu}} + M_{\nu s}\frac{\partial \psi}{\partial \boldsymbol{s}} \, \mathrm{d}S.$$

Inserting this relation into (6.91), we obtain

$$- \int_\Omega M_{11}\frac{\partial^2 \psi}{\partial x_1^2} + 2M_{12}\frac{\partial^2 \psi}{\partial x_1 \partial x_2} + M_{22}\frac{\partial^2 \psi}{\partial x_2^2} \, \mathrm{d}\boldsymbol{x} \tag{6.92}$$
$$+ \int_{\partial\Omega} M_\nu \frac{\partial \psi}{\partial \boldsymbol{\nu}} + M_{\nu s}\frac{\partial \psi}{\partial \boldsymbol{s}} \, \mathrm{d}S. - \int_{\Gamma_{tr}} Q_\nu \psi \, \mathrm{d}S = \int_\Omega f\psi \, \mathrm{d}\boldsymbol{x}.$$

Last, in order to prepare the weak formulation for the incorporation of the boundary data $V^* = Q_\nu^* + \partial M_{\nu s}^*/\partial s$ (part of the traction and soft support boundary conditions), we need to include the quantity $\partial M_{\nu s}/\partial s$ into the boundary integrals explicitly. This operation requires the application of Green's theorem along the boundary $\partial\Omega$, which in turn brings up some extra considerations about the smoothness of the boundary $\partial\Omega$.

When the boundary $\partial\Omega$ is a smooth curve and when the function $M_{\nu s} = M_{\nu s}(s)$ is continuously differentiable in the interval $[a, b]$ which is supposed to correspond to the part Γ_{tr} of the boundary $\partial\Omega$, it holds

$$\int_{\Gamma_{tr}} M_{\nu s}\frac{\partial \psi}{\partial \boldsymbol{s}} \, \mathrm{d}S = - \int_{\Gamma_{tr}} \frac{\partial M_{\nu s}}{\partial \boldsymbol{s}} \psi \, \mathrm{d}S + [M_{\nu s}\psi]_a^b.$$

Similarly, if $M_{\nu s}$ is continuously differentiable in the whole $\partial\Omega$, then

$$- \int_{\partial\Omega} M_{\nu s}\frac{\partial \psi}{\partial \boldsymbol{s}} \, \mathrm{d}S = \int_{\partial\Omega} \frac{\partial M_{\nu s}}{\partial \boldsymbol{s}} \psi \, \mathrm{d}S = \int_{\Gamma_{tr}} \frac{\partial M_{\nu s}}{\partial \boldsymbol{s}} \psi \, \mathrm{d}S.$$

However, in reality the boundary $\partial\Omega$ often has corners. Suppose that the boundary $\partial\Omega$ is defined parameterically by means of the parameter $s \in (0, l)$, and that there are N_c corner points $0 < s_i < l$, $i = 1, 2, \ldots, N_c$. In general the function $M_{\nu s}(s)$ has jumps at these points since the normal vector $\boldsymbol{\nu}$ varies discontinuously. The integration by parts with a piecewise continuous function $M_{\nu s}(s)$ gives (we assume $M_{\nu s}(0) = M_{\nu s}(l)$):

$$-\int_{\partial\Omega} M_{\nu s}\frac{\partial\psi}{\partial s}\,\mathrm{d}S = -\int_0^l M_{\nu s}\frac{\partial\psi}{\partial s}\,\mathrm{d}s$$

$$= \int_0^l \frac{\partial M_{\nu s}}{\partial s}\psi\,\mathrm{d}s + \sum_{i=1}^{N_c}(M_{\nu s}(s_i+0) - M_{\nu s}(s_i-0))\psi(s_i).$$

Suppose, moreover, that the given twisting moment $M^*_{\nu s}(s)$ has jumps at the corner points $s = s_j$, where $0 < s_j < b$, $j = 1, 2, \ldots, p$ ($p \le N_c, b \le l, s \in (0, b)$ on Γ_{tr}). Then

$$\int_{\Gamma_{tr}} M^*_{\nu s}\frac{\partial\psi}{\partial s}\,\mathrm{d}S = \int_0^b M^*_{\nu s}\frac{\partial\psi}{\partial s}\,\mathrm{d}s$$

$$= -\int_0^b \frac{\partial M^*_{\nu s}}{\partial s}\psi\,\mathrm{d}s - \sum_{j=1}^{p}(M^*_{\nu s}(s_j+0) - M^*_{\nu s}(s_j-0))\psi(s_j) + [M^*_{\nu s}\psi]_0^b.$$

Writing $M^*_{\nu s}(s_j+0) - M^*_{\nu s}(s_j-0) = h_j$, we see that the corner discontinuities in $M_{\nu s}$ and $M^*_{\nu s}$ produce the following terms,

$$\sum_{j=1}^{N_c}(M_{\nu s}(s_i+0) - M_{\nu s}(s_i-0))\psi(s_i) - \sum_{j=1}^{p} h_j\psi(s_j),$$

which are not present when the boundary $\partial\Omega$ is smooth. Taking this fact into account, from (6.92) we obtain

$$-\int_\Omega M_{11}\frac{\partial^2\psi}{\partial x_1^2} + 2M_{12}\frac{\partial^2\psi}{\partial x_1\partial x_2} + M_{22}\frac{\partial^2\psi}{\partial x_2^2}\,\mathrm{d}\boldsymbol{x} \tag{6.93}$$

$$= \int_\Omega f\psi\,\mathrm{d}\boldsymbol{x} + \int_{\Gamma_{tr}}\left(Q^*_\nu + \frac{\partial M^*_{\nu s}}{\partial s}\right)\psi - M^*_\nu\frac{\partial\psi}{\partial\boldsymbol{\nu}}\,\mathrm{d}S - \int_{\Gamma_{ss}} M^*_\nu\frac{\partial\psi}{\partial\boldsymbol{\nu}}\,\mathrm{d}S + \sum_{j=1}^{p} h_j\psi(s_j).$$

The boundary integral over Γ_{ss} only contains M^*_ν, because $\psi \equiv 0$ on Γ_{ss}, and no boundary integral over Γ_{cl} is present since both ψ and $\partial\psi/\partial\boldsymbol{\nu}$ are zero on Γ_{cl}. The Dirichlet lift $G(\boldsymbol{x})$ is implemented into the left-hand side in the usual way, by decomposing the moments M_{ij} into

$$M_{ij}(w) = M_{ij}(W + G) = M_{ij}(W) + M_{ij}(G),\ \ i, j = 1, 2,$$

and leading all integrals containing $M_{ij}(G)$ over to the right-hand side. The weak formulation of equation (6.90) reads: Find a function $W \in V(\Omega)$ such that

$$a(W, \psi) = l(\psi)\ \text{ for all } \psi \in V, \tag{6.94}$$

where

$$a(W, \psi) = -\int_\Omega M_{11}(W)\frac{\partial^2\psi}{\partial x_1^2} + 2M_{12}(W)\frac{\partial^2\psi}{\partial x_1\partial x_2} + M_{22}(W)\frac{\partial^2\psi}{\partial x_2^2}\,\mathrm{d}\boldsymbol{x},$$

and

$$
\begin{aligned}
l(\psi) = \int_{\Omega} f\psi \,\mathrm{d}\boldsymbol{x} + \int_{\Gamma_{tr}} \left(Q_{\nu}^{*} + \frac{\partial M_{\nu s}^{*}}{\partial s}\right)\psi - M_{\nu}^{*}\frac{\partial \psi}{\partial \boldsymbol{\nu}}\,\mathrm{d}S \\
- \int_{\Gamma_{ss}} M_{\nu}^{*}\frac{\partial \psi}{\partial \boldsymbol{\nu}}\,\mathrm{d}S + \sum_{j=1}^{p} h_j \psi(s_j) \\
+ \int_{\Omega} M_{11}(G)\frac{\partial^2 \psi}{\partial x_1^2} + 2M_{12}(G)\frac{\partial^2 \psi}{\partial x_1 \partial x_2} + M_{22}(G)\frac{\partial^2 \psi}{\partial x_2^2}\,\mathrm{d}\boldsymbol{x}.
\end{aligned}
$$

The weak solution satisfying both the essential boundary conditions (6.86) and the natural boundary conditions (6.87) is, as usual, $w = W + G$.

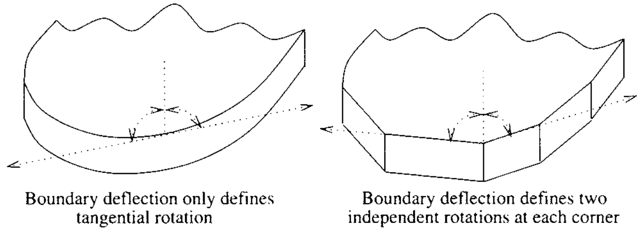
The important question is now whether the bilinear form $a(\cdot,\cdot) : V(\Omega) \times V(\Omega) \to \mathbb{R}$ is bounded and V-elliptic, and whether $l : V(\Omega) \to \mathbb{R}$ is a bounded linear form.

Theorem 6.1 (Unique solvability) *By $C^0(\Omega)$ denote the space of continuous functions with compact support lying in Ω and let $[C^0(\Omega)]'$ be its dual (i.e., the space of all continuous linear forms over $C^0(\Omega)$). Suppose that $f \in [C^0(\Omega)]'$, $Q_{\nu}^{*} + \partial M_{\nu s}^{*}/\partial s \in L^1(\Gamma_{tr})$ and $M_{\nu}^{*} \in L^q(\Gamma_{ss} \cup \Gamma_{tr})$, $1 < q < \infty$. Let at least one of Γ_{cl} and Γ_{ss} be nonempty, and if Γ_{cl} is empty, then let Γ_{ss} not be contained in a single straight line. Then there exists a unique weak solution to (6.94).*

Proof: The proof requires to verify the boundedness of the forms a and l, and to prove the V-ellipticity of the form a in $V(\Omega) \times V(\Omega)$. This is done via the Korn inequalities, which lie beyond the scope of this introductory text. See, e.g., the nice monograph by Nečas and Hlaváček [86], Theorem 4.1, for the proof. ■

6.5.5 Babuška's paradox of thin plates

Let Γ be a nonempty subset of $\partial\Omega$. As mentioned above, the thin plate assumption (P5) implies that the prescription of w on Γ defines ϕ_s on Γ. Thus whenever Γ contains a corner and w is prescribed, this yields two independent rotations ϕ_s^- and ϕ_s^+ at both sides of the corner, which define both ϕ_ν and ϕ_s (and consequently ϕ_ν, ϕ_s and a fixed boundary condition at the corner). The situation is depicted in Figure 6.26.


Boundary deflection only defines tangential rotation

Boundary deflection defines two independent rotations at each corner

Figure 6.26 Nonsmooth approximation of a smooth boundary changes the physics of the thin plate problem when the deflection w is prescribed.

The approximation of a smooth boundary Γ by means of a nonsmooth curve Γ_h changes simply-supported boundary to fixed boundary at the corners. Consequently, the numerical scheme does not converge to the exact solution to the original problem. For more details see [7] and [8].

6.6 DISCRETIZATION BY H^2-CONFORMING ELEMENTS

The weak formulation of plate bending problems takes place in the Sobolev space $H^2(\Omega_h)$. In this section we discuss the H^2-conforming Argyris elements which are the most natural choice for their discretization. For alternative mixed methods leading to simpler (but not necessarily more efficient) discretizations based on standard H^1-conforming elements see, e.g., [18, 95] and [124].

6.6.1 Lowest-order (quintic) Argyris element, unisolvency

For spatial reasons let us restrict our discussion to triangular elements. The basic Argyris triangle is a quintic element (K, P, Σ), where K is a triangular domain, $P = P^5(K)$ and the set $\Sigma = \{L_1, L_2, \dots, L_{21}\}$ comprises the degrees of freedom depicted in Figure 6.27.

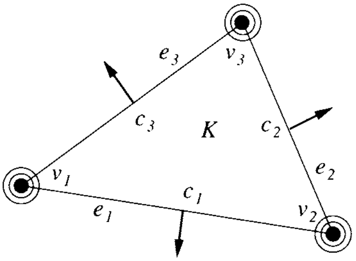


Figure 6.27 Twenty-one DOF on the lowest-order (quintic) Argyris triangle.

The black dots stand for Lagrange DOF associated with function values at the vertices. Each inner circle surrounding a black dot represents a pair of Hermite DOF associated with first directional derivatives at the vertices (we choose $\partial/\partial x_1$ and $\partial/\partial x_2$). Further, each outer circle stands for three Argyris DOF corresponding to second directional derivatives (we choose $\partial^2/\partial x_1^2$, $\partial^2/\partial x_1 x_2$ and $\partial^2/\partial x_2^2$). The arrows indicate the DOF associated with the normal derivatives at the edge midpoints. The partial derivatives can be exchanged for the derivatives in the directions of the edges, analogously to Hermite triangles (see Figure 6.17).

Lemma 6.6 *The Argyris element $(K_t, P^5(K_t), \Sigma)$, where Σ consists of the 21 above-defined degrees of freedom, is unisolvent.*

Proof: Take an arbitrary $g \in P^5(K)$ such that $L_j(g) = 0$ for all $j = 1, 2, \dots, 21$. We need to show that necessarily $g = 0$. First, g restricted to the edge e_1 is a fifth-degree polynomial that vanishes at the endpoints of e_1 together with its first and second derivatives. Since these six independent parameters define a unique one-dimensional fifth-degree polynomial, it follows that $g = 0$ on e_1. Analogously $g = 0$ on the remaining two edges e_2 and e_3. Thus g vanishes on the whole boundary of the element, and it can be written as a product

$$g(\boldsymbol{x}) = \tilde{g}(\boldsymbol{x})\lambda_1(\boldsymbol{x})\lambda_2(\boldsymbol{x})\lambda_3(\boldsymbol{x}), \quad \tilde{g} \in P^2(K),$$

where λ_k, $k = 1, 2, 3$, are the barycentric coordinates in K. Recall that $\lambda_1 + \lambda_2 + \lambda_3 = 1$ and λ_k is an affine function that vanishes on the edge e_k and attains a value of one at the opposite vertex. There is a one-to-one relation between $\lambda_1, \lambda_2, \lambda_3$ and $x \in K$, and therefore we can express both g and $\tilde{g}$ in terms of the barycentric coordinates,

$$g(\lambda_1, \lambda_2, \lambda_3) = \tilde{g}(\lambda_1, \lambda_2, \lambda_3)\lambda_1\lambda_2\lambda_3. \tag{6.95}$$

The partial derivative $\partial/\partial\lambda_k$ is the derivative in the normal direction to the edge e_k. Using (6.95), from the zero second derivative DOF at v_1 one obtains

$$0 = \frac{\partial^2 g}{\partial\lambda_1\partial\lambda_3}(v_1) = \frac{\partial^2 \tilde{g}}{\partial\lambda_1\partial\lambda_3}(v_1)\underbrace{\lambda_1(v_1)}_{=0}\lambda_2(v_1)\underbrace{\lambda_3(v_1)}_{=0}$$

$$+\frac{\partial\tilde{g}}{\partial\lambda_1}(v_1)\underbrace{\lambda_1(v_1)}_{=0}\lambda_2(v_1) + \frac{\partial\tilde{g}}{\partial\lambda_3}(v_1)\lambda_2(v_1)\underbrace{\lambda_3(v_1)}_{=0}+\tilde{g}(v_1)\underbrace{\lambda_2(v_1)}_{=1} = \tilde{g}(v_1).$$

Analogously it is $\tilde{g}(v_2) = \tilde{g}(v_3) = 0$. For the normal derivative at the midpoint c_1 of the edge e_1 we have

$$0 = \frac{\partial g}{\partial\lambda_1}(c_1) = \frac{\partial\tilde{g}}{\partial\lambda_1}(c_1)\underbrace{\lambda_1(c_1)}_{=0}\lambda_2(c_1)\lambda_3(c_1) + \tilde{g}(c_1)\underbrace{\lambda_2(c_1)\lambda_3(c_1)}_{\neq 0},$$

and it follows that $\tilde{g}(c_1) = 0$. Analogously we conclude that $\tilde{g}(c_2) = \tilde{g}(c_3) = 0$. Finally, from the unisolvency of a second-order Lagrange element with the nodes $v_1, v_2, v_3, c_1, c_2, c_3$, we obtain that necessarily $\tilde{g} = 0$. Therefore $g = 0$ and the triangular quintic Argyris element is unisolvent. ■

6.6.2 Local interpolant, conformity

Consider an element $K_i \in \mathcal{T}_{h,p}$. The quintic Argyris element K_i is endowed with the standard nodal interpolant (3.28),

$$\mathcal{I}_{K_i}(g) = \sum_{i=1}^{N_P} L_i(g)\theta_i,$$

$N_P = 21$. Here the nodal basis functions θ_i of the space $P^5(K_i)$, meeting the delta property (3.4), are constructed in the standard way as described in Paragraph 3.1.1. Nodal shape functions on the reference triangular domain K_t will be calculated in Paragraph 6.6.3. As usual, the local interpolant exists if all linear forms L_i, $i = 1, 2, \ldots, N_P$ are defined for g.

Conformity to $H^2(\Omega_h)$ Let $\mathcal{T}_{h,p} = \{K_1, K_2, \ldots, K_M\}$ comprise M quintic triangular Argyris elements. According to Definition 3.6, the global nodal interpolant is defined elementwise as

$$\mathcal{I}(g)|_{K_i} = \mathcal{I}_{K_i}(g) \qquad \text{for all } i = 1, 2, \ldots, M.$$

Since $C^2(\Omega_h)$ is dense in $H^2(\Omega_h)$, according to Definition 3.7 the finite element mesh $\mathcal{T}_{h,p}$ conforms to the space $H^2(\Omega_h)$ if the following implication holds:

$$g \in C^2(\Omega_h) \quad \Rightarrow \quad \mathcal{I}(g) \in H^2(\Omega_h).$$

The piecewise polynomial interpolant $\mathcal{I}(g)$ belongs to $H^2(\Omega_h)$ if and only if it is once continuously differentiable (see Section A.4), i.e., if and only if it is smooth across all element interfaces and at all grid vertices.

Lemma 6.7 *Every regular finite element mesh $\mathcal{T}_{h,p}$ consisting of triangular quintic Argyris elements conforms to the space $H^2(\Omega_h)$.*

Proof: Consider a pair of elements $K_1, K_2 \in \Omega_h$ that share an edge e, as depicted in Figure 6.28.

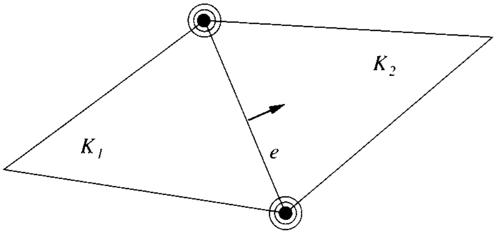


Figure 6.28 Smoothness of the global Argyris interpolant across the edge e. The edge e is equipped with a unique unit normal vector ν_e, to which the normal derivative DOF on both elements are related.

By $g_{K_1,e}$ and $g_{K_2,e}$ denote the restrictions of $\mathcal{I}(g)$ to the edge e on the elements K_1 and K_2, respectively. These are one-dimensional fifth-degree polynomials whose values as well as first and second derivatives agree at the endpoints of e. Since a unique fifth-degree polynomial is determined via these six parameters, $g_{K_1,e} = g_{K_2,e}$ and $\mathcal{I}(g)$ is continuous across e.

Next by $g'_{K_1,e}$ and $g'_{K_2,e}$ denote the derivative of $\mathcal{I}(g)$ in the unique normal direction ν_e to the edge e on the elements K_1 and K_2, respectively. These one-dimensional fourth-degree polynomials coincide at the endpoints of e due to the agreement of the derivative DOF. Analogously their first derivatives at the endpoints of e coincide due to the agreement of the second derivative DOF. Finally, their values at the midpoint of e are the same because of the agreement of the normal derivative DOF. Thus $g'_{K_1,e} = g'_{K_2,e}$. Obviously the tangential derivatives of $g'_{K_1,e}$ and $g'_{K_2,e}$ along the edge e are the same, and therefore the global interpolant is continuously differentiable across e. The smoothness of $\mathcal{I}(g)$ at all grid vertices is obvious, and therefore we can conclude that the mesh $\mathcal{T}_{h,p}$ conforms to the space $H^2(\Omega_h)$. ■

6.6.3 Nodal shape functions on the reference domain

The twenty-one nodal shape functions of the quintic Argyris element on the reference triangle K_t can be obtained using the procedure described in Paragraph 3.1.1, i.e., choosing a suitable basis of the space $P^5(K_t)$, $\mathcal{B} = \{g_1, g_2, \dots, g_{21}\}$ (for example the monomial basis), inverting the Vandermonde matrix $L_i(g_j)$ and reading the coefficients of the nodal shape functions from the columns. The resulting unique nodal basis is shown in Figures 6.29–6.35 (the graphs have different scaling).

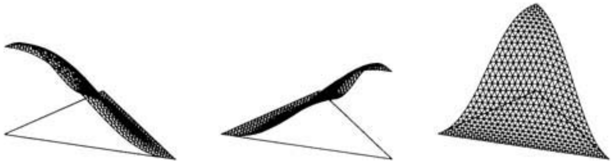


Figure 6.29 Nodal basis of the quintic Argyris element; shape functions $\varphi_t^{v_1}, \varphi_t^{v_2}$, and $\varphi_t^{v_3}$, representing the function values at the vertices v_1, v_2, and v_3.

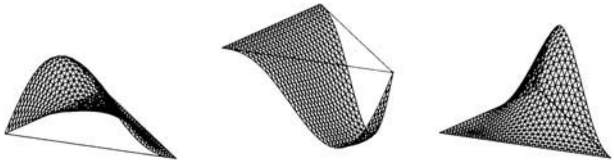


Figure 6.30 Nodal basis of the quintic Argyris element; shape functions $\varphi_t^{v_1,1}, \varphi_t^{v_2,1}$, and $\varphi_t^{v_3,1}$, representing $\partial/\partial x_1$ at the vertices v_1, v_2, and v_3.

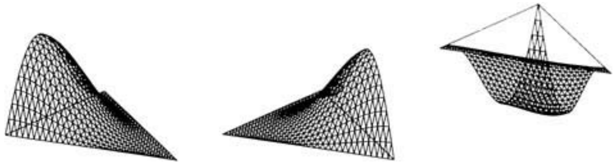


Figure 6.31 Nodal basis of the quintic Argyris element; shape functions $\varphi_t^{v_1,2}, \varphi_t^{v_2,2}$, and $\varphi_t^{v_3,2}$, representing $\partial/\partial x_2$ at the vertices v_1, v_2, and v_3.

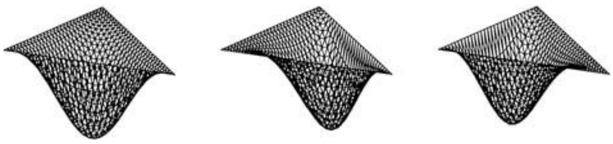


Figure 6.32 Nodal basis of the quintic Argyris element; shape functions $\varphi_t^{e_1,n}, \varphi_t^{e_2,n}$, and $\varphi_t^{e_3,n}$, representing the normal derivatives at the midpoints of the edges e_1, e_2, and e_3.

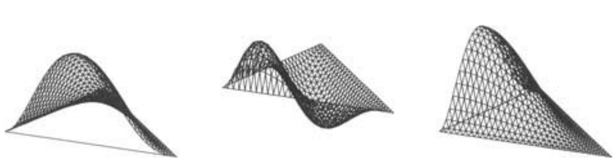


Figure 6.33 Nodal basis of the quintic Argyris element; shape functions $\varphi_t^{v_1,1,1}, \varphi_t^{v_1,1,2}$, and $\varphi_t^{v_1,2,2}$, representing $\partial^2/\partial x_1^2$, $\partial^2/\partial x_1 \partial x_2$, and $\partial^2/\partial x_2^2$ at the vertex v_1.

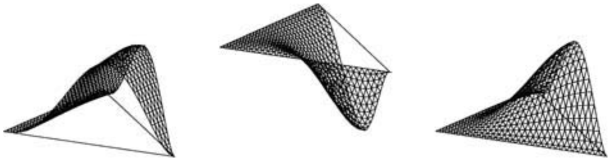


Figure 6.34 Nodal basis of the quintic Argyris element; shape functions $\varphi_t^{v_2,1,1}$, $\varphi_t^{v_2,1,2}$, and $\varphi_t^{v_2,2,2}$, representing $\partial^2/\partial x_1^2$, $\partial^2/\partial x_1\partial x_2$, and $\partial^2/\partial x_2^2$ at the vertex v_2.

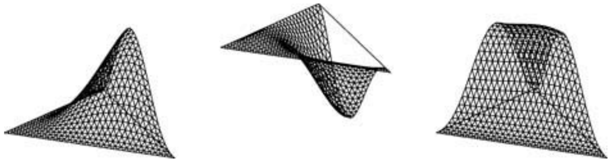


Figure 6.35 Nodal basis of the quintic Argyris element; shape functions $\varphi_t^{v_3,1,1}$, $\varphi_t^{v_3,1,2}$, and $\varphi_t^{v_3,2,2}$, representing $\partial^2/\partial x_1^2$, $\partial^2/\partial x_1\partial x_2$, and $\partial^2/\partial x_2^2$ at the vertex v_3.

6.6.4 Transformation to reference domains

Next the integrals involved in the weak formulation (6.94) are transformed from a mesh triangle K to the reference triangular domain K_t via the standard affine reference map $\boldsymbol{x}_K : K_t \longrightarrow K$. The procedure is slightly more technical, but otherwise similar to what was done for second-order elliptic problems in Paragraph 4.1.4. Without loss of generality, we assume that the constant Jacobian J_K of the map $\boldsymbol{x}_K$ is positive. By

$$\tilde{w} = w \circ \boldsymbol{x}_K \tag{6.96}$$

we denote the transformation of a function $w \in P^5(K)$ to K_t. In Paragraph 4.1.4 we learned how to express the first partial derivatives $\partial w/\partial x_j$, $j = 1, 2$, by means of the partial derivatives $\partial \tilde{w}/\partial \xi_j$ and the Jacobi matrix $\boldsymbol{J}_K = \mathrm{D}\boldsymbol{x}_K/\mathrm{D}\boldsymbol{\xi}$. The rule (4.20) for the transformation of gradients yields

$$\nabla w(\boldsymbol{x}) = \boldsymbol{J}_K^{-T} \nabla \tilde{w}(\boldsymbol{\xi}), \quad \boldsymbol{x} = \boldsymbol{x}_K(\boldsymbol{\xi}).$$

Now we have to do the same for the second partial derivatives of $\tilde{w}$. It follows from (6.96) that

$$\begin{aligned}
\frac{\partial^2 \tilde{w}}{\partial \xi_1^2} &= \left(\frac{\partial^2 w}{\partial x_1^2} J_{11} + \frac{\partial^2 w}{\partial x_1 \partial x_2} J_{21}\right) J_{11} + \left(\frac{\partial^2 w}{\partial x_1 \partial x_2} J_{11} + \frac{\partial^2 w}{\partial x_2^2} J_{21}\right) J_{21}, \qquad (6.97)\\
\frac{\partial^2 \tilde{w}}{\partial \xi_1 \partial \xi_2} &= \left(\frac{\partial^2 w}{\partial x_1^2} J_{12} + \frac{\partial^2 w}{\partial x_1 \partial x_2} J_{22}\right) J_{11} + \left(\frac{\partial^2 w}{\partial x_1 \partial x_2} J_{12} + \frac{\partial^2 w}{\partial x_2^2} J_{22}\right) J_{21},\\
\frac{\partial^2 \tilde{w}}{\partial \xi_2^2} &= \left(\frac{\partial^2 w}{\partial x_1^2} J_{12} + \frac{\partial^2 w}{\partial x_1 \partial x_2} J_{22}\right) J_{12} + \left(\frac{\partial^2 w}{\partial x_1 \partial x_2} J_{12} + \frac{\partial^2 w}{\partial x_2^2} J_{22}\right) J_{22},
\end{aligned}$$

where J_{ij} is the ijth entry of the Jacobi matrix $\boldsymbol{J}_K$. Thus the second derivatives of w depend on the second derivatives of $\tilde{w}$ linearly,

$$\begin{pmatrix} \dfrac{\partial^2 \tilde{w}}{\partial \xi_1^2} \\ \dfrac{\partial^2 \tilde{w}}{\partial \xi_1 \partial \xi_2} \\ \dfrac{\partial^2 \tilde{w}}{\partial \xi_2^2} \end{pmatrix} = \boldsymbol{A} \begin{pmatrix} \dfrac{\partial^2 w}{\partial x_1^2} \\ \dfrac{\partial^2 w}{\partial x_1 \partial x_2} \\ \dfrac{\partial^2 w}{\partial x_2^2} \end{pmatrix},$$

where the constant matrix $\boldsymbol{A}$ has the form

$$\boldsymbol{A} = \begin{pmatrix} J_{11}^2 & 2J_{11}J_{21} & J_{21}^2 \\ J_{11}J_{12} & (J_{11}J_{22} + J_{12}J_{21}) & J_{21}J_{22} \\ J_{12}^2 & 2J_{12}J_{22} & J_{22}^2 \end{pmatrix}.$$

This matrix is invertible since $\det(\boldsymbol{A}) = \det^3(\boldsymbol{J}_K)$, and its inverse has the form

$$\boldsymbol{A}^{-1} = \frac{1}{\det^2(\boldsymbol{J}_K)} \begin{pmatrix} J_{22}^2 & -2J_{21}J_{22} & J_{21}^2 \\ -J_{12}J_{22} & (J_{11}J_{22} + J_{12}J_{21}) & -J_{11}J_{21} \\ J_{12}^2 & -2J_{11}J_{12} & J_{11}^2 \end{pmatrix}.$$

Hence, for the second partial derivatives of w it holds

$$\begin{pmatrix} \dfrac{\partial^2 w}{\partial x_1^2} \\ \dfrac{\partial^2 w}{\partial x_1 \partial x_2} \\ \dfrac{\partial^2 w}{\partial x_2^2} \end{pmatrix} = \boldsymbol{A}^{-1} \begin{pmatrix} \dfrac{\partial^2 \tilde{w}}{\partial \xi_1^2} \\ \dfrac{\partial^2 \tilde{w}}{\partial \xi_1 \partial \xi_2} \\ \dfrac{\partial^2 \tilde{w}}{\partial \xi_2^2} \end{pmatrix}.$$

This allows us to replace the second partial derivatives of w in (6.94) with terms containing the second partial derivatives of $\tilde{w}$ and the entries of the Jacobi matrix $\boldsymbol{J}_K$, as we wanted. An additional multiplication with $\det(\boldsymbol{J}_K)$ dictated by the Substitution Theorem accomplishes the transformation of the integrals to the reference domain K_t.

6.6.5 Design of basis functions

The last ingredient needed for the assembly of the stiffness matrix and of the load vector is a suitable basis of the space $V_{h,p}$, consisting in this case of globally smooth piecewise-polynomial functions defined in the domain Ω_h. The basis functions are designed by means of the shape functions from Paragraph 6.6.3 and the reference maps $\boldsymbol{x}_K$ and their derivatives, analogously to what we did for Hermite elements in Paragraph 6.4.3. The new interesting aspect of the Argyris elements is the presence of the DOF associated with the second derivatives at the grid vertices and the DOF related to the normal derivatives at edge midpoints.

Consider a bounded polygonal domain $\Omega_h \subset \mathbb{R}^2$ covered with a finite element mesh $\mathcal{T}_{h,p}$ consisting of $M \geq 1$ triangular quintic Argyris elements. The finite element space $V_{h,p}$ (which, for simplicity, is not constrained with any essential boundary conditions) has the form

$$V_{h,p} = \{v \in C^1(\Omega_h);\ v|_{K_i} \in P^p(K_i),\ i = 1, 2, \ldots, M\}. \tag{6.98}$$

Proposition 6.2 *The dimension of the space $V_{h,p}$ is*

$$N = dim(V_{h,p}) = 6M_v + M_e, \tag{6.99}$$

where M_v is the number of grid vertices and M_e the number of mesh edges.

Proof: There are 6 degrees of freedom associated with every grid vertex and one degree of freedom for the midpoint of every edge. Each of these degrees of freedom is represented by one basis function in the basis of $V_{h,p}$. ■

The basis of the space $V_{h,p}$ consists of three types of basis functions:

- Lagrange basis functions associated with the function values at the grid vertices,
- Hermite basis functions representing the first partial derivatives ($\partial/\partial x_1$ and $\partial/\partial x_2$ at the grid vertices and the normal derivatives at edge midpoints),
- Argyris basis functions related to the second partial derivatives $\partial^2/\partial x_1^2, \partial^2/\partial x_1 x_2$ and $\partial^2/\partial x_2^2$ at the grid vertices.

Lagrange vertex functions

Figure 6.29 shows the three Lagrange vertex shape functions which are present in the basis of the polynomial space $P^5(K_t)$. The shape functions are smooth, and therefore the fact that $\partial/\partial\xi_1 = \partial/\partial\xi_2 = 0$ at the vertices of K_t implies that any first directional derivative at any vertex of K_t is zero. Analogously, since the second partial derivatives at the vertices are zero, any second directional derivative at any vertex is zero. Therefore one could be tempted to construct the Lagrange basis functions in Ω_h as usual, by composing the Lagrange shape functions with the inverse of the reference maps $\boldsymbol{x}_K : K_t \to K$ (this was done for the Lagrange–Fekete elements in Paragraph 4.3.6 and for the Hermite–Fekete elements in Paragraph 6.4.3). However, in such case the resulting basis functions would not be smooth at the midpoints of element interfaces, since the reference map $\boldsymbol{x}_K$ does not preserve the normal vectors at the edge midpoints. Fortunately this can be cured by means of the shape functions $\varphi_t^{e_1,n}, \varphi_t^{e_2,n}$, and $\varphi_t^{e_3,n}$:

Consider a grid vertex $\boldsymbol{x}_i$ and the vertex patch (4.14) of all elements adjacent to $\boldsymbol{x}_i$,

$$S(i) = \bigcup_{k \in N(i)} \overline{K}_k,$$

where

$$N(i) = \{k;\ K_k \in \mathcal{T}_{h,p},\ \boldsymbol{x}_i \text{ is a vertex of } K_k\}.$$

The Lagrange vertex basis function $v_i^{(v)}$ corresponding to the grid vertex $\boldsymbol{x}_i$ is smooth in the whole domain Ω_h and it vanishes in $\Omega_h \setminus S(i)$. For any element $K \in S(i)$ the restriction of $v_i^{(v)}$ to K is defined as

$$v_i^{(v)}|_K = \varphi^{(v)} \circ \boldsymbol{x}_K^{-1},$$

where the function $\varphi^{(v)}$ is defined on K_t by

$$\varphi^{(v)}(\boldsymbol{\xi}) = \varphi_t^{v_m}(\boldsymbol{\xi}) + \beta_1\varphi_t^{e_1,n}(\boldsymbol{\xi}) + \beta_2\varphi_t^{e_2,n}(\boldsymbol{\xi}) + \beta_3\varphi_t^{e_3,n}(\boldsymbol{\xi}).$$

Here v_m is a vertex of K_t such that $\boldsymbol{x}_K(v_m) = \boldsymbol{x}_i$ and the shape functions were defined in Paragraph 6.6.3. Analytical formulae for the unknown real coefficients β_k are obtained from the conditions

$$\begin{aligned} \frac{\partial v_i^{(v)}}{\partial \boldsymbol{\nu}_e}(\boldsymbol{x}_e) &= 0, \\ \frac{\partial v_i^{(v)}}{\partial \boldsymbol{\nu}_f}(\boldsymbol{x}_f) &= 0, \\ \frac{\partial v_i^{(v)}}{\partial \boldsymbol{\nu}_g}(\boldsymbol{x}_g) &= 0. \end{aligned} \tag{6.100}$$

The symbols e, f, and g stand for the edges of K such that $e = \boldsymbol{x}_K(e_1)$, $f = \boldsymbol{x}_K(e_2)$, and $g = \boldsymbol{x}_K(e_3)$, and the points $\boldsymbol{x}_e, \boldsymbol{x}_f$, and $\boldsymbol{x}_g$ are the midpoints of the edges e, f, and g, respectively.

Hermite vertex basis functions

Now we combine the technique developed for the Hermite elements with the trick introduced in the previous step. For every grid vertex $\boldsymbol{x}_i$ there is a pair of Hermite vertex basis functions representing $\partial/\partial x_1$ and $\partial/\partial x_2$ at $\boldsymbol{x}_i$, say, $v_i^{(1)}$ and $v_i^{(2)}$. Both these functions are smooth in the whole domain Ω_h and they vanish in $\Omega_h \setminus S(i)$. Their first partial derivatives $\partial/\partial x_1$ and $\partial/\partial x_2$ vanish at all boundary vertices of the patch $S(i)$, and their first normal derivatives vanish at the midpoints of all grid edges. At the vertex $\boldsymbol{x}_i$ these functions satisfy

$$\frac{\partial}{\partial x_j} v_i^{(k)}(\boldsymbol{x}_i) = \delta_{jk}, \quad j, k = 1, 2.$$

The second partial derivatives $\partial^2/\partial x_1^2$ $\partial^2/\partial x_1 x_2$ and $\partial^2/\partial x_2^2$ of these functions vanish at all grid vertices.

Let us begin with the restriction of $v_i^{(1)}$ to an element $K \in S(i)$. Consider the pair of Hermite vertex functions $\varphi_t^{v_m,1}$ and $\varphi_t^{v_m,2}$ associated with the vertex v_m of K_t, such that $\boldsymbol{x}_K(v_m) = \boldsymbol{x}_i$. We look for the vertex function in the form

$$v_i^{(1)} = \varphi^{(1)} \circ \boldsymbol{x}_K^{-1},$$

where

$$\varphi^{(1)}(\boldsymbol{\xi}) = \alpha_1 \varphi_t^{v_m,1}(\boldsymbol{\xi}) + \alpha_2 \varphi_t^{v_m,2}(\boldsymbol{\xi}) + \beta_1 \varphi_t^{e_1,n}(\boldsymbol{\xi}) + \beta_2 \varphi_t^{e_2,n}(\boldsymbol{\xi}) + \beta_3 \varphi_t^{e_3,n}(\boldsymbol{\xi}).$$

Now the analytical formulae for the unknown coefficients are obtained from the conditions

$$\begin{aligned} \frac{\partial}{\partial x_j} v_i^{(1)}(\boldsymbol{x}_i) &= \delta_{1j}, \quad j = 1, 2, \\ \frac{\partial v_i^{(1)}}{\partial \boldsymbol{\nu}_e}(\boldsymbol{x}_e) &= 0, \end{aligned} \tag{6.101}$$

$$\begin{aligned}
\frac{\partial v_i^{(1)}}{\partial \boldsymbol{\nu}_f}(\boldsymbol{x}_f) &= 0, \\
\frac{\partial v_i^{(1)}}{\partial \boldsymbol{\nu}_g}(\boldsymbol{x}_g) &= 0,
\end{aligned}$$

where the symbols $\boldsymbol{x}_e, \boldsymbol{x}_f, \boldsymbol{x}_g, \boldsymbol{\nu}_e, \boldsymbol{\nu}_f$, and $\boldsymbol{\nu}_g$ have the same meaning as in (6.100). The other Hermite vertex basis function $v_i^{(2)}$, representing $\partial/\partial x_2$ at $\boldsymbol{x}_i$, is constructed analogously.

Argyris vertex basis functions

Next let us construct the vertex basis functions associated with the second partial derivatives. For every grid vertex $\boldsymbol{x}_i$ there are three Argyris vertex basis functions representing $\partial^2/\partial x_1^2$, $\partial^2/\partial x_1 x_2$ and $\partial^2/\partial x_2^2$ at $\boldsymbol{x}_i$, let us call them, e.g., $v_i^{(1,1)}$, $v_i^{(1,2)}$ and $v_i^{(2,2)}$. These functions are smooth in the whole domain Ω_h and they vanish in $\Omega_h \setminus S(i)$. Their first partial derivatives $\partial/\partial x_1$ and $\partial/\partial x_2$ vanish at all grid vertices, and their normal derivatives vanish at the midpoints of all grid edges. Their second partial derivatives vanish at all grid vertices except for $\boldsymbol{x}_i$, where they satisfy

$$\begin{aligned}
\frac{\partial^2}{\partial x_1^2} v_i^{(1,1)}(\boldsymbol{x}_i) &= 1, \\
\frac{\partial^2}{\partial x_1 x_2} v_i^{(1,1)}(\boldsymbol{x}_i) &= 0, \\
\frac{\partial^2}{\partial x_2^2} v_i^{(1,1)}(\boldsymbol{x}_i) &= 0, \\
\frac{\partial^2}{\partial x_1^2} v_i^{(1,2)}(\boldsymbol{x}_i) &= 0, \\
\frac{\partial^2}{\partial x_1 x_2} v_i^{(1,2)}(\boldsymbol{x}_i) &= 1, \\
\frac{\partial^2}{\partial x_2^2} v_i^{(1,2)}(\boldsymbol{x}_i) &= 0, \\
\frac{\partial^2}{\partial x_1^2} v_i^{(2,2)}(\boldsymbol{x}_i) &= 0, \\
\frac{\partial^2}{\partial x_1 x_2} v_i^{(2,2)}(\boldsymbol{x}_i) &= 0, \\
\frac{\partial^2}{\partial x_2^2} v_i^{(2,2)}(\boldsymbol{x}_i) &= 1.
\end{aligned}$$

Consider the three Argyris vertex shape functions $\varphi_t^{v_m,1,1}$, $\varphi_t^{v_m,1,2}$ and $\varphi_t^{v_m,2,2}$ associated with the vertex v_m of K_t, such that $\boldsymbol{x}_K(v_m) = \boldsymbol{x}_i$. The vertex function $v_i^{(1,1)}$ in any element $K \in S(i)$ is sought in the form

$$v_i^{(1,1)} = \varphi^{(1,1)} \circ \boldsymbol{x}_K^{-1},$$

where

$$
\begin{aligned}
\varphi^{(1,1)}(\boldsymbol{\xi}) = \alpha_1 \varphi_t^{v_m,1,1}(\boldsymbol{\xi}) + \alpha_2 \varphi_t^{v_m,1,2}(\boldsymbol{\xi}) + \alpha_3 \varphi_t^{v_m,2,2}(\boldsymbol{\xi}) \\
+ \beta_1 \varphi_t^{e_1,n}(\boldsymbol{\xi}) + \beta_2 \varphi_t^{e_2,n}(\boldsymbol{\xi}) + \beta_3 \varphi_t^{e_3,n}(\boldsymbol{\xi}).
\end{aligned}
$$

The analytical formulae for the six unknown coefficients are obtained from the conditions

$$
\begin{aligned}
\frac{\partial^2}{\partial x_1^2} v_i^{(1,1)}(\boldsymbol{x}_i) &= 1, \\
\frac{\partial^2}{\partial x_1 x_2} v_i^{(1,1)}(\boldsymbol{x}_i) &= 0, \\
\frac{\partial^2}{\partial x_2^2} v_i^{(1,1)}(\boldsymbol{x}_i) &= 0, \\
\frac{\partial v_i^{(1,1)}}{\partial \boldsymbol{\nu}_e}(\boldsymbol{x}_e) &= 0, \\
\frac{\partial v_i^{(1,1)}}{\partial \boldsymbol{\nu}_f}(\boldsymbol{x}_f) &= 0, \\
\frac{\partial v_i^{(1,1)}}{\partial \boldsymbol{\nu}_g}(\boldsymbol{x}_g) &= 0,
\end{aligned}
$$

where the symbols $\boldsymbol{x}_e, \boldsymbol{x}_f, \boldsymbol{x}_g, \boldsymbol{\nu}_e, \boldsymbol{\nu}_f$ and $\boldsymbol{\nu}_g$ have the same meaning as in (6.100). The remaining Argyris vertex basis functions are constructed analogously.

Hermite basis functions associated with normal derivatives at edge midpoints

The design of the remaining set of basis functions, representing the normal derivatives at the midpoints of mesh edges, involves the orientation issue analogous to the one encountered in the design of higher-order Lagrange edge functions in Paragraph 4.3.6. This time a unique unit normal vector $\boldsymbol{\nu}_{s_j}$ needs to be assigned to every edge s_j in the mesh $\mathcal{T}_{h,p}$. This is equivalent to assigning a unique global orientation to mesh edges (see Paragraph 4.3.6). Consider the edge element patch (4.49),

$$
S_e(j) = \bigcup_{k \in N_e(j)} \overline{K_k},
$$

where

$$
N_e(j) = \{k;\ K_k \in \mathcal{T}_{h,p},\ s_j \text{ is an edge of } K_k\}.
$$

To every mesh edge s_j there is one Hermite edge basis function $v_{s_j}^{(n)}$ whose normal derivative at $\boldsymbol{x}_{s_j}$ satisfies

$$
\frac{\partial}{\partial \boldsymbol{\nu}_{s_j}} v_{s_j}^{(n)}(\boldsymbol{x}_{s_j}) = 1.
$$

This function is smooth in the whole domain Ω_h and it vanishes in $\Omega_h \setminus S_e(j)$. Its first partial derivatives $\partial/\partial x_1$ and $\partial/\partial x_2$ vanish at all grid vertices, and its normal derivative vanishes at the midpoints of all grid edges except for s_j. The second partial derivatives $\partial^2/\partial x_1^2$ $\partial^2/\partial x_1 x_2$ and $\partial^2/\partial x_2^2$ vanish at all grid vertices.

Let us begin with the restriction of $v_{s_j}^{(n)}$ to any element $K \in S_e(j)$. Consider the Hermite edge shape functions $\varphi_t^{e_1,n}$, $\varphi_t^{e_2,n}$ and $\varphi_t^{e_3,n}$. The edge basis function associated with s_j is sought in the form

$$v_{s_j}^{(n)} = \varphi^{(n)} \circ \boldsymbol{x}_K^{-1},$$

where

$$\varphi^{(n)} = \beta_1 \varphi_t^{e_1,n}(\boldsymbol{\xi}) + \beta_2 \varphi_t^{e_2,n}(\boldsymbol{\xi}) + \beta_3 \varphi_t^{e_3,n}(\boldsymbol{\xi}).$$

Now the analytical formulae for the unknown coefficients are obtained from the conditions

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\nu}_{s_j}} v_{s_j}^{(n)}(\boldsymbol{x}_{s_j}) &= o(s_j, K), \\ \frac{\partial}{\partial \boldsymbol{\nu}_{s_k}} v_{s_j}^{(n)}(\boldsymbol{x}_{s_k}) &= 0, \\ \frac{\partial}{\partial \boldsymbol{\nu}_{s_l}} v_{s_j}^{(n)}(\boldsymbol{x}_{s_l}) &= 0, \end{aligned} \tag{6.102}$$

where s_k, s_l are the remaining two edges of K and $\boldsymbol{x}_{s_k}, \boldsymbol{x}_{s_l}$ their midpoints. The orientation flag $o(s_j, K) = 1$ if $\boldsymbol{\nu}_{s_j}$ points outside of the element K and $o(s_j, K) = -1$ in the opposite case. Herewith the construction of the basis of the space $V_{h,p}$ from (6.98) is accomplished.

6.6.6 Higher-order nodal Argyris–Fekete elements

In this section we comment on the extension of the quintic nodal Argyris element to a general polynomial degree $p \geq 5$. Consider a triagular element $K \in \mathcal{T}_{h,p}$, and recall that the dimension of the space $P^p(K)$ is $N_P = (p+1)(p+2)/2$. The higher-order element inherits the three Lagrange DOF associated with the function values at the vertices, the six Hermite DOF related to the first partial derivatives at the vertices and the nine Argyris DOF corresponding to the second partial derivatives at the vertices. Thus $N_P - 18$ degrees of freedom remain to be defined on the edges and in the element interior.

Recall that with one Hermite DOF per edge, representing the normal derivative at the midpoint, the quintic Argyris element was both unisolvent and conforming to the space $H^2(\Omega_h)$. In the general case one needs $p-5$ Lagrange DOF and $p-4$ Hermite DOF per edge in order to satisfy the conformity requirements of the space H^2. For example, the edge-interior Fekete points corresponding to $p-4$ and $p-3$ can be used to define these Lagrange and Hermite DOF, respectively.

This means that

$$\frac{(p+1)(p+2)}{2} - 18 - 3(p-5) - 3(p-4) = \frac{(p-4)(p-5)}{2}$$

Lagrange degrees of freedom remain to be defined in the element interior. Their number suggests to choose the interior Fekete points corresponding to $p-3$.

The sixth- and seventh-order Argyris–Fekete elements on the reference triangular domain K_t are illustrated in Figure 6.36.

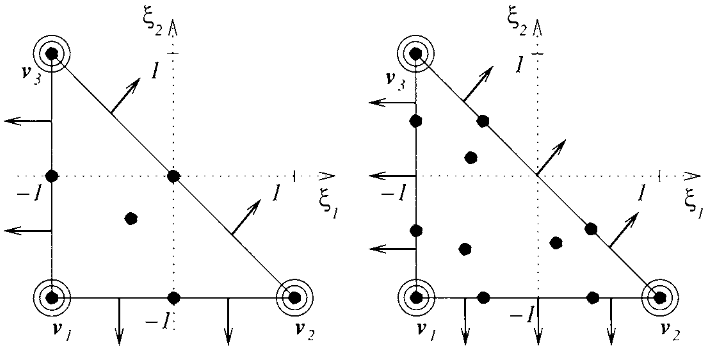


Figure 6.36 The sixth- and seventh-order Argyris–Fekete elements on K_t.

Unisolvency, conformity Both the unisolvency of Argyris–Fekete elements and their conformity to $H^2(\Omega_h)$ can be checked analogously to the quintic case:

Lemma 6.8 *The Argyris–Fekete element $(K_t, P^p(K_t), \Sigma)$, where $p \geq 5$ and Σ consists of the $(p+1)(p+2)/2$ above-defined degrees of freedom, is unisolvent.*

Proof: The proof is analogous to the proof of Lemma 6.6. ■

Lemma 6.9 *Every regular finite element mesh $\mathcal{T}_{h,p}$ consisting of triangular Argyris–Fekete elements of a uniform polynomial degree $p \geq 5$ conforms to the space $H^2(\Omega_h)$.*

Proof: The proof is a straightforward generalization of the proof of Lemma 6.7. ■

It is worth mentioning that the finite element space $V_{h,p}$ on a finite element mesh $\mathcal{T}_{h,p}$ comprising M triangular Argyris–Fekete elements (where for simplicity no degrees of freedom are constrained by boundary conditions) has the form

$$V_{h,p} = \{v \in C^1(\Omega_h);\ v|_{K_i} \in P^p(K_i)\}. \tag{6.103}$$

Proposition 6.3 *The dimension of the space $V_{h,p}$ is*

$$N = dim(V_{h,p}) = 6M_v + (2p-9)M_e + \frac{(p-4)(p-5)}{2}M, \tag{6.104}$$

where M_v is the number of grid vertices and M_e the number of mesh edges.

Proof: There are 6 DOF per grid vertex, $(p-4)+(p-5)$ in the interior of every mesh edge and $(p-4)(p-5)/2$ in every element interior. ■

A basis of the space $V_{h,p}$ can be constructed analogously to the quintic case (see Paragraph 6.6.5).

6.7 EXERCISES

Exercise 6.1 *Consider an interval $\Omega = (a,b)$ and a clamped prismatic beam with $u(a) = u(b) = \nabla u(a) = \nabla u(b) = 0$. Assume that $E(x)I(x) = 1$ and $f(x) = F_0$ for all $x \in \Omega$. Calculate the exact solution to the Euler–Bernoulli model (6.7).*

Exercise 6.2 *Return to Exercise 6.1 and consider a prismatic beam of a constant square cross-section* $h_0 \times h_0$ *and a constant modulus of elasticity* E. *Calculate the exact solution to the Euler–Bernoulli beam model. Hint: Use relation (6.4) to calculate* $I(x)$.

Exercise 6.3 *Extend the problem from Exercise 6.1 to a cantilever beam with* $u(a) = \nabla u(a) = 0$ *and* $M(b) = M_b$, $F_s(b) = F_b$. *Calculate the exact solution to the Euler–Bernoulli beam model.*

Exercise 6.4 *Prove Lemma 6.1 for nonconstant, strictly positive* $b \in L^{\infty}(\Omega)$.

Exercise 6.5 *Write the weak formulation of the Euler–Bernoulli model (6.7) in the simply supported case (Figure 6.4). Prescribe the deflection* $u_a = u_b = 0$ *and bending moments* $M_a = M_b = M$ *at both ends.*

Exercise 6.6 *Write the weak formulation of the Euler–Bernoulli model (6.7) in the cantilever case (Figure 6.5). Prescribe the deflection* $u_a = 0$ *and slope* $\nabla u = 0$ *at the clamped end, and prescribe a moment* M_b *and shear force* F_b *at the free end.*

Exercise 6.7 *Show in detail that (6.25) holds. Hint: Subtract the exact and approximate weak formulations.*

Exercise 6.8 *Show in detail how (6.25) implies that the approximate solution* $u_{h,p} \in V_{h,p} \subset V$ *does not depend on the choice of the basis* $\{v_1, v_2, \ldots, v_N\}$ *of the space* $V_{h,p}$.

Exercise 6.9 *Consider a bounded one-dimensional domain* $\Omega = (a, b)$, *problem (6.7) with the boundary conditions (6.11), and a space* $V_{h,p} \subset H_0^2(\Omega)$ *consisting of smooth, piecewise-quadratic functions over a mesh* $\mathcal{T}_{h,p} = \{K_1, K_2, \ldots, K_M\}$, $K_i = (x_{i-1}, x_i)$.

1. *What is the dimension* N *of the space* $V_{h,p}$?
2. *Design* N *basis functions of the space* $V_{h,p}$, *whose supports do not extend over more than three elements.*

Exercise 6.10 *Verify that the shape functions (6.29), (6.30) satisfy the delta property (6.28),*

$$L_j\left(\theta_k^{(i)}\right) = \delta_{jk} \qquad \textit{for all } 1 \le j, k \le 4.$$

Exercise 6.11 *Calculate the dimension of the space* $V_{h,p}$ *in (6.39).*

Exercise 6.12 *Construct a fifth-order Hermite element on the reference domain* $K_a = (-1, 1)$ *using two interior Gauss–Lobatto points* $\pm 0.44721359549995793928183 47$.

Exercise 6.13 *Consider the cubic, fourth-order and fifth-order Hermite elements on the reference domain K_a. Construct and plot the corresponding Hermite interpolants of the function $g(x) = \arctan(10x)$.*

Exercise 6.14 *Consider the biharmonic problem*

$$\Delta^2 u(x) = 8\pi^4 \left(2\sin^2(\pi x) - 1\right)$$

in the interval $\Omega = (a, b) = (0, 10)$, equipped with the boundary conditions

$$u(a) = u(b) = u'(a) = u'(b) = 0.$$

1. *Calculate the exact solution u.*
2. *Consider a mesh consisting of $M = 2, 5, 10, 15, 20, 30, 40$ and 50 cubic Hermite elements.*
   (a) *Calculate the approximate solution $u_{h,p}$ for each M. Present the plots of u and $u_{h,p}$ for $M = 2, 5, 10$ and 15.*
   (b) *Present a plot of the error in the $H^2(\Omega)$-norm, $\|u - u_{h,p}\|_{2,2}$. Use a decimal logarithmic scale. As usual, put the number of unknowns on the horizontal axis.*
3. *Do the same for fourth-order and fifth-order Hermite elements.*
4. *Compare the convergence curves. Which scheme was most efficient?*
5. *Guess the speed of convergence in all three cases.*

Exercise 6.15 *Prove Lemma 6.3.*

Exercise 6.16 *Show that the cubic triangular Hermite elements from Figure 6.17 are equivalent. Hint: Establish a one-to-one relation between the degrees of freedom associated with the pairs of the directional derivatives at the vertices.*

Exercise 6.17 *Verify that the nodal shape functions (6.59) satisfy the delta property (3.4).*

Exercise 6.18 *Show that $p - 3$ pairwise distinct Lagrange degrees of freedom placed symmetrically into the interior of each element edge are enough to ensure the global continuity of approximation for Hermite elements of the order $p \geq 3$ (this statement holds generally for mixed meshes consisting of Hermite triangles and quads of a uniform polynomial degree).*

Exercise 6.19 *Calculate the dimension of the space $V_{h,p}$ in (6.61).*

Exercise 6.20 *Perform in detail the proof of Lemma 6.8.*

Exercise 6.21 *Verify in detail the conformity of the triangular Argyris–Fekete elements of the general order $p \geq 5$ following the proof of Lemma 6.7.*

CHAPTER 7

EQUATIONS OF ELECTROMAGNETICS

In this chapter we introduce the basic quantities of electromagnetics, formulate their relations in terms of partial differential equations, and show how these equations can be solved via the finite element method. Emphasis is given to potential equations and to the Maxwell's equations, with particular interest in the time-harmonic field. We do not attempt to cover all interesting aspects of theoretical and computational electromagnetics: It is our goal to provide a sufficiently informative introduction that (a) should allow the reader to start solving practical problems and (b) prepare her/him for the study of more specialized literature. To mention just two books, [83] can be recommended to mathematically oriented readers who are especially interested in time-harmonic Maxwell's equations, and [102] addresses engineering audience.

Section 7.1 presents important basic facts about the macroscopic (continuous) model of the electromagnetic field, such as the four basic laws of electromagnetics, the Maxwell's equations in the integral and differential forms, media characteristics, basic properties of conductors, dielectrics and magnetic materials, and interface conditions. With an appropriate insight, many typical problems of electromagnetics can be formulated in terms of potentials and solved by means of the standard continuous finite elements. The scalar electric potential and the scalar and vector magnetic potentials are introduced in Section 7.2. The equations for the field vectors and the time-harmonic Maxwell's equations are derived in Section 7.3.

The rest of the chapter is devoted to the weak formulation and finite element analysis of the time-harmonic Maxwell's equations by means of edge elements. In Section 7.4 we define the Hilbert space $\boldsymbol{H}(\mathrm{curl})$, derive the weak formulation of the equations, show how various types of boundary conditions are incorporated into the sesquilinear weak form, and prove the existence and uniqueness of the weak solution in a simplified setting.

In Section 7.5 we perform the standard series of steps involved in the finite element method: We introduce the lowest-order Whitney element and the general higher-order edge element of Nédélec on the reference domain, use appropriate transformation to construct the basis functions in physical mesh elements, and transform the weak formulation of the Maxwell's equations to the reference domain. At the end the interpolation on higher-order nodal edge elements is discussed.

7.1 ELECTROMAGNETIC FIELD AND ITS BASIC CHARACTERISTICS

The macroscopic theory of the electromagnetic field is based on the following four vector quantities:

- electric field strength $\boldsymbol{E} = \boldsymbol{E}(\boldsymbol{x}, t)$,
- electric flux density $\boldsymbol{D} = \boldsymbol{D}(\boldsymbol{x}, t)$,
- magnetic field strength $\boldsymbol{H} = \boldsymbol{H}(\boldsymbol{x}, t)$,
- magnetic flux density $\boldsymbol{B} = \boldsymbol{B}(\boldsymbol{x}, t)$.

Based on empirical experience, it is reasonable to assume that these quantities are continuous and continuously differentiable almost everywhere in the computational domain, except for sets of zero measure such as interfaces separating materials with different electromagnetic properties. The points where the field is continuous are called regular, the others are singular. The electromagnetic field may be classified with respect to a number of various properties and characteristics, for example:

- field sources (electric charges, currents, permanent magnets),
- dimensionality given by the lowest number of coordinates that fully describe the field distribution (1D, 2D, 3D models),
- boundedness (fields bounded in a finite domain or open-boundary fields),
- time evolution of the field quantities [static (stationary) fields, time-harmonic fields, general time dependencies],
- types of media (homogeneous or inhomogeneous, linear or nonlinear, isotropic or anisotropic, disperse or indisperse),
- motion of sources or media,

and others.

7.1.1 Integration along smooth curves

In this chapter we will need to integrate both scalar and vector fields along smooth curves. Without loss of generality, we can assume that a smooth curve $\mathcal{C} \subset \mathbb{R}^d$ always can be parameterized from the interval $(0, 1)$, i.e.,

$$C = C(s) = (C_1, \dots, C_d)(s), \qquad s \in (0,1),$$

as shown in Figure 7.1.

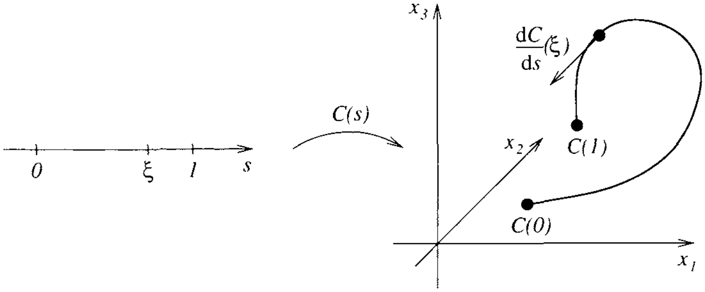


Figure 7.1 Parameterization of a smooth curve and its derivative.

For simplicity we use the same symbol C for the curve and its parameterization. A curve C is called smooth if the derivatives $C_i'(s) = (dC_i/ds)(s)$ of all of its components are continuous in $(0,1)$. Without loss of generality, we assume that the parameterization $C(s)$ of a smooth curve C satisfies the condition

$$|C'(s)| \neq 0 \quad \text{for all } s \in (0,1).$$

Then for every $\xi \in (0,1)$ the derivative $(dC/ds)(\xi)$ is a vector tangential to the curve C at the point $C(\xi) \in \mathbb{R}^d$. A curve C is said to be closed if it is defined in $[0,1]$ and $C(0) = C(1)$.

A scalar field $\varphi : \mathbb{R}^d \to \mathbb{R}$ is integrated along the curve C using the standard formula

$$\int_C \varphi \, dC = \int_0^1 \varphi(C(s))|C'(s)| \, ds,$$

where $|C'(s)|$ is the magnitude of the derivative $C'(s)$,

$$|C'(s)| = \sqrt{\left(\frac{dC_1}{ds}\right)^2 + \ldots + \left(\frac{dC_d}{ds}\right)^2}.$$

For example, the length of C is obtained by integrating the function $\varphi(s) = 1$,

$$\int_C 1 \, dC = \int_0^1 \sqrt{\left(\frac{dC_1}{ds}\right)^2 + \ldots + \left(\frac{dC_d}{ds}\right)^2} \, ds.$$

Vector fields $\boldsymbol{F} : \mathbb{R}^d \to \mathbb{R}^d$ are integrated along the curve C using another standard formula,

$$\int_C \boldsymbol{F} \cdot dC = \int_0^1 \boldsymbol{F}(C(s)) \cdot C'(s) \, ds.$$

7.1.2 Maxwell's equations in integral form

The mathematical model of the electromagnetic field, that nowadays is known as the Maxwell's equations, first appeared in the *Treatise on Electricity and Magnetism* by James Clerk Maxwell in 1873. These equations are assumed to be one of the greatest achievements of the 19th-century mathematics. Among Maxwell's other remarkable contributions were (a) the observation that light is an electromagnetic phenomenon (around 1862) and (b) the development of the Maxwell–Boltzmann kinetic theory of gases, which he published independently of Ludwig Boltzmann in 1866.

Figure 7.2 James Clerk Maxwell (1831–1879).

The Maxwell's equations consist of Ampère's law, Faraday's law of induction, and Gauss' laws for electricity and magnetism. Consider a planar simply-connected area $\mathcal{A}$ whose boundary C is a closed smooth curve. **Ampère's law**,

$$\int_C \boldsymbol{H} \cdot \mathrm{d}C = I + \frac{\mathrm{d}\Psi}{\mathrm{d}t},$$

postulates that the line integral of the tangential component of the magnetic field strength $\boldsymbol{H}$ along C is proportional to the total current passing through the area $\mathcal{A}$ in the normal direction. This current is given by the sum of the conductive current I and displacement current $\mathrm{d}\Psi/\mathrm{d}t$. The conductive current I is a scalar quantity defined by

$$I = \int_{\mathcal{A}} \boldsymbol{J} \cdot \boldsymbol{\nu}\, \mathrm{d}S,$$

where $\boldsymbol{J}$ stands for the vector-valued density of conductive currents. The dielectric flux Ψ is defined by

$$\Psi = \int_{\mathcal{A}} \boldsymbol{D} \cdot \boldsymbol{\nu}\, \mathrm{d}S,$$

where $\boldsymbol{D}$ is the electric flux density and the symbol $\boldsymbol{\nu}$ stands for the unit normal vector to $\mathcal{A}$, oriented positively with respect to the orientation of the curve C (right-hand rule). **Faraday's law of induction,**

$$\int_C \boldsymbol{E} \cdot \mathrm{d}C = -\frac{\mathrm{d}\Phi}{\mathrm{d}t}.$$

represents an analogous rule for the electric field strength $\boldsymbol{E}$: The line integral of the tangential component of the electric field $\boldsymbol{E}$ along any closed smooth planar loop C is equal to the negative of the rate of temporal change of the magnetic flux Φ through the corresponding area $\mathcal{A}$ in the normal direction. The magnetic flux Φ is defined by

$$\Phi = \int_{\mathcal{A}} \boldsymbol{B} \cdot \boldsymbol{\nu}\, \mathrm{d}S.$$

Gauss' law for electricity,

$$\Psi = \int_S \boldsymbol{D} \cdot \boldsymbol{\nu}\, \mathrm{d}S = Q. \tag{7.1}$$

says that the total dielectric flux Ψ out of any (simply-connected) volume $\mathcal{V}$ with a sufficiently regular boundary S is equal to the total electric charge Q contained in the volume $\mathcal{V}$. The total electric charge Q is defined by

$$Q = \int_{\mathcal{V}} \varrho\, \mathrm{d}\boldsymbol{x}.$$

where ϱ is the electric charge density. The symbol $\boldsymbol{\nu}(\boldsymbol{x})$ stands for the outer normal vector to the surface S at a point $\boldsymbol{x} \in S$. Finally, **Gauss' law for magnetism**,

$$\int_S \boldsymbol{B} \cdot \boldsymbol{\nu}\, \mathrm{d}S = 0. \tag{7.2}$$

postulates that the magnetic flux Φ out of any volume $\mathcal{V}$ with a boundary S is zero, or, in other words, that the magnetic field is divergence-free (solenoidal).

The main advantage of the integral form of the Maxwell's equations is that it provides a good idea about the relations between the field sources and field quantities. Its computational application, however, is limited to rather simple problems, characterized by trivial geometries and linear material properties. For practical purposes it is desirable to transform the Maxwell's equations (7.1.2)–(7.2) into partial differential equations.

7.1.3 Maxwell's equations in differential form

The transformation of equations (7.1.2)–(7.2) into partial differential equations is done by means of Stokes' and Gauss' theorems of calculus (see, e.g., [36]). Let us begin with Ampère's law: Applying Stokes' theorem to (7.1.2), we obtain

$$\int_{\mathcal{A}} (\nabla \times \boldsymbol{H}) \cdot \boldsymbol{\nu}\, \mathrm{d}S = \int_{\mathcal{A}} \left(\boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t} \right) \cdot \boldsymbol{\nu}\, \mathrm{d}S,$$

where $\boldsymbol{\nu}$ is the outer normal unit vector to the area $\mathcal{A}$. From the fact that the area $\mathcal{A}$ is arbitrary it follows that

$$\nabla \times \boldsymbol{H} = \boldsymbol{J} + \frac{\partial \boldsymbol{D}}{\partial t}. \tag{7.3}$$

Analogously, Faraday's law (7.1.2) leads to

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}, \tag{7.4}$$

and Gauss' theorems (7.1) and (7.2) yield

$$\nabla \cdot \boldsymbol{D} = \varrho \tag{7.5}$$

and

$$\nabla \cdot \boldsymbol{B} = 0. \tag{7.6}$$

Let us remark that the density of conductive currents $\boldsymbol{J}$ may include both source currents and eddy currents. The above equations hold exactly only at the regular points of the domain, on interfaces one has to impose special interface conditions (to be formulated later in Paragraph 7.1.8).

Most methods of computational electromagnetics (both analytical and numerical) are based on the differential form of the Maxwell's equations. The main advantage of the PDE model is its ability to include nonlinearities, anisotropy and other nontrivial aspects of field computations. Next let us formulate the constitutive relations between the field vectors and physical properties of involved media, which form an indivisible part of the electromagnetic field model.

7.1.4 Constitutive relations and the equation of continuity

The field vectors $\boldsymbol{E}$, $\boldsymbol{D}$, $\boldsymbol{H}$, and $\boldsymbol{B}$ are coupled with the media via the relations

$$\begin{aligned} \boldsymbol{D} &= \epsilon \boldsymbol{E}, \qquad (7.7)\\ \boldsymbol{B} &= \mu \boldsymbol{H}, \qquad (7.8)\\ \boldsymbol{J} &= \gamma(\boldsymbol{E} + \boldsymbol{E}_{\mathrm{v}}). \qquad (7.9) \end{aligned}$$

The symbols ϵ, μ, and γ denote the permittivity, magnetic permeability, and electric conductivity, respectively. The material parameters are generally tensors that may either be constant, or functions of the position, direction, local values of the field, frequency, or state variables (such as temperature or pressure). It is worth mentioning that in the special isotropic case when a tensor is diagonal with equal diagonal entries, the tensor-vector product can formally be replaced with the corresponding product of the vector and the diagonal entry. The quantity $\boldsymbol{E}_{\mathrm{v}}$ is the intensity of applied forces of, for instance, electrochemical, photovoltaic, or thermoelectric origin. For further reference by

$$\boldsymbol{J}_a = \gamma \boldsymbol{E}_{\mathrm{v}}$$

we denote the applied current density.

Equation of continuity Taking the divergence of Ampère's law (7.3) and using Gauss' law for electricity (7.5), under sufficient regularity assumptions on all involved quantities, one obtains the continuity equation for the conductive current,

$$\frac{\partial \varrho}{\partial t} + \nabla \cdot \boldsymbol{J} = 0. \tag{7.10}$$

This equation, analogously to the Maxwell's equations in the differential form, only holds where $\boldsymbol{J}$ is smooth.

7.1.5 Media and their characteristics

From the point of view of their electromagnetic properties, media can be split into three basic categories: conductors, dielectrics, and magnetic materials. We find it useful to describe these three material types in more detail in Paragraphs 7.1.6–7.1.7, after mentioning some of their more general attributes:

- A medium is called homogeneous when its parameters (permittivity, permeability, electric conductivity, and others) are independent of the position. In the opposite case the medium is inhomogeneous. An example of a homogeneous medium is a copper conductor, while imperfectly mixed electrolyte represents an inhomogeneous medium. A medium is called homogeneous by parts if it consists of several homogeneous subdomains with different material constants.
- A medium is called linear when its parameters are independent of the electromagnetic field. This property is typical for air, various gases, many liquids, and nonmagnetic metals, such as aluminum, copper, or stainless steel. In nonlinear media, some of the parameters are field-dependent (such as, for example, the magnetic permeability of iron).
- A medium is called isotropic when its physical properties do not depend on the direction of the electromagnetic field. As mentioned above, in such case the material parameters can be observed as scalar quantities (in the general case they are tensors). To give some examples of anisotropic materials, let us mention cold rolled oriented steel sheets for magnetic cores and piezoelectric materials.
- A medium is called disperse when its physical parameters are dependent on the frequency of the electromagnetic field. Media independent of the frequency are called indisperse.

7.1.6 Conductors and dielectrics

Perfect conductors are supposed to contain an unlimited amount of free charges. An external electric field produces motion of these charges to an equilibrium position characterized by zero internal field in the material (the field due to free charges in the conductor is exactly opposite to the original external field). The time necessary for such a redistribution of charges in good conductors (silver, copper, aluminum, etc.) is in normal conditions of the order of 10^{-18} s. That is why we can consider this redistribution practically instantaneous except for modeling extremely high frequency effects.

7.1.6.1 Dielectrics The atoms and molecules of materials with no free charges contain bound charges. Therefore an external electric field $\boldsymbol{E}(\boldsymbol{x}, t)$ turns them into elementary electric dipoles that generate another electric field in the opposite direction. This effect, which is called electric polarization, may be quantified by the polarization vector $\boldsymbol{P}(\boldsymbol{x}, t)$ that gives the volume density of the moments of the elementary dipoles. This vector may be expressed in terms of the electric field $\boldsymbol{E}$,

$$\boldsymbol{P} = \epsilon_0 \chi_e \boldsymbol{E} \tag{7.11}$$

where $\epsilon_0 \doteq 10^{-9}/36\pi$ [F/m] is the permittivity of vacuum. The susceptibility of the material χ_e may exhibit scalar or tensorial character. When, for simplicity, χ_e is a scalar, the electric flux density $\boldsymbol{D}(\boldsymbol{x}, t)$ in the material consists of the applied flux density and the polarization vector. This can be expressed as

$$\boldsymbol{D} = \epsilon_0 \boldsymbol{E} + \epsilon_0 \chi_e \boldsymbol{E} = \epsilon_0 (1 + \chi_e) \boldsymbol{E} = \epsilon_0 \epsilon_r \boldsymbol{E} = \epsilon \boldsymbol{E}\,. \tag{7.12}$$

The quantity

$$\epsilon_r = 1 + \chi_e,$$

which is a tensor in the general case, is referred to as the relative permittivity of the material. According to the relations between the vectors $\boldsymbol{E}$, $\boldsymbol{P}$, and $\boldsymbol{D}$, we split media into

- dielectrically linear and nonlinear: In dielectrically linear materials the relations of $\boldsymbol{E}$, $\boldsymbol{P}$ and $\boldsymbol{D}$ are linear and vice versa.
- dielectrically soft and hard: In dielectrically soft materials both $\boldsymbol{P} = \boldsymbol{0}$ and $\boldsymbol{D} = \boldsymbol{0}$ when $\boldsymbol{E} = \boldsymbol{0}$. Dielectrically hard materials exhibit nonzero polarization and/or electric flux density even with no electric field $\boldsymbol{E}$ present.
- dielectrically isotropic and anisotropic: In dielectrically isotropic materials all three vectors $\boldsymbol{E}$, $\boldsymbol{P}$ and $\boldsymbol{D}$ are collinear and vice versa.

Most dielectric materials are linear and perfectly soft. Some of them, called pyroelectrics, exhibit within specific temperature ranges spontaneous polarization (while $\boldsymbol{E} = \boldsymbol{0}$). The polarization also can be affected by mechanical strains and stresses or various state variables. In various applications, electrically conductive materials are modeled by sufficiently high value of ϵ_r (the higher the polarization, the lower the electric field inside them).

7.1.7 Magnetic materials

Similarly, an external magnetic field $\boldsymbol{H}(\boldsymbol{x}, t)$ influences the motion of electrons in particular atoms and, consequently, their magnetic moment. According to the value of the moment, we split materials into diamagnetic, paramagnetic and ferromagnetic. *Diamagnetic materials* exhibit no magnetic moment in the absence of external field $\boldsymbol{H}$. When such a field is applied, it affects the motion of electrons and a new magnetic field, acting against the original field $\boldsymbol{H}$, is induced. Consequently, the original field $\boldsymbol{H}$ is weakened. Particles (atoms, ions, molecules) in *paramagnetic materials* are characterized by a nonzero magnetic moment even with zero external field $\boldsymbol{H}$. After applying a nonzero magnetic field $\boldsymbol{H}$, the microscopic moments orient themselves in its direction, causing its moderate strengthening. *Ferromagnetic materials* contain, in addition to nonzero magnetic moments analogous to paramagnetic materials, so-called Weiss' domains in which particular moments exhibit the same direction. These directions generally differ from one Weiss' domain to another, so that their effects are mutually compensated. An external magnetic field $\boldsymbol{H}$ orients the microscopic moments in individual domains according to its direction, causing its significant strenghtening.

The described effects are modeled in terms of a vector quantity $\boldsymbol{M}(\boldsymbol{x}, t)$ referred to as magnetization. The basic relation between vectors $\boldsymbol{H}$, $\boldsymbol{M}$ and the magnetic flux density $\boldsymbol{B}$ is given by the formula

$$\boldsymbol{B} = \mu_0(\boldsymbol{H} + \boldsymbol{M}), \tag{7.13}$$

where $\mu_0 \doteq 4\pi 10^{-7}$ [H/m] is the magnetic permeability of vacuum. The magnetization $\boldsymbol{M}$ is a function of the field $\boldsymbol{H}$,

$$\boldsymbol{M} = \chi_{\mathrm{m}}\boldsymbol{H}. \tag{7.14}$$

Here χ_{m} denotes the magnetic susceptibility that, again, is of either scalar or tensor character. Substituting (7.14) into (7.13), one obtains

$$\boldsymbol{B} = \mu_0(1 + \chi_{\mathrm{m}})\boldsymbol{H} = \mu_0\mu_r\boldsymbol{H} = \mu\boldsymbol{H}. \tag{7.15}$$

The quantity $\mu_r = 1 + \chi_{\mathrm{m}}$ is called relative magnetic permeability of the material.

Analogously to dielectrics, also magnetic materials are split into linear and nonlinear, soft and hard, and isotropic and anisotropic. In the rest of this paragraph let us say a few words about ferromagnetic materials, which are of great practical importance.

Ferromagnetics Ferromagnetics are nonlinear materials in which the fields $\boldsymbol{M}$ and $\boldsymbol{B}$ are functions of both the magnetic field $\boldsymbol{H}$ and the past history of the material. The magnetization $\boldsymbol{M}$ initially grows with $\boldsymbol{H}$, but from some given magnitude of $\boldsymbol{H}$, which is typical for the given material, the magnitude of $\boldsymbol{M}$ practically does not change anymore. Then we say that the ferromagnetic material is saturated (the microscopic magnetic moments in all internal Weiss' domains are oriented according to the direction of the external field $\boldsymbol{H}$). This behavior, moreover, significantly depends on the temperature of the material. After exceeding the Curie's point the originally ferromagnetic material becomes paramagnetic.

The steady-state dependence of $\boldsymbol{B}$ on $\boldsymbol{H}$ is given by the hysteresis curve. This curve is narrow in the case of soft ferromagnetics and wide for hard ferromagnetics. For the sake of simplicity, however, we often approximate at least narrow hysteresis curves by magnetization curves that are obtained for the first magnetization of the material. In general, the modeling of hysteresis curves is very difficult.

7.1.8 Conditions on interfaces

The differential form of the Maxwell's equations is not defined on material interfaces where the partial derivatives of field quantities are generally discontinuous. Therefore the PDEs have to be completed by suitable interface conditions, which are derived from the original integral form of the equations.

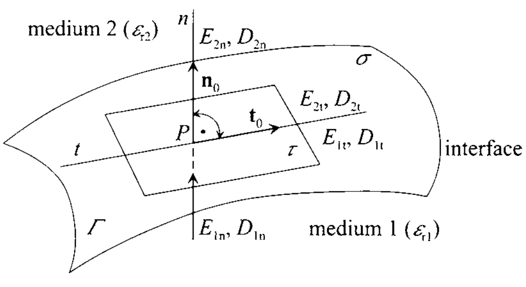


Figure 7.3 Electric field on a media interface.

Consider an interface Γ, shown in Figure 7.3, which separates two media of different relative permittivities ϵ_{r1} and ϵ_{r2}, and some point $P \in \Gamma$ where Γ is smooth. By $\sigma(P)$ denote the surface density of the electric charge at the point P. Let τ be the tangential plane to the interface Γ at the point P. Consider the line n passing through P in the normal direction to Γ, and another line t passing through P in any direction tangential to Γ. The symbols $\boldsymbol{t}_0$ and $\boldsymbol{\nu}_0$ represent the unitary vectors in the directions t and n, respectively.

The interface conditions for the electric field at the point P follow from the integral equations (7.1.2) and (7.1). The tangential component of the electric field $\boldsymbol{E}$ is continuous at P, and the normal component of the electric flux density $\boldsymbol{D}$ has a jump of the magnitude $\sigma(P)$:

$$E_{1t}(P) = E_{2t}(P), \;\; D_{2n}(P) - D_{1n}(P) = \sigma(P). \tag{7.16}$$

Consider an analogous arrangement (Figure 7.4) with two materials of different relative magnetic permeabilities μ_{r1} and μ_{r2}. The interface carries an electric current of the surface density K_t.

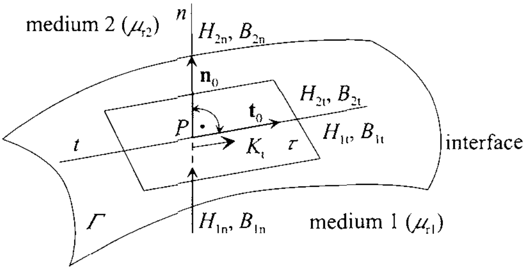


Figure 7.4 Magnetic field on a media interface.

The interface conditions for the magnetic field follow from the integral equations (7.1.2) and (7.2). The normal component of the magnetic flux density $\boldsymbol{B}$ is continuous at P, while the tangential component of the magnetic field $\boldsymbol{H}$ has at P a jump of the magnitude $K_t(P)$:

$$B_{1n}(P) = B_{2n}(P), \;\; H_{2t}(P) - H_{1t}(P) = K_t(P), \tag{7.17}$$

Finally consider an interface of two media with different electric conductivities γ_{r1} and γ_{r2} (Figure 7.5). Assume that an electric current crosses the interface.

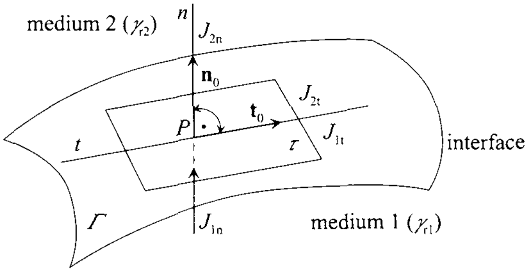


Figure 7.5 Current field on a media interface.

It follows from the continuity equation (7.10) that the normal component of the current density $\boldsymbol{J}$ is continuous across Γ,

$$J_{1\mathrm{n}}(P) = J_{2\mathrm{n}}(P). \tag{7.18}$$

As we shall see further, the weak formulation of the Maxwell's equations used in this text takes care about these conditions automatically.

7.2 POTENTIALS

The finite element approximation of the field vectors $\boldsymbol{E}$ and $\boldsymbol{H}$ requires the application of special vector-valued finite elements (edge elements). These elements are more difficult to deal with than the standard continuous elements. For example, the electric field $\boldsymbol{E}$ is discontinuous on material interfaces where the scalar potential φ is continuous. At reentrant corners, where the scalar potential φ remains continuous and bounded, the electric field $\boldsymbol{E}$ often diverges to infinity.

Therefore we find it useful to mention situations when the Maxwell's equations reduce to simpler problems solvable by means of the standard continuous elements.

7.2.1 Scalar electric potential

It is well known that every smooth vector field $\boldsymbol{F}$ that is irrotational,

$$\nabla \times \boldsymbol{F} = \boldsymbol{0},$$

is the gradient of some scalar function ϕ,

$$\boldsymbol{F} = \nabla(\phi + C),$$

where C is an arbitrary constant. The function ϕ is called the potential of $\boldsymbol{F}$. In a stationary electric field ($\boldsymbol{E} = \boldsymbol{E}(\boldsymbol{x})$ and $\boldsymbol{D} = \boldsymbol{D}(\boldsymbol{x})$), Faraday's law (7.4) reduces to

$$\nabla \times \boldsymbol{E} = \boldsymbol{0}, \tag{7.19}$$

which means that $\boldsymbol{E}$ can be written in the form

$$\boldsymbol{E} = -\nabla(\varphi_e + C), \tag{7.20}$$

where φ_e is referred to as the electric potential. The minus sign in (7.20) is a standard convention, corresponding to the fact that (positive) work has to be done when a charge is moved toward a field produced by charge(s) of the same sign. The electric potential may be interpreted as the work needed to move a unit charge from one point of the electric field to another point. The constant C in the electric potential may be determined according to various criteria, for example, from the requirement $\varphi_e(\boldsymbol{x}) \to 0$ as $|\boldsymbol{x}| \to \infty$.

It follows from (7.20) that for any two points $A, B \in \mathbb{R}^d$ that are connected through a smooth curve $\mathcal{C} : (0, 1) \to \mathbb{R}^d$, the following holds:

$$\int_C \boldsymbol{E} \cdot \mathrm{d}C = \int_0^1 \boldsymbol{E}(C(s)) \cdot C'(s)\,\mathrm{d}s = -\int_0^1 \nabla\varphi_e(C(s)) \cdot C'(s)\,\mathrm{d}s = \varphi_e(A) - \varphi_e(B). \tag{7.21}$$

The difference of the electric potentials at points A and B is called voltage and denoted by u_{AB}. If the loop C is closed, the following holds:

$$\int_C \boldsymbol{E} \cdot \mathrm{d}C = 0 \tag{7.22}$$

(fields with this property are called conservative).

Point sets with the same potential (curves in 2D and surfaces in 3D) are called equipotentials. In 2D an equipotential curve $C \subset \mathbb{R}^2$ starting from a point $A \in \mathbb{R}^2$ can be constructed easily (numerically) using the relation

$$\varphi_e(C(s)) = \text{const} \Leftrightarrow \nabla\varphi_e(C(s)) \cdot C'(s) = 0 \Leftrightarrow \boldsymbol{E}(C(s)) \cdot C'(s) = 0 \tag{7.23}$$

[i.e., C is perpendicular to the field vector $\boldsymbol{E}$ at every its point $C(s)$]. The construction of equipotential surfaces in 3D is more difficult (and it may not a bad idea to leave this task to a visualization software).

Lines orthogonal to equipotentials are called force lines. Both in 2D and 3D they can be calculated easily via the relation

$$\nabla\varphi_e(C(s)) \times C'(s) = \mathbf{0} \Leftrightarrow \boldsymbol{E}(C(s)) \times C'(s) = \mathbf{0} \tag{7.24}$$

[i.e., the field vector $\boldsymbol{E}$ is tangential to C at every its point $C(s)$]. The force lines connect different potential levels and, indeed, are not closed curves.

Equation for φ_e Putting together Gauss' law for electricity (7.5), the constitutive relation (7.7), and the gradient expression (7.20) for the stationary electric field $\boldsymbol{E}$, we obtain a second-order elliptic partial differential equation

$$-\nabla \cdot (\epsilon\nabla\varphi_e) = \varrho. \tag{7.25}$$

This equation attains an especially simple form in the isotropic homogeneous case,

$$-\Delta\varphi_e = \frac{\varrho}{\epsilon}. \tag{7.26}$$

Equation (7.25) is considered in some bounded domain $\Omega \subset \mathbb{R}^d$ and equipped with standard boundary conditions for second-order elliptic problems (see Paragraphs 1.2.5 and 1.2.6). The Dirichlet conditions represent a prescribed potential (voltage). Homogeneous Neumann conditions are prescribed on the line/plane of symmetry in the case of symmetric problems, and nonhomogeneous Neumann conditions generally on the part of the boundary where the normal component $\boldsymbol{E} \cdot \boldsymbol{\nu}$ of the electric field (which is equal to $-\partial\varphi_e/\partial\boldsymbol{\nu}$) is given. Homogeneous Neumann boundary condition may also be used, for example, far from the source where it is reasonable to assume that the field does not change anymore.

Variational formulation and unique solvability A variational formulation of the form (1.66) is obtained in the standard way. It is worth mentioning that it requires the components of ϵ to be L^∞-functions. Thus piecewise discontinuous coefficients corresponding to various materials are indeed possible, and the resulting potential still is a H^1-function. The existence and uniqueness of solution is a consequence of the Lax–Milgram lemma, as it was described in Paragraph 1.2.8 (under the assumption that the part Γ_D of $\partial\Omega$ corresponding to the Dirichlet boundary conditions is not empty).

Calculation of E After calculating the (continuous, elementwise-polynomial) distribution of the scalar electric potential φ_e in the computational domain Ω_h, the electric field $\boldsymbol{E}$ is obtained via the relation (7.20). It is interesting to observe that the tangential component of $\boldsymbol{E} = -\nabla\varphi_e$ is continuous, i.e., $\boldsymbol{E}$ lies in the desired Hilbert space $\boldsymbol{H}(\mathrm{curl})$.

7.2.2 Scalar magnetic potential

For a stationary electromagnetic field Ampère's law (7.3) reduces to $\nabla \times \boldsymbol{H} = \boldsymbol{J}$ ($\partial \boldsymbol{D}/\partial t = \boldsymbol{0}$ is frequently assumed also for nonstationary fields with sufficiently slow time-variation). In domains where $\boldsymbol{J} = \boldsymbol{0}$, such as in the air and other insulators, the field $\boldsymbol{H}$ is irrotational,

$$\nabla \times \boldsymbol{H} = \boldsymbol{0}. \tag{7.27}$$

Then one can introduce the scalar magnetic potential φ_m such that

$$\boldsymbol{H} = -\nabla(\varphi_m + C), \tag{7.28}$$

where C is an arbitrary constant. This constant can be defined, for example, by requesting $\varphi_m = 0$ somewhere. Gauss' law for magnetism (7.6) together with the constitutive relation (7.8) yield a second-order elliptic equation

$$-\nabla \cdot (\mu \nabla \varphi_m) = 0,$$

which is analogous to the potential equation (7.25). The properties of the magnetic potential φ_m are analogous to the electric potential φ. It is worth mentioning that the above model does not cover a conductor–insulator interface. On such interfaces one has to consider interface conditions from Paragraph 7.1.8. Equipotentials and force lines are defined in the same way as for the scalar electric potential φ.

7.2.3 Vector potential and gauge transformations

After introducing the scalar potentials φ_e and φ_m for the fields $\boldsymbol{E}$ and $\boldsymbol{H}$ in the stationary case in Paragraphs 7.2.1 and 7.2.2, let us proceed to the general time-dependent case. It is well known that any sufficiently regular vector field $\boldsymbol{f}$ that is divergence-free (solenoidal),

$$\nabla \cdot \boldsymbol{f} = 0,$$

can be written in the form

$$\boldsymbol{f} = \nabla \times \boldsymbol{F},$$

where the field $\boldsymbol{F}$ is called the vector potential of $\boldsymbol{f}$. Gauss' law for magnetism (7.6) yields that the (nonstationary) divergence-free magnetic flux density $\boldsymbol{B}(\boldsymbol{x}, t)$ can be expressed by means of a vector magnetic potential $\boldsymbol{A}(\boldsymbol{x}, t)$,

$$\boldsymbol{B} = \nabla \times \boldsymbol{A}. \tag{7.29}$$

Faraday's law (7.4) yields

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t} = -\frac{\partial(\nabla \times \boldsymbol{A})}{\partial t},$$

and therefore (if all partial derivatives of $\boldsymbol{A}$ are continuous),

$$\nabla \times \left(\boldsymbol{E} + \frac{\partial \boldsymbol{A}}{\partial t} \right) = \boldsymbol{0}.$$

This irrotational vector field can be expressed as the gradient of a scalar function $\varphi(\boldsymbol{x}, t)$,

$$\boldsymbol{E} + \frac{\partial \boldsymbol{A}}{\partial t} = -\nabla\varphi,$$

where φ is a time-dependent generalization of the scalar electric potential φ_e from (7.20). Thus the electric field $\boldsymbol{E}$ has the form

$$\boldsymbol{E} = -\nabla\varphi - \frac{\partial \boldsymbol{A}}{\partial t}. \tag{7.30}$$

The potentials $\boldsymbol{A}$ and φ are not unique: Many different pairs of $\boldsymbol{A}$ and φ generate the same fields $\boldsymbol{B}$ and $\boldsymbol{E}$. It is easy to verify that equations (7.29) and (7.30) are invariant under the transformations

$$\tilde{\varphi} = \varphi + \frac{\partial \varphi}{\partial t} + C, \tag{7.31}$$

$$\tilde{\boldsymbol{A}} = \boldsymbol{A} - \nabla\varphi, \tag{7.32}$$

where C is an arbitrary real constant. Transformations (7.31) and (7.32) are called gauge transformations.

Coulomb and Lorentz gauges While the constant C may be made unique by requesting $\varphi(\boldsymbol{x}) \to 0$ as $|\boldsymbol{x}| \to \infty$, various uniqueness conditions may be imposed on $\boldsymbol{A}$ and φ. The most frequently used condition in the stationary case is the Coulomb gauge

$$\nabla \cdot \boldsymbol{A} = 0. \tag{7.33}$$

In the nonstationary case, one often uses the more general Lorentz gauge

$$\nabla \cdot \boldsymbol{A} + \frac{1}{c^2}\frac{\partial \varphi}{\partial t} = 0, \tag{7.34}$$

where $c^2 = 1/(\epsilon_0\mu_0)$ is the square of the speed of light in the vacuum. The Lorentz gauge is naturally motivated in the potential formulation of the Maxwell's equations, which we discuss in the next paragraph.

7.2.4 Potential formulation of Maxwell's equations

The scalar and vector potentials introduced in Paragraph 7.2.3 can be used to highlight the wave structure of the Maxwell's equations. For simplicity let us stay in an isotropic homogeneous material of permittivity ϵ_0 and permeability μ_0.

Begin with Ampère's law (7.3) and use the constitutive relation (7.7) to obtain

$$\nabla \times \boldsymbol{H} = \boldsymbol{J} + \epsilon_0 \frac{\partial \boldsymbol{E}}{\partial t}.$$

Substituting further for $\boldsymbol{E}$ from (7.30) and using the constitutive relation (7.8), we have

$$\nabla \times \boldsymbol{B} = \mu_0 \boldsymbol{J} + \epsilon_0 \mu_0 \frac{\partial}{\partial t} \left(-\nabla \varphi - \frac{\partial \boldsymbol{A}}{\partial t} \right).$$

Equation (7.29) together with a standard vector identity for the curl–curl operator yields

$$\nabla \times \boldsymbol{B} = \nabla \times (\nabla \times \boldsymbol{A}) = \nabla (\nabla \cdot \boldsymbol{A}) - \Delta \boldsymbol{A}.$$

Putting together the last two equations and using $c^2 = 1/(\epsilon_0 \mu_0)$, we obtain

$$\frac{1}{c^2} \frac{\partial^2 \boldsymbol{A}}{\partial t^2} - \Delta \boldsymbol{A} = \mu_0 \boldsymbol{J} - \nabla \left(\nabla \cdot \boldsymbol{A} + \frac{1}{c^2} \frac{\partial \varphi}{\partial t} \right). \tag{7.35}$$

At this point it becomes clear why the Lorentz gauge was chosen in the form (7.34). With the Lorentz gauge, (7.35) simplifies to a wave equation for the vector potential $\boldsymbol{A}$,

$$\frac{\partial^2 \boldsymbol{A}}{\partial t^2} - c^2 \Delta \boldsymbol{A} = \mu_0 c^2 \boldsymbol{J}. \tag{7.36}$$

Here the Laplace operator Δ is applied to every component of $\boldsymbol{A}$. Second-order hyperbolic equations, boundary conditions, weak formulation, and the existence and uniqueness of their solution were discussed in detail in Section 1.4.

7.2.5 Other wave equations

Equation (7.36) is not the only wave equation that can be derived from the Maxwell's equations. Staying with an isotropic homogeneous material, Gauss' law for electricity (7.5) together with the constitutive relation (7.7) yield

$$\nabla \cdot \boldsymbol{E} = \frac{\varrho}{\epsilon_0}.$$

Substituting for $\boldsymbol{E}$ from (7.30), we obtain

$$\nabla \cdot \left(-\nabla \varphi - \frac{\partial \boldsymbol{A}}{\partial t} \right) = \frac{\varrho}{\epsilon_0}.$$

If the partial derivatives of $\boldsymbol{A}$ are continuous, they can be interchanged and one obtains

$$-\Delta \varphi - \frac{\partial (\nabla \cdot \boldsymbol{A})}{\partial t} = \frac{\varrho}{\epsilon_0}.$$

In the stationary case this reduces to the Poisson equation (7.26). In the nonstationary case the application of the Lorentz gauge (7.34) yields a wave equation for the scalar potential φ,

$$\frac{\partial^2\varphi}{\partial t^2} - c^2\Delta\varphi = \frac{c^2\varrho}{\epsilon_0}. \tag{7.37}$$

Wave equations can also be formulated directly for the field vectors $\boldsymbol{E}$ and $\boldsymbol{H}$. Assume an empty space where

$$\varrho = 0, \ \boldsymbol{J} = \boldsymbol{0}.$$

Using Ampère's law (7.3) together with the constitutive relations (7.7) and (7.8), we obtain

$$\nabla \times \boldsymbol{B} = \epsilon_0\mu_0\frac{\partial \boldsymbol{E}}{\partial t}. \tag{7.38}$$

Taking the curl of Faraday's law (7.4), under regularity assumptions sufficient for the interchange of the temporal derivative with the curl operator on the right-hand side, we have

$$\nabla \times (\nabla \times \boldsymbol{E}) = -\frac{\partial}{\partial t}(\nabla \times \boldsymbol{B}). \tag{7.39}$$

Substituting from (7.38) into (7.39) and using the identity

$$\nabla \times (\nabla \times \boldsymbol{E}) = \nabla(\nabla\cdot\boldsymbol{E}) - \Delta\boldsymbol{E}, \tag{7.40}$$

we can write

$$\nabla(\nabla\cdot\boldsymbol{E}) - \Delta\boldsymbol{E} = -\epsilon_0\mu_0\frac{\partial^2\boldsymbol{E}}{\partial t^2}. \tag{7.41}$$

Since $\varrho = 0$, from (7.5) it follows that

$$\nabla\cdot\boldsymbol{E} = 0,$$

and thus (7.41) finally yields

$$\frac{\partial^2\boldsymbol{E}}{\partial t^2} - c^2\Delta\boldsymbol{E} = \boldsymbol{0}. \tag{7.42}$$

An identical equation holds for $\boldsymbol{B}$.

7.3 EQUATIONS FOR THE FIELD VECTORS

In the following we turn our attention to the original Maxwell's equations expressed in terms of the field vectors $\boldsymbol{E}$ and $\boldsymbol{H}$. For our purposes it is not practical to cover the material properties of the involved media in their most general form (such as anisotropy, dependence on state variables, etc.). We assume that the permittivity ϵ, permeability μ and the electric conductivity γ are scalar functions depending on the position in space only.

7.3.1 Equation for the electric field

Consider Faraday's law (7.4) divided by the permeability μ. Using the constitutive relation (7.8) and applying the curl operator, we obtain

$$\nabla \times \left(\mu^{-1} \nabla \times \boldsymbol{E}\right) = -\nabla \times \frac{\partial \boldsymbol{H}}{\partial t}. \tag{7.43}$$

If the partial derivatives of $\boldsymbol{H}$ are continuous, they can be interchanged and we obtain

$$\nabla \times \left(\mu^{-1} \nabla \times \boldsymbol{E}\right) = -\frac{\partial}{\partial t}(\nabla \times \boldsymbol{H}). \tag{7.44}$$

Substituting for $\nabla \times \boldsymbol{H}$ from Ampère's law (7.3) and using the constitutive relations (7.7) and (7.9), we can write

$$\nabla \times \left(\mu^{-1} \nabla \times \boldsymbol{E}\right) = -\frac{\partial}{\partial t}\left[\gamma \boldsymbol{E} + \boldsymbol{J}_a + \epsilon \frac{\partial \boldsymbol{E}}{\partial t}\right], \tag{7.45}$$

and finally,

$$\nabla \times \left(\mu^{-1} \nabla \times \boldsymbol{E}\right) + \gamma \frac{\partial \boldsymbol{E}}{\partial t} + \epsilon \frac{\partial^2 \boldsymbol{E}}{\partial t^2} = -\frac{\partial \boldsymbol{J}_a}{\partial t}. \tag{7.46}$$

In practice the third term on the left-hand side sometimes is neglected when dealing with lower frequencies (typically less than 1 MHz).

7.3.2 Equation for the magnetic field

Taking the curl of Ampère's law (7.3) and substituting from the constitutive relations (7.7) and (7.9), we obtain

$$\nabla \times (\nabla \times \boldsymbol{H}) = \nabla \times (\gamma \boldsymbol{E} + \boldsymbol{J}_a) + \nabla \times \frac{\partial(\epsilon \boldsymbol{E})}{\partial t}. \tag{7.47}$$

Under regularity assumptions sufficient for the interchange of the temporal and spatial derivatives, we can write

$$\nabla \times (\nabla \times \boldsymbol{H}) = \nabla \times (\gamma \boldsymbol{E}) + \nabla \times \boldsymbol{J}_a + \frac{\partial(\nabla \times \epsilon \boldsymbol{E})}{\partial t}. \tag{7.48}$$

In the case of piecewise-constant parameters γ and ϵ we can substitute for $\nabla \times \boldsymbol{E}$ from Faraday's law (7.4) and for $\boldsymbol{B}$ from the constitutive relation (7.8) to obtain

$$\mu^{-1}\nabla \times (\nabla \times \boldsymbol{H}) + \gamma \frac{\partial \boldsymbol{H}}{\partial t} + \epsilon \frac{\partial^2 \boldsymbol{H}}{\partial t^2} = \mu^{-1}\nabla \times \boldsymbol{J}_a. \tag{7.49}$$

It is easy to see that (7.49) is equivalent to (7.46) in the case of constant material parameters.

7.3.3 Interface and boundary conditions

The partial differential equations (7.46) and (7.49) only are defined at regular points of the computational domain Ω. At singular points such as material interfaces, additional conditions have to be supplemented (see Paragraph 7.1.8). Suitable weak formulation of the Maxwell's equations (to be introduced in Section 7.4) takes care about interior interfaces automatically, while conditions on external interfaces are imposed as boundary conditions.

Interface conditions First let S be an internal interface in a computational domain Ω, separating two subregions $\Omega_1, \Omega_2 \subset \Omega$ with generally different material properties, as shown in Figure 7.6.

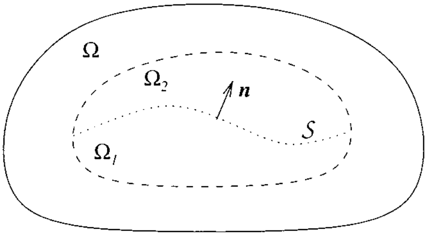


Figure 7.6 Internal interface separating regions with different material properties.

By ν denote the unit normal vector to S, defined almost everywhere at S, pointing in the direction from Ω_1 to Ω_2. By $\epsilon_i, \mu_i, \gamma_i$ denote the permittivity, permeability and electric conductivity in Ω_i. On S, equations (7.16) yield the conditions

$$(\boldsymbol{E}_1 - \boldsymbol{E}_2) \times \boldsymbol{\nu} = \boldsymbol{0}, \tag{7.50}$$

$$(\epsilon_1 \boldsymbol{E}_1 - \epsilon_2 \boldsymbol{E}_2) \cdot \boldsymbol{\nu} = \sigma \tag{7.51}$$

for the electric field strength $\boldsymbol{E}$, and analogously the relations (7.17) can be rewritten into the conditions

$$(\boldsymbol{H}_1 - \boldsymbol{H}_2) \times \boldsymbol{\nu} = \boldsymbol{K}_{\mathrm{t}}, \tag{7.52}$$

$$(\mu_1 \boldsymbol{H}_1 - \mu_2 \boldsymbol{H}_2) \cdot \boldsymbol{\nu} = 0 \tag{7.53}$$

for the magnetic field strength $\boldsymbol{H}$. Finally, (7.18) may be rearranged to

$$(\boldsymbol{J}_1 - \boldsymbol{J}_2) \times \boldsymbol{\nu} = \boldsymbol{0}. \tag{7.54}$$

Truncation boundary conditions Some problems take place in unbounded domains (air, vacuum, etc.). The easiest way to solve them is to restrict the electromagnetic field to a sufficiently large bounded domain Ω by imposing artificial boundary conditions of the form

$$\boldsymbol{E} \cdot \boldsymbol{\nu} = 0 \tag{7.55}$$

and

$$\boldsymbol{H} \cdot \boldsymbol{\nu} = 0 \tag{7.56}$$

on S. (In the case of boundary conditions the surface S forms part of the boundary $\partial\Omega$.) These conditions determine that the fields $\boldsymbol{E}$ and $\boldsymbol{H}$ are tangential to the boundary $\partial\Omega$.

Perfect conductor boundary conditions When the material in the outer domain Ω_{ext} is a perfect conductor with $\gamma_{ext} \to \infty$, it follows from the constitutive relation (7.9) that the electric field $\boldsymbol{E}_{ext}$ must vanish in Ω_{ext} for the current $\boldsymbol{J}_{ext}$ to remain finite. Then the interface condition (7.50) reduces to

$$\boldsymbol{E} \times \boldsymbol{\nu} = \mathbf{0}. \tag{7.57}$$

Imperfect conductor (impedance) boundary conditions In practice we use various boundary conditions to model imperfect conductors. One of the standard ways is to exploit the impedance Z, which quantifies the manner a material resists the flow of electric current if a given voltage is applied. The impedance differs from simple resistance in that it takes into account possible phase offset. In our case, if Ω_{ext} consists of such material, we restrict ourselves to a basic impedance boundary condition of the form

$$\boldsymbol{\nu} \times \boldsymbol{H} - Z(\boldsymbol{\nu} \times \boldsymbol{E}) \times \boldsymbol{\nu} = \mathbf{0}. \tag{7.58}$$

The impedance Z is a positive material-dependent function defined on the interface S.

Symmetry boundary conditions The impedance condition (7.58) is used with $Z = 0$ to model interfaces of symmetry,

$$\boldsymbol{H} \times \boldsymbol{\nu} = \mathbf{0}. \tag{7.59}$$

We shall see later that in the time-harmonic case, via Faraday's law (7.4), condition (7.59) yields

$$(\nabla \times \underline{\boldsymbol{E}}) \times \boldsymbol{\nu} = \mathbf{0} \tag{7.60}$$

for the phasor of the electric field (see below).

7.3.4 Time-harmonic Maxwell's equations

Assume that all time-varying quantities of the electromagnetic field are harmonic with a frequency $\omega > 0$,

$$\begin{aligned} \boldsymbol{E}(\boldsymbol{x},t) &= \mathbb{R}\mathrm{e}(\underline{\boldsymbol{E}}(\boldsymbol{x})e^{-j\omega t}), \\ \boldsymbol{D}(\boldsymbol{x},t) &= \mathbb{R}\mathrm{e}(\underline{\boldsymbol{D}}(\boldsymbol{x})e^{-j\omega t}), \\ \boldsymbol{H}(\boldsymbol{x},t) &= \mathbb{R}\mathrm{e}(\underline{\boldsymbol{H}}(\boldsymbol{x})e^{-j\omega t}), \\ \boldsymbol{B}(\boldsymbol{x},t) &= \mathbb{R}\mathrm{e}(\underline{\boldsymbol{B}}(\boldsymbol{x})e^{-j\omega t}). \end{aligned} \tag{7.61}$$

Here $\mathbb{R}\mathrm{e}(\cdot)$ denotes the real part of a complex number, j is the imaginary unit, $j^2 = -1$, and the underlined quantities are called phasors. In the language of phasors, equation (7.46) turns into

$$\nabla \times \left(\mu^{-1}\nabla \times \underline{\boldsymbol{E}}\right) - \omega(j\gamma + \epsilon\omega)\underline{\boldsymbol{E}} = j\omega\underline{\boldsymbol{J}}_a, \tag{7.62}$$

and equation (7.49) attains the form

$$\mu^{-1}\nabla \times (\nabla \times \underline{\boldsymbol{H}}) - \omega(j\gamma + \epsilon\omega)\underline{\boldsymbol{H}} = -\mu^{-1}(\nabla \times \underline{\boldsymbol{J}}_a). \tag{7.63}$$

7.3.5 Helmholtz equation

The Helmholtz equation of electromagnetics is a special case of the wave equation for a harmonic electromagnetic field. There are several versions associated with various wave equations (see Paragraphs 7.2.4 and 7.2.5). First consider an empty space characterized by the material parameters $\epsilon = \epsilon_0$, $\mu = \mu_0$, $\gamma = 0$, zero electric charge density $\varrho = 0$ and zero conductive current density $\boldsymbol{J} = \boldsymbol{0}$. When substituting from (7.61) into the wave equation (7.42), we immediately obtain

$$\Delta\underline{\boldsymbol{E}} + k^2\underline{\boldsymbol{E}} = 0, \tag{7.64}$$

where the symbol $k = \omega/c = \omega\sqrt{\epsilon_0\mu_0}$ stands for the wave number. One can obtain the same result from the time-harmonic Maxwell's equations (7.62): With (7.61), (7.7) and $\varrho = 0$ Gauss' law for electricity (7.5) reduces to

$$\nabla \cdot \underline{\boldsymbol{E}} = 0.$$

Therefore identity (7.40) yields

$$\nabla \times (\nabla \times \underline{\boldsymbol{E}}) = \nabla(\nabla \cdot \underline{\boldsymbol{E}}) - \Delta\underline{\boldsymbol{E}} = -\Delta\underline{\boldsymbol{E}}.$$

Putting this into (7.62) and using $\gamma = 0$, we obtain (7.64) again.

Helmholtz equation for $\underline{\boldsymbol{A}}$ Consider a more general case with a nonzero harmonic conductive current density

$$\boldsymbol{J}(\boldsymbol{x}, t) = \mathbb{R}\mathrm{e}(\underline{\boldsymbol{J}}(\boldsymbol{x})e^{-j\omega t}),$$

and assume the vector potential $\boldsymbol{A}$ in the harmonic form

$$\boldsymbol{A}(\boldsymbol{x}, t) = \mathbb{R}\mathrm{e}(\underline{\boldsymbol{A}}(\boldsymbol{x})e^{-j\omega t}).$$

The wave equation (7.36) immediately yields

$$\Delta\underline{\boldsymbol{A}} + k^2\underline{\boldsymbol{A}} = -\mu_0\underline{\boldsymbol{J}}, \tag{7.65}$$

$k = \omega/c$.

Helmholtz equation for $\underline{\varphi}$ Last consider a nonzero electric charge density ϱ of a harmonic form

$$\varrho(\boldsymbol{x}, t) = \mathbb{R}\mathrm{e}(\underline{\varrho}(\boldsymbol{x})e^{-j\omega t}),$$

and assume a harmonic scalar potential φ in the harmonic form

$$\varphi(\boldsymbol{x}, t) = \mathbb{R}\mathrm{e}(\underline{\varphi}(\boldsymbol{x})e^{-j\omega t}).$$

The wave equation (7.37) reduces to

$$\Delta\underline{\varphi} + k^2\underline{\varphi} = -\frac{\underline{\varrho}}{\epsilon_0}, \tag{7.66}$$

where $k = \omega/c$.

7.4 TIME-HARMONIC MAXWELL'S EQUATIONS

In Sections 7.2 and 7.3 we formulated partial differential equations governing the electromagnetic field either directly or via its potentials. Due to the limited length of this text we do not address in more detail the wave and Helmholtz equations, whose weak formulation and discretization take place in the Sobolev space H^1. Instead, in the rest of this chapter we focus on the time-harmonic Maxwell's equation (7.62). This equation contains the curl–curl operator which exhibits new challenges from the points of view of both mathematical analysis and finite element discretization.

Vector operations such as the cross-product of two vectors or the curl of a vector are native in 3D. For example, the cross-product of two linearly independent vectors lying in the x_1x_2-plane is a vector normal to this plane. Therefore, the 3D setting is more natural for the mathematical analysis of the Maxwell's equations. We formulate a sufficiently general model problem in Paragraph 7.4.2, derive its variational formulation in Paragraph 7.4.3, and show the existence and uniqueness of its solution in Paragraph 7.4.4. In order to simplify the analysis, in what follows we assume piecewise-isotropic materials, so that the tensor material parameters ϵ and μ can be treated as scalars. See, e.g., [83] and the references therein for the discussion of the general tensor case.

We will return back to the 2D setting for the finite element discretization in Section 7.5. This step is justified by the fact that every 2D problem is equivalent to a 3D problem whose solution does not depend on the x_3-variable, i.e., such that the resulting field has the form

$$\boldsymbol{E} = (E_1(x_1, x_2), E_2(x_1, x_2), 0)^T.$$

7.4.1 Normalization

The time-harmonic Maxwell's equation (7.62) is normalized to a relative form that is more suitable for the numerical solution. We warn the reader that for the rest of the chapter we stop underlining phasors, and rescale $\underline{\boldsymbol{E}}$, $\underline{\boldsymbol{H}}$, and $\underline{\boldsymbol{J}}_a$ following [32] to

$$\sqrt{\epsilon_0}\underline{\boldsymbol{E}} \to \boldsymbol{E}, \quad \sqrt{\mu_0}\underline{\boldsymbol{H}} \to \boldsymbol{H}, \quad \frac{\underline{\boldsymbol{J}}_a}{\sqrt{\epsilon_0}} \to \boldsymbol{J}_a. \tag{7.67}$$

Let us define the relative permittivity ϵ_r and relative permeability μ_r by

$$\begin{aligned} \epsilon_r &= \frac{1}{\epsilon_0}\left(\epsilon + \frac{j\gamma}{\omega}\right), \\ \mu_r &= \frac{\mu}{\mu_0}. \end{aligned} \tag{7.68}$$

Notice that $\epsilon_r = \mu_r = 1$ in vacuum. Multiplying (7.62) with μ_0 and redefining $\boldsymbol{E}$ according to (7.67), we obtain

$$\nabla \times \left(\mu_r^{-1} \nabla \times \boldsymbol{E}\right) - k^2 \epsilon_r \boldsymbol{E} = \boldsymbol{\Phi}, \tag{7.69}$$

where the right-hand side $\boldsymbol{\Phi}$ has the form

$$\boldsymbol{\Phi} = jk\sqrt{\mu_0}\frac{\boldsymbol{J}_a}{\sqrt{\epsilon_0}} = jk\sqrt{\mu_0}\boldsymbol{J}_a,$$

and the wave number $k = \omega\sqrt{\epsilon_0 \mu_0} = \omega / c$ was defined before.

7.4.2 Model problem

Assume a bounded simply-connected domain $\Omega \subset \mathbb{R}^3$ with a Lipschitz-continuous boundary $\partial\Omega$ that consists of two disjoint open parts Γ_P and Γ_I.

The part Γ_P represents an interface to a perfect conductor equipped with the boundary condition (7.57),

$$\boldsymbol{E} \times \boldsymbol{\nu} = \boldsymbol{0} \qquad \text{on } \Gamma_P, \tag{7.70}$$

and Γ_I represents an impedance boundary associated with the boundary condition (7.58). For a time-harmonic field, with regard to the normalization (7.67), the impedance condition attains the form

$$\mu_r^{-1}(\nabla \times \boldsymbol{E}) \times \boldsymbol{\nu} - jk\lambda \boldsymbol{E}_T = \boldsymbol{g} \qquad \text{on } \Gamma_I. \tag{7.71}$$

Here the impedance

$$\lambda = Z\sqrt{\frac{\mu_0}{\epsilon_0}} \tag{7.72}$$

(where Z is a material parameter) is a positive function defined on Γ_I, and the symbol

$$\boldsymbol{E}_T = (\boldsymbol{\nu} \times \boldsymbol{E}) \times \boldsymbol{\nu}$$

stands for the tangential projection of the phasor $\boldsymbol{E}$ to the boundary Γ_I.

The data λ and $\boldsymbol{g}$ are zero on parts of Γ_I representing surfaces of symmetry, where the impedance condition (7.71) reduces to (7.59),

$$(\nabla \times \boldsymbol{E}) \times \boldsymbol{\nu} = \boldsymbol{0}. \tag{7.73}$$

Precise assumptions on the coefficients and data, as needed for the existence and uniqueness theorem, will be given in Paragraph 7.4.4.

7.4.3 Weak formulation

Every inner product $(\boldsymbol{a}, \boldsymbol{b})_V$ must satisfy

$$(\boldsymbol{a}, \boldsymbol{a})_V = \|\boldsymbol{a}\|_V^2$$

(see Lemma A.32), where $\|\cdot\|_V$ is the norm induced by the inner product. Since the norm is a real-valued function, the "dot product" in $\mathbb{C}^n$ requires one of the vectors to be complex-conjugate,

$$(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{a} \cdot \overline{\boldsymbol{b}} = \sum_{i=1}^{n} a_i \overline{b}_i$$

(the complex-conjugate $\overline{z}$ of a complex number $z = a + bi$ is $\overline{z} = a - bi$).

The variational identity Testing equation (7.69) by a sufficiently smooth complex vector-valued test function

$$\boldsymbol{F}(\boldsymbol{x}) = (F_1(x_1, x_2), F_2(x_1, x_2), 0)$$

and integrating over Ω, we obtain

$$\int_\Omega \left[\nabla \times (\mu_r^{-1} \nabla \times \boldsymbol{E}) \cdot \overline{\boldsymbol{F}} - k^2 \epsilon_r \boldsymbol{E} \cdot \overline{\boldsymbol{F}}\right] \mathrm{d}\boldsymbol{x} = \int_\Omega \boldsymbol{\Phi} \cdot \overline{\boldsymbol{F}} \, \mathrm{d}\boldsymbol{x}.$$

The minimum regularity of $\boldsymbol{F}$, as usual, will be determined later from the integrals in the weak formulation. Using Green's theorem together with the identities

$$\nabla \cdot (\boldsymbol{a} \times \boldsymbol{b}) = (\nabla \times \boldsymbol{a}) \cdot \boldsymbol{b} - \boldsymbol{a} \cdot (\nabla \times \boldsymbol{b})$$

and

$$\boldsymbol{a} \cdot (\boldsymbol{b} \times \boldsymbol{c}) = (\boldsymbol{a} \times \boldsymbol{b}) \cdot \boldsymbol{c}$$

(all operations being performed in 3D), we obtain

$$\int_\Omega \left[(\mu_r^{-1} \nabla \times \boldsymbol{E}) \cdot (\nabla \times \overline{\boldsymbol{F}}) - k^2 \epsilon_r \boldsymbol{E} \cdot \overline{\boldsymbol{F}}\right] \mathrm{d}\boldsymbol{x} + \int_{\partial\Omega} \boldsymbol{\nu} \times (\mu_r^{-1} \nabla \times \boldsymbol{E}) \cdot \overline{\boldsymbol{F}}_T \, \mathrm{d}S = \int_\Omega \boldsymbol{\Phi} \cdot \overline{\boldsymbol{F}} \mathrm{d}\boldsymbol{x}, \tag{7.74}$$

where

$$\boldsymbol{F}_T = (\boldsymbol{\nu} \times \boldsymbol{F}) \times \boldsymbol{\nu}$$

stands for the tangential projection of the vector $\boldsymbol{F}$ to the boundary $\partial\Omega$.

Next let us incorporate the boundary conditions (7.70) and (7.71) into (7.74). The perfect conductor boundary condition states that the field $\boldsymbol{E}$ is normal to the boundary Γ_P, and (7.70) implies that

$$\boldsymbol{F}_T = \boldsymbol{0} \quad \text{on } \Gamma_P.$$

This choice eliminates the Γ_P-portion of the surface integral in (7.74). Applying the impedance boundary condition (7.71) on Γ_I, we obtain

$$\int_\Omega \left[(\mu_r^{-1} \nabla \times \boldsymbol{E}) \cdot (\nabla \times \overline{\boldsymbol{F}}) - k^2 \epsilon_r \boldsymbol{E} \cdot \overline{\boldsymbol{F}}\right] \mathrm{d}\boldsymbol{x} - \int_{\Gamma_I} jk\lambda \boldsymbol{E}_T \cdot \overline{\boldsymbol{F}}_T \, \mathrm{d}S \tag{7.75}$$

$$= \int_\Omega \boldsymbol{\Phi} \cdot \overline{\boldsymbol{F}}\, \mathrm{d}\boldsymbol{x} + \int_{\Gamma_I} \boldsymbol{g} \cdot \overline{\boldsymbol{F}}_T \mathrm{d}S.$$

The space for $\boldsymbol{E}$ We see from (7.75) that the appropriate space for $\boldsymbol{E}$ is

$$V = \{\boldsymbol{E} \in \boldsymbol{H}(\mathrm{curl}, \Omega);\ \boldsymbol{\nu} \times \boldsymbol{E} = \boldsymbol{0} \text{ on } \Gamma_P\}, \tag{7.76}$$

where the Hilbert space $\boldsymbol{H}(\mathrm{curl}, \Omega)$ consists of vector-valued L^2-functions whose curl lies in $(L^2(\Omega))^3$,

$$\boldsymbol{H}(\mathrm{curl}, \Omega) = \{\boldsymbol{E} \in (L^2(\Omega))^3;\ \nabla \times \boldsymbol{E} \in (L^2(\Omega))^3\}.$$

The space V, when equipped with the inner product

$$(\boldsymbol{E}, \boldsymbol{F})_V = \int_\Omega \boldsymbol{E} \cdot \overline{\boldsymbol{F}}\, \mathrm{d}\boldsymbol{x} + \int_\Omega (\nabla \times \boldsymbol{E}) \cdot (\nabla \times \overline{\boldsymbol{F}})\, \mathrm{d}\boldsymbol{x} + \int_{\Gamma_I} \boldsymbol{E}_T \cdot \overline{\boldsymbol{F}}_T \,\mathrm{d}S,$$

i.e.,

$$(\boldsymbol{E}, \boldsymbol{F})_V = (\boldsymbol{E}, \boldsymbol{F})_\Omega + (\nabla \times \boldsymbol{E}, \nabla \times \boldsymbol{F})_\Omega + (\boldsymbol{E}_T, \boldsymbol{F}_T)_{\Gamma_I},$$

is a Hilbert space. Indeed this inner product induces a norm $\|\boldsymbol{E}\|_V^2 = (\boldsymbol{E}, \boldsymbol{E})_V$. Before writing down the weak formulation of problem (7.69), (7.70), (7.71), let us list appropriate assumptions on the domain, coefficients and data.

Assumptions on the domain, coefficients and data Recall from Paragraph 7.4.2 that the domain $\Omega \subset \mathbb{R}^3$ is assumed to be bounded and simply-connected, with a Lipschitz-continuous boundary $\partial\Omega$ consisting of two disjoint relatively open parts Γ_P and Γ_I, $\partial\Omega = \overline{\Gamma}_P \cup \overline{\Gamma}_I$. In order to incorporate various materials, the domain Ω is allowed to be split into several disjoint open simply-connected subdomains $\Omega_1, \Omega_2, \ldots, \Omega_n$ with a Lipschitz-continuous boundary, such that $\overline{\Omega} = \bigcup_{i=1}^n \overline{\Omega}_i$. The parameters ϵ_r and μ_r are allowed to be generally discontinuous, but smooth in each subdomain Ω_i. For reasons that will become clear later, the parameter ϵ_r requires two more conditions to hold:

1. The restriction of ϵ_r to each subdomain Ω_i is a H^3-function (then $\epsilon_r \in C^1(\overline{\Omega}_i)$ and it is possible to extend it smoothly to the whole Ω).
2. There exists a positive constant $C_\epsilon > 0$ such that for each subdomain Ω_i either $\mathrm{Im}(\epsilon_r) \geq C_\epsilon$ or $\mathrm{Im}(\epsilon_r) = 0$, $i = 1, 2, \ldots, n$.

The positive impedance function λ is assumed to lie in $L^\infty(\Gamma_I)$. The right-hand sides $\boldsymbol{\Phi}$ and $\boldsymbol{g}$ are required to lie in $(L^2(\Omega))^3$ and $(L^2(\Gamma_I))^3$, respectively.

Weak formulation Under the above assumptions on the coefficients and data, the weak formulation of the model problem (7.69), (7.70), (7.71) reads:

Find the electric field phasor $\boldsymbol{E} \in V$ satisfying

$$a(\boldsymbol{E}, \boldsymbol{F}) = l(\boldsymbol{F}) \quad \text{for all } \boldsymbol{F} \in V. \tag{7.77}$$

where the sesquilinear form $a(\cdot, \cdot)$ is defined on $V \times V$ by

$$a(\boldsymbol{e}, \boldsymbol{f}) = (\mu_r^{-1}\nabla \times \boldsymbol{e}, \nabla \times \boldsymbol{f})_\Omega - k^2(\epsilon_r \boldsymbol{e}, \boldsymbol{f})_\Omega - jk(\lambda \boldsymbol{e}_T, \boldsymbol{f}_T)_{\Gamma_I}$$

and the linear form $l(\cdot)$ is defined on V as

$$l(\boldsymbol{f}) = (\boldsymbol{\Phi}, \boldsymbol{f})_\Omega + (\boldsymbol{g}, \boldsymbol{f}_T)_{\Gamma_I}.$$

7.4.4 Existence and uniqueness of solution

Under the above assumptions, there exists a unique solution to problem (7.77).

Theorem 7.1 (Existence and uniqueness of $\boldsymbol{E}$) *Consider the assumptions on the domain Ω, boundary parts Γ_P and Γ_I, and coefficients and data ϵ_r, μ_r, λ, $\boldsymbol{\Phi}$ and $\boldsymbol{g}$, listed in Paragraph 7.4.3. Moreover, assume that at least one of the following conditions holds:*

1. *The impedance boundary Γ_I is not empty.*
2. *The imaginary part $Im(\epsilon_r) > 0$ in some open subdomain $\Omega_+ \subset \Omega$.*

Then for any wave number $k > 0$, problem (7.77) has a unique solution $\boldsymbol{E} \in V$. In addition, there exists a constant C_k independent of $\boldsymbol{E}$, $\boldsymbol{\Phi}$ and $\boldsymbol{g}$ (but depending on k) such that

$$||\boldsymbol{E}||_V \le C_k(||\boldsymbol{F}||_{(L^2(\Omega))^3} + ||\boldsymbol{g}||_{(L^2(\Gamma_I))^3}).$$

Outline of proof In order to prove the existence and uniqueness of solution, one has to overcome the following basis difficulties:

1. The curl operator contains a large null space (all functions $\boldsymbol{e} \in \boldsymbol{H}(\text{curl})$ such that $\boldsymbol{e} = \nabla\varphi$, $\varphi \in H^1(\Omega)$). This null space has to be removed using the Helmholtz decomposition.

2. Because of the term $-k^2(\epsilon_r \boldsymbol{E}, \boldsymbol{F})_\Omega$ the sesquilinear form $a(\cdot, \cdot)$ is not V-elliptic, which excludes an application of the Lax–Milgram lemma. The Fredholm alternative (Theorem A.17), which is used instead, requires an operator reformulation of the problem into the form

$$(I + K)\boldsymbol{e} = \boldsymbol{f}, \tag{7.78}$$

where I is the identity operator and K a compact operator.

3. The Fredholm alternative requires a separate proof of the uniqueness of solution to (7.78) (which is equivalent to proving that the homogeneous equation $(I + K)\boldsymbol{e} = \boldsymbol{0}$ only has a trivial solution). Then it implies the existence of solution to (7.78) for every right-hand side $\boldsymbol{f}$ from the underlying Hilbert space.

Let us discuss all these steps in more detail, following [83]. For simplicity we consider the case of μ_r piecewise constant in the subdomains $\Omega_1, \Omega_2, \ldots, \Omega_n$, and assume Γ_P and Γ_I to be connected.

Helmholtz decomposition

Lemma 7.1 *Under the assumptions from Paragraph 7.4.3 let $\boldsymbol{e} \in V$ such that $\boldsymbol{e}_T = \boldsymbol{0}$ on the impedance boundary Γ_I and $\nabla \times \boldsymbol{e} = \boldsymbol{0}$ in Ω. Then the scalar potential φ such that $\boldsymbol{e} = \nabla\varphi$ lies in the space*

$$S = \{\varphi \in H^1(\Omega);\ \varphi = 0 \text{ on } \Gamma_I,\ \varphi = const. \text{ on } \Gamma_P\}.$$

Proof: The proof follows easily from the fact that the tangential component

$$\boldsymbol{e}_T = (\boldsymbol{\nu} \times \boldsymbol{e}) \times \boldsymbol{\nu} = (\nabla\varphi)_T.$$

Thus $\boldsymbol{e}_T$ is constant on each component Γ_I and Γ_P. Without loss of generality, the constant on one component can be chosen to be zero. ■

Theorem 7.2 (Helmholtz decomposition) *The space*

$$\nabla S = \{\nabla\varphi;\ \varphi \in S\} \subset V,$$

is a closed subspace of V. Define

$$V_0 = \{\boldsymbol{e} \in V;\ (\epsilon_r \boldsymbol{e}, \nabla\varphi)_\Omega = 0 \text{ for all } \varphi \in S\}.$$

Then V is the direct sum of the subspaces V_0 and ∇S,

$$V = V_0 \oplus \nabla S. \tag{7.79}$$

Proof: This lemma was proved in [72]. The situation is simple when ϵ_r is real, since the bilinear form $(\epsilon_r \boldsymbol{e}, \boldsymbol{f})_\Omega$ is an inner product in $(L^2(\Omega))^3$, and the result follows immediately from the basic projection theorem for Hilbert spaces (Theorem A.14 in Paragraph A.3.5).

The complex case is not difficult either. Since S is closed in $H^1(\Omega)$, also ∇S is closed in V. Define a sesquilinear form

$$\tilde{a}(\boldsymbol{e}, \boldsymbol{f}) = (\nabla \times \boldsymbol{e}, \nabla \times \boldsymbol{f})_\Omega + (\epsilon_r \boldsymbol{e}, \boldsymbol{f})_\Omega + (\boldsymbol{e}_T, \boldsymbol{f}_T)_{\Gamma_I}, \quad \boldsymbol{e}, \boldsymbol{f} \in V.$$

Since there exist positive constants C_1 and C_2 such that $C_1 \leq \mathrm{Re}(\epsilon_r) \leq C_2$ in Ω, the following holds:

1. There exists a constant C independent of $\boldsymbol{e}$ such that

$$|\tilde{a}(\boldsymbol{e}, \boldsymbol{e})| \geq C\|\boldsymbol{e}\|_V^2 \tag{7.80}$$

for all $\boldsymbol{e} \in V$. (This is verified easily when taking the real part of ϵ_r).

2. There exists a constant C independent of $\boldsymbol{e}$ and $\boldsymbol{f}$ such that

$$|\tilde{a}(\boldsymbol{e}, \boldsymbol{e})| \leq C\|\boldsymbol{e}\|_V \|\boldsymbol{f}\|_V \tag{7.81}$$

for all $\boldsymbol{e}, \boldsymbol{f} \in V$.

According to the Lax–Milgram lemma, for every $\boldsymbol{e} \in V$ there exists a unique function $P\boldsymbol{e} \in \nabla S$ such that

$$\tilde{a}(P\boldsymbol{e}, \boldsymbol{f}) = (\epsilon_r \boldsymbol{e}, \boldsymbol{f}) \quad \text{for all } \boldsymbol{f} \in \nabla S.$$

The operator $P : V \to \nabla S$ is linear, bounded and indeed $P\boldsymbol{e} = \boldsymbol{e}$ if $\boldsymbol{e} \in \nabla S$. Thus P is a projection and any function $\boldsymbol{e} \in V$ can be written uniquely as

$$\boldsymbol{e} = P\boldsymbol{e} + (I - P)\boldsymbol{e}.$$

The proof is accomplished by realizing that $(I - P)\boldsymbol{e} \in V_0$ since

$$(\epsilon_r(I - P)\boldsymbol{e}, \nabla\varphi)_\Omega = \tilde{a}((I - P)\boldsymbol{e}, \nabla\varphi) = 0$$

for all $\varphi \in S$. ■

Fredholm operator equation Using Theorem 7.2, every solution $\boldsymbol{E} \in V$ can be decomposed uniquely into

$$\boldsymbol{E} = \boldsymbol{E}_0 + \nabla\varphi, \tag{7.82}$$

where $\boldsymbol{E}_0 \in V_0$ and $\varphi \in S$. Substituting (7.82) into (7.77) and using the facts that $\nabla \times \nabla\varphi = \boldsymbol{0}$ and $(\nabla\varphi) \times \boldsymbol{\nu} = \boldsymbol{0}$ on $\partial\Omega$, we obtain

$$\begin{aligned}(\mu_r^{-1}\nabla \times \boldsymbol{E}_0, \nabla \times \boldsymbol{F})_\Omega - k^2(\epsilon_r(\boldsymbol{E}_0 + \nabla\varphi), \boldsymbol{F})_\Omega - jk(\lambda \boldsymbol{E}_{0,T}, \boldsymbol{F}_T)_{\Gamma_I} \\ = (\boldsymbol{\Phi}, \boldsymbol{F})_\Omega + (\boldsymbol{g}, \boldsymbol{F}_T)_{\Gamma_I}\end{aligned} \tag{7.83}$$

for all $\boldsymbol{F} \in V$. Choosing now $\boldsymbol{F} = \nabla\psi$ for some $\psi \in S$, (7.83) simplifies to

$$-k^2(\epsilon_r(\boldsymbol{E}_0 + \nabla\varphi), \nabla\psi))_\Omega = (\boldsymbol{\Phi}, \nabla\psi)_\Omega.$$

Now, since $\boldsymbol{E}_0 \in V_0$, the potential φ satisfies

$$-k^2(\epsilon_r\nabla\varphi, \nabla\psi))_\Omega = (\boldsymbol{\Phi}, \nabla\psi)_\Omega \quad \text{for all } \psi \in S. \tag{7.84}$$

Using the assumptions for ϵ_r, it is not difficult to show that the variational problem (7.84) has a unique solution that moreover satisfies the estimate

$$\|\nabla\varphi\|_{(L^2(\Omega))^3} \leq C\|\boldsymbol{\Phi}\|_{(L^2(\Omega))^3},$$

where C is some positive constant independent of $\boldsymbol{\Phi}$. From here it is clear that determining $\boldsymbol{E}$ is equivalent to determining $\boldsymbol{E}_0$.

Therefore in the following let us look for $\boldsymbol{E}_0 \in V_0$ such that

$$\begin{aligned}(\mu_r^{-1}\nabla \times \boldsymbol{E}_0, \nabla \times \boldsymbol{F})_\Omega - k^2(\epsilon_r\boldsymbol{E}_0, \boldsymbol{F})_\Omega - jk(\lambda \boldsymbol{E}_{0,T}, \boldsymbol{F}_T)_{\Gamma_I} \\ = (\boldsymbol{\Phi}, \boldsymbol{F})_\Omega + (\boldsymbol{g}, \boldsymbol{F}_T)_{\Gamma_I} + k^2(\epsilon_r\nabla\varphi, \boldsymbol{F})_\Omega\end{aligned} \tag{7.85}$$

for all $\boldsymbol{F} \in V_0$. (We can restrict ourselves to the test functions from V_0 since $V_0 \subset V$.)

The analysis of (7.85) is more demanding than the analysis of (7.84) was, and this is where the Fredholm alternative comes into play. The idea of transformation of equation (7.85) into an operator equation is as follows:

$$\begin{aligned}&\underbrace{(\mu_r^{-1}\nabla \times \boldsymbol{E}_0, \nabla \times \boldsymbol{F})_\Omega + k^2(\epsilon_r \boldsymbol{E}_0, \boldsymbol{F})_\Omega - jk(\lambda \boldsymbol{E}_{0,T}, \boldsymbol{F}_T)_{\Gamma_I}}_{s(\boldsymbol{E}_0, \boldsymbol{F})}\\ &\underbrace{-2k^2(\epsilon_r \boldsymbol{E}_0, \boldsymbol{F})_\Omega}_{s(K\boldsymbol{E}_0, \boldsymbol{F})}\\ &= \underbrace{(\boldsymbol{\Phi}, \boldsymbol{F})_\Omega + (\boldsymbol{g}, \boldsymbol{F}_T)_{\Gamma_I} + k^2(\epsilon_r \nabla\varphi, \boldsymbol{F})_\Omega}_{s(\mathcal{F}, \boldsymbol{F})}\end{aligned} \tag{7.86}$$

for all $\boldsymbol{F} \in V_0$. Hence, define a sesquilinear form

$$s(\boldsymbol{e}, \boldsymbol{f}) = (\mu_r^{-1}\nabla \times \boldsymbol{e}, \nabla \times \boldsymbol{f})_\Omega + k^2(\epsilon_r \boldsymbol{e}, \boldsymbol{f})_\Omega - jk(\lambda \boldsymbol{e}_T, \boldsymbol{f}_T)_{\Gamma_I}$$

for all $\boldsymbol{e}, \boldsymbol{f} \in V$. Postponing the analysis of the form $s(\cdot,\cdot)$ to Lemmas 7.2 and 7.3, let us define an operator $K : (L^2(\Omega))^3 \to V_0 \subset (L^2(\Omega))^3$ by

$$s(K\boldsymbol{f}, \boldsymbol{F}) = -2k^2(\epsilon_r \boldsymbol{f}, \boldsymbol{F}) \quad \text{for all } \boldsymbol{F} \in V_0.$$

and a right-hand side $\mathcal{F} \in V_0$ by

$$s(\mathcal{F}, \boldsymbol{F}) = (\boldsymbol{\Phi}, \boldsymbol{F})_\Omega + (\boldsymbol{g}, \boldsymbol{F}_T)_{\Gamma_I} + k^2(\epsilon_r \nabla\varphi, \boldsymbol{F}) \quad \text{for all } \boldsymbol{F} \in V_0.$$

Using the operator K and the right-hand side $\mathcal{F}$, problem (7.85) can be written in the form of a Fredholm operator equation,

$$(I + K)\boldsymbol{E}_0 = \mathcal{F}. \tag{7.87}$$

The next step consists in showing that both K and $\mathcal{F}$ are well-defined and that K is a compact operator.

Verification of Fredholm assumptions Let us verify that indeed equation (7.87) is well-defined and that the operator K satisfies the assumptions of the Fredholm alternative (Theorem A.17). We begin by showing that $s(\cdot,\cdot)$ is V-elliptic.

Lemma 7.2 *There exists a constant C_s independent of $\boldsymbol{e}$ (but depending on $\epsilon_r, \mu_r, \lambda$ and k) such that*

$$|s(\boldsymbol{e}, \boldsymbol{e})| \geq C_s \|\boldsymbol{e}\|_V^2 \quad \textit{for all } \boldsymbol{e} \in V. \tag{7.88}$$

Proof: Again the situation is simple when ϵ_r is real-valued. The definition of $s(\cdot,\cdot)$ yields

$$\begin{aligned}|s(\boldsymbol{e}, \boldsymbol{e})|^2 = &\left(\|\mu_r^{-1/2}\nabla \times \boldsymbol{e}\|_{(L^2(\Omega))^3}^2 + k^2\|\mathbb{R}\mathrm{e}(\epsilon_r)^{1/2}\boldsymbol{e}\|_{(L^2(\Omega))^3}^2\right)^2\\ &+ \left(k^2\|\mathrm{Im}(\epsilon_r)^{1/2}\boldsymbol{e}\|_{(L^2(\Omega))^3}^2 - k\|\lambda^{1/2}\boldsymbol{e}_T\|_{(L^2(\Gamma_I))^3}^2\right)^2.\end{aligned}$$

Expanding the expression on the right-hand side and using the modified Young inequality (A.49) with $p = q = 2$ we find that for any $\delta > 0$ it is

$$|s(e, e)|^2 \geq \|\mu_r^{-1/2}\nabla \times e\|^4_{(L^2(\Omega))^3} + k^4\|\mathrm{Re}(\epsilon_r)^{1/2}e\|^4_{(L^2(\Omega))^3}$$
$$+k^4\frac{\delta - 1}{\delta}\|\mathrm{Im}(\epsilon_r)^{1/2}e\|^4_{(L^2(\Omega))^3} + k^2(1-\delta)\|\lambda^{1/2}e_T\|_{(L^2(\Gamma_I))^3}.$$

From the assumptions on the coefficients there exist constants $c_{Re} > 0$ and $c_{Im} \geq 0$ such that $\mathrm{Re}(\epsilon_r) \geq c_{Re}$ and $\mathrm{Im}(\epsilon_r) \leq c_{Im}$ in Ω. Choosing $\delta < 1$, we can estimate

$$\|\mathrm{Re}(\epsilon_r)^{1/2}e\|^4_{(L^2(\Omega))^3} + \frac{\delta-1}{\delta}\|\mathrm{Im}(\epsilon_r)^{1/2}e\|^4_{(L^2(\Omega))^3} \geq \left(c^2_{Re} + \frac{\delta-1}{\delta}c^2_{Im}\right)\|e\|^4_{(L^2(\Omega))^3}.$$

Thus is we choose δ such that

$$\frac{c^2_{Im}}{c^2_{Re} + c^2_{Im}} < \delta < 1,$$

inequality (7.88) follows. ■

The operator K and the right-hand side $\mathcal{F}$ have the following properties:

Lemma 7.3 *The operator $K : (L^2(\Omega))^3 \to V_0$ is bounded and compact, and there exists a constant $C > 0$ such that*

$$\|Kf\|_V \leq C\|f\|_{(L^2(\Omega))^3}. \tag{7.89}$$

The right-hand side $\mathcal{F}$ is well-defined, and

$$\|\mathcal{F}\|_V \leq C(\|F\|_{(L^2(\Omega))^3} + \|g\|_{(L^2(\Gamma_I))^3} + \|\nabla\varphi\|_{(L^2(\Omega))^3}).$$

Proof: The V-ellipticity of the form $s(\cdot,\cdot)$, required by the Lax–Milgram lemma, was shown in Lemma 7.2. To verify boundedness, use the Cauchy–Schwarz inequality,

$$|s(e, f)| \leq C(\|\nabla \times e\|_{(L^2(\Omega))^3}\|\nabla \times f\|_{(L^2(\Omega))^3} + \|e\|_{(L^2(\Omega))^3}\|f\|_{(L^2(\Omega))^3}$$
$$+\|e_T\|_{(L^2(\Gamma_I))^3}\|f_T\|_{(L^2(\Gamma_I))^3}).$$

Here the constant C depends on the lower and upper bounds for ϵ_r, μ_r and λ. Thus by the Lax–Milgram lemma Kf is well-defined and inequality (7.89) holds.

It remains to be shown that K is compact. Consider a bounded sequence $\{u_n\}_{n=1}^{\infty} \subset (L^2(\Omega))^3$. By (7.89) the sequence $\{Ku_n\}_{n=1}^{\infty} \subset (L^2(\Omega))^3$ is bounded in $V_0 \subset (L^2(\Omega))^3$. It follows from the compact embedding of V_0 in $(L^2(\Omega))^3$ (see, e.g., [83], Theorem 4.7 for details) that there exists a subsequence that converges strongly in $(L^2(\Omega))^3$. Therefore the operator K is compact. The rest of the proof for $\mathcal{F}$ is analogous. ■

Since the operator K and the right-hand side $\mathcal{F}$ satisfy the assumptions of the Fredholm alternative (Theorem A.17), we obtain the existence of a unique solution to (7.87) if we can prove that the homogeneous equation

$$(I + K)\boldsymbol{E}_0 = \boldsymbol{0} \tag{7.90}$$

only has a trivial solution.

Uniqueness of the trivial solution to (7.90) This most difficult part of the proof is decomposed into two steps. First, using the assumptions on the coefficient ϵ_r and boundary Γ_I we show that the solution is unique either in the region where $\mathrm{Im}(\epsilon_r) > 0$ or on Γ_I. Next a unique continuation result is applied to show that the solution is unique everywhere.

Let us begin with introducing a basic continuation result for real-valued functions.

Lemma 7.4 *Let Ω be a connected domain in $\mathbb{R}^3$ and suppose $\boldsymbol{f} \in (H^2(\Omega))^3$, where $\boldsymbol{f} = (f_1, f_2, f_3)^T$ is a real-valued function that satisfies*

$$|\Delta \boldsymbol{f}| \le C \sum_{r=1}^{3} (|f_r| + |\nabla f_r|)$$

almost everywhere in Ω, where C is a positive constant. Let $\boldsymbol{x}_0 \in \Omega$ be such that $\boldsymbol{f} \equiv \boldsymbol{0}$ in some open neighborhood $B(\boldsymbol{x}_0) \subset \Omega$. Then $\boldsymbol{f} \equiv \boldsymbol{0}$ in Ω.

Proof: This result was proved in [33] for $\boldsymbol{f} \in (C^2(\Omega))^3$. Since it only relies on the fact that $\Delta \boldsymbol{f}$ is well-defined in $L^2(\Omega)$, it can be extended to $(H^2(\Omega))^3$. ■

In the next step the result of Lemma 7.4 is extended to the vector-valued complex case.

Lemma 7.5 *Suppose that $\Omega \subset \mathbb{R}^3$ is an open connected domain. Let ϵ_r be real-valued and smooth in $\overline{\Omega}$ and μ_r be real-valued and constant in Ω. Let $\boldsymbol{e}, \boldsymbol{f} \in \boldsymbol{H}(\mathrm{curl}, \Omega)$ satisfy*

$$\begin{aligned} jk\epsilon_r \boldsymbol{e} + \nabla \times \boldsymbol{f} &= 0, \\ jk\mu_r \boldsymbol{f} - \nabla \times \boldsymbol{e} &= \boldsymbol{0} \end{aligned} \tag{7.91}$$

in Ω and that $\boldsymbol{e}$ vanishes in an open subdomain of Ω. Then $\boldsymbol{e} \equiv \boldsymbol{0}$ and $\boldsymbol{f} \equiv \boldsymbol{0}$ in Ω.

Remark 7.1 *The result of Lemma 7.5 was proved more generally in [121], under the assumptions that ϵ_r and μ_r are symmetric, real-valued, uniformly positive definite, and bounded matrix functions of the spatial variable in $(L^\infty(\Omega))^{3\times 3}$.*

Proof: By (7.91) we have

$$\nabla \times \epsilon_r^{-1} \nabla \times \boldsymbol{f} - k^2 \mu_r \boldsymbol{f} = \boldsymbol{0}$$

in Ω. Taking the divergence of this equation and using the fact that μ_r is constant, we have

$$\nabla \cdot \boldsymbol{f} = 0$$

in Ω. Rewriting

$$\nabla \times \epsilon_r^{-1} \nabla \times \boldsymbol{f} = \epsilon_r^{-1} \nabla \times (\nabla \times \boldsymbol{f}) + (\nabla \epsilon_r^{-1}) \times \nabla \times \boldsymbol{f}$$

and using the identity

$$\nabla \times \nabla \times \boldsymbol{f} = \nabla(\nabla \cdot \boldsymbol{f}) - \Delta \boldsymbol{f},$$

we obtain

$$\Delta \boldsymbol{f} = \epsilon_r(\nabla \epsilon_r^{-1}) \times \nabla \times \boldsymbol{f} - k^2 \mu_r \boldsymbol{f}.$$

Thus $\Delta \boldsymbol{f} \in L^2(\Omega)$, and by standard interior elliptic regularity results (see, e.g., [76]) it holds that for any compact $\Omega_0 \subset \Omega$, $\boldsymbol{f} \in (H^2(\Omega_0))^3$. Applying Lemma 7.4 to the real and imaginary part of $\boldsymbol{f}$ separately, we conclude that $\boldsymbol{f} \equiv \mathbf{0}$ in Ω_0. Since the domain Ω_0 was arbitrary, $\boldsymbol{f} \equiv \mathbf{0}$ in Ω. ■

The one-before-last step required to finish the proof consists in introducing the Calderon extension theorem.

Theorem 7.3 (Calderon extension theorem) *Let Ω be a bounded domain in $\mathbb{R}^N$ with Lipschitz-continuous boundary. Let $s \geq 1$ be an integer number and $1 < p < \infty$. Then there exists a continuous linear extension operator*

$$\Pi : W^{s,p}(\Omega) \to W^{s,p}(\mathbb{R}^N)$$

such that

$$(\Pi u)(\boldsymbol{x}) = u(\boldsymbol{x}) \quad \textit{for all } \boldsymbol{x} \in \Omega \ \textit{ and } \ u \in W^{s,p}(\Omega).$$

In the special case of $p = 2$ the operator Π exists for all $s \geq 0$.

Proof: See, e.g., [1]. ■

Finally we can formulate and prove the desired input for the Fredholm alternative, i.e., that the homogeneous equation $(I + K)\boldsymbol{E}_0 = \mathbf{0}$ only has a trivial solution $\boldsymbol{E}_0 = \mathbf{0}$.

Theorem 7.4 *Recall the assumptions on the coefficients and data listed earlier in Paragraph 7.4.3. Further, suppose that Im$(\epsilon_r) \geq C_\epsilon > 0$ in some open subdomain of Ω or Γ_I is not empty. Then the homogeneous equation*

$$a(\boldsymbol{E}, \boldsymbol{F}) = (\mu_r^{-1} \nabla \times \boldsymbol{E}, \nabla \times \boldsymbol{F})_\Omega - k^2(\epsilon_r \boldsymbol{E}, \boldsymbol{F})_\Omega - jk(\lambda \boldsymbol{E}_T, \boldsymbol{F}_T)_{\Gamma_I} = 0 \ \textit{ for all } \boldsymbol{F} \in V \tag{7.92}$$

only has a trivial solution $\boldsymbol{E} = \mathbf{0}$.

Proof: Evidently $\boldsymbol{e}_0 \equiv \mathbf{0}$ is a solution to (7.92), but it is not quite clear whether it is unique. Therefore consider any function $\boldsymbol{e} \in V$ that satisfies (7.92). Using $\boldsymbol{F} = \boldsymbol{e}$ and taking the imaginary part of the resulting equation, we have

$$k^2(\mathrm{Im}(\epsilon_r)\boldsymbol{e}, \boldsymbol{e})_\Omega + k(\lambda \boldsymbol{e}_T, \boldsymbol{e}_T)_{\Gamma_I} = 0. \tag{7.93}$$

Assuming that λ is real and positive, this yields $\boldsymbol{e}_T = \mathbf{0}$ on Γ_I and $\boldsymbol{e} = \mathbf{0}$ in any subdomain of Ω on which Im(ϵ_r) is positive. If this happens to be true in the whole Ω, then the proof is finished. Otherwise consider some subdomain $\Omega_p \subset \Omega$ where Im(ϵ_r) is positive. We see that $\boldsymbol{e} = \mathbf{0}$ on all subdomains where Im$(\epsilon_r) \neq 0$. Let Ω_q be subdomain of Ω on which Im$(\epsilon_r) = 0$ and

1. $\overline{\Omega}_p \cap \overline{\Omega}_q$ is a Lipschitz surface with nonempty interior,

2. ϵ_r is real and smooth on Ω_q.

The Calderon extension Theorem 7.3 (using the assumption $\epsilon_r \in H^3(\Omega_q)$) allows us to extend ϵ_r smoothly from Ω_p to $\Omega_q \cup \Omega_q$. Also μ_r is (constantly) extended to $\Omega_q \cup \Omega_q$.

Let $B_r(\boldsymbol{x}_0)$ be an open ball of a sufficiently small radius r centered at a point $\boldsymbol{x}_0 \in \overline{\Omega}_p \cap \overline{\Omega}_q$ such that $B_r(\boldsymbol{x}_0) \subset (\overline{\Omega}_p \cup \overline{\Omega}_q)$ and ϵ_r is positive in $\Omega_q \cup B_r(\boldsymbol{x}_0)$. Since $\boldsymbol{e} = \boldsymbol{0}$ on Ω_p, it is

$$\nabla \times \mu_r^{-1} \nabla \times \boldsymbol{e} - k^2 \epsilon_r \boldsymbol{e} = \boldsymbol{0}$$

in $\Omega_q \cup B_r(\boldsymbol{x}_0)$ and $\boldsymbol{e}$ vanishes in $\Omega_q \cup B_r(\boldsymbol{x}_0)$. Now, since both ϵ_r and μ_r are real-valued, we use Lemma 7.4 to conclude that $\boldsymbol{e} = \boldsymbol{0}$ in $\Omega_q \cup B_r(\boldsymbol{x}_0)$ and therefore also in $\Omega_p \cup \Omega_q$. In this way we may continue until all subdomains where ϵ_r is real are reached, and we conclude that $\boldsymbol{e} = \boldsymbol{0}$ in Ω.

If ϵ_r is real in the whole domain Ω, then we need to use the assumption that Γ_I is not empty. By (7.93) we know that $\boldsymbol{e}_T = \boldsymbol{0}$ on Γ_I. We proceed analogously to the previous case. Let Ω_q be a subdomain of Ω such that $\overline{\Omega}_q \cap \Gamma_I$ contains an open subdomain of Γ_I, and such that ϵ_r is smooth in Ω_q. We can extend ϵ_r smoothly to $\mathbb{R}^3$. Since this function is positive on Ω_q, there exists an open ball $B_r(\boldsymbol{x}_0)$ centered at a point $\boldsymbol{x}_0 \in \overline{\Omega}_q \cap \Gamma_I$ such that ϵ_r is positive on $\Omega_q \cup B_r(\boldsymbol{x}_0)$ and $(B_r(\boldsymbol{x}_0) \cap \Omega) \subset \Omega_q$. When extending $\boldsymbol{e}$ by zero to $B_r(\boldsymbol{x}_0) \setminus \overline{\Omega}_q$, we have that

$$\int_{\Omega_q \cup B_r(\boldsymbol{x}_0)} \mu_r^{-1} \nabla \times \boldsymbol{e} \cdot \nabla \times \overline{\boldsymbol{F}} - k^2 \epsilon_r \boldsymbol{e} \cdot \overline{\boldsymbol{F}} \, \mathrm{d}\boldsymbol{x} = 0$$

for all $\boldsymbol{F} \in \boldsymbol{H}_0(\mathrm{curl}, \Omega_q \cup B_r(\boldsymbol{x}_0))$. Thus $\boldsymbol{e}$ is a weak solution of the Maxwell's equations there and $\boldsymbol{e}$ vanishes in $B_r(\boldsymbol{x}_0) \setminus \overline{\Omega}_q$. Hence by Lemma 7.4, $\boldsymbol{e}$ vanishes in $\Omega_q \cup B_r(\boldsymbol{x}_0)$ and thus also in Ω_q. We conclude that $\boldsymbol{e} = \boldsymbol{0}$ in Ω by crossing boundaries of subdomains Ω_p on which ϵ_r is differentiable. ■

7.5 EDGE ELEMENTS

In the early era of finite element methods for the Maxwell's equations it was generally assumed that $[H^1(\Omega_h)]^d$ was the correct space for the discretization of the electric field $\boldsymbol{E}$. However, the globally continuous discretizations exhibited spurious waves and other unwanted phenomena, the origin of which was not known (see, e.g., [61, 85] and [115]). Later it was realized that the space $\boldsymbol{H}(\mathrm{curl}, \Omega_h)$ was larger than $[H^1(\Omega_h)]^d$: The space $\boldsymbol{H}(\mathrm{curl}, \Omega_h)$ admits discontinuous functions and functions with stronger singularities than $[H^1(\Omega_h)]^d$. Solutions lying in $\boldsymbol{H}(\mathrm{curl}, \Omega_h) \setminus [H^1(\Omega_h)]^d$ thus cannot be approximated in finite element subspaces of $[H^1(\Omega_h)]^d$. One such example is presented in Paragraph B.2.8.

This discovery initiated the development of discontinuous vector-valued elements conforming to the space $\boldsymbol{H}(\mathrm{curl}, \Omega_h)$. Since both in 2D and 3D the degrees of freedom on the lowest-order $\boldsymbol{H}(\mathrm{curl}, \Omega_h)$-conforming elements were associated with the element edges, these elements were called edge elements.

The lowest-order edge elements were first introduced by Whitney [123] in a different context of geometrical integration theory. Later the lowest-order edge elements were independently rediscovered and applied to the Maxwell's equations by several authors (see, e.g., [2, 10] and [12]). In this section we introduce the reader to the concept of nodal edge elements, more precisely to the first family of Nédélec elements [87].

We begin with formulating the conformity requirements of the space $\boldsymbol{H}(\mathrm{curl})$ in Paragraph 7.5.1. Lowest-order Whitney elements and suitable reference maps are introduced in Paragraph 7.5.2. Higher-order Nédélec edge elements are discussed in Paragraph 7.5.3. The transformation of the Maxwell's equations from a general triangular element to the reference domain is described in Paragraph 7.5.4. Interpolation on edge elements is discussed in Paragraph 7.5.5. The finite element discretization is presented in two spatial dimensions.

7.5.1 Conformity requirements of the space $\boldsymbol{H}(\mathrm{curl})$

In this paragraph we formulate the conformity requirements of the space $\boldsymbol{H}(\mathrm{curl}, \Omega_h)$ that dictate the structure of the edge elements. Because of the vector operations used, again it is natural to begin in three spatial dimensions. The two-dimensional case is addressed in Remark 7.2 following Lemma 7.6.

Lemma 7.6 *Consider a polygonal domain $\Omega_h \subset \mathbb{R}^d$ covered with a finite element mesh $\mathcal{T}_{h,p}$, and a function $\boldsymbol{E} : \Omega_h \to \mathbb{R}^d$, $d = 3$, such that*

1. *$\boldsymbol{E}|_K \in [H^1(K)]^d$ for each element $K \in \mathcal{T}_{h,p}$,*
2. *for each element interface $f = \overline{K}_1 \cap \overline{K}_2$, $K_1, K_2 \in \mathcal{T}_{h,p}$ the traces of the tangential components $\boldsymbol{\nu}_f \times \boldsymbol{E}|_{K_1}$ and $\boldsymbol{\nu}_f \times \boldsymbol{E}|_{K_2}$ on f are the same, where $\boldsymbol{\nu}_f$ is a unit normal vector to f.*

Then $\boldsymbol{E} \in \boldsymbol{H}(\mathrm{curl}, \Omega_h)$. On the other hand, if $\boldsymbol{E} \in \boldsymbol{H}(\mathrm{curl}, \Omega_h)$ and condition 1. holds, then condition 2. is satisfied.

Proof: Let $\boldsymbol{E}|_K \in [H^1(K)]^d$ for each element $K \in \mathcal{T}_{h,p}$. For every $K \in \mathcal{T}_{h,p}$ define

$$\boldsymbol{w}_K = \nabla \times (\boldsymbol{E}|_K).$$

Clearly the function

$$\boldsymbol{w} = \sum_{K \in \mathcal{T}_{h,p}} \boldsymbol{w}_K \chi_K,$$

where χ_K is the characteristic function of K ($\chi_K = 1$ in K and it vanishes outside of K), is defined almost everywhere in Ω_h and lies in the space $[L^2(\Omega_h)]^d$. Further, consider an arbitrary function

$$\boldsymbol{\varphi} \in [C_0^\infty(\Omega_h)]^d = \mathcal{D},$$

and use Green's theorem to calculate

$$\begin{aligned}
(\nabla \times \boldsymbol{E}, \boldsymbol{\varphi}) &= -\int_{\Omega_h} \boldsymbol{E} \cdot \nabla \times \boldsymbol{\varphi} \,\mathrm{d}\boldsymbol{x} = -\sum_{K \in \mathcal{T}_{h,p}} \int_K (\boldsymbol{E}|_K) \cdot \nabla \times \boldsymbol{\varphi} \,\mathrm{d}\boldsymbol{x} \\
&= \sum_{K \in \mathcal{T}_{h,p}} \int_K \nabla \times (\boldsymbol{E}|_K) \cdot \boldsymbol{\varphi} \,\mathrm{d}\boldsymbol{x} \\
&\quad - \sum_{f, f = \overline{K}_1 \cap \overline{K}_2, K_1, K_2 \in \mathcal{T}_{h,p}} \int_f (\boldsymbol{E}|_{K_1} \times \boldsymbol{\nu}_f - \boldsymbol{E}|_{K_2} \times \boldsymbol{\nu}_f) \cdot \boldsymbol{\varphi} \,\mathrm{d}S \\
&= \int_{\Omega_h} \boldsymbol{w} \cdot \boldsymbol{\varphi} \,\mathrm{d}\boldsymbol{x}
\end{aligned}$$

and therefore $\nabla \times \boldsymbol{E} = \boldsymbol{w}$ and $\boldsymbol{E} \in \boldsymbol{H}(\mathrm{curl}, \Omega_h)$.

Conversely, if $\boldsymbol{E} \in \boldsymbol{H}(\mathrm{curl}, \Omega_h)$, define $\boldsymbol{w} = \nabla \times \boldsymbol{E}$. Since $\boldsymbol{E}|_K \in [H^1(\Omega_h)]^d$, the trace on f is well defined and we obtain

$$\sum_{f, f=\overline{K}_1 \cap \overline{K}_2, K_1, K_2 \in \mathcal{T}_{h,p}} \int_f (\boldsymbol{\nu}_f \cdot \boldsymbol{E}|_{K_1} - \boldsymbol{\nu}_f \cdot \boldsymbol{E}|_{K_2})\varphi = 0$$

for all $\varphi \in \mathcal{D}$. Hence *1.* holds. ■

Remark 7.2 *When interpreting Lemma 7.6 properly for three-dimensional vector fields of the form $\boldsymbol{E} = (E_1(x_1, x_2), E_2(x_1, x_2), 0)^T$, it is easy to see that condition* 2. *attains the following form for two-dimensional approximations: For each element interface $f = \overline{K}_1 \cap \overline{K}_2$, $K_1, K_2 \in \mathcal{T}_{h,p}$ the traces of the tangential components $\boldsymbol{t}_f \cdot \boldsymbol{E}|_{K_1}$ and $\boldsymbol{t}_f \cdot \boldsymbol{E}|_{K_2}$ on f are the same, where $\boldsymbol{t}_f = (-\nu_{f,2}, \nu_{f,1})^T$ is a unit tangential vector to f.*

7.5.2 Lowest-order (Whitney) edge elements

We have shown in Paragraph 7.5.1 that the finite element approximation has to have continuous tangential components on all mesh edges in order to conform to the space $\boldsymbol{H}(\mathrm{curl}, \Omega)$. The lowest-order approximations that satisfy this requirement are with continuous and constant tangential components on the edges. Let us stay on the reference domain K_t first. The two-dimensional space $[P^0(K_t)]^2$ is too small to generate three linearly independent constant tangential components on the edges of K_t. Therefore we need to take one higher degree polynomial from the space $[P^1(K_t)]^2$.

Hence the lowest-order element $(K_t, \hat{\boldsymbol{P}}, \hat{\Sigma})$ on the reference triangular domain K_t is equipped with the polynomial space

$$\hat{\boldsymbol{P}} = \left\{ \hat{\boldsymbol{E}} \in [P^1(K_t)]^2;\ \hat{\boldsymbol{E}} \cdot \hat{\boldsymbol{t}}_j|_{e_j} \text{ is constant, } j = 1, \ldots, 3 \right\}, \tag{7.94}$$

where $\hat{\boldsymbol{t}}_j$ stands for the unit tangential vector to the edge e_j of K_t. The orientation of the edges is shown in Figure 7.7.

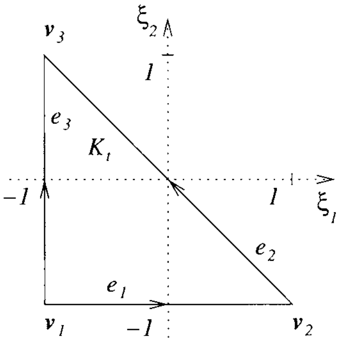


Figure 7.7 Orientation of the edges on the reference domain K_t.

For future reference let us write the unit tangential vectors to the edges explicitly,

$$\hat{t}_1 = (1,0)^T, \quad \hat{t}_2 = \left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)^T, \quad \hat{t}_3 = (0,1)^T.$$

It is left to the reader as an easy exercise to verify that

$$N_P = \dim(\hat{\boldsymbol{P}}) = 3.$$

Accordingly the set of degrees of freedom contains three linear forms, $\hat{\Sigma} = \{\hat{L}_{e_1,0}, \hat{L}_{e_2,0}, \hat{L}_{e_3,0}\}$, where $\hat{L}_{e_j,0} : \hat{\boldsymbol{P}} \to \mathbb{R}$ is defined as the integral

$$\hat{L}_{e_j,0}(\hat{\boldsymbol{E}}) = \int_{e_j} \hat{\boldsymbol{E}} \cdot \hat{\boldsymbol{t}}_j \, \mathrm{d}\xi \quad \text{for all } \hat{\boldsymbol{E}} \in \hat{\boldsymbol{P}} \tag{7.95}$$

of the tangential component of the field $\hat{\boldsymbol{E}}$ on the edge e_j.

Lemma 7.7 (Unisolvency) *The finite element* $(K_t, \hat{\boldsymbol{P}}, \hat{\Sigma})$ *is unisolvent.*

Proof: According to Definition 3.2 we have to show that the following implication holds:

$$\hat{L}_{e_1,0}(\boldsymbol{g}) = \hat{L}_{e_2,0}(\boldsymbol{g}) = \hat{L}_{e_3,0}(\boldsymbol{g}) = 0 \Rightarrow \boldsymbol{g} = \boldsymbol{0} \quad \text{for all } \boldsymbol{g} \in \hat{\boldsymbol{P}}. \tag{7.96}$$

Let us find some basis in the space $\hat{\boldsymbol{P}}$ first. A general polynomial $\boldsymbol{g} \in [P^1(K_t)]^2$ has the form

$$\boldsymbol{g}(\xi_1, \xi_2) = (a_0 + a_1\xi_1 + a_2\xi_2, b_0 + b_1\xi_1 + b_2\xi_2)^T.$$

The condition $\boldsymbol{g} \cdot \hat{\boldsymbol{t}}_1 = \text{const.}$ on the edge e_1, where

$$\boldsymbol{g} \cdot \hat{\boldsymbol{t}}_1 = \boldsymbol{g}(\xi_1, \xi_2) \cdot (1,0)^T = a_0 + a_1\xi_1 + a_2\xi_2 = a_0 + a_1\xi_1 - a_2,$$

implies that $a_1 = 0$. Similarly the condition $\boldsymbol{g} \cdot (0,1)^T = \text{const.}$ on the edge e_3 yields $b_2 = 0$, and the last condition $\boldsymbol{g} \cdot (-1,1)^T = \text{const.}$ on the edge e_2 means that $a_2 = -b_1$. Hence any polynomial $\boldsymbol{g} \in \hat{\boldsymbol{P}}$ has the form

$$\boldsymbol{g}(\xi_1, \xi_2) = (a_0 + a_2\xi_2, b_0 - a_2\xi_1)^T,$$

and, for example, the set

$$B = \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} \xi_2 \\ -\xi_1 \end{pmatrix} \right\} = \{\boldsymbol{g}_1, \boldsymbol{g}_2, \boldsymbol{g}_3\} \tag{7.97}$$

is a basis in $\hat{\boldsymbol{P}}$. Now any $\boldsymbol{g} \in \hat{\boldsymbol{P}}$ can be expressed uniquely as

$$\boldsymbol{g} = \alpha_1\boldsymbol{g}_1 + \alpha_2\boldsymbol{g}_2 + \alpha_3\boldsymbol{g}_3,$$

and the left-hand side of the implication (7.96) can be written as

$$\begin{aligned}
\alpha_1\hat{L}_{e_1,0}(\boldsymbol{g}_1) + \alpha_2\hat{L}_{e_1,0}(\boldsymbol{g}_2) + \alpha_3\hat{L}_{e_1,0}(\boldsymbol{g}_3) &= 0, \\
\alpha_1\hat{L}_{e_2,0}(\boldsymbol{g}_1) + \alpha_2\hat{L}_{e_2,0}(\boldsymbol{g}_2) + \alpha_3\hat{L}_{e_2,0}(\boldsymbol{g}_3) &= 0, \\
\alpha_1\hat{L}_{e_3,0}(\boldsymbol{g}_1) + \alpha_2\hat{L}_{e_3,0}(\boldsymbol{g}_2) + \alpha_3\hat{L}_{e_3,0}(\boldsymbol{g}_3) &= 0.
\end{aligned}$$

Since the coefficient matrix of this equation,

$$\boldsymbol{L} = \{\hat{L}_{e_i,0}(\boldsymbol{g}_k)\}_{i,k=1}^{N_P} = \begin{pmatrix} 2 & 0 & -2 \\ -2 & 2 & 0 \\ 0 & 2 & 2 \end{pmatrix}, \tag{7.98}$$

is nonsingular, we conclude that $\alpha_1 = \alpha_2 = \alpha_3 = 0$. ■

Nodal basis of the element $(\boldsymbol{K_t}, \hat{\boldsymbol{P}}, \hat{\Sigma})$ Next let us apply the standard procedure from Paragraph 3.1.1 to construct the nodal basis satisfying the delta property (3.2). Using the basis (7.97) as the underlying basis, the generalized Vandermonde matrix (3.7) has the form (7.98). The inverse of $\boldsymbol{L}$,

$$\boldsymbol{L}^{-1} = \frac{1}{4}\begin{pmatrix} 1 & -1 & 1 \\ 1 & 1 & 1 \\ -1 & -1 & 1 \end{pmatrix},$$

determines that the nodal shape functions $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$ have the form

$$\begin{aligned} \hat{\theta}_1(\boldsymbol{\xi}) &= \frac{1}{4}(\boldsymbol{g}_1(\boldsymbol{\xi}) + \boldsymbol{g}_2(\boldsymbol{\xi}) - \boldsymbol{g}_3(\boldsymbol{\xi})) = \frac{1}{4}\begin{pmatrix} 1-\xi_2 \\ 1+\xi_1 \end{pmatrix}, \\ \hat{\theta}_2(\boldsymbol{\xi}) &= \frac{1}{4}(-\boldsymbol{g}_1(\boldsymbol{\xi}) + \boldsymbol{g}_2(\boldsymbol{\xi}) - \boldsymbol{g}_3(\boldsymbol{\xi})) = \frac{1}{4}\begin{pmatrix} -1-\xi_2 \\ 1+\xi_1 \end{pmatrix}, \\ \hat{\theta}_3(\boldsymbol{\xi}) &= \frac{1}{4}(\boldsymbol{g}_1(\boldsymbol{\xi}) + \boldsymbol{g}_2(\boldsymbol{\xi}) + \boldsymbol{g}_3(\boldsymbol{\xi})) = \frac{1}{4}\begin{pmatrix} 1+\xi_2 \\ 1-\xi_1 \end{pmatrix}. \end{aligned} \tag{7.99}$$

Another equivalent expression of the Whitney shape functions (7.99) is

$$\begin{aligned} \hat{\theta}_1(\boldsymbol{\xi}) &= \frac{1}{|e_1|}\left(\frac{\hat{\lambda}_3(\boldsymbol{\xi})\hat{\boldsymbol{\nu}}_2}{\hat{\boldsymbol{\nu}}_2 \cdot \hat{\boldsymbol{t}}_1} + \frac{\hat{\lambda}_2(\boldsymbol{\xi})\hat{\boldsymbol{\nu}}_3}{\hat{\boldsymbol{\nu}}_3 \cdot \hat{\boldsymbol{t}}_1}\right), \\ \hat{\theta}_2(\boldsymbol{\xi}) &= \frac{1}{|e_2|}\left(\frac{\hat{\lambda}_1(\boldsymbol{\xi})\hat{\boldsymbol{\nu}}_3}{\hat{\boldsymbol{\nu}}_3 \cdot \hat{\boldsymbol{t}}_2} + \frac{\hat{\lambda}_3(\boldsymbol{\xi})\hat{\boldsymbol{\nu}}_1}{\hat{\boldsymbol{\nu}}_1 \cdot \hat{\boldsymbol{t}}_2}\right), \\ \hat{\theta}_3(\boldsymbol{\xi}) &= \frac{1}{|e_3|}\left(\frac{\hat{\lambda}_2(\boldsymbol{\xi})\hat{\boldsymbol{\nu}}_1}{\hat{\boldsymbol{\nu}}_1 \cdot \hat{\boldsymbol{t}}_3} + \frac{\hat{\lambda}_1(\boldsymbol{\xi})\hat{\boldsymbol{\nu}}_2}{\hat{\boldsymbol{\nu}}_2 \cdot \hat{\boldsymbol{t}}_3}\right), \end{aligned} \tag{7.100}$$

where $\hat{\boldsymbol{\nu}}_i$ is the unit outer normal vector to the edge e_i,

$$\hat{\boldsymbol{\nu}}_1 = (0, -1)^T, \quad \hat{\boldsymbol{\nu}}_2 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)^T, \quad \hat{\boldsymbol{\nu}}_3 = (-1, 0)^T,$$

and the barycentric coordinates $\hat{\lambda}_i(\boldsymbol{\xi})$ are affine functions satisfying

$$\hat{\lambda}_i(\boldsymbol{\xi}) = \begin{cases} 0 & \text{on the edge } e_i, \\ 1 & \text{at the remaining vertex of } K_t \text{ not lying at } e_i. \end{cases} \tag{7.101}$$

For K_t the barycentric coordinates have the form

$$\hat{\lambda}_1(\boldsymbol{\xi}) = \frac{1+\xi_2}{2}, \quad \hat{\lambda}_2(\boldsymbol{\xi}) = \frac{-\xi_1-\xi_2}{2}, \quad \hat{\lambda}_3(\boldsymbol{\xi}) = \frac{1+\xi_1}{2}. \tag{7.102}$$

A third way to express the Whitney shape functions $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$ is

$$\begin{aligned} \hat{\theta}_1 &= \hat{\lambda}_2\nabla\hat{\lambda}_3 - \hat{\lambda}_3\nabla\hat{\lambda}_2, \\ \hat{\theta}_2 &= \hat{\lambda}_3\nabla\hat{\lambda}_1 - \hat{\lambda}_1\nabla\hat{\lambda}_3, \\ \hat{\theta}_3 &= \hat{\lambda}_2\nabla\hat{\lambda}_1 - \hat{\lambda}_1\nabla\hat{\lambda}_2. \end{aligned} \tag{7.103}$$

The equivalence of relations (7.99), (7.100), and (7.103) is left to the reader as an easy exercise.

Whitney element on a triangular domain $K \in \mathcal{T}_{k,p}$ Proper treatment of orientation of the tangential vectors to mesh edges is essential when working with edge elements. Assume a mesh edge s_j with endpoints $\boldsymbol{x}_{i_1}$ and $\boldsymbol{x}_{i_2}$. Define the global orientation of this edge as $s_j = \boldsymbol{x}_{i_1}\boldsymbol{x}_{i_2}$ if $i_1 < i_2$ and $s_j = \boldsymbol{x}_{i_2}\boldsymbol{x}_{i_1}$ otherwise.

Consider a triangular element $K \in \mathcal{T}_{h,p}$ and the affine reference map $\boldsymbol{x}_K : K_t \to K$ with $J_K = \det(\mathrm{D}\boldsymbol{x}_K/\mathrm{D}\boldsymbol{\xi}) > 0$ defined in (3.21). By a_1, a_2, and a_3 denote the edges of K so that $a_1 = \boldsymbol{x}_K(e_1)$, $a_2 = \boldsymbol{x}_K(e_2)$, and $a_3 = \boldsymbol{x}_K(e_3)$. Locally on K, each edge a_i is assigned a unique orientation flag $o_{K,i}$,

$$o_{K,i} = \begin{cases} 1 & \text{if the orientations of } a_i \text{ and } \boldsymbol{x}_K(e_i) \text{ are the same,} \\ -1 & \text{otherwise.} \end{cases}$$

Then the Whitney edge element on K is defined as a triad $(K, \boldsymbol{P}_K, \Sigma_K)$, where

$$\boldsymbol{P}_K = \left\{ \boldsymbol{E} \in [P^1(K)]^2;\ \boldsymbol{E}\cdot\boldsymbol{t}_j|_{a_j} \text{ is constant, } j = 1,\dots,3 \right\}. \tag{7.104}$$

The symbol $\boldsymbol{t}_j$ stands for the unit tangential vector to the edge a_j that corresponds to its unique global orientation in $\mathcal{T}_{h,p}$. The set of degrees of freedom Σ_K comprises three linear forms $L_{a_1,0}, L_{a_2,0}$ and $L_{a_3,0}$ defined by

$$L_{a_j,0}(\boldsymbol{E}) = \int_{a_j} \boldsymbol{E}\cdot o_{K,j}\boldsymbol{t}_j \,\mathrm{d}\zeta \quad \text{for all } \boldsymbol{E} \in \boldsymbol{P}_K. \tag{7.105}$$

Lemma 7.8 (Unisolvency) *The finite element* $(K, \boldsymbol{P}_K, \Sigma_K)$ *is unisolvent.*

Proof: Analogous to the proof of Lemma 7.7. ■

Nodal basis of the element $(K, \boldsymbol{P}_K, \Sigma_K)$ The unique nodal basis satisfying the delta property (3.2) can be designed routinely using the inverse of the Vandermonde matrix (3.7), analogously to what was done above for the Whitney element $(K_t, \hat{\boldsymbol{P}}, \hat{\Sigma})$ on the reference domain.

Alternatively, for the Whitney element $(K, \boldsymbol{P}_K, \Sigma_K)$ the delta property $L_{a_i,0}(\theta_j) = \delta_{ij}$ is equivalent to the condition

$$\theta_i \cdot o_{K,j}\boldsymbol{t}_j = \frac{\delta_{ij}}{|a_j|} \quad \text{for all } 1 \le i,j \le 3. \tag{7.106}$$

Thus, for example, the formulae (7.100) can naturally be extended to

$$\begin{aligned} \theta_1(\boldsymbol{x}) &= \frac{o_{K,1}}{|a_1|}\left(\frac{\lambda_3(\boldsymbol{x})\boldsymbol{\nu}_2}{\boldsymbol{\nu}_2\cdot\boldsymbol{t}_1}+\frac{\lambda_2(\boldsymbol{x})\boldsymbol{\nu}_3}{\boldsymbol{\nu}_3\cdot\boldsymbol{t}_1}\right), \\ \theta_2(\boldsymbol{x}) &= \frac{o_{K,2}}{|a_2|}\left(\frac{\lambda_1(\boldsymbol{x})\boldsymbol{\nu}_3}{\boldsymbol{\nu}_3\cdot\boldsymbol{t}_2}+\frac{\lambda_3(\boldsymbol{x})\boldsymbol{\nu}_1}{\boldsymbol{\nu}_1\cdot\boldsymbol{t}_2}\right), \\ \theta_3(\boldsymbol{x}) &= \frac{o_{K,3}}{|a_3|}\left(\frac{\lambda_2(\boldsymbol{x})\boldsymbol{\nu}_1}{\boldsymbol{\nu}_1\cdot\boldsymbol{t}_3}+\frac{\lambda_1(\boldsymbol{x})\boldsymbol{\nu}_2}{\boldsymbol{\nu}_2\cdot\boldsymbol{t}_3}\right), \end{aligned} \tag{7.107}$$

where $\boldsymbol{t}_i$ and $\boldsymbol{\nu}_i$ are the unit tangential and normal vectors to the edge a_i of K, respectively, and $\lambda_j(\boldsymbol{x})$ are the corresponding barycentric coordinates on K defined analogously to (7.101).

Equivalence of the elements $(K_t, \hat{\boldsymbol{P}}, \hat{\Sigma})$ ***and*** $(K, \boldsymbol{P}_K, \Sigma_K)$ Before the discretization can be performed in an element-by-element fashion on the reference domain K_t, as usual we need to find a suitable linear operator $\boldsymbol{\Phi}_K : \hat{\boldsymbol{P}} \to \boldsymbol{P}_K$ so that the equivalence of the elements $(K_t, \hat{\boldsymbol{P}}, \hat{\Sigma})$ and $(K, \boldsymbol{P}_K, \Sigma_K)$ according to Definition 3.8 can be established. At this point it is customary to use the De Rham diagram to make a quick argument leading directly to the correct map $\boldsymbol{\Phi}_K$. However, let us save this for later and first show in Example 7.1 why a straightforward extension of the usual operator $\Phi_K(\hat{g}) = \hat{g} \circ \boldsymbol{x}_K^{-1}$, which was used to establish the equivalence of Lagrange elements, does not work for edge elements.

■ EXAMPLE 7.1 (Trying the map $\boldsymbol{\Psi}(\hat{\boldsymbol{E}}) = \hat{\boldsymbol{E}} \circ \boldsymbol{x}_K^{-1}$)

By $\boldsymbol{\Psi}_K : \hat{\boldsymbol{P}} \to \boldsymbol{P}_K$ denote the linear operator from Definition 3.8, and suppose for a moment that it has the form

$$\boldsymbol{\Psi}_K(\hat{\boldsymbol{E}}) = \hat{\boldsymbol{E}} \circ \boldsymbol{x}_K^{-1}. \tag{7.108}$$

Let K be an element with the vertices $[0,0], [0,1], [-1,0]$ and $\boldsymbol{x}_K : K_t \to K$ the corresponding affine reference map,

$$\boldsymbol{x}_K(\boldsymbol{\xi}) = \frac{1}{2}\begin{pmatrix} -\xi_2 - 1 \\ \xi_1 + 1 \end{pmatrix},$$

as shown in Figure 7.8.

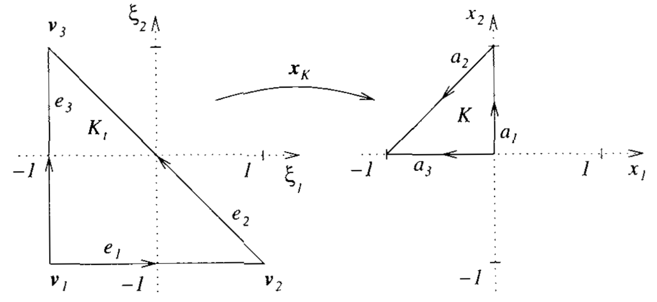


Figure 7.8 Affine transformation $\boldsymbol{x}_K : K_t \to K$.

With the edges of K oriented as shown in Figure 7.8, it is $o_{K,1} = o_{K,2} = o_{K,3} = 1$ and by (7.107) the basis function θ_1 on K has the form

$$\theta_1 = \begin{pmatrix} -x_2 \\ x_1 + 1 \end{pmatrix}.$$

However, relation (7.108) yields a different result,

$$\boldsymbol{\Psi}_K(\hat{\theta}_1) = \hat{\theta}_1 \circ \boldsymbol{x}_K^{-1} = \frac{1}{2}\begin{pmatrix} x_1 + 1 \\ x_2 \end{pmatrix},$$

which has nonconstant tangential components on all edges a_1, a_2 and a_3, and thus does not lie in the polynomial space $\boldsymbol{P}_K$!

The reason for this incompatibility is that the map $\boldsymbol{\Psi}_K$ transformed both vector components of $\hat{\theta}_1$ on all edges from K_t to K exactly, but the direction of the unit tangential vectors to the edges changed.

The way to solve the problem encountered in Example 7.1 is to define

$$\boldsymbol{\Phi}_K = \boldsymbol{\Psi}_K \circ \boldsymbol{\Theta},$$

where the linear transformation $\boldsymbol{\Theta} : \mathbb{R}^2 \to \mathbb{R}^2$ adjusts the field on the reference domain K_t so that

$$\frac{|a_j|}{|e_j|}\int_{e_j} \boldsymbol{\Theta}(\hat{\boldsymbol{E}}) \cdot o_{K,j}\boldsymbol{t}_j \,\mathrm{d}\xi = \int_{e_j} \hat{\boldsymbol{E}} \cdot \hat{\boldsymbol{t}}_j \,\mathrm{d}\xi, \quad j = 1, 2, 3, \tag{7.109}$$

where $o_{K,j}\boldsymbol{t}_j$ is the unit tangential vector to the edge a_j of K oriented compatibly with $\hat{\boldsymbol{t}}_j$ through the map $\boldsymbol{x}_K$.

The tangential vectors $\hat{\boldsymbol{t}}_j$ are transformed by $\boldsymbol{x}_K$ according to the relation

$$\left(\frac{\mathrm{D}\boldsymbol{x}_K}{\mathrm{D}\boldsymbol{\xi}}\right) |e_j| \hat{\boldsymbol{t}}_j = |a_j| o_{K,j}\boldsymbol{t}_j. \tag{7.110}$$

Let the matrix $\boldsymbol{T}$ of the type 2×2 represent the transformation $\boldsymbol{\Theta}$. Then, after substituting for $o_{K,j}\boldsymbol{t}_j$ from (7.110), relation (7.109) becomes

$$\frac{|a_j|}{|e_j|}\int_{e_j} \boldsymbol{T}\hat{\boldsymbol{E}} \cdot \frac{|e_j|}{|a_j|}\left(\frac{\mathrm{D}\boldsymbol{x}_K}{\mathrm{D}\boldsymbol{\xi}}\right)\hat{\boldsymbol{t}}_j \,\mathrm{d}\xi = \int_{e_j} \hat{\boldsymbol{E}} \cdot \hat{\boldsymbol{t}}_j \,\mathrm{d}\xi, \quad j = 1, 2, 3,$$

which in turn is equivalent to

$$\int_{e_j} \boldsymbol{T}\left(\frac{\mathrm{D}\boldsymbol{x}_K}{\mathrm{D}\boldsymbol{\xi}}\right)^T \hat{\boldsymbol{E}} \cdot \hat{\boldsymbol{t}}_j \,\mathrm{d}\xi = \int_{e_j} \hat{\boldsymbol{E}} \cdot \hat{\boldsymbol{t}}_j \,\mathrm{d}\xi, \quad j = 1, 2, 3, \tag{7.111}$$

To satisfy equation (7.111), the matrix $\boldsymbol{T}$ has to have the form

$$\boldsymbol{T} = \left(\frac{\mathrm{D}\boldsymbol{x}_K}{\mathrm{D}\boldsymbol{\xi}}\right)^{-T}.$$

Thus finally the correct transformation relation is

$$\boldsymbol{E} = \boldsymbol{\Phi}_K(\hat{\boldsymbol{E}}), \tag{7.112}$$

where

$$\boldsymbol{E}(\boldsymbol{x}) = \left(\frac{\mathrm{D}\boldsymbol{x}_K}{\mathrm{D}\boldsymbol{\xi}}\right)^{-T} \hat{\boldsymbol{E}}(\boldsymbol{\xi}), \qquad \boldsymbol{x} = \boldsymbol{x}_K(\boldsymbol{\xi}). \tag{7.113}$$

Lemma 7.9 *The finite elements $(K_t, \hat{\boldsymbol{P}}, \hat{\Sigma})$ and $(K, \boldsymbol{P}_K, \Sigma_K)$ are equivalent under the transformation $\boldsymbol{\Phi}_K : \hat{\boldsymbol{P}} \to \boldsymbol{P}_K$ defined in (7.112), (7.113).*

Proof: According to Definition 3.8, we need to verify that

$$\boldsymbol{\Phi}_K(\hat{\boldsymbol{P}}) = \boldsymbol{P}_K, \tag{7.114}$$

and

$$\hat{L}_{e_j,0}(\hat{\boldsymbol{E}}) = L_{a_j,0}(\boldsymbol{\Phi}_K(\hat{\boldsymbol{E}})) \quad \text{for all } \hat{\boldsymbol{E}} \in \hat{\boldsymbol{P}} \text{ and } j = 1, 2, 3. \tag{7.115}$$

However, (7.114) is clear from the linearity of the transformation Θ and affinity of the reference map $\boldsymbol{x}_K$. To verify relation (7.115), calculate

$$\begin{aligned}
\hat{L}_{e_j,0}(\hat{\boldsymbol{E}}) &= \int_{e_j} \hat{\boldsymbol{E}} \cdot \hat{\boldsymbol{t}}_j \, \mathrm{d}\xi \\
&= \int_{e_j} \hat{\boldsymbol{E}} \cdot \left[\left(\frac{\mathrm{D}\boldsymbol{x}_K}{\mathrm{D}\boldsymbol{\xi}}\right)^{-1} \frac{|a_j|}{|e_j|} o_{K,j} \boldsymbol{t}_j\right] \mathrm{d}\xi \\
&= \int_{a_j} \left[\frac{|e_j|}{|a_j|} \left(\frac{\mathrm{D}\boldsymbol{x}_K}{\mathrm{D}\boldsymbol{\xi}}\right)^{-T} \hat{\boldsymbol{E}} \circ \boldsymbol{x}_K^{-1}\right] \cdot \left[\frac{|a_j|}{|e_j|} o_{K,j} \boldsymbol{t}_j\right] \mathrm{d}\zeta \\
&= \int_{a_j} \boldsymbol{\Phi}_K(\hat{\boldsymbol{E}}) \cdot [o_{K,j} \boldsymbol{t}_j] \, \mathrm{d}\zeta \\
&= L_{a_j,0}(\boldsymbol{\Phi}_K(\hat{\boldsymbol{E}})).
\end{aligned}$$

■

Design of basis functions Let a polygonal domain $\Omega_h \subset \mathbb{R}^2$ be covered with a finite element mesh $\mathcal{T}_{h,p}$ consisting of M triangular Whitney elements $K_1, K_2, \ldots, K_M$. Then the Galerkin subspace $V_{h,p}$ of the space $V = \boldsymbol{H}(\mathrm{curl}, \Omega_h)$ has the form

$$\begin{aligned}
V_{h,p} = \{&\boldsymbol{E}_{h,p} \in V;\ \boldsymbol{E}_{h,p} \in [P^1(K_i)]^2 \text{ for all } K_i \in \mathcal{T}_{h,p}, \\
&\boldsymbol{E}_{h,p} \cdot \boldsymbol{t}_{s_j}|_{s_j} = \text{const. for every mesh edge } s_j\}
\end{aligned}$$

(we do not consider essential boundary conditions at this point, the incorporation of boundary conditions will be discussed later). In this case the dimension of the space $V_{h,p}$ is

$$\dim(V_{h,p}) = M_e,$$

where M_e is the number of unconstrained edges in the mesh $\mathcal{T}_{h,p}$. By unconstrained, as before, we mean an edge where degrees of freedom are present.

Assume such edge s_j in the mesh, along with the corresponding element patch,

$$S_e(j) = \bigcup_{k \in N_e(j)} \overline{K}_k, \tag{7.116}$$

where

$$N_e(j) = \{k;\ K_k \in \mathcal{T}_{h,p},\ s_j \text{ is an edge of } K_k\}, \tag{7.117}$$

as shown in Figure 7.9.

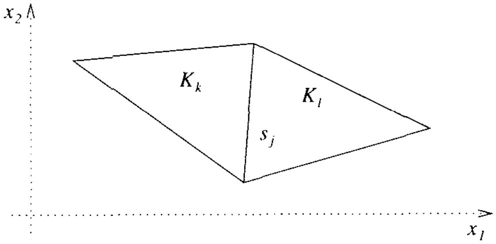


Figure 7.9 Element patch $S_e(j)$ corresponding to an interior mesh edge s_j.

For each element $K_k \in S_e(j)$, by e_m denote the edge of the reference domain K_t, such that $\boldsymbol{x}_{K_k}(e_m) = s_j$. Define $o_{K_k,m} = 1$ if the orientations of $\boldsymbol{x}_{K_k}(e_m)$ and s_j are the same, and $o_{K_k,m} = -1$ otherwise . The lowest-order (Whitney) basis function $\boldsymbol{E}_0^{s_j}$ associated with the edge s_j is zero in $\Omega_h \setminus S_e(j)$, and in $S_e(j)$ it is defined by

$$\boldsymbol{E}_0^{s_j}(\boldsymbol{x}) = o_{K_k,m} \left(\frac{\mathrm{D}\boldsymbol{x}_{K_k}}{\mathrm{D}\boldsymbol{\xi}} \right)^{-T} \hat{\theta}_m \circ \boldsymbol{x}_{K_k}^{-1}, \quad K_k \subset S_e(j),$$

where $\hat{\theta}_m$ is the Whitney shape function on the reference domain K_t corresponding to the edge e_m. Let us remark that as usual no explicit inversion of the maps $\boldsymbol{x}_{K_i}$ is needed for the assembling algorithm.

7.5.3 Higher-order edge elements of Nédélec

Next let us generalize the lowest-order edge elements from Paragraph 7.5.2 to the first family of Nédélec elements [87]. For this we need a special polynomial space on the reference triangular domain K_t. We begin with defining a space of scalar homogeneous polynomials of degree k,

$$\tilde{P}^k = \operatorname{span}\left\{\xi_1^i \xi_2^j;\ i + j = k;\ \boldsymbol{\xi} \in K_t\right\},$$

and a special subspace of homogeneous vector polynomials of degree k,

$$\hat{\boldsymbol{S}}^k = \left\{\boldsymbol{p} \in (\tilde{P}^k)^2;\ \boldsymbol{\xi} \cdot \boldsymbol{p}(\boldsymbol{\xi}) = 0\right\}, \tag{7.118}$$

where $\boldsymbol{\xi} \cdot \boldsymbol{p}(\boldsymbol{\xi}) = \xi_1 p_1(\xi_1, \xi_2) + \xi_2 p_2(\xi_1, \xi_2)$. Let us calculate the dimension of $\hat{\boldsymbol{S}}^k$: The space (7.118) is the nullspace of the linear transformation $\boldsymbol{p} \in (\tilde{P}^k)^2 \to \boldsymbol{\xi} \cdot \boldsymbol{p} \in \tilde{P}^{k+1}$, which is a surjection (for every $q \in \tilde{P}^{k+1}$ there exists a $\boldsymbol{p} \in (\tilde{P}^k)^2$ such that $q = \boldsymbol{\xi} \cdot \boldsymbol{p}$). It follows from Lemma A.9 that the dimension of $\hat{\boldsymbol{S}}^k$ is

$$\begin{aligned} \dim(\hat{\boldsymbol{S}}^k) &= \dim((\tilde{P}^k)^2) - \dim(\tilde{P}^{k+1}) \\ &= 2\dim(\tilde{P}^k) - \dim(\tilde{P}^{k+1}) = 2(k+1) - (k+2) = k. \end{aligned}$$

The following lemma gives a geometrical characterization of polynomials in the space $\hat{\boldsymbol{S}}^k$:

Lemma 7.10 *Let* $\boldsymbol{p} \in \hat{\boldsymbol{S}}^k$. *Then the tangential component of* $\boldsymbol{p}$ *along any straight line is a* $(k-1)$*th-degree polynomial.*

Proof: Any straight line $\boldsymbol{\omega}$ in $\mathbb{R}^2$ can be written as $\boldsymbol{\omega}(s) = (q_1 + sv_1, q_2 + sv_2)^T$, where $\boldsymbol{q} = (q_1, q_2)^T$ is a point in $\mathbb{R}^2$, $\boldsymbol{v} = (v_1, v_2)^T$ a unitary directional vector (tangential to $\boldsymbol{\omega}$) and s a real parameter. The condition $\boldsymbol{\xi} \cdot \boldsymbol{p}(\boldsymbol{\xi}) = 0$ on $\boldsymbol{\omega}$ yields

$$\begin{aligned} 0 &= \boldsymbol{\xi} \cdot \boldsymbol{p}(\boldsymbol{\xi})|_{\boldsymbol{\omega}} \\ &= \boldsymbol{\omega}(s) \cdot \boldsymbol{p}(\boldsymbol{\omega}(s)) \\ &= (q_1 + sv_1)p_1(\boldsymbol{\omega}(s)) + (q_2 + sv_2)p_2(\boldsymbol{\omega}(s)) \\ &= \underbrace{q_1 p_1(\boldsymbol{\omega}(s)) + q_2 p_2(\boldsymbol{\omega}(s))}_{\in P^k(\mathbb{R})} + s\underbrace{[v_1 p_1(\boldsymbol{\omega}(s)) + v_2 p_2(\boldsymbol{\omega}(s))]}_{\boldsymbol{v} \cdot \boldsymbol{p}(\boldsymbol{\omega}(s))}. \end{aligned}$$

Since $s\boldsymbol{v} \cdot \boldsymbol{p}(\boldsymbol{\omega}(s)) \in P^k(\mathbb{R})$, necessarily it is $\boldsymbol{v} \cdot \boldsymbol{p}(\boldsymbol{\omega}(s)) \in P^{k-1}(\mathbb{R})$, which concludes the proof. ■

The polynomial space on the general Nédélec element is defined as

$$\hat{\boldsymbol{P}}^k = [P^{k-1}(K_t)]^2 \oplus \hat{\boldsymbol{S}}^k. \tag{7.119}$$

The basis (7.97) confirms that $\hat{\boldsymbol{P}}^1$ indeed is the space on the lowest-order Whitney element. The dimension is calculated easily,

$$N_P = \dim(\hat{\boldsymbol{P}}^k) = 2\dim(P^{k-1}(K_t)) + \dim(\hat{\boldsymbol{S}}^k) = 2\frac{k(k+1)}{2} + k = k(k+2).$$

It follows from Lemma 7.10 that the traces of the tangential components of $\hat{\boldsymbol{P}}^k$-functions to the edges of K_t are polynomials of the degree less than or equal to $k-1$. The Whitney space (7.94) obeyed the same rule.

Lemma 7.11 *The space* $\hat{\boldsymbol{P}}^k$ *is a part of an algebraic decomposition*

$$[P^k(K_t)]^2 = \hat{\boldsymbol{P}}^k \oplus \nabla \tilde{P}^{k+1}.$$

Proof: Let $\hat{\boldsymbol{E}} \in \hat{\boldsymbol{P}}^k \cap \nabla \tilde{P}^{k+1}$. Then there is some homogeneous scalar polynomial φ in $\tilde{P}^{k+1}$ such that $\hat{\boldsymbol{E}} = \nabla\varphi$. The facts that $\nabla\varphi \in \tilde{P}^k$ and $\nabla\varphi \in \hat{\boldsymbol{P}}^k$ imply that $\nabla\varphi \in \hat{\boldsymbol{S}}^k$. From here it follows that $\boldsymbol{\xi} \cdot \nabla\varphi(\boldsymbol{\xi}) = 0$. Since $\varphi \in \tilde{P}^{k+1}$, it satisfies

$$\varphi(\boldsymbol{\xi}) = \frac{\boldsymbol{\xi} \cdot \nabla\varphi(\boldsymbol{\xi})}{k+1},$$

and therefore $\varphi = 0$. Thus $\hat{\boldsymbol{P}}^k \cap \nabla\tilde{P}^{k+1} = \{0\}$. Since $\dim([P^k(K_t)]^2) = (k+1)(k+2)$, $\dim(\hat{\boldsymbol{P}}^k) = k(k+2)$ and $\dim(\nabla\tilde{P}^{k+1}) = k+2$, it is

$$\dim([P^k(K_t)]^2) = \dim(\hat{\boldsymbol{P}}^k) + \dim(\nabla\tilde{P}^{k+1}),$$

which concludes the proof. ■

The following result is needed for the unisolvency proof of higher-order nodal edge elements:

Lemma 7.12 *Let $\hat{\boldsymbol{E}} \in \hat{\boldsymbol{P}}^k$ be such that*

$$\nabla \times \hat{\boldsymbol{E}} = 0$$

(where $\nabla \times \hat{\boldsymbol{E}} = \partial\hat{E}_2/\partial x_1 - \partial\hat{E}_1/\partial x_2$ is the surface curl). Then there exists $\varphi \in P^k(K_t)$ such that $\hat{\boldsymbol{E}} = \nabla\varphi$.

Proof: It follows from the De Rham diagram (see, e.g., [111]) that for $\hat{\boldsymbol{E}} \in \hat{\boldsymbol{P}}^k$ such that $\nabla \times \hat{\boldsymbol{E}} = 0$ there exists a scalar potential $\varphi \in H^1(K_t)$ such that $\hat{\boldsymbol{E}} = \nabla\varphi$. It follows from $\hat{\boldsymbol{E}} \in [P^k(K_t)]^2$ that $\varphi \in P^{k+1}(K_t)$. Let us write $\varphi = \varphi_1 + \varphi_2$ where $\varphi_1 \in P^k(K_t)$ and $\varphi_2 \in \tilde{P}^{k+1}$. Since $\hat{\boldsymbol{E}} \in \hat{\boldsymbol{P}}^k$, Lemma 7.11 implies that $\nabla\varphi_2 = 0$. The fact that φ_2 is a homogeneous polynomial implies that $\varphi_2 = 0$, which concludes the proof. ■

Nédélec element on the reference domain K_t For a given $k \geq 1$ the Nédélec element of degree k on the reference triangular domain K_t is defined as a triad $(K_t, \hat{\boldsymbol{P}}^k, \hat{\Sigma}^k)$, where the set of degrees of freedom $\hat{\Sigma}^k$ comprises $N_P = k(k+2)$ linear forms associated with the edges and interior of K_t. To begin with, for each edge e_i there are k degrees of freedom of the form

$$\hat{L}_{e_i,j}(\hat{\boldsymbol{E}}) = \int_{e_i} \hat{\boldsymbol{E}} \cdot \hat{\boldsymbol{t}}_i \, \hat{q}_j^{(i)} \, \mathrm{d}\xi \qquad \text{for all } j = 0, 1, \ldots, k-1, \tag{7.120}$$

where the functions $\hat{q}_j^{(i)}$ are the Legendre polynomials L_j, transformed to the edge e_i. For $j = 0$ one obtains the degrees of freedom $L_{e_i,0}$ on the lowest-order (Whitney) element (7.95). The $(k-1)k$ interior (bubble) degrees of freedom are defined by

$$\hat{L}_{b,j}(\hat{\boldsymbol{E}}) = \int_{K_t} \hat{\boldsymbol{E}} \cdot \hat{\boldsymbol{q}}_j \, \mathrm{d}\boldsymbol{\xi}, \tag{7.121}$$

where $\hat{\boldsymbol{q}}_j$, $j = 1, 2, \ldots, (k-1)k$, is a basis of the space $[P^{k-2}(K_t)]^2$. The reason why these degrees of freedom are called interior is that the traces of the corresponding nodal shape functions vanish on the whole boundary of K_t (to be shown in detail in Paragraph 7.5.5).

The edge and bubble degrees of freedom (7.120) and (7.121) together constitute the set $\hat{\Sigma}^k$. Now the unisolvency result from Lemma 7.7 can be extended to the general polynomial degree $k \geq 1$:

Lemma 7.13 (Unisolvency) *The finite element $(K_t, \hat{\boldsymbol{P}}^k, \hat{\Sigma}^k)$ is unisolvent.*

Proof: Let $\hat{\boldsymbol{E}} \in \hat{\boldsymbol{P}}^k$ be arbitrary such that all the $k(k+2)$ degrees of freedom (7.120) and (7.121) vanish on $\hat{\boldsymbol{E}}$. It is our aim to show that necessarily $\hat{\boldsymbol{E}} = \boldsymbol{0}$. By Lemma 7.10 it is $\hat{\boldsymbol{E}} \cdot \hat{\boldsymbol{t}}_i \in P^{k-1}(e_i)$ for all $1 \leq i \leq 3$. Since

$$\int_{e_i} \hat{\boldsymbol{E}} \cdot \hat{\boldsymbol{t}}_i \hat{q}_j^{(i)} \, \mathrm{d}\xi = 0$$

for all basis functions $\hat{q}_j^{(i)} \in P^{k-1}(e_i)$, the tangential component $\hat{\boldsymbol{E}} \cdot \hat{\boldsymbol{t}}_i$ is L^2-orthogonal to the whole space $P^{k-1}(e_i)$, and therefore it has to be zero on e_i. This holds for all three edges of K_t. Hence Stokes' theorem of calculus (see, e.g., [36]) yields

$$\int_{K_t} (\nabla \times \hat{\boldsymbol{E}}) \hat{q} \, \mathrm{d}\boldsymbol{\xi} = \int_{K_t} \hat{\boldsymbol{E}} \cdot (\bar{\nabla} \times \hat{q}) \, \mathrm{d}\boldsymbol{\xi} \tag{7.122}$$

for all $\hat{q} \in P^{k-1}(K_t)$. Here, similarly to Lemma 7.12, we prefer to use the surface curl $\nabla \times \hat{\boldsymbol{E}} = \partial \hat{E}_2/\partial x_1 - \partial \hat{E}_1/\partial x_2$ and the vector-valued curl of a scalar function, $\bar{\nabla} q = (-\partial q/\partial \xi_2, \partial q/\partial \xi_1)^T$, over going with the curl operator to 3D as we did in Section 7.4.

Since $\bar{\nabla} \times \hat{q} \in [P^{k-2}(K_t)]^2$ and all the volume degrees of freedom (7.121) vanish, by (7.122) we have that

$$\int_{K_t} (\nabla \times \hat{\boldsymbol{E}}) \hat{q} \, \mathrm{d}\boldsymbol{\xi} = 0 \qquad \text{for all} \quad \hat{q} \in P^{k-1}(K_t),$$

and thus $\nabla \times \hat{\boldsymbol{E}} = 0$ in K_t. By Lemma 7.12 there exists $\varphi \in P^k(K_t)$ such that $\hat{\boldsymbol{E}} = \nabla \varphi$. Since the tangential component of $\hat{\boldsymbol{E}}$ on ∂K_t is zero, φ is constant on ∂K_t. Without loss of generality, we can assume that this constant is zero. Herewith the proof is finished for $k \leq 2$. For higher polynomial degrees $k \geq 3$ the function φ can be expressed using the barycentric coordinates (7.102), $\varphi = \hat{\lambda}_1 \hat{\lambda}_2 \hat{\lambda}_3 \psi$, where $\psi \in P^{k-3}(K_t)$. Since all the volume degrees of freedom (7.121) are zero, it is $\psi = 0$ and consequently $\hat{\boldsymbol{E}} = \boldsymbol{0}$, which concludes the proof. ■

The unique nodal basis of the space $\hat{\boldsymbol{P}}^k$ can be constructed routinely via the generalized Vandermonde matrix (3.7). In the following let us design the Nédélec element on a general triangular domain and discuss the affine equivalence of Nédélec elements.

Nédélec element on a triangular domain $K \in \mathcal{T}_{h,p}$ Let every mesh edge be equipped with a unique global orientation given by the global enumeration of vertices. Consider a triangular element $K \in \mathcal{T}_{h,p}$ and the affine reference map $\boldsymbol{x}_K : K_t \to K$ with a positive Jacobian $J_K = \det(\mathrm{D}\boldsymbol{x}_K/\mathrm{D}\boldsymbol{\xi}) > 0$ defined in (3.21). By a_1, a_2 and a_3 denote the edges of K so that $a_1 = \boldsymbol{x}_K(e_1)$, $a_2 = \boldsymbol{x}_K(e_2)$ and $a_3 = \boldsymbol{x}_K(e_3)$. Locally on K, each edge a_i is assigned a unique orientation flag $o_{K,i}$ analogously to the lowest-order case. The Nédélec edge element on K is defined as a triad $(K, \boldsymbol{P}_K^k, \Sigma_K^k)$, where

$$\boldsymbol{P}_K^k = [P^{k-1}(K)]^2 \oplus \boldsymbol{S}^k, \tag{7.123}$$

and the subspace $\boldsymbol{S}^k$ of homogeneous vector polynomials is defined analogously to (7.118). The set of degrees of freedom Σ^k consists of $N_P = k(k+2)$ linear forms associated with the edges and interior of K. The edge degrees of freedom have the form

$$L_{a_i,j}(\boldsymbol{E}) = \int_{a_i} \boldsymbol{E} \cdot o_{K,i} \boldsymbol{t}_i \, q_j^{(i)} \, \mathrm{d}\zeta \tag{7.124}$$

where $j = 0, 1, \ldots, k-1$. For each $j = 0, 1, \ldots, k-1$ the function $q_j^{(i)}$ again is chosen to be the Legendre polynomial $L_0, L_1, \ldots, L_{k-1}$, transformed to the edge a_i. The $(k-1)k$ interior degrees of freedom have the form

$$L_{b,j}(\boldsymbol{E}) = \int_K \boldsymbol{E} \cdot \boldsymbol{q}_j \, \mathrm{d}\boldsymbol{x}, \tag{7.125}$$

where $\boldsymbol{q}_j$, $j = 1, 2, \ldots, (k-1)k$, is the basis of the space $[P^{k-2}(K)]^2$ defined using the basis $\hat{\boldsymbol{q}}_j$, $j = 1, 2, \ldots, (k-1)k$, of $[P^{k-2}(K_t)]^2$ and the reference map $\boldsymbol{x}_K$ as follows,

$$\boldsymbol{q}_j(\boldsymbol{x}) = \frac{1}{J_K} \left(\frac{\mathrm{D}\boldsymbol{x}_K}{\mathrm{D}\boldsymbol{\xi}} \right) \hat{\boldsymbol{q}}_j(\boldsymbol{\xi}), \qquad \boldsymbol{x} = \boldsymbol{x}_K(\boldsymbol{\xi}). \tag{7.126}$$

It is left to the reader as an easy exercise to prove that the functions $\boldsymbol{q}_j$, $j = 1, 2, \ldots, (k-1)k$, indeed constitute a basis in the space $[P^{k-2}(K)]^2$.

Lemma 7.14 (Unisolvency) *The finite element $(K, \boldsymbol{P}_K^k, \Sigma_K^k)$ is unisolvent.*

Proof: Analogous to the proof of Lemma 7.13. ■

Equivalence of the elements $(K_t, \hat{\boldsymbol{P}}^k, \hat{\Sigma}^k)$ and $(K, \boldsymbol{P}_K^k, \Sigma_K^k)$ At this point it remains to be shown that the general Nédélec elements are equivalent under transformation (7.112), (7.113): $\boldsymbol{E} = \boldsymbol{\Phi}_K(\hat{\boldsymbol{E}})$,

$$\boldsymbol{E}(\boldsymbol{x}) = \left(\frac{\mathrm{D}\boldsymbol{x}_K}{\mathrm{D}\boldsymbol{\xi}} \right)^{-T} \hat{\boldsymbol{E}}(\boldsymbol{\xi}), \qquad \boldsymbol{x} = \boldsymbol{x}_K(\boldsymbol{\xi}), \tag{7.127}$$

which was derived for the lowest-order elements.

Lemma 7.15 *The finite elements $(K_t, \hat{\boldsymbol{P}}^k, \hat{\Sigma}^k)$ and $(K, \boldsymbol{P}_K^k, \Sigma_K^k)$ are equivalent under the transformation $\boldsymbol{\Phi}_K : \hat{\boldsymbol{P}}^k \to \boldsymbol{P}_K^k$ defined in (7.127).*

Proof: It follows from the definition (7.119) of the space $\hat{\boldsymbol{P}}^k$ and the definition (7.127) of the transformation $\boldsymbol{\Phi}_K$ that $\boldsymbol{\Phi}_K(\hat{\boldsymbol{P}}) = \boldsymbol{P}_K$. Recall the transformation relation (7.110) for the unit tangential vectors $\hat{\boldsymbol{t}}_i$ to the edges of K_t,

$$\hat{\boldsymbol{t}}_i = o_{K,i} \frac{|a_i|}{|e_i|} \left(\frac{\mathrm{D}\boldsymbol{x}_K}{\mathrm{D}\boldsymbol{\xi}} \right)^{-1} \boldsymbol{t}_i. \tag{7.128}$$

For the edge degrees of freedom $\hat{L}_{e_i,j}$, $0 \leq j \leq k-1$ and $1 \leq i \leq 3$, we have

$$\begin{aligned}
\hat{L}_{e_i,j}(\hat{\boldsymbol{E}}) &= \int_{e_i} \hat{\boldsymbol{E}} \cdot \hat{\boldsymbol{t}}_i \, \hat{q}_j^{(i)} \, \mathrm{d}\xi \\
&= \int_{e_i} \hat{\boldsymbol{E}} \cdot \left[o_{K,i} \frac{|a_i|}{|e_i|} \left(\frac{\mathrm{D}\boldsymbol{x}_K}{\mathrm{D}\boldsymbol{\xi}} \right)^{-1} \boldsymbol{t}_i \right] \hat{q}_j^{(i)} \, \mathrm{d}\xi \\
&= \int_{a_i} \frac{|e_i|}{|a_i|} \left[\left(\frac{\mathrm{D}\boldsymbol{x}_K}{\mathrm{D}\boldsymbol{\xi}} \right)^{-T} \hat{\boldsymbol{E}} \circ \boldsymbol{x}_K^{-1} \right] \cdot \left[o_{K,i} \frac{|a_i|}{|e_i|} \boldsymbol{t}_i \right] q_j^{(i)} \, \mathrm{d}\zeta \\
&= \int_{a_i} \boldsymbol{\Phi}_K(\hat{\boldsymbol{E}}) \cdot o_{K,i} \boldsymbol{t}_i q_j^{(i)} \, \mathrm{d}\zeta \\
&= L_{a_i,j}(\boldsymbol{\Phi}_K(\hat{\boldsymbol{E}})).
\end{aligned}$$

Using (7.126) and (7.127), we find out that also the bubble degrees of freedom satisfy

$$\begin{aligned}
\hat{L}_{b,j}(\hat{\boldsymbol{E}}) &= \int_{K_t} \hat{\boldsymbol{E}} \cdot \hat{\boldsymbol{q}}_j \, \mathrm{d}\boldsymbol{\xi} \\
&= \int_{K_t} \left[J_K \left(\frac{\mathrm{D}\boldsymbol{x}_K}{\mathrm{D}\boldsymbol{\xi}} \right)^{-T} \hat{\boldsymbol{E}} \right] \cdot \left[\frac{1}{J_K} \left(\frac{\mathrm{D}\boldsymbol{x}_K}{\mathrm{D}\boldsymbol{\xi}} \right) \hat{\boldsymbol{q}}_j \right] \mathrm{d}\boldsymbol{\xi} \\
&= \int_{K} \left[\left(\frac{\mathrm{D}\boldsymbol{x}_K}{\mathrm{D}\boldsymbol{\xi}} \right)^{-T} \hat{\boldsymbol{E}} \circ \boldsymbol{x}_K^{-1} \right] \cdot \left[\frac{1}{J_K} \left(\frac{\mathrm{D}\boldsymbol{x}_K}{\mathrm{D}\boldsymbol{\xi}} \right) \hat{\boldsymbol{q}}_j \circ \boldsymbol{x}_K^{-1} \right] \mathrm{d}\boldsymbol{x} \\
&= \int_{K} \boldsymbol{E} \cdot \boldsymbol{q}_j \, \mathrm{d}\boldsymbol{x} \\
&= L_{b,j}(\boldsymbol{\Phi}_K(\hat{\boldsymbol{E}}))
\end{aligned}$$

for all $j = 1, 2, \ldots, (k-1)k$. ■

7.5.4 Transformation of weak forms to the reference domain

In this paragraph we transform the integrals involved in the weak formulation (7.77) of the model problem to the reference domain, as required by the element-by-element assembling procedure. We focus on triangular elements, but we will point out where the quadrilateral case differs. Recall from Paragraph 7.4.3 the weak formulation: Find $\boldsymbol{E} \in V$ such that

$$a(\boldsymbol{E}, \boldsymbol{F}) = l(\boldsymbol{F}) \qquad \text{for all } \boldsymbol{F} \in V, \tag{7.129}$$

where the sesquilinear form $a(\cdot, \cdot)$ is defined on $V \times V$ by

$$a(\boldsymbol{e}, \boldsymbol{f}) = (\mu_r^{-1} \nabla \times \boldsymbol{e}, \nabla \times \boldsymbol{f})_\Omega - k^2 (\epsilon_r \boldsymbol{e}, \boldsymbol{f})_\Omega - jk(\lambda \boldsymbol{e}_T, \boldsymbol{f}_T)_{\Gamma_I} \tag{7.130}$$

and the linear form $l(\cdot)$ is defined on V as

$$l(\boldsymbol{f}) = (\boldsymbol{\Phi}, \boldsymbol{f})_\Omega + (\boldsymbol{g}, \boldsymbol{f}_T)_{\Gamma_I}. \tag{7.131}$$

The Hilbert space V was defined in (7.76),

$$V = \{\boldsymbol{E} \in \boldsymbol{H}(\mathrm{curl}, \Omega);\ \boldsymbol{\nu} \times \boldsymbol{E} = \mathbf{0} \text{ on } \Gamma_P\}.$$

Let $K \in \mathcal{T}_{h,p}$ be a triangular mesh element and $\boldsymbol{x}_K : K_t \to K$ the corresponding affine reference map. Recall that in the quadrilateral case the reference map is biaffine, and the element K must be convex so that the map is a bijection. The determinant of the Jacobi matrix of the reference map $\boldsymbol{x}_K(\boldsymbol{\xi})$ is denoted by $J_K(\boldsymbol{\xi})$. Without loss of generality, we assume that $J_K(\boldsymbol{\xi}) > 0$ in K_t. Moreover, J_K is constant in the triangular case. When the field $\hat{\boldsymbol{E}}$ transforms from K_t to K according to the rule (7.127), its curl changes to

$$\nabla \times \boldsymbol{e}(\boldsymbol{x}) = J_K^{-1}(\boldsymbol{\xi})\hat{\nabla} \times \hat{\boldsymbol{e}}(\boldsymbol{\xi}), \qquad \boldsymbol{x} = \boldsymbol{x}_K(\boldsymbol{\xi})$$

(see, e.g., [45, 83] and [111]). For clarity, we use the symbol $\hat{\nabla}$ for the nabla operator in the reference coordinates $\boldsymbol{\xi}$ on K_t.

The first part of the form (7.130), restricted to the element K, transforms as

$$\begin{aligned}(\mu_r^{-1}\nabla \times \boldsymbol{e}, \nabla \times \boldsymbol{f})_K &= \int_K \mu_r^{-1}(\boldsymbol{x})[\nabla \times \boldsymbol{e}(\boldsymbol{x})] \cdot [\nabla \times \overline{\boldsymbol{f}}(\boldsymbol{x})]\,\mathrm{d}\boldsymbol{x} \\ &= \int_{K_t} \mu_r^{-1}(\boldsymbol{x}_K(\boldsymbol{\xi}))\left[J_K^{-1}\hat{\nabla} \times \hat{\boldsymbol{e}}(\boldsymbol{\xi})\right] \cdot \left[J_K^{-1}\hat{\nabla} \times \overline{\hat{\boldsymbol{f}}}(\boldsymbol{\xi})\right] J_K\,\mathrm{d}\boldsymbol{\xi} \\ &= \int_{K_t} \mu_r^{-1}(\boldsymbol{x}_K(\boldsymbol{\xi}))J_K^{-1}\left[\hat{\nabla} \times \hat{\boldsymbol{e}}(\boldsymbol{\xi})\right] \cdot \left[\hat{\nabla} \times \overline{\hat{\boldsymbol{f}}}(\boldsymbol{\xi})\right]\,\mathrm{d}\boldsymbol{\xi}.\end{aligned}$$

Although the existence and uniqueness analysis in Section 7.4 was restricted to piecewise-isotropic materials, generally the relative permittivity ϵ_r is a tensor,

$$\epsilon_r(\boldsymbol{x}) = \begin{pmatrix} \epsilon_{r,11}(\boldsymbol{x}) & \epsilon_{r,12}(\boldsymbol{x}) \\ \epsilon_{r,21}(\boldsymbol{x}) & \epsilon_{r,22}(\boldsymbol{x}) \end{pmatrix}.$$

For the second term on the right-hand side of (7.130) we obtain

$$\begin{aligned}k^2(\epsilon_r\boldsymbol{e}, \boldsymbol{f})_K &= k^2\int_K [\epsilon_r(\boldsymbol{x})\boldsymbol{e}(\boldsymbol{x})] \cdot \overline{\boldsymbol{f}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \\ &= k^2\int_{K_t} J_K\left[\epsilon_r(\boldsymbol{x}_K(\boldsymbol{\xi}))\left(\frac{\mathrm{D}\boldsymbol{x}_K}{\mathrm{D}\boldsymbol{\xi}}\right)^{-T}\hat{\boldsymbol{e}}(\boldsymbol{\xi})\right] \cdot \left[\left(\frac{\mathrm{D}\boldsymbol{x}_K}{\mathrm{D}\boldsymbol{\xi}}\right)^{-T}\overline{\hat{\boldsymbol{f}}}(\boldsymbol{\xi})\right]\,\mathrm{d}\boldsymbol{\xi}.\end{aligned}$$

The last volume integral to be transformed is the first term on the right-hand side of (7.131),

$$(\boldsymbol{\Phi}, \boldsymbol{f})_K = \int_K \boldsymbol{\Phi}(\boldsymbol{x}) \cdot \overline{\boldsymbol{f}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{K_t} J_K\boldsymbol{\Phi}(\boldsymbol{x}_K(\boldsymbol{\xi})) \cdot \left[\left(\frac{\mathrm{D}\boldsymbol{x}_K}{\mathrm{D}\boldsymbol{\xi}}\right)^{-T}\overline{\hat{\boldsymbol{f}}}(\boldsymbol{\xi})\right]\,\mathrm{d}\boldsymbol{\xi}.$$

Next assume an edge a of the element K that lies on the impedance boundary Γ_I. Let e_i be the edge of K_t such that $a = \boldsymbol{x}_K(e_i)$. It follows from (7.128) that

$$\boldsymbol{t}_a = o_{K,i}\frac{|e_i|}{|a|}\left(\frac{\mathrm{D}\boldsymbol{x}_K}{\mathrm{D}\boldsymbol{\xi}}\right)\hat{\boldsymbol{t}}_i.$$

As before, the symbol $\hat{\boldsymbol{t}}_i$ stands for the unit tangential vector to the edge e_i, oriented as shown in Figure 7.7. It follows from (7.127) and (7.128) that the tangential components are transformed via the relation

$$\boldsymbol{e}(\boldsymbol{x}_K(\boldsymbol{\xi})) \cdot \boldsymbol{t}_a(\boldsymbol{x}_K(\boldsymbol{\xi})) = o_{K,i} \frac{|e_i|}{|a|} \hat{\boldsymbol{e}}(\boldsymbol{\xi}) \cdot \hat{\boldsymbol{t}}_i(\boldsymbol{\xi}). \tag{7.132}$$

Using this identity, the reader can transform all boundary integrals involved in the forms (7.130) and (7.131) easily. The only point where one has to be careful is to keep the global orientation of the boundary edges on Γ_I consistent with the outer normal vector to Γ_I, $\boldsymbol{\nu} \times \boldsymbol{f} = \nu_1 f_2 - \nu_2 f_1 = \boldsymbol{t} \cdot \boldsymbol{f}$ on Γ_I. In other words, the direction of the boundary edges lying on Γ_I cannot be chosen arbitrarily, since the tangential vector is determined by the outer normal vector, $\boldsymbol{t} = (\nu_1, -\nu_2)^T$. In the next paragraph let us briefly mention the interpolation on the nodal edge elements.

7.5.5 Interpolation on edge elements

The interpolation on the edge elements of Nédélec exactly fits into the general framework of interpolation on nodal elements (see Paragraph 3.3.1). It is sufficient to discuss the situation on the reference domain K_t, since the field from an arbitrary triangular mesh element $K \in \mathcal{T}_{h,p}$ can be transformed to $\hat{K}$, and the interpolant back to K, using the relation (7.113) in both directions.

The set of degrees of freedom on the kth-degree edge element $(K_t, \hat{\boldsymbol{P}}^k, \hat{\Sigma}^k)$ contains the $3k$ linear forms (7.120),

$$\hat{L}_{e_i,j}(\hat{\boldsymbol{E}}) = \int_{e_i} \hat{\boldsymbol{E}} \cdot \hat{\boldsymbol{t}}_i \, \hat{q}_j^{(i)} \, \mathrm{d}\xi, \qquad j = 0, 1, \ldots, k-1, \tag{7.133}$$

where $\hat{q}_0^{(i)}, \hat{q}_1^{(i)}, \ldots, \hat{q}_{k-1}^{(i)}$ are the Legendre polynomials $L_0, L_1, \ldots, L_{k-1}$, transformed to the edge e_i. The $(k-1)k$ interior degrees of freedom (7.121) have the form

$$\hat{L}_{b,j}(\hat{\boldsymbol{E}}) = \int_{K_t} \hat{\boldsymbol{E}} \cdot \hat{\boldsymbol{q}}_j \, \mathrm{d}\boldsymbol{\xi}, \tag{7.134}$$

where $\hat{\boldsymbol{q}}_j$, $j = 1, 2, \ldots, (k-1)k$, is a suitable basis of the space $[P^{k-2}(K_t)]^2$. The choice of this basis influences the conditioning of the discrete problem, but we will not discuss this issue at the moment.

As we said before, the unique set of nodal shape functions can be constructed routinely by inverting the generalized Vandermonde matrix (see Paragraph 3.1.2). Explicit formulae of the lowest-order (Whitney) shape functions were introduced in (7.103). For simplicity, by $\hat{\theta}_{e_i,j}$ and $\hat{\theta}_{b,j}$ we denote the nodal shape functions corresponding to the edge degrees of freedom (7.133) and the bubble degrees of freedom (7.134), respectively.

Let $\hat{\boldsymbol{E}} \in \boldsymbol{H}(\mathrm{curl}, K_t)$ for which all the degrees of freedom (7.133), (7.134) are defined. Then the local nodal interpolant is given by (3.28),

$$\mathcal{I}_{K_t}(\hat{\boldsymbol{E}}) = \sum_{i=1}^{3} \sum_{j=0}^{k-1} \hat{L}_{e_i,j}(\hat{\boldsymbol{E}}) \hat{\theta}_{e_i,j}(\boldsymbol{\xi}) + \sum_{j=1}^{(k-1)k} \hat{L}_{b,j}(\hat{\boldsymbol{E}}) \hat{\theta}_{b,j}(\boldsymbol{\xi}). \tag{7.135}$$

The edge degrees of freedom (7.133) are not defined for all functions from the space $\boldsymbol{H}(\mathrm{curl}, \Omega_h)$. The question of finding largest function spaces where all the linear forms are defined is discussed, e.g., in [39, 40] and [83].

7.5.6 Conformity of edge elements to the space $\boldsymbol{H}(\mathrm{curl})$

Let $\mathcal{T}_{h,p} = \{K_1, K_2, \ldots, K_M\}$ be a finite element mesh over a polygonal domain $\Omega_h \subset \mathbb{R}^2$, consisting of M Nédélec edge elements of the same polynomial degree $k \geq 1$. In order to verify the conformity to the space $V = \boldsymbol{H}(\mathrm{curl}, \Omega_h)$, one performs the following steps:

1. Consider an arbitrary function $g \in V$ such that all degrees of freedom $L_{e_i,j}(g)$ and $L_{b,j}$ on all elements K_m, $1 \leq m \leq M$, are defined.
2. Construct the local interpolant $\mathcal{I}_{K_m}$ for each element K_m using (7.135).
3. Construct the global interpolant $\mathcal{I}$ by "glueing together" the local interpolants $\mathcal{I}_{K_m}$, $1 \leq m \leq M$ (this operation is described exactly in Definition 3.6).
4. Check whether the piecewise-polynomial function $\mathcal{I}$ lies in the space V.

We know from Paragraph 7.5.1 that the conformity requirement of the space $\boldsymbol{H}(\mathrm{curl}, \Omega_h)$ is the continuity of the tangential component of the global interpolant $\mathcal{I}$ on all element interfaces. The desired conformity result is based on the properties of the nodal shape functions defined on the reference domain:

Lemma 7.16 *Let $(K_t, \hat{\boldsymbol{P}}^k, \hat{\Sigma}^k)$ be the Nédélec edge element of the degree $k \geq 1$ on the reference domain K_t, equipped with the polynomial space (7.119) and the edge and bubble degrees of freedom (7.120) and (7.121), respectively. Let $\hat{\theta}_{e_i,j}$, $j = 0, 1, \ldots, k-1$, and $\hat{\theta}_{b,j}$, $j = 1, 2, \ldots, (k-1)k$ be the unique set of nodal shape functions satisfying the delta property*

$$
\begin{aligned}
\hat{L}_{e_i,r}(\hat{\theta}_{e_i,s}) &= \delta_{rs} && \textit{for all } 0 \leq r, s \leq k-1, && (7.136)\\
\hat{L}_{e_i,r}(\hat{\theta}_{b,s}) &= 0 && \textit{for all } 0 \leq r \leq k-1,\ 1 \leq s \leq (k-1)k, && (7.137)\\
\hat{L}_{b,r}(\hat{\theta}_{e_i,s}) &= 0 && \textit{for all } 1 \leq r \leq (k-1)k,\ 0 \leq s \leq k-1, &&\\
\hat{L}_{b,r}(\hat{\theta}_{b,s}) &= \delta_{rs} && \textit{for all } 1 \leq r, s \leq (k-1)k. &&
\end{aligned}
$$

Then for every $i = 1, 2, 3$ and $j = 0, 1, \ldots, k-1$ the trace of the tangential component $\hat{\theta}_{e_i,j} \cdot \hat{\boldsymbol{t}}_i$ of the edge function $\hat{\theta}_{e_i,j}$ to the edge e_i is the Legendre polynomial L_j, transformed to the edge e_i. The trace of the tangential component of $\hat{\theta}_{e_i,j}$ vanishes on the remaining two edges of K_t. For every $j = 1, 2, \ldots, (k-1)k$ the trace of the tangential component of the bubble function $\hat{\theta}_{b,j}$ vanishes on the whole boundary of K_t.

Proof: It follows from the definition of the edge degrees of freedom (7.120) and the delta property (7.136) that

$$\delta_{rs} = \hat{L}_{e_i,r}(\hat{\boldsymbol{E}}) = \int_{e_i} \hat{\theta}_{e_i,s} \cdot \hat{\boldsymbol{t}}_i \, \hat{q}_r^{(i)} \, \mathrm{d}\xi \qquad \text{for all } 0 \leq r, s \leq k-1.$$

Since the transformed Legendre polynomials $\hat{q}_r^{(i)}$ form an L^2-orthonormal system in the space $P^{k-1}(e_i)$, for every $0 \leq s \leq k-1$ the trace of the function $\hat{\theta}_{e_i,s} \cdot \hat{\boldsymbol{t}}_i \in P^{k-1}(e_i)$ to the edge e_i necessarily is the transformed Legendre polynomial $\hat{q}_s^{(i)}$. By the same token it follows from (7.137) that the traces of the bubble functions $\hat{\theta}_{b,s}$, $1 \leq r \leq (k-1)k$ vanish on all edges e_i, $i = 1, 2, 3$. Similarly, the trace of the tangential component of $\hat{\theta}_{e_i,s}$ vanishes on the remaining two edges of K_t. ■

It is easy to see that these properties translate to a general mesh element $K \in \mathcal{T}_{h,p}$. From here it follows that the bubble part of the local nodal interpolants (7.135) cannot influence the tangential component of the global interpolant on element interfaces. We know from Paragraph 7.5.1 that the tangential components of $\boldsymbol{H}(\mathrm{curl}, \Omega_h)$-functions are continuous on element interfaces. Consider an edge a_i shared by a pair of mesh elements, say, K_r and K_s. Taking into account that the traces of the tangential components of the local interpolants $\mathcal{I}_{K_r}$ and $\mathcal{I}_{K_s}$ to the edge a are the transformed Legendre polynomials, and that the edge degrees of freedom (7.133) perform the L^2-projection on the edge a of a function that is the same for both elements K_r and K_s, we conclude that the trace of the tangential component of the global interpolant $\mathcal{I}$ necessarily is continuous on the edge a_i. This means that the Nédélec elements conform to the space $\boldsymbol{H}(\mathrm{curl}, \Omega_h)$.

7.6 EXERCISES

Exercise 7.1 *Consider a continuous, piecewise-polynomial approximation $\varphi_{e,h,p}$ of the scalar electric potential φ_e. Show that the approximate electric field $\boldsymbol{E}_{h,p} = -\nabla\varphi_{e,h,p}$ lies in the space $\boldsymbol{H}(\mathrm{curl})$.*

Exercise 7.2 *Verify that equations (7.29) and (7.30) are invariant under the gauge transformations (7.31), (7.32).*

Exercise 7.3 *Show that (7.49) is equivalent to (7.46) in the case of piecewise-constant material parameters.*

Exercise 7.4 *Verify inequalities (7.80) and (7.81) in the proof of Theorem 7.2.*

Exercise 7.5 *Show that dim$(\hat{\boldsymbol{P}}) = 3$ for the lowest-order edge elements defined in (7.94).*

Exercise 7.6 *Check the equivalence of definitions (7.99), (7.100), and (7.103).*

Exercise 7.7 *Construct the unique nodal basis of the Whitney element $(K, \boldsymbol{P}_K, \Sigma_K)$, where $K \in \mathcal{T}_h, p$ is a triangular domain, $\hat{\boldsymbol{P}}$ the polynomial space (7.104) and Σ_K the set of degrees of freedom (7.105).*

Exercise 7.8 *Construct the unique nodal basis of the edge element $(K_t, \hat{\boldsymbol{P}}^2, \hat{\Sigma}^2)$ on the reference domain.*

Exercise 7.9 *Show that if the functions $\hat{\boldsymbol{q}}_j$, $j = 0, 1, \ldots, (k-1)k - 1$, constitute a basis in the space $[P^{k-2}(K_t)]^2$, then the functions $\boldsymbol{q}_j$, $j = 0, 1, \ldots, (k-1)k - 1$, obtained by (7.126), constitute a basis in the space $[P^{k-2}(K)]^2$.*

Exercise 7.10 *Let $K \in \mathcal{T}_{h,p}$ be a triangular element whose edge a lies on the boundary Γ_I, and let e_i be the corresponding edge of the reference domain K_t such that $\boldsymbol{x}_K(e_i) = a$. Use relation (7.132) to transform to the reference domain the corresponding part of the boundary integrals $(\lambda \boldsymbol{e}_T, \boldsymbol{f}_T)_{\Gamma_I}$ and $(\boldsymbol{g}, \boldsymbol{f}_T)_{\Gamma_I}$ which are involved in the weak formulation (7.130), (7.131).*

APPENDIX A

BASICS OF FUNCTIONAL ANALYSIS

This chapter presents elementary linear functional analysis which is needed for a first course in PDEs and modern numerical methods. Linear spaces are presented in increasing order of complexity, as shown in Figure A.1.

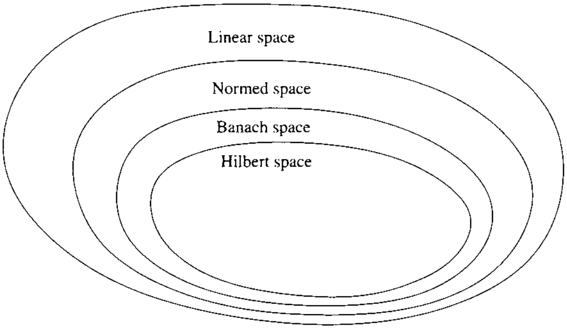


Figure A.1 Structure of linear spaces discussed in this chapter.

This text is not a traditional course in functional analysis. It assumes less at the beginning and does not address all abstract concepts of a standard functional-analytic course. On the other hand, topics needed for the study of PDEs and numerical methods, such as the L^p and Sobolev spaces, are discussed in more detail, and many examples are provided.

A.1 LINEAR SPACES

In the first section let us refresh the knowledge of linear algebra and show its application to finite-dimensional spaces of functions.

A.1.1 Real and complex linear space

A linear space V usually is defined over a general commutative body (field) $\mathcal{B}$. However, the real line $\mathbb{R}$ and the complex plane $\mathbb{C}$ are the only commutative bodies of practical importance for most applications related to partial differential equations and numerical methods.

Definition A.1 (Linear space) *Let $\mathcal{B} = \mathbb{R}$ or $\mathcal{B} = \mathbb{C}$. A nonempty abstract set V endowed with two binary operations '$+$': $V \times V \to V$ (addition) and '$\cdot$': $\mathcal{B} \times V \to V$ (multiplication by scalars) is (real or complex)* linear space *if and only if the following ten conditions are satisfied for all $a, b \in \mathcal{B}$ and u, v and $w \in V$:*

1. *$v + w$ belongs to V. (Closure of V under addition.)*

2. *$u + (v + w) = (u + v) + w$. (Associativity of addition in V.)*

3. *There exists a neutral element 0 in V, such that for all elements v in V, $v + 0 = v$. (Existence of an additive identity element in V.)*

4. *For all v in V, there exists an element w in V, such that $v + w = 0$. (Existence of additive inverses in V.)*

5. *$v + w = w + v$. (Commutativity of addition in V.)*

6. *$a \cdot v$ belongs to V. (Closure of V under scalar multiplication.)*

7. *$a \cdot (b \cdot v) = (ab) \cdot v$. (Associativity of scalar multiplication in V.)*

8. *If 1 denotes the multiplicative identity of the commutative body $\mathcal{B}$, then $1 \cdot v = v$. (Neutrality of one.)*

9. *$a \cdot (v + w) = a \cdot v + a \cdot w$. (Distributivity with respect to addition.)*

10. *$(a + b) \cdot v = a \cdot v + b \cdot v$. (Distributivity with respect to scalar addition.)*

The multiplication by scalars $a \cdot u$ usually is abbreviated to au, and $u + (-1)v$ is written shortly as $u - v$. Definition A.1 only imposes linearity to some set of abstract objects, without limiting the properties of the objects in any other way. For example, a linear space V may contain real or complex numbers, while another linear space W may consist of real or complex vectors, matrices, infinite real sequences, functions, etc. In what follows, by space we always mean linear space. The type of objects contained in a space always will be clear from the context. Most of the time we shall simply say "the set V is a linear space" when the binary operations $'+'$ and $'\cdot'$ are clear from the context.

Definition A.1 says nothing about the size of the objects lying in a linear space. The notion of size will first be introduced in Section A.2 in the context of normed spaces.

A.1.2 Checking whether a set is a linear space

Frequently one needs to decide whether a set V is or is not a linear space. Before verifying all properties listed in Definition A.1, it is useful to ask the following two simpler questions:

- Does the set V contain a zero element?
- Is the set V closed under linear combination?

Negative answer to any of them prevents V from being a linear space. Try to verify the following assertions:

■ **EXAMPLE A.1**

1. The set $\mathbb{R}$ of all real numbers is a linear space.
2. The set $\mathbb{R}^+$ of all positive real numbers is not a linear space.
3. The set of all natural numbers is not a linear space.
4. Let n be a natural number. The set $\mathbb{R}^n$ of all real vectors with n components (n-dimensional Euclidean space) is a linear space.
5. The set $\mathcal{V}_0^n$ of all real vectors in $\mathbb{R}^n$ with zero average of entries is a linear space.
6. The set $\mathcal{V}_1^n$ of all real vectors in $\mathbb{R}^n$ whose average of entries equals one is not a linear space.
7. The set $\mathcal{V}_{00}^n$ of all real vectors in $\mathbb{R}^n$ whose both the first and the last entries are zero is a linear space.
8. Let m and n be natural numbers. The set $\mathcal{M}^{n\times m}$ of all real $n \times m$ matrices is a linear space.
9. The set $\mathcal{M}_0^{n\times n}$ of all real $n \times n$ matrices whose diagonal only contains zeros is a linear space.
10. The set $\mathcal{M}_2^{n\times n}$ of all real $n \times n$ matrices whose diagonal only contains the number 2 is not a linear space.
11. The set $\mathcal{M}_{00}^{n\times n}$ of all real $n \times n$ matrices whose sum of all entries is zero is a linear space.
12. The set $F(a, b)$ of all real-valued functions defined in a bounded real interval (a, b) is a linear space.
13. The set $F^-(a, b)$ of all real-valued functions which are negative in (a, b) is not a linear space.
14. The set $F_0(a, b)$ of all real-valued integrable functions whose integral mean value in (a, b) is zero is a linear space.
15. The set $F_1(a, b)$ of all real-valued integrable functions whose integral mean value in (a, b) is one is not a linear space.
16. The set $\mathcal{O}(a, b)$ containing the zero function only is a linear space.

17. The set $C(a, b)$ of all real-valued functions continuous in (a, b) is a linear space.

18. The set $D(a, b)$ of all real-valued functions containing at least one discontinuity in (a, b) is not a linear space.

19. In the closed interval $[a, b]$, the set $C_{1,b} = \{u \in C([a, b]);\ u(b) = 1\}$ is not a linear space (see Figure A.2).

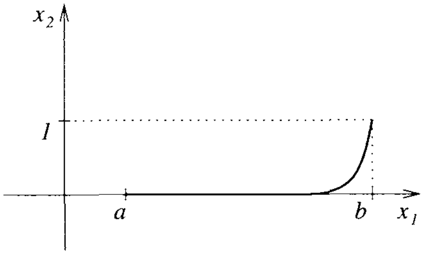


Figure A.2 The set $C_{1,b}$ does not contain the zero function, therefore it cannot be a linear space.

20. The set $C_{0,a,b}$ of all real-valued continuous functions which vanish at both endpoints of $[a, b]$ is a linear space.

21. The set $C^k(a, b)$ of all real-valued functions in (a, b) which are k-times continuously differentiable is a linear space.

22. The set $P^k(a, b)$ of all polynomials of the degree k or lower in (a, b), is a linear space.

23. The set $P^k(a, b) \setminus P^{k-1}(a, b)$ of all polynomials of degree exactly k in (a, b) is not a linear space.

24. The set of all infinite real sequences

$$\mathcal{S} = \{\{x_i\}_{i=1}^{\infty}\},$$

endowed with the binary operations

$$ax = \{ax_i\}_{i=1}^{\infty}, \qquad x + y = \{x_i + y_i\}_{i=1}^{\infty},$$

is a linear space.

25. The set $\mathcal{S}_0$ of all real sequences whose eleventh entry is zero is a linear space.

26. The set $\mathcal{S}_1$ of all real sequences whose first entry is one is not a linear space.

27. The set $\mathcal{S}_{50}$ of all real sequences such that the sum of the first fifty entries is zero is a linear space.

A.1.3 Intersection and union of subspaces

Next let us define the subspace of a linear space and introduce basic operations with subspaces, such as their intersection, union, sum, and direct sum.

Definition A.2 (Subspace of a linear space) *Let $\mathcal{B} = \mathbb{R}$ or $\mathcal{B} = \mathbb{C}$ and V a (real or complex) linear space. A nonempty subset $W \subset V$ is a* subspace *of V if*

1. $u, v \in W \Rightarrow u + v \in W$,

2. $a \in \mathcal{B}$, $u \in W \Rightarrow au \in W$,

i.e., when W is a linear space itself.

■ **EXAMPLE A.2 (Subspaces)**

1. Let $V = \mathbb{R}^2$ be the two-dimensional Euclidean space and $(0,0)^T \neq w \in V$. The space

$$W = \{\alpha w;\ \alpha \in \mathbb{R}\},$$

which is a line passing through the origin, is a subspace of V (Figure A.3).

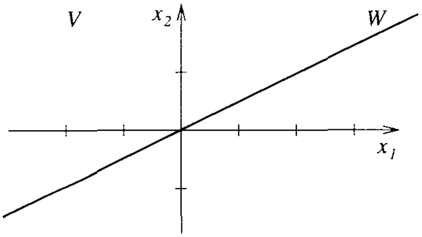


Figure A.3 Subspace W corresponding to the vector $w = (2,1)^T$.

2. Also for $w = (0,0)^T$, the space $W = \{\alpha w;\ \alpha \in \mathbb{R}\} = \{(0,0)^T\}$ is a (trivial) subspace of V.

3. Let $V = \mathbb{R}^3$ and u, v be a pair of nonzero vectors in V which do not lie on the same line. Then the space

$$W_1 = \{\alpha u + \beta v;\ \alpha, \beta \in \mathbb{R}\},$$

which is a plane passing through the origin, is a subspace of V. The space

$$W_2 = \{\alpha u;\ \alpha \in \mathbb{R}\},$$

is a subspace of V and also a subspace of W_1.

4. The space $W = \mathcal{D}^{n\times n}$ of all diagonal $n \times n$ matrices (where the zero matrix also is considered diagonal) is a subspace of the space $V = \mathcal{M}^{n\times n}$ of all $n \times n$ matrices.

5. Consider the space $V = P^n(a, b)$ of polynomials of degree less or equal to n in some interval $(a, b) \subset \mathbb{R}$. For any $0 \leq m \leq n$ the polynomial space $W = P^m(a, b)$ is a subspace of V.

6. The space $W = C^n(a, b)$, $n \geq 0$ of n-times continuously differentiable functions is a subspace of $V = C(a, b) = C^0(a, b)$.

When checking whether a subset W of a linear space V is a subspace of V, a good first question to ask is whether W contains the zero entry. If the answer is negative, then W cannot be a linear space. Otherwise we need to verify the above two properties of subspaces. The intersection of subspaces of a linear space always is a linear space:

Lemma A.1 *Every intersection $W = W_1 \cap W_2 \cap \ldots \cap W_k$, $k \geq 2$, of subspaces W_i of a linear space V is a linear space.*

Proof: The zero element $0 \in V$ lies in all subspaces $W_1, W_2, \ldots, W_k$ and therefore also in W. Every pair of elements $u, v \in W$ lies in all linear spaces $W_1, W_2, \ldots, W_k$. Therefore also $u + v$ lies in all linear spaces $W_1, W_2, \ldots, W_k$ and consequently in W. Implication 2. of Definition A.2 can be verified similarly. ■

■ EXAMPLE A.3 (Intersection of subspaces)

1. Let $V = \mathbb{R}^2$ and w_1, w_2 a pair of nonzero vectors that do not lie on the same line. Figure A.4 shows that the intersection of the spaces

$$W_1 = \{\alpha w_1;\ \alpha \in \mathbb{R}\}$$

and

$$W_2 = \{\alpha w_2;\ \alpha \in \mathbb{R}\}$$

is the trivial linear space $W = \{0\}$.

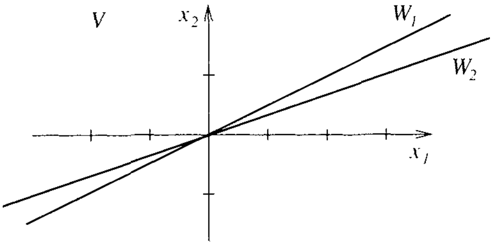


Figure A.4 Intersection of subspaces W_1 and W_2 given by the vectors $w_1 = (2, 1)^T$ and $w_2 = (3, 1)^T$, respectively.

2. Let $W_1 = \mathcal{S}^{n\times n}$ and $W_2 = \mathcal{A}^{n\times n}$ be the spaces of symmetric and antisymmetric real $n \times n$ matrices, respectively. Both W_1 and W_2 are subspaces of $V = \mathcal{M}^{n\times n}$. The intersection $W = W_1 \cap W_2 = \{0\}$, where 0 stands for the zero matrix. Indeed W is a linear space and subspace of V.

3. Consider the polynomial spaces $V = P^n(a,b)$, $W_1 = P^r(a,b)$ and $W_2 = P^s(a,b)$, where $0 \le r \le s \le n$ and $(a,b) \subset \mathbb{R}$. Then both W_1 and W_2 are subspaces of V, and so is their intersection, which is W_1.

One has to be more careful with the union of subspaces, since it is not necessarily a linear space:

■ EXAMPLE A.4 (Union of subspaces)

1. Consider the linear space $V = \mathbb{R}^2$ and a pair of vectors $w_1 = (-2,1)^T, w_2 = (3,1)^T$. Define the subspaces $W_1 = \{\alpha w_1;\ \alpha \in \mathbb{R}\}$ and $W_2 = \{\alpha w_2;\ \alpha \in \mathbb{R}\}$. By W denote the union $W_1 \cup W_2$. Evidently both the vectors w_1 and w_2 lie in W but $w = w_1 + w_2 \notin W$ (Figure A.5). Therefore $W = W_1 \cup W_2$ is not a linear space.

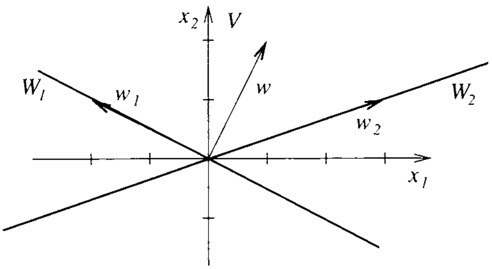


Figure A.5 Union of subspaces W_1 and W_2 given by the vectors $w_1 = (-2,1)^T$ and $w_2 = (3,1)^T$.

2. Consider the linear space $V = C(a,b)$, $a, b \in \mathbb{R}$, $a < b$ of continuous functions defined in the interval (a,b). Choose $c, d \in (a,b)$, $c \neq d$. Define linear spaces

$$W_1 = \{f \in V;\ f(c) = 0\},\quad W_2 = \{g \in V;\ g(d) = 0\}.$$

Obviously $W_1, W_2 \subset V$. The union of W_1 and W_2 is defined as

$$W = W_1 \cup W_2 = \{h \in V;\ h(c) = 0 \text{ or } h(d) = 0\}.$$

Choose now some functions $f_1, f_2 \in V$ such that $f_1(c) = 0 \neq f_1(d)$ and $g_1(c) \neq 0 = g_1(d)$. Then $f_1, g_1 \in W$ but $f_1 + g_1 \notin W$. Therefore W is not a linear space.

Proposition A.1 *Let $W_2 \subset W \subset V$ be linear spaces. In this case the union $W \cup W_2 = W$, which is a subspace of V.*

A.1.4 Linear combination and linear span

Definition A.3 (Linear combination) *Let V be a real or complex linear space, $v_1, v_2, \ldots, v_k$ elements of V, and $a_1, a_2, \ldots, a_k$ real or complex coefficients. The element*

$$v = \sum_{i=1}^{k} a_i v_i$$

of V is said to be a linear combination *of the elements $v_1, v_2, \ldots, v_k$ with the coefficients $a_1, a_2, \ldots, a_k$.*

Definition A.4 (Linear span) *Let S be a subset of a linear space V (not necessarily a subspace of V). The* linear span *of S, usually denoted by $[S]$ or* span*(S), is defined to be the intersection of all subspaces of the space V that contain the set S.*

Recall that the zero element always lies in a linear span since it is contained in every linear space.

■ EXAMPLE A.5 (Linear span)

Consider an interval $(a, b) \subset \mathbb{R}$, the linear space $V = C(a, b)$ of continuous functions in (a, b), and the set $S = \{1, x, x^2\} \subset V$. The linear span of S,

$$[S] = \{a_0 + a_1 x + a_2 x^2;\ a_0, a_1, a_2 \in \mathbb{R}\},$$

is nothing else than the space of quadratic polynomials, i.e., $[S] = P^2(a, b)$.

Lemma A.2 *Let S be a subset of a linear space V. Then the linear span $[S]$ is the smallest subspace of V containing the set S with respect to inclusion. In other words, there is no subspace W of V such that $W \subset [S]$, $W \neq [S]$, and $S \subset W$.*

Proof: Defined as intersection of subspaces of V, the linear span $[S]$ is a subspace of V (Lemma A.1). Definition A.4 further says that $[S]$ is subset of every subspace $W \subset V$ such that $S \subset W$. ■

Lemma A.3 *Let S be a subset of a linear space V. The linear span $[S]$ is identical with the set of all linear combinations of elements of S.*

Proof: By W let us denote the set of all linear combinations of elements of S. The zero element (trivial linear combination) lies in W. For all $u, v \in W$ the sum $u + v$ (another linear combination) lies in W. Similarly, for all $a \in \mathcal{B}$ and $u \in W$ the product au lies in W. Hence, according to Definition A.2, W is a subspace of V. Obviously $S \subset W$. According to Lemma A.2, $[S] \subset W$. Conversely, it is easy to see that W is subset of every subspace $Z \subset V$ such that $S \subset Z$. Therefore $W \subset [S]$. ■

Lemma A.4 *Let W, W_2 be subsets of a linear space V (not necessarily subspaces). Then*

1. *$W \subset [W]$,*
2. *if $W \subset W_2$, then $[W]$ is a subspace of $[W_2]$,*
3. *$[[W]] = [W]$,*
4. *if $W = \emptyset$, then $[W] = \{0\}$ (not an empty set!),*
5. *if $W \subset W_2 \subset [W]$, then $[W] = [W_2]$.*

Proof: All the above properties follow easily from Definition A.4. ■

A.1.5 Sum and direct sum of subspaces

Definition A.5 (Sum and direct sum) *Let W_1 and W_2 be subspaces of a linear space V. By the* sum $W_1 + W_2$ *we mean the linear span of the union of W_1 and W_2, i.e., $[W_1 \cup W_2]$.*
We say that V is a direct sum *of its subspaces W_1 and W_2 (written as $V = W_1 \oplus W_2$) if*

1. $V = W_1 + W_2$,

2. $W_1 \cap W_2 = \{0\}$.

If $V = W_1 \oplus W_2$ then W_2 is direct complement *of W_1 and vice versa. We also say that $V = W_1 \oplus W_2$ is* direct decomposition *of V into subspaces W_1, W_2.*

Lemma A.5 *A linear space V is a direct sum of its subspaces W_1, W_2 if and only if every element $v \in V$ can be expressed uniquely as $v = w_1 + w_2$, where $w_1 \in W_1$ and $w_2 \in W_2$.*

Proof: If $V = W_1 \oplus W_2$, it follows from property *1.* of Definition A.5 that every element $v \in V$ can be expressed as $v = w_1 + w_2$ with $w_1 \in W_1$ and $w_2 \in W_2$. Assume that moreover $v = v_1 + v_2$, where $v_1 \in W_1$ and $v_2 \in W_2$. Then from $w_1 + w_2 = v_1 + v_2$ it follows that the element $w_1 - v_1 = v_2 - w_2$ lies in the intersection $W_1 \cap W_2$. Property *2.* of Definition A.5 implies that $v_1 = w_1$ and $v_2 = w_2$, and thus the decomposition of the element v is unique.

Now assume that every element $v \in V$ can be decomposed uniquely into a sum $v = w_1 + w_2$ with $w_1 \in W_1$ and $w_2 \in W_2$. This means that $V = W_1 + W_2$. It remains to be verified that $W_1 \cap W_2 = \{0\}$. Every element $u \in W_1 \cap W_2$ can be written in the form $u = u + 0 = 0 + u$. Uniqueness of the decomposition yields that $u = 0$. ■

Both Definition A.5 and Lemma A.5 can be naturally extended to a finite and countable infinite number of subspaces.

■ EXAMPLE A.6 (Sums and direct sums)

1. Consider the linear space $V = \mathbb{R}^2$ and a pair of vectors $v_1 = (-1, 1)^T$, $v_2 = (-1, -1)^T$. Define the subspaces $W_1 = \{\alpha v_1;\ \alpha \in \mathbb{R}\}$ and $W_2 = \{\alpha v_2;\ \alpha \in \mathbb{R}\}$. It is $V = W_1 + W_2$ and moreover $W_1 \cap W_2 = \{(0, 0)^T\}$, therefore $V = W_1 \oplus W_2$. According to Lemma A.5 this is equivalent to the fact that every vector $v \in V$ can be written uniquely as $v = w_1 + w_2$, where $w_1 \in W_1$ and $w_2 \in W_2$ (Figure A.6).

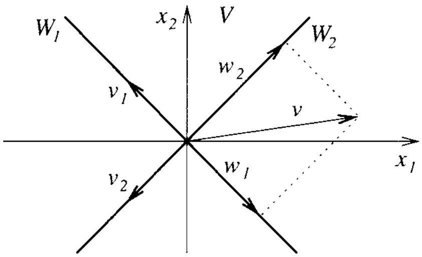


Figure A.6 Unique decomposition of a vector in a direct sum of subspaces.

2. Consider the space $V = \mathcal{M}^{n\times n}$ of real $n \times n$ matrices and its subspaces
$$\begin{aligned} V_1 &= \mathcal{M}_L^{n\times n} = \{M \in V;\ m_{ij} = 0 \text{ if } j \geq i\}, \\ V_2 &= \mathcal{M}_D^{n\times n} = \{M \in V;\ m_{ij} = 0 \text{ if } j \neq i\}, \\ V_3 &= \mathcal{M}_U^{n\times n} = \{M \in V;\ m_{ij} = 0 \text{ if } j \leq i\} \end{aligned}$$
of lower-diagonal, diagonal, and upper-diagonal matrices, respectively. Clearly, it is $V = V_1 + V_2 + V_3$. Since, in addition, $V_1 \cap V_2 = \{0\}$, $V_1 \cap V_3 = \{0\}$, and $V_2 \cap V_3 = \{0\}$, it is $V = V_1 \oplus V_2 \oplus V_3$. Accordingly, every matrix $M \in V$ can be decomposed uniquely into $M = M_1 + M_2 + M_3$, where $M_i \in V_i$, $i = 1, \ldots, 3$.

3. Consider an interval $(a, b) \subset \mathbb{R}$ and the space of continuous functions $V = C(a, b)$ with its subspaces $W_1 = \{w_1 \in C(a, b);\ w_1(a) = 0\}$ and $W_2 = \{w_2 \in C(a, b);\ w_2(b) = 0\}$. Clearly it is $V = W_1 + W_2$. Since
$$W_1 \cap W_2 = C_0(a, b) = \{w \in C(a, b);\ w(a) = w(b) = 0\} \neq \{0\},$$
according to Definition A.5 the space V cannot be direct sum of W_1 and W_2.

There is a one-to-one relation between direct sums and idempotent linear operators. (An operator $P : V \to V$ is said to be idempotent if $P^2 = P$.) These operators are called projections, and we will study them in more detail in Paragraph A.3.5.

A.1.6 Linear independence, basis, and dimension

Next let us introduce the notion of linear independence, basis, and dimension of a linear space.

Definition A.6 (Linear independence) *Let V be a real or complex linear space and let $v_1, v_2, \ldots, v_k \in V$. These elements are said to be* linearly dependent *if there exists a nontrivial set of real or complex coefficients $a_1, a_2, \ldots, a_k$, respectively, such that*
$$\sum_{i=0}^{k} a_i v_i = 0$$
(by nontrivial *we mean that at least one coefficient a_j is nonzero). In the opposite case the elements $v_1, v_2, \ldots, v_k$ are said to be* linearly independent. *Sometimes a subset $S \subset V$ is called linearly dependent/independent if all its elements are linearly dependent/independent.*

■ **EXAMPLE A.7 (Linear independence)**

1. In the space $V = \mathbb{R}^3$ consider three vectors $v_1 = (1, 0, 0)^T$, $v_2 = (1, 1, 1)^T$, and $v_3 = (4, -2, -2)^T$. These vectors are linearly dependent since
$$-6v_1 + 2v_2 + v_3 = (0, 0, 0)^T.$$

2. Let us decide if the functions $w_1 = x$, $w_2 = 2 - 3x$, and $w_3 = 1 + x^2$ in the space $V = P^2(-1, 1)$ are linearly independent. If

$$a_1 w_1 + a_2 w_2 + a_3 w_3 = 0,$$

then clearly $a_3 = 0$ since the square of x in w_3 cannot be eliminated by any linear combination of the functions w_1 and w_2. Thus

$$a_1 w_1 + a_2 w_2 = 0.$$

The constant 2 in w_2 cannot be eliminated by w_1, and therefore $a_2 = 0$. Therefore also $a_1 = 0$ and the functions w_1, w_2, and w_3 are linearly independent.

The following lemma gives a useful characterization of linear independence.

Lemma A.6 *Consider a subset S of a linear space V. Then*

1. *S is linearly independent if and only if none of its elements can be expressed as a linear combination of its remaining elements.*
2. *S is linearly independent if and only if the following implication holds: If $R \subset S$ and $[R] = [S]$ then $R = S$.*
3. *Let $v \in V$. If S is linearly independent and $S + \{v\}$ linearly dependent, then $v \in [S]$.*

Proof: The proof is a simple exercise using Definition A.6. ■

Definition A.7 (Basis of a linear space) *Let V be a linear space. Every linearly independent subset $S \subset V$ such that $[S] = V$ is said to be a* basis *of the space V.*

A linear space V may have many different bases: Any set of linearly independent elements of V that generate the whole space is a basis. This is illustrated in Example A.8.

■ EXAMPLE A.8 (Nonuniqueness of basis)

1. Let $V = \mathbb{R}^3$. The set

$$\mathcal{B} = \{(1,0,0)^T, (0,1,0)^T, (0,0,1)^T\}$$

is a basis of V (canonical basis). Any other set of vectors

$$\mathcal{B}_{\alpha,\beta,\gamma} = \{(\alpha,0,0)^T, (0,\beta,0)^T, (0,0,\gamma)^T\},$$

where α, β, γ are nonzero real numbers, also is a basis of V. Herewith the list is not complete, since obviously the basis vectors can have more than just one nonzero components.

2. Consider the polynomial space $V = P^1(0,1)$ and its monomial basis

$$\mathcal{B}_1 = \{1, x\}.$$

Another example of a basis in this space is, e.g.,

$$\mathcal{B}_2 = \{x, 1-x\}.$$

3. Let S be the space of all real sequences from Example A.1. The infinite set $\mathcal{B}_1 = \{r_1, r_2, r_3, \ldots\}$, where r_i is a real sequence whose entries are all zero except for the ith entry which is one, is a basis of S. Another infinite set $\mathcal{B}_2 = \{s_1, s_2, \ldots\}$, where s_i is a real sequence whose entries are all equal to one except for the ith entry which is zero, also is a basis of S.

After presenting a few concrete examples of bases, the following Theorem A.1 guarantees that every linear space has at least one.

Theorem A.1 (Existence of basis) *Every linear space V has a basis.*

Proof: Let V be a linear space and $\mathcal{E}$ the set of all its linearly independent subsets. Obviously the empty set lies in $\mathcal{E}$ thus $\mathcal{E}$ is not empty. It is easy to see that the set $\mathcal{E}$ is partially ordered by inclusion (the union of a chain of linearly independent subsets of V is again a linearly independent subset of V). Hence the Zorn's Lemma implies the existence of a maximal element S in $\mathcal{E}$. Assume that there exists an element $v \in V$ such that $u \notin S$. The maximality of S implies that $S \cup \{u\}$ is linearly dependent. According to Lemma A.6, assertion *3.*, $u \in [S]$. Thus $[S] = V$ and therefore S is a basis of the linear space V. ■

For future reference lat us introduce the notion of separable space.

Definition A.8 (Separable space) *A linear space V is called* separable *if there exists a finite or a countable infinite basis of V.*

The following Lemmas A.7 and A.8 have a technical nature, but they are useful for the definition of the dimension of a linear space.

Lemma A.7 *Let V be a linear space. Any set of n linearly independent elements u_1, u_2, ..., $u_n \in V$ cannot be expressed by linear combinations of any $n-1$ elements $v_1, v_2, \ldots,$ $v_{n-1} \in V$.*

Proof: Let us proceed by induction. Obviously the assertion is valid for $n = 1$. Assume that the assertion is valid for n and not valid for $n+1$. Thus it is possible to express

$$\begin{aligned} u_1 &= a_{1,1}v_1 + \ldots + a_{1,n}v_n \\ u_2 &= a_{2,1}v_1 + \ldots + a_{2,n}v_n \\ &\vdots \\ u_n &= a_{n,1}v_1 + \ldots + a_{n,n}v_n \\ u_{n+1} &= a_{n+1,1}v_1 + \ldots + a_{n+1,n}v_n. \end{aligned}$$

The elements $u_1, u_2, \ldots, u_{n+1}$ are linearly independent. Hence $u_{n+1} \neq 0$ and at least one of the coefficients in the last equation is nonzero. Assume for example that $a_{n+1,n} \neq 0$. Let us calculate v_n from the last equation and insert it to the first n equations. We obtain a smaller system of n equations of the form

$$\begin{aligned} u_1 + c_1 u_{n+1} &= b_{1,1}v_1 + \ldots + b_{1,n}v_n \\ u_2 + c_2 u_{n+1} &= b_{2,1}v_1 + \ldots + b_{2,n}v_n \\ &\vdots \\ u_n + c_n u_{n+1} &= b_{n,1}v_1 + \ldots + b_{n,n}v_n \end{aligned}$$

This is a contradiction with the assumption for n if we show that the n elements $u_1 + c_1u_{n+1}, u_2 + c_2u_{n+1}, \ldots, u_n + c_nu_{n+1}$ are linearly independent. Assume that they are not, i.e., that there exists a nontrivial linear combination

$$\begin{aligned} 0 &= d_1(u_1 + c_1u_{n+1}) + d_2(u_2 + c_2u_{n+1}) + \ldots + d_n(u_n + c_nu_{n+1}) \\ &= d_1u_1 + d_2u_2 + \ldots + d_nu_n + (d_1c_1 + d_2c_2 + \ldots + d_nc_n)u_{n+1}. \end{aligned}$$

Since $u_1, u_2, \ldots, u_{n+1}$ are linearly independent, necessarily $d_1 = d_2 = \ldots = d_n = 0$ which is a contradiction. ■

Lemma A.8 *All bases of a linear space V have the same cardinality.*

Proof: The assertion follows straightforward from Lemma A.7 both when the cardinality of the basis is finite and infinite. ■

Now, using the assertion of Lemma A.8, we finally can define the dimension of a linear space.

Definition A.9 (Dimension of a linear space) *Cardinality of any basis of a linear space V is said to be the* dimension *of V, denoted by dim(V).*

Before we approach linear operators, let us define expansion coefficients of elements of linear spaces in terms of a basis.

Definition A.10 (Expansion coefficients) *Let V be a linear space of dimension n and $B = \{v_1, v_2, \ldots, v_n\}$ its basis. Every element $u \in V$ can be written uniquely as*

$$u = \sum_{i=1}^{n} c_iv_i.$$

The coefficients $c_1, c_2, \ldots, c_n$ are called expansion coefficients *of u with respect to the basis B, and we write them in a vector form*

$$\langle u\rangle_B = (c_1, c_2, \ldots, c_n)^T.$$

■ EXAMPLE A.9 (Expansion coefficients)

1. The set $B_V = \{v_1, v_2, v_3\}$, where $v_1 = (1,1,0)^T$, $v_2 = (1,0,1)^T$ and $v_3 = (0,1,1)^T$, is a basis of the space $V = \mathbb{R}^3$. The expansion coefficients of the element $v = (1,2,3)^T$ are obtained from the vector equation

$$c_1v_1 + c_2v_2 + c_3v_3 = v.$$

Componentwise, this yields a system of three linear algebraic equations. After solving the system, we obtain

$$\langle v\rangle_{B_V} = (0,1,2)^T.$$

2. The set $B_W = \{w_1, w_2, w_3\}$, where $w_1 = x^2 - x$, $w_2 = x^2 - 1$ and $w_3 = x^2 + x$ is a basis in the space $P^2(-1, 1)$. The expansion coefficients of the function $w = 3x^2 + 2x + 1$ are obtained from the equation

$$c_1 w_1 + c_2 w_2 + c_3 w_3 = w.$$

Comparing the coefficients of monomials of the same degree, we obtain a system of three linear algebraic equations which yields

$$\langle w \rangle_{B_W} = (1, -2, 3)^T.$$

A.1.7 Linear operator, null space, range

In this paragraph let us introduce linear operators and mention some of their basic properties, including the one-to-one relation to matrices in finite-dimensional spaces.

Definition A.11 (Linear operator, null space, range) *Let U and V be real or complex linear spaces. A map $f : U \to V$ is said to be* linear operator *if and only if*

1. *$f(u + v) = f(u) + f(v)$ for all $u, v \in U$,*
2. *$f(au) = af(u)$ for all $u \in U$ and all coefficients a.*

The null space *$N(f)$ of the linear operator f is the set*

$$N(f) = \{u \in U;\ f(u) = 0\}.$$

The range *$R(f)$ of the operator f is defined as*

$$R(f) = \{v \in V;\ \text{there exists } u \in U \text{ such that } f(u) = v\}.$$

The linear operator f is said to be an injection *if $N(f) = \{0\}$,* surjection *if $R(f) = V$ and* bijection (one-to-one) *if it is both injection and surjection.*

Basic properties of the null space and range of linear operators are summarized in the following lemma.

Lemma A.9 *Let U and V be real or complex linear spaces and $f : U \to V$ a linear operator. Then the following holds:*

1. *$N(f)$ is a subspace of U.*
2. *$R(f)$ is a subspace of V.*
3. *If dim(U) and dim(V) are finite then*

$$\dim(U) = \dim(N(f)) + \dim(R(f)).$$

Proof: These assertions follow easily from the linearity of f. Let us begin with the first one:

$$u, v \in N(f) \Rightarrow f(\alpha u + \beta v) = \alpha f(u) + \beta f(v) = 0 \Rightarrow \alpha u + \beta v \in N(f)$$

for all coefficients α, β. Ad 2: For every $v_1, v_2 \in R(f)$ there exist $u_1, u_2 \in U$ such that $f(u_1) = v_1$ and $f(u_2) = v_2$. Then

$$\alpha v_1 + \beta v_2 = \alpha f(u_1) + \beta f(u_2) = f(\underbrace{\alpha u_1 + \beta u_2}_{\in U}),$$

and therefore $\alpha v_1 + \beta v_2 \in R(f)$.

Ad 3: Since $N(f)$ is a linear space, it has a basis $\{w_1, w_2, \ldots, w_k\}$. Let $\{v_1, v_2, \ldots, v_m\}$ be a basis of $R(f)$. For every $1 \leq j \leq m$ let $\hat{v}_j \in U$ be such that $f(\hat{v}_j) = v_j$. It is sufficient to show that the set

$$\mathcal{B}_U = \{w_1, w_2, \ldots, w_k, \hat{v}_1, \hat{v}_2, \ldots, \hat{v}_m\}$$

is a basis of U. First let us see that its elements are linearly independent: Assume a set of coefficients such that

$$\sum_{i=1}^{k} \alpha_i w_i + \sum_{j=1}^{m} \beta_j \hat{v}_j = 0. \tag{A.1}$$

Applying the linear operator f to both sides, we obtain that

$$\sum_{i=1}^{k} \alpha_i f(w_i) + \sum_{j=1}^{m} \beta_j f(\hat{v}_j) = \sum_{j=1}^{m} \beta_j v_j = 0.$$

Since the elements $v_1, v_2, \ldots, v_m$ are linearly independent, it is $\beta_1 = \beta_2 = \ldots = \beta_m = 0$. It follows from (A.1) that also $\alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$. It remains to be shown that every $u \in U$ can be represented by the elements of $\mathcal{B}_U$: Since $f(u)$ lies in $R(f)$, we can express

$$f(u) = \sum_{j=1}^{m} \gamma_j v_j.$$

The coefficients γ_j can be used to define

$$u_R = \sum_{j=1}^{m} \gamma_j \hat{v}_j \in U.$$

The proof is finished by realizing that $u - u_R \in N(f)$. ■

■ EXAMPLE A.10 (Linear operators)

1. Consider the spaces $V = \mathbb{R}^3$ and $W = \mathbb{R}^3$, along with a map $f : V \rightarrow W$ defined by

   $$f(u) = 2u \quad \text{for all } u \in V.$$

   For every $u, v \in V$ it is $f(u + v) = 2(u + v) = 2u + 2v = f(u) + f(v)$ and moreover $f(\alpha u) = 2\alpha u = \alpha 2u = \alpha f(u)$ for every $u \in V$ and $\alpha \in \mathbb{R}$. Therefore f

is a linear operator. It is easy to see that $N(f) = \{(0,0,0)^T\}$. Moreover, since for every vector $w \in W$ we can define a vector $v = w/2 \in V$ such that $f(v) = w$, it is $R(f) = W$. Therefore f is a bijection.

2. Next consider the polynomial spaces $V = P^4(-1,1)$ and $W = P^3(-1,1)$, and a map $D : V \rightarrow W$ defined by

$$D(u) = \frac{\mathrm{d}u}{\mathrm{d}x} = u' \quad \text{for all } u \in V.$$

For every $u, v \in V$ it is $D(u+v) = (u+v)' = u' + v' = D(u) + D(v)$ and moreover $D(\alpha u) = (\alpha u)' = \alpha u' = \alpha D(u)$ for every $u \in V$ and $\alpha \in \mathbb{R}$. Therefore D is a linear operator. Since $D(u) = 0$ if and only if u is constant, we have $N(D) = P^0(-1,1)$. Further, for every $w \in V$ we can find a $v \in V$ so that $w = D(v)$ (take some primitive function of w), and thus $R(D) = W$. Hence D is a surjection but not a bijection.

3. Let $V = P^2(-1,1)$, $W = \mathbb{R}$, and $A : V \rightarrow W$ be defined by

$$A(u) = \int_{-1}^{1} u(x)\,\mathrm{d}x \quad \text{for all } u \in V.$$

The linearity of A easily follows from the linearity of the integral. In this case $N(A) = \{u \in V;\ \int_{-1}^{1} u(x)\,\mathrm{d}x = 0\}$, which is the space of all quadratic polynomials with zero integral mean value in the interval $(-1,1)$. Since for any given number $w \in \mathbb{R}$ we can define a constant function $v = w/2 \in V$ such that $A(v) = w$, we have $R(V) = W$. Hence the operator A is a surjection but not a bijection.

Next let us define the linear space $\mathcal{L}(V,W)$ of all linear operators $L : V \rightarrow W$.

Definition A.12 (Space of linear operators) *Let V and W be real or complex linear spaces. Let $f, g : V \rightarrow W$ be linear operators. We define*

$$\begin{aligned} (f+g)(v) &= f(v) + g(v), \\ (af)(v) &= af(v) \end{aligned} \qquad \text{(A.2)}$$

for all $v \in V$ and all coefficients a. With these operations we can define the linear space of all linear operators from V to W and denote it by $\mathcal{L}(V,W)$.

Lemma A.10 (Inverse operator) *Let V and W be real or complex linear spaces and let $f \in \mathcal{L}(V,W)$ be a bijection. Then f is invertible and $f^{-1} \in \mathcal{L}(W,V)$ is a bijection.*

Proof: Follows easily from Definition A.11. ■

Next, Definition A.13 introduces the matrix representation of linear operators in finite-dimensional spaces. The equivalence of linear operators and matrices is established in Lemma A.11 and illustrated in Example A.11.

Definition A.13 (Matrix representation in finite-dimensional spaces) *Let V and W be finite-dimensional real or complex linear spaces of dimensions $dim(V) = m$ and $dim(W) = n$. Let $B_V = \{v_1, v_2, \ldots, v_m\}$ be a basis of V, $B_W = \{w_1, w_2, \ldots, w_n\}$ a basis of W, and $f \in \mathcal{L}(V,W)$. For every element $v_j \in B_V$ the element $f(v_j)$ lies in W, and thus we can express it in terms of the basis B_W with a unique set of coefficients $m_{1j}, m_{2j}, \ldots, m_{nj}$,*

$$f(v_j) = \sum_{i=1}^{n} m_{ij} w_i.$$

The $n \times m$ *matrix* M_f, $(M_f)_{ij} = m_{ij}$, $1 \le i \le n$, $1 \le j \le m$, *is said to be the* matrix representation *of the linear operator* f *with respect to the bases* B_V *and* B_W. *Or, we just say that* M_f *is the matrix of the linear operator* f.

In other words, the matrix M_f of a linear operator f is constructed by taking basis vectors of V, expressing their images through f by means of the basis vectors in W, and writing the sets of the corresponding expansion coefficients as columns of the matrix M_f.

Lemma A.11 *Let* V, W *be real or complex linear spaces of dimensions* m, n *and let* B_V, B_W *be their bases, respectively. Let* $f \in \mathcal{L}(V, W)$. *The* $n \times m$ *matrix* M_f *represents the linear operator* f *if and only if it holds*

$$\langle f(v) \rangle_{B_W} = M_f \langle v \rangle_{B_V} \qquad \textit{for all } v \in V. \tag{A.3}$$

Proof: First assume that M_f is the matrix of the linear operator f. For every $v \in V$ we denote $\langle v \rangle_{B_V} = (b_1, b_2, \dots, b_m)^T$ and calculate

$$f(v) = f\left(\sum_{j=1}^{m} b_j v_j\right) = \sum_{j=1}^{m} b_j f(v_j) = \sum_{j=1}^{m} b_j \sum_{i=1}^{n} m_{ij} w_i = \sum_{i=1}^{n} \left(\sum_{j=1}^{m} m_{ij} b_j\right) w_i.$$

Thus the expansion coefficients of $f(v)$ to the basis B_W are

$$\langle f(v) \rangle_{B_W} = \left(\sum_{j=1}^{m} m_{1j} b_j, \sum_{j=1}^{m} m_{2j} b_j, \dots, \sum_{j=1}^{m} m_{nj} b_j\right)^T = M_f \langle v \rangle_{B_V}.$$

Now the opposite implication. Assume some $n \times m$ matrix $\tilde{M}$ such that

$$\langle f(v) \rangle_{B_W} = \tilde{M} \langle v \rangle_{B_V} \qquad \text{for all } v \in V. \tag{A.4}$$

Let M_f be the matrix of the linear operator f from Definition A.13. By the implication that we already proved, $\langle f(v) \rangle_{B_W} = M_f \langle v \rangle_{B_V}$ for all $v \in V$. Relation (A.4) yields $\tilde{M} \langle v \rangle_{B_V} = M_f \langle v \rangle_{B_V}$ for all $v \in V$. The choice $\langle v \rangle_{B_V} = (1, 0, 0, \dots, 0)^T$ now yields that the first column of $\tilde{M}$ is identical to the first column of M_f, and so on. ■

■ EXAMPLE A.11 (Matrix representation of linear operators)

1. Let us begin with a linear operator f that rotates vectors in the space $V = \mathbb{R}^2$ counterclockwise by a given angle $\alpha \in \mathbb{R}$. Hence $W = V$.

   Using the canonical bases $B_V = \{(1, 0)^T, (0, 1)^T\} = \{v_1, v_2\}$ and $B_W = \{(1, 0)^T, (0, 1)^T\} = \{w_1, w_2\}$, and Figure A.7, it is easy to see that

$$\langle f(v_1) \rangle_{B_W} = \langle (\cos\alpha, \sin\alpha)^T \rangle_{B_W} = (\cos\alpha, \sin\alpha)^T$$

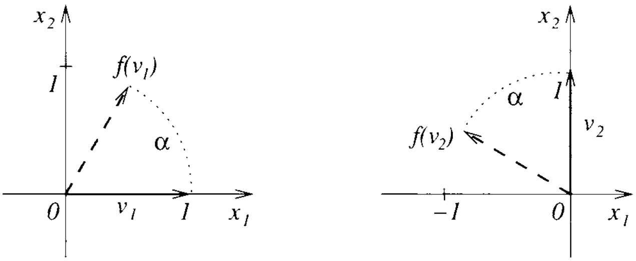


Figure A.7 Linear operator in $\mathbb{R}^2$ (rotation of vectors).

and

$$\langle f(v_2)\rangle_{B_W} = \langle(-\sin\alpha, \cos\alpha)^T\rangle_{B_W} = (-\sin\alpha, \cos\alpha)^T.$$

Therefore f is represented by the matrix

$$M_f = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}.$$

Choosing now an arbitrary two-dimensional vector $v = (a, b)^T \in V$, for $f(v)$ we have

$$f(v) = \langle f(v)\rangle_{B_W} = M_f\langle v\rangle_{B_V} = M_f v = \begin{pmatrix} a\cos\alpha - b\sin\alpha \\ a\sin\alpha + b\cos\alpha \end{pmatrix}.$$

2. Let us return to the derivative operator $D \in \mathcal{L}(V, W)$, $D(u) = u'$, from Example A.10. In the spaces $V = P^4(-1, 1)$ and $W = P^3(-1, 1)$ we consider the monomial bases $B_V = \{1, x, x^2, x^3, x^4\} = \{v_1, v_2, v_3, v_4, v_5\}$ and $B_W = \{1, x, x^2, x^3\} = \{w_1, w_2, w_3, w_4\}$, respectively. It is easy to calculate

$$\begin{aligned} \langle D(v_1)\rangle_{B_W} &= \langle 0\rangle_{B_W} = (0, 0, 0, 0)^T, \\ \langle D(v_2)\rangle_{B_W} &= \langle 1\rangle_{B_W} = (1, 0, 0, 0)^T, \\ \langle D(v_3)\rangle_{B_W} &= \langle 2x\rangle_{B_W} = (0, 2, 0, 0)^T, \\ \langle D(v_4)\rangle_{B_W} &= \langle 3x^2\rangle_{B_W} = (0, 0, 3, 0)^T, \\ \langle D(v_5)\rangle_{B_W} &= \langle 4x^3\rangle_{B_W} = (0, 0, 0, 4)^T. \end{aligned}$$

Hence the derivative operator $D(u) = u'$ is represented by the 4×5 matrix

$$M_D = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{pmatrix}.$$

This means that now we can perform derivatives of fourth-degree polynomials using a matrix-vector multiplication. Take, e.g., $v = 3 + 2x^2 - 3x^3 + x^4 \in V$, whose expansion with respect to the basis B_V is

$$\langle v\rangle_{B_V} = (3, 0, 2, -3, 1)^T.$$

For the derivative $v' = D(v)$ we have

$$\langle D(v)\rangle_{B_W} = M_D\langle v\rangle_{B_V} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} 3 \\ 0 \\ 2 \\ -3 \\ 1 \end{pmatrix} = (0, 4, -9, 4)^T.$$

Thus $v' = D(v) = 4x - 9x^2 + 4x^3$.

In what follows, by the symbol I we denote the $n \times n$ identity matrix, $I = \mathrm{diag}(1, 1, \ldots, 1)$ (the dimension n will always be clear from the context).

Definition A.14 (Nonsingular, singular, and inverse matrix) *Let M be a real or complex $n \times n$ matrix. Then M is said to be* nonsingular *if its n columns are linearly independent vectors. Otherwise M is* singular. *The matrix M^{-1} such that $MM^{-1} = M^{-1}M = I$ is said to be the* inverse *of M.*

In Definition A.14 one can use linearly independent rows instead of linearly independent columns.

Lemma A.12 *Let V and W be real or complex linear spaces of the same dimension n and let $f \in \mathcal{L}(V, W)$ be represented by a matrix M_f. The matrix M_f is nonsingular if and only if $N(f) = \{0\}$ (i.e., if f is bijection).*

Proof: Follows easily from Lemma A.11. ■

A.1.8 Composed operators and change of basis

Composition of linear operators is analogous to composition of functions in real analysis:

Lemma A.13 (Composition of linear operators) *Let U, V, and W be real or complex linear spaces, $f \in \mathcal{L}(U, V)$ and $g \in \mathcal{L}(V, W)$. Then the composition $g \circ f \in \mathcal{L}(U, W)$.*

Proof: Follows easily from Definition A.11. ■

Lemma A.14 (Representation of composed operators) *Let U, V, and W be finite-dimensional real or complex linear spaces with bases B_U, B_V, and B_W, respectively. Let $f \in \mathcal{L}(U, V)$ be represented by matrix M_f with respect to the bases B_U and B_V, and let $g \in \mathcal{L}(V, W)$ be represented by matrix M_g with respect to the bases B_V and B_W. Then the composition $g \circ f \in \mathcal{L}(U, W)$ is represented by the matrix M_gM_f.*

Proof: Follows easily from Lemma A.11. ■

Corollary A.1 (Inverse operator & inverse matrix) *Let V and W be real or complex linear spaces of dimension n and let $f \in \mathcal{L}(V, W)$, represented by matrix M_f, be a bijection. Then the matrix M_f is nonsingular and $f^{-1} \in \mathcal{L}(W, V)$ is represented by the inverse matrix M_f^{-1}.*

Definition A.15 (Transition matrix) *Let V be a real or complex linear space of dimension n and let B_1 and B_2 be its (different) bases. By* transition matrix *from the basis B_1 to the basis B_2 we mean the $n \times n$ matrix M representing the identity $\mathcal{I} \in \mathcal{L}(V, V)$,*

$$\langle v \rangle_{B_2} = M \langle v \rangle_{B_1}$$

for all $v \in V$.

■ **EXAMPLE A.12 (Change of basis)**

1. Let $V = \mathbb{R}^3$. It is our aim to construct the transition matrix from the basis B_1 to the basis B_2, where

$$B_1 = \{(1,0,0)^T, (0,1,0)^T, (0,0,1)^T\} = \{e_1, e_2, e_3\}$$

and

$$B_2 = \{(1,1,0)^T, (1,0,1)^T, (0,1,1)^T\} = \{w_1, w_2, w_3\}. \tag{A.5}$$

The two bases are depicted in Figure A.8.

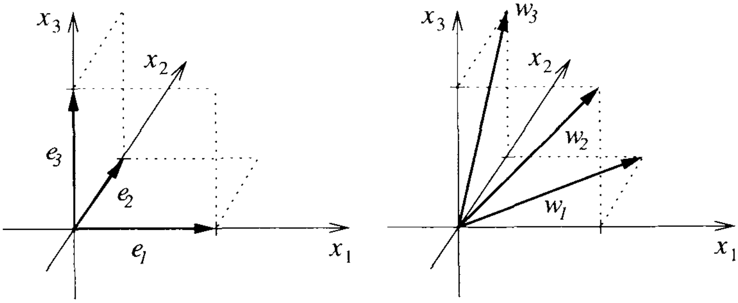


Figure A.8 Canonical basis of $\mathbb{R}^3$.

In this case it is convenient to construct the transition matrix M^{-1} from B_2 to B_1 first, since it just contains the vectors w_1, w_2, and w_3 in its columns,

$$M^{-1} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.$$

Inverting the matrix M^{-1}, we obtain the desired transition matrix M,

$$M = \begin{pmatrix} 1/2 & 1/2 & -1/2 \\ 1/2 & -1/2 & 1/2 \\ -1/2 & 1/2 & 1/2 \end{pmatrix}.$$

2. Next consider the space $V = P^3(0,1)$ equipped with the monomial basis $B_1 = \{1, x, x^2, x^3\}$ and another basis

$$B_2 = \{x, 1-x, x(1-x), x(1-x)(2x-1)/2\}$$

(which is better, e.g., for finite element approximation). Let us construct a transition matrix from B_1 to B_2. Since $1 = x + (1 - x)$ it is $\langle v_1 \rangle_{B_2} = (1, 1, 0, 0)^T$. The second element of B_1 is identical to the first element of B_2, and therefore $\langle v_2 \rangle_{B_2} = (1, 0, 0, 0)^T$. Using the identity $x^2 = x - (x - x^2)$, we obtain

$$\langle v_3 \rangle_{B_2} = (1, 0, -1, 0)^T,$$

and finally,

$$\langle v_4 \rangle_{B_2} = (1, 0, -3/2, -1)^T.$$

Hence the transition matrix M has the form

$$M = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & -\frac{3}{2} \\ 0 & 0 & 0 & -1 \end{pmatrix}.$$

A.1.9 Determinants, eigenvalues, and eigenvectors

Eigenvalues and eigenvectors (eigenfunctions) play an important role in computational engineering and science. On the practical side, they often are connected with vibrations, resonance, or related phenomena. One also needs them for theoretical purposes in numerical linear algebra, analysis of partial differential equations, numerical methods, and other fields.

Definition A.16 (Permutation and its sign) *By S_n, $n > 0$, we denote the set of all bijections of the set $\{1, 2, \ldots, n\}$ into itself. Every $P \in S_n$ is called* permutation *on the set $\{1, 2, \ldots, n\}$. For a permutation $P \in S_n$ let m be the number of pairs $(i, j) \subset \{1, 2, \ldots, n\}$, $i < j$, such that $P(i) > P(j)$. We define*

$$\mathrm{sgn}(P) = (-1)^m$$

and call P even *or* odd *if $sgn(P) = 1$ or $sgn(P) = -1$, respectively.*

Definition A.17 (Determinant) *Let M be a real or complex $n \times n$ matrix.* Determinant *of M is defined as*

$$\det(M) = \sum_{P \in S_n} \mathrm{sgn}(P) \prod_{j=1}^{n} m_{P(j)j}.$$

Lemma A.15 (Basic rules for determinants) *Let M be a real or complex $n \times n$ matrix. Then*

1. $\det(M^T) = \det(M)$.
2. *Let $\tilde{M}$ be a matrix obtained by performing a permutation $Q \in S_n$ to the rows (or columns) of M. Then* $\det(\tilde{M}) = sgn(Q) \cdot \det(M)$.
3. *The matrix M is nonsingular if and only if $det(M) \neq 0$.*

4. *Let $\tilde{M}$ be a matrix obtained by multiplying a row (or column) of M with a coefficient c. Then* $\det(\tilde{M}) = c\det(M)$.

5. $\det(cM) = c^n \det(M)$.

6. *For all* $1 \leq j \leq n$ *it is*

$$\det(M) = \sum_{i=1}^{n} (-1)^{i+j} m_{ij} \det(M_{ij})$$

where the matrix M_{ij} is obtained by leaving out ith row and jth column from the matrix M. Here, $(-1)^{i+j}\det(M_{ij})$ *is the algebraic complement of the entry m_{ij}.*

7. *Let $\tilde{M}$ be matrix of the same type as M. Then* $\det(\tilde{M}M) = \det(\tilde{M})\det(M)$.

8. *Let M be nonsingular and M^{-1} its inverse. Then det$(M^{-1}) = 1/$det(M).*

Proof: Proofs of these assertions can be found, e.g., in [75]. ■

Definition A.18 (Eigenvalue, spectrum) *Let M a real or complex $n \times n$ matrix. The* characteristic matrix *of M is the polynomial matrix $\lambda I - M$. The* characteristic polynomial *of M is the determinant of $\lambda I - M$. The* eigenvalues *of M are the roots of its characteristic polynomial. The* spectrum $\sigma(M)$ *of M is the file of all of its eigenvalues. By* (algebraic) multiplicity *of an eigenvalue one means its multiplicity as a root of the characteristic polynomial.*

■ EXAMPLE A.13 (Complex and real eigenvalues)

Consider a matrix

$$M = \begin{pmatrix} 1 & 0 & -1 \\ -1 & 0 & -2 \\ 2 & 0 & -1 \end{pmatrix}.$$

The corresponding characteristic matrix has the form

$$\lambda I - M = \begin{pmatrix} \lambda - 1 & 0 & 1 \\ 1 & \lambda & 2 \\ -2 & 0 & \lambda + 1 \end{pmatrix}$$

and the characteristic polynomial is

$$\det(\lambda I - M) = \lambda(\lambda^2 + 1).$$

The roots of $\det(\lambda I - M)$ are $0, i, -i$. Hence, as a real matrix, M has a single eigenvalue $\lambda_1 = 0$. As a complex matrix, it has three eigenvalues $\lambda_1 = 0$, $\lambda_2 = i$ and $\lambda_3 = -i$.

Lemma A.16 *Let M be a real or complex $n \times n$ matrix. There exists a polynomial $g(\lambda)$ of the degree n^2 or lower such that $g(M) = 0$.*

Proof: The linear space V of all $n \times n$ matrices has the dimension n^2. Hence the $n^2 + 1$ matrices $M^{n^2}, M^{n^2-1}, \ldots, M, I \in V$ must be linearly dependent and there exists a nontrivial set of coefficients $a_0, a_1, \ldots, a_{n^2}$ such that

$$a_0 M^{n^2} + a_1 M^{n^2-1} + \ldots + a_{n^2-1} M + a_{n^2} I = 0.$$

Thus

$$a_0 \lambda^{n^2} + a_1 \lambda^{n^2-1} + \ldots + a_{n^2-1}\lambda + a_{n^2}$$

is the sought polynomial. ■

Lemma A.16 can be strengthened to the following famous theorem.

Theorem A.2 (Cayley–Hamilton) *Every matrix is a root of its characteristic polynomial.*

Proof: This proof is slightly more technical and we refer, e.g., to [75]. ■

Definition A.19 (Eigenvector) *Let M be a real or complex $n \times n$ matrix and λ one of its eigenvalues. Any vector $u \neq 0$ such that*

$$Mu = \lambda u$$

is said to be eigenvector *of M corresponding to the eigenvalue λ.*

The following proposition is introduced for future reference:

Proposition A.2 *Let M be a $n \times n$ matrix. There exists at least one eigenvector to every eigenvalue $\lambda \in \sigma(M)$.*

Proof: Since $\det(\lambda I - M) = 0$, by Lemma A.12 the matrix $\lambda I - M$ is singular. Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be the linear operator represented by the matrix $\lambda I - M$. Then $N(f) \neq \{0\}$ by the same lemma. Thus there exists a nontrivial vector $v \in N(f)$ such that $Mv = \lambda v$.■

A.1.10 Hermitian, symmetric, and diagonalizable matrices

Definition A.20 (Diagonalizable matrix) *Let M be a real or complex $n \times n$ matrix. M is* diagonalizable *if there exists a nonsingular matrix C (real or complex, respectively) such that the matrix $D = C^{-1}MC$ is diagonal.*

Generally, any two matrices A and B satisfying the above relation $B = C^{-1}AC$ with some nonsingular matrix C are called similar. Thus a matrix M is diagonalizable if it is similar to a diagonal matrix D.

Theorem A.3 (Diagonalization theorem) *Let M be a real or complex $n \times n$ matrix. M is diagonalizable if and only if it has n linearly independent eigenvectors.*

Theorem A.3 states that the matrix M is diagonalizable if and only if it is possible to construct a basis in $\mathbb{R}^n$ that consists of the eigenvectors of M. Then the matrix C from Definition A.20 has the eigenvectors in its columns, and it is identical to the transition matrix from the eigenvector basis to the canonical basis. The matrix C^{-1} represents the transition matrix back to the eigenvector basis. Thus the diagonal matrix D represents the same linear operator as M, expressed with respect to the eigenvector bases in $\mathbb{R}^n$ instead of the canonical ones.

Proof: Assume that M is diagonalizable, i.e., there exists a diagonal matrix D such that $D = C^{-1}MC$. As explained above, the matrix C can be interpreted as a transition matrix from some basis B of $\mathbb{R}^n$ to the canonical basis of $\mathbb{R}^n$ and analogously C^{-1} as a transition matrix from the canonical basis back to B. According to Definition A.15, the columns c_i are the basis vectors of B expressed with respect to the canonical basis. It remains to be verified that the vectors c_i, $1 \leq i \leq n$, are eigenvectors of M. However, this is exactly what relation $MC = CD$ says (look at it column-wise), and moreover it says that diagonal entries of the matrix D are eigenvalues of M.

Now assume that there exists a basis B consisting of eigenvectors $c_1, c_2, \dots, c_n$ of the matrix M. By λ_i, $1 \leq i \leq n$ denote eigenvalues of M such that $Mc_i = \lambda_i c_i$. Putting these relations together for all i, we obtain a matrix equation

$$MC = CD,$$

where the matrix D =diag$(\lambda_1, \lambda_2, \dots, \lambda_n)$. Hence,

$$C^{-1}MC = D$$

and M is diagonalizable. ■

Definition A.21 (Hermitian and symmetric matrices) *Let M be a complex $n \times n$ matrix. M is* Hermitian *if $m_{ij} = \overline{m}_{ji}$ for all $1 \leq i, j \leq n$ (the symbol $\overline{m}_{ji}$ stands for the complex conjugate of m_{ji}). Let M be a real $n \times n$ matrix. M is* symmetric *if $m_{ij} = m_{ji}$ for all $1 \leq i, j \leq n$.*

Lemma A.17 *All eigenvalues of Hermitian matrices are real.*

This lemma obviously covers symmetric real matrices.

Proof: Let M be a Hermitian $n \times n$ matrix and $\lambda \in \sigma(M)$ any of its eigenvalues. By Proposition A.2 there exists an eigenvector v of M such that $Mv = \lambda v$. The ith row of this vector identity has the form

$$\sum_{j=1}^{n} m_{ij} v_j = \lambda v_i,$$

where v_i, v_j are the ith and jth components of the vector v. Multiplying with $\overline{v}_i$ and summing over $i = 1, 2, \dots, n$, we obtain

$$\sum_{i=1}^{n} \sum_{j=1}^{n} m_{ij} v_j \overline{v}_i = \lambda \sum_{i=1}^{n} |v_i|^2.$$

Since the sum on the right-hand side obviously is real, it is sufficient to verify that the left-hand side is a real number. Indeed this is true since

$$\overline{\sum_{i=1}^{n} \sum_{j=1}^{n} m_{ij} v_j \overline{v}_i} = \sum_{i=1}^{n} \sum_{j=1}^{n} \overline{m}_{ij} \overline{v}_j v_i = \sum_{i=1}^{n} \sum_{j=1}^{n} m_{ji} v_i \overline{v}_j = \sum_{i=1}^{n} \sum_{j=1}^{n} m_{ij} v_j \overline{v}_i.$$

■

A.1.11 Linear forms, dual space, and dual basis

Linear forms are special linear operators with values in $\mathbb{R}$ or $\mathbb{C}$. They play an essential role in the theory of partial differential equations and finite element methods.

Definition A.22 (Linear form and dual space) *Let V be a real or complex linear space. Linear operator f from V to $\mathbb{R}$ or $\mathbb{C}$ is called* linear form. *The space of all linear forms over the space V is called* dual space *and denoted by V'.*

■ EXAMPLE A.14 (Linear forms)

1. Let $V = \mathbb{R}^n$. The operator $f : V \to \mathbb{R}$ defined as the average of vector entries,

$$f(v) = \frac{1}{n}\sum_{i=1}^{n} v_i \quad \text{for all } v \in V,$$

is a linear form over V, i.e., $f \in V'$.

2. Let $V = C(a,b)$ be the space of continuous functions in some interval (a,b). The integral operator $A : V \to \mathbb{R}$ defined by

$$A(f) = \int_a^b f(x)\,\mathrm{d}x \quad \text{for all } f \in V,$$

is a linear form over V.

3. Again let $V = C(a,b)$, and let c be some point in the interior of the interval (a,b). The operator $g_c : V \to \mathbb{R}$, associated with the function value at c,

$$g_c(f) = f(c) \quad \text{for all } f \in V,$$

is a linear form over V.

Lemma A.18 *Let V be a real or complex linear space of dimension n. Then the dual space V' has the same dimension n.*

Proof: We leave this to the reader as an exercise. Use Definitions A.6 and A.22. ■

Definition A.23 (Dual basis) *Let V be a real or complex linear space of dimension n and $B = \{v_1, v_2, \ldots, v_n\}$ a basis in V. The basis $B' = \{f_1, f_2, \ldots, f_n\}$ of the space V' is said to be the* dual basis *to B if*

$$f_i(v_j) = \delta_{ij} \tag{A.6}$$

for all $1 \le i, j \le n$ (the symbol δ_{ij} is the Kronecker delta, i.e., $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise).

Let us prove the existence of the dual basis in the following Lemma A.19 and give some examples in Example A.15.

Lemma A.19 (Existence of dual basis) *Let V be a real or complex linear space of dimension n. To every basis B of V there exists a dual basis B' of V'.*

Proof: Let $B = \{v_1, v_2, \ldots, v_n\}$ be a basis of V. First we define the operators $f_1, f_2, \ldots, f_n$ on the basis elements only,

$$f_i(v_j) = \delta_{ij}. \tag{A.7}$$

Next we extend them to the whole space V by defining

$$f_i(u) = \sum_{j=1}^{n} a_j f_i(v_j).$$

Here the coefficients $a_1, a_2, \ldots, a_n$ are the unique expansion coefficients of the element u with respect to the basis B,

$$\langle u \rangle_B = (a_1, a_2, \ldots, a_n)^T.$$

Obviously the forms $f_1, f_2, \ldots, f_n$ are linear, and it is sufficient to show that they also are linearly independent. By contradiction suppose that they are linearly dependent. Then it is possible to express one of them (for example f_1) as a nontrivial linear combination of the others, i.e.,

$$f_1 = c_2 f_2 + c_3 f_3 + \ldots + c_n f_n.$$

However, by (A.7) we obtain

$$1 = f_1(v_1) = \sum_{i=2}^{n} c_i \underbrace{f_i(v_1)}_{=0},$$

which finishes the proof. ■

■ **EXAMPLE A.15 (Dual basis)**

In the space $V = P^2(-1, 1)$ consider the basis

$$B = \left\{ \frac{x(x-1)}{2}, 1 - x^2, \frac{x(x+1)}{2} \right\} = \{v_1, v_2, v_3\},$$

shown in Figure A.9.

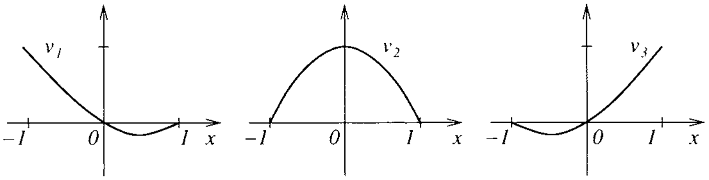


Figure A.9 Basis $B = \{v_1, v_2, v_3\}$.

The dual basis $B' = \{f_1, f_2, f_3\}$ consists of linear forms $f_i : V \to \mathbb{R}$ such that

$$f_i(v) = v(x_i) \quad \text{for all } v \in V,$$

where $x_1 = -1$, $x_2 = 0$, and $x_3 = 1$. The reader can easily verify that the delta property (A.6) indeed holds true.

Now, since $B' = \{f_1, f_2, f_3\}$ is a basis in V', any linear form in V' can be expressed uniquely in terms of the basis functions f_1, f_2, f_3. Let us try, for example, the linear form $A \in V'$ associated with the integral

$$A(g) = \int_{-1}^{1} g(x)\, \mathrm{d}x.$$

We look for coefficients $\beta_1, \beta_2, \beta_3$ such that

$$A = \beta_1 f_1 + \beta_2 f_2 + \beta_3 f_3. \tag{A.8}$$

Applying A to the basis B, by (A.8) and the delta property $f_i(v_j) = \delta_{ij}$, we have

$$A(v_1) = \beta_1, \quad A(v_2) = \beta_2, \quad A(v_3) = \beta_3.$$

Calculating the integrals of the basis functions v_1, v_2 and v_3, we obtain $\beta_1 = 1/3$, $\beta_2 = 4/3$, $\beta_3 = 1/3$, and thus

$$A = \frac{1}{3}f_1 + \frac{4}{3}f_2 + \frac{1}{3}f_3. \tag{A.9}$$

Thus by (A.9) we can integrate all quadratic polynomials using their function values at $-1, 0$ and 1,

$$A(g) = \frac{1}{3}f_1(g) + \frac{4}{3}f_2(g) + \frac{1}{3}f_3(g) = \frac{1}{3}g(-1) + \frac{4}{3}g(0) + \frac{1}{3}g(1)$$

[this is the Simpson's rule for the interval $(-1, 1)$].

A.1.12 Exercises

Exercise A.1 *Consider the set $\mathcal{M}^{n\times n}$ of all real $n \times n$ matrices.*

1. *Show that $\mathcal{M}^{n\times n}$ is a linear space.*

2. *Show that the set $\mathcal{D}^n$ of diagonal $n \times n$ matrices is a subspace of $\mathcal{M}^{n\times n}$.*

Exercise A.2 *Let $S = \{u \in C(0,1);\ u(1) = a\}$, $a \in \mathbb{R}$. For what values of a is S linear space?*

Exercise A.3 *Prove in detail all assertions in Example A.1.2.*

Exercise A.4 *Let S be the set of all twice continuously differentiable functions satisfying the differential equation*

$$u''(x) + u(x) = 0.$$

1. *Prove that S is a linear space. Hint: The equation does not need to be solved.*

2. *Is the set $\tilde{S}$ of all solutions to the differential equation*

$$u''(x) + u(x) + 1 = 0$$

a linear space as well?

Exercise A.5 *Prove in detail Proposition A.1.*

Exercise A.6 *Let $V = \mathbb{R}^5$, $W_1 = \{v \in V; v_1 = 0\}$ and $W_2 = \{v \in V; v_2 = 0\}$. Show that W_1 and W_2 are subspaces of V and construct the space $W_1 \cap W_2$.*

Exercise A.7 *Prove Lemma A.4.*

Exercise A.8 *Let $I = (0,1) \times (0,1) \subset \mathbb{R}^2$. In I consider the sets S_1 of all polynomials of degree exactly one, S_2 of all polynomials of degree exactly three, and S_3 of polynomials that in one variable are of degree exactly one and that are of degree three or less in the other variable. Construct the linear span $S = [S_1 \cup S_2 \cup S_3]$ in the linear space $P^4(I)$ of polynomials of degree lower or equal to four. Hint: The degree of a bivariate polynomial f is the highest sum $k + l$ of powers among all monomials $x^k y^l$ appearing in f. What is the dimension of S?*

Exercise A.9 *Consider the linear space $V = \mathcal{M}^{n\times n}$ from Exercise A.1.*

1. *Show that the set $\mathcal{S}^{n\times n}$ of symmetric $n \times n$ matrices is a subspace of V.*

2. *Show that the set $\mathcal{A}^{n\times n}$ of antisymmetric $n \times n$ matrices is a subspace of V.*

3. *Show that $V = \mathcal{S} \oplus \mathcal{A}$. Hint: Symmetric part of a matrix $M \in V$ is defined as $M_s = (M + M^T)/2 \in \mathcal{S}^{n\times n}$. For all $M \in V$, the transpose M^T is defined in a standard way as $(M^T)_{i,j} \equiv M_{j,i}$, $1 \le i,j \le n$.*

Exercise A.10 *Prove Lemma A.8, case of finite cardinality.*

Exercise A.11 *Consider $V = \mathbb{R}^4$ and its subspaces $V_1 = [(-3,0,2,0)^T]$ and $V_2 = [(1,0,2,-3)^T, (3,2,1,-5)^T, (-1,2,1,-2)^T]$. Compute $dim(V_1 + V_2)$ and $dim(V_1 \cap V_2)$. Hint: Select a basis among the vectors generating V_2. Check whether the vector generating V_1 lies in V_2.*

Exercise A.12 *Consider the linear space $P^3(-1,1)$ of real-valued polynomials defined in the interval $(-1,1)$. Consider the bases*

$$B_1 = \left\{1, \frac{1+x}{2}, \frac{(1+x)^2}{4}, \frac{(1+x)^3}{8}\right\},$$

and

$$B_2 = \left\{\frac{1+x}{2}, \frac{1-x}{2}, \frac{1-x^2}{4}, \frac{x(1-x^2)}{8}\right\}.$$

Construct the transition matrix M from B_1 to B_2.

Exercise A.13 *Prove Lemma A.10.*

Exercise A.14 *Let $V = \mathbb{R}^5$. Consider the canonical basis*

$$B_1 = \{(1,0,0,0,0)^T, (0,1,0,0,0)^T, (0,0,1,0,0)^T, (0,0,0,1,0)^T, (0,0,0,0,1)^T\}$$

and another basis

$$B_2 = \{(1,1,0,0,0)^T, (0,1,1,0,0)^T, (0,1,0,1,0)^T, (0,0,1,1,0)^T, (0,0,0,1,1)^T\}.$$

Construct the transition matrix M from B_1 to B_2.

Exercise A.15 *Prove Lemma A.12.*

Exercise A.16 *Consider the polynomial spaces $V = P^6(-1,1)$ and $W = P^4(-1,1)$ equipped with the monomial bases $B_V = \{1, x, x^2, x^3, x^4, x^5, x^6\}$ and $B_W = \{1, x, x^2, x^3, x^4\}$.*

1. *Write the matrix representation of the linear operator $\varphi : V \longrightarrow W$,*

$$\varphi(v) = \frac{\mathrm{d}^2 v}{\mathrm{d}x^2}$$

   with respect to the bases B_V and B_W.

2. *Determine $N(\varphi)$ and $R(\varphi)$. Is φ a bijection?*

Exercise A.17 *Prove Lemma A.13.*

Exercise A.18 *Consider the polynomial spaces*

$$V = \{v \in P^4(-1,1), v(x) = v(-x) \textit{ for all } x \in (-1,1)\}$$

and

$$W = \{w \in P^5(-1,1), w(x) = -w(-x) \textit{ for all } x \in (-1,1)\},$$

equipped with the bases $B_V = \{1, x^2, x^4\}$ and $B_W = \{x, x^3, x^5\}$.

1. *Write the matrix representation M_ψ of the linear operator $\psi : V \longrightarrow W$, $\psi(f) = F$, where F is a primitive function to f such that $F(0) = 0$.*

2. *Determine $N(\psi)$ and $R(\psi)$. Is ψ a bijection?*

3. *If ψ is a bijection then find the inverse operator $\psi^{-1} : W \longrightarrow V$, write its matrix representation $M_{\psi^{-1}}$ with respect to the bases B_W and B_V, and verify that the matrix $M_{\psi^{-1}}$ is inverse to M_ψ.*

Exercise A.19 *Prove Lemma A.14.*

Exercise A.20 *Consider the real matrices*

$$M_1 = \begin{pmatrix} 0 & x & y & z \\ x & 0 & z & y \\ y & z & 0 & x \\ z & y & x & 0 \end{pmatrix}, \quad M_2 = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & z^2 & y^2 \\ 1 & z^2 & 0 & x^2 \\ 1 & y^2 & x^2 & 0 \end{pmatrix}, \quad x, y, z \in \mathbb{R}.$$

Without evaluating the determinants, use rules from Lemma A.15 to show that $\det(M_1) = \det(M_2)$.

Exercise A.21 *Prove Lemma A.18.*

Exercise A.22 *Let* $A \in \mathbb{R}^{n\times n}$ *be a regular* $n \times n$ *matrix.*

1. *Use its characteristic polynomial and the Cayley–Hamilton theorem to derive an explicit formula for its inverse* A^{-1}*, based on the following matrix operations: (a) matrix multiplication and (b) the sum of a matrix with a diagonal matrix.*

2. *Consider the matrix*

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & -2 & 3 \end{pmatrix}.$$

Write its inverse using the above-defined operations (a) and (b).

A.2 NORMED SPACES

Normed spaces are linear spaces endowed with norm (size). The notion of norm allows us to define convergence and limit for sequences of vectors, matrices, functions, linear operators, and other objects. In normed spaces we can distinguish between open and closed sets, and introduce the notion of completeness. Through complete normed (Banach) spaces we arrive at the Lebesgue L^p-spaces, which are essential for the study of partial differential equations and finite element methods.

A.2.1 Norm and seminorm

Definition A.24 (Norm and normed space) *Let* V *be a real or complex linear space. A* norm *on* V *is any function* $\|\cdot\|_V : V \to \mathbb{R}$ *with the following properties:*

1. $\|v\|_V \geq 0$ *for all* $v \in V$*, and* $\|v\|_V = 0$ *if and only if* $v = 0$*,*
2. $\|av\|_V = |a|\|v\|_V$ *for all* $v \in V$ *and all coefficients* $a \in \mathbb{R}$ *(or* $\mathbb{C}$*),*
3. $\|u + v\|_V \leq \|u\|_V + \|v\|_V$ *for all* $u, v \in V$.

A linear space V *endowed with a norm* $\|\cdot\|_V$ *is said to be* normed space.

Remark A.1

1. *In a normed space* $\|v\|_V < \infty$ *for all* $v \in V$.

2. *The subscript V in the symbol $\|\cdot\|_V$ is often omitted when the meaning of $\|\cdot\|$ is clear from the context.*

3. *The last condition in Definition A.24 (triangular inequality) usually is the most difficult one to prove.*

For future reference, let us mention the triangular inequality for real numbers:

■ EXAMPLE A.16 (Triangular inequality in $\mathbb{R}$)

Let $a, b \in \mathbb{R}$. Then

$$|a+b| \le |a| + |b|.$$

This easily follows from the analysis of four possible cases: $a > 0$ & $b > 0$, $a \le 0$ & $b > 0$, $a > 0$ & $b \le 0$ and $a \le 0$ & $b \le 0$.

A linear space may be endowed with many different norms, in which case one obtains different normed spaces. Several examples of norms are collected in Example A.17. In what follows the symbols (a, b) and $[a, b]$ will stand for nonempty bounded intervals.

■ EXAMPLE A.17 (Norms)

1. Let $V = \mathbb{R}$. The absolute value function $\|\cdot\| : V \to \mathbb{R}$, $\|u\| = |u|$, clearly satisfies

   (a) $\|v\| = |v| > 0$ for all $0 \ne v \in V$,

   (b) $\|av\| = |a||v| = |a|\|v\|$ for all $v \in V$ and all coefficients a,

   (c) $\|u + v\| = |u + v| \le |u| + |v| = \|u\| + \|v\|$ for all $u, v \in V$,

   and thus it is a norm in V.

2. Let $V = \mathbb{R}^n$. The function

$$\|v\|_\infty = \max_{1 \le i \le n} |v_i| \tag{A.10}$$

   is a norm in V (discrete maximum norm). The first two properties of Definition A.24 are obvious, and the triangular inequality is verified as follows:

$$\begin{aligned}\|u + v\|_\infty &= \max_{1 \le i \le n} |u_i + v_i| \le \max_{1 \le i \le n} (|u_i| + |v_i|) \\ &\le \max_{1 \le i \le n} |u_i| + \max_{1 \le i \le n} |v_i| = \|u\|_\infty + \|v\|_\infty.\end{aligned}$$

3. Let $V = \mathbb{R}^n$. The function

$$\|v\|_1 = \sum_{i=1}^{n} |v_i|, \tag{A.11}$$

where v_i are the components of v (discrete integral norm), is a norm in V. The proof is analogous to the previous case. For $1 \leq p < \infty$ this norm generalizes to the discrete p-norm,

$$\|v\|_p = \left(\sum_{i=1}^{n} |v_i|^p \right)^{\frac{1}{p}}. \tag{A.12}$$

4. With $p = 2$ the discrete p-norm (A.12) yields the Euclidean norm

$$\|v\|_2 = \sqrt{\sum_{i=1}^{n} v_i^2}. \tag{A.13}$$

In this case the proof of the triangular inequality is trivial for $n = 1$. Extension to general $n \geq 1$ can be done by induction.

5. The norms defined in $\mathbb{R}^n$ can be extended naturally to $n \times n$ matrices, i.e., to the space $V = \mathbb{R}^{n \times n}$. For example, the discrete maximum norm (A.10) can be extended to

$$\|M\|_\infty = \max_{1 \leq i,j \leq n} |m_{ij}|. \tag{A.14}$$

Another extension of the norm (A.10), which will be useful later, is

$$\|M\| = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |m_{ij}|. \tag{A.15}$$

The Euclidean norm in $\mathbb{R}^n$ is extended to the Frobenius norm,

$$\|M\|_f = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} m_{ij}^2}, \tag{A.16}$$

Here m_{ij} is the entry of the matrix M at the position i, j.

6. Let $V = C([a, b])$ be the space of functions continuous in a closed interval $[a, b]$. The function

$$\|v\|_\infty = \max_{x \in [a,b]} |v(x)| \tag{A.17}$$

is a norm in V (maximum norm).

7. Let $V = P^k(a, b)$. The function

$$\|v\|_1 = \int_a^b |v(x)| \, \mathrm{d}x \tag{A.18}$$

is a norm in V. For $1 \leq p < \infty$ this norm can be extended to the p-norm,

$$\|v\|_p = \left(\int_a^b |v(x)|^p \, \mathrm{d}x \right)^{\frac{1}{p}} . \tag{A.19}$$

Here the subscripts in the norms $\| \cdot \|_\infty$ and $\| \cdot \|_p$ are originated in the Lebesgue spaces L^∞ and L^p which will be discussed in Paragraph A.2.9. For later use let us define seminorm:

Definition A.25 (Seminorm) *Let V be a real or complex linear space. A* seminorm *on V is any function $| \cdot |_V : V \longrightarrow \mathbb{R}$ with the following properties:*

1. $|v|_V \geq 0$ *for all $v \in V$,*

2. $|av|_V = |a||v|_V$ *for all $v \in V$ and all coefficients a,*

3. $|u + v|_V \leq |u|_V + |v|_V$.

■ EXAMPLE A.18 (Seminorm)

In a real interval (a, b) consider the space of smooth functions with bounded derivative,

$$V = \{v \in C^1(a, b); \sup_{x \in (a,b)} |v'(x)| < \infty\}.$$

The function

$$|v| = \sup_{x \in (a,b)} |v'(x)| \tag{A.20}$$

is a seminorm in V. All three properties of Definition A.25 are easy to verify. The only difference between norm and seminorm is that while

$$\|v\|_V = 0 \quad \text{implies that} \quad v = 0,$$

it can be

$$|v|_V = 0 \quad \text{where} \quad v \neq 0,$$

as it was with constant functions in (A.20).

The advantage of seminorms is that they are easier to evaluate than full norms. When restricted to a subspace $W \subset V$ which does not contain the nonzero elements of V where the seminorm vanishes, the seminorm becomes a full norm. This is shown in Example A.19.

■ **EXAMPLE A.19 (Seminorm which is norm in a subspace)**

Consider the space V from Example A.18. The functions preventing the seminorm (A.20) from being a full norm are nonzero constants. For example, in the subspace $W \subset V$ of functions antisymmetric with respect to the midpoint of the interval (a, b),

$$W = \left\{ w \in V;\ w\left(\frac{a+b}{2} - x\right) = -w\left(\frac{a+b}{2} + x\right) \text{ for all } x \in (a, b) \right\},$$

these functions are not present, and therefore the function $\|\cdot\| = |\cdot|$ is a norm in W. The space W is a normed space according to Definition A.24.

On the other hand, sometimes it is practical to create a full norm by adjusting a seminorm so that it does not vanish on nonzero functions. This is shown in Example A.20.

■ **EXAMPLE A.20 (Changing seminorm to a norm)**

Consider the space V from Example A.18 again and adjust the seminorm (A.20) to

$$\|v\| = \sup_{x \in (a,b)} |v'(x)| + \left| v\left(\frac{a+b}{2}\right) \right|. \tag{A.21}$$

Now the nonzero constants make $\|\cdot\|$ no longer vanish, and therefore it is a norm in V. The first two properties of Definition A.24 are obvious, and the triangular inequality also is easy to show:

$$\begin{aligned} \|u + v\| &= \sup_{x \in (a,b)} |u'(x) + v'(x)| + |u\left(\frac{a+b}{2}\right) + v\left(\frac{a+b}{2}\right)| \\ &\le \sup_{x \in (a,b)} (|u'(x)| + |v'(x)|) + (|u\left(\frac{a+b}{2}\right)| + |v\left(\frac{a+b}{2}\right)|) \\ &\le \sup_{x \in (a,b)} |u'(x)| + |u\left(\frac{a+b}{2}\right)| + \sup_{x \in (a,b)} |v'(x)| + |v\left(\frac{a+b}{2}\right)| = \|u\| + \|v\|. \end{aligned}$$

The space V endowed with the norm (A.21) is a normed space according to Definition A.24.

A.2.2 Convergence and limit

This paragraph is devoted to the asymptotic behavior of infinite sequences in normed spaces. Let us begin by introducing the notion of boundedness:

Definition A.26 (Bounded sequence) *Let V be a normed space. The sequence $\{u_n\}_{n=1}^{\infty} \subset V$ is said to be* bounded *in V if there exists a $C > 0$ such that $\|u_n\| \le C$ for all n.*

Next let us use the notion of norm to define the convergence and limit of a sequence:

Definition A.27 (Convergence and limit of a sequence) *Let V be a normed space. The sequence $\{u_n\}_{n=1}^{\infty} \subset V$ is said to be* convergent *in V if there exists an element $v \in V$ such* *that for every* $0 < \epsilon \in \mathbb{R}$ *there exists an index* n_ϵ *such that* $\|v - u_n\| < \epsilon$ *for all* $n > n_\epsilon$. *The element* v *is the* limit *of the sequence* $\{u_n\}_{n=1}^\infty$. *We usually write*

$$\lim_{n\to\infty} u_n = v$$

or

$$u_n \to v \quad \text{as} \quad n \to \infty.$$

Lemma A.20 summarizes basic facts about convergent sequences.

Lemma A.20 *Let* V *be a normed space,* $\{u_n\}_{n=1}^\infty$ *a sequence in* V *and* $v \in V$. *The following holds:*

1.

$$\lim_{n\to\infty} u_n = v \quad \text{\textit{if and only if}} \quad \lim_{n\to\infty} \|u_n - v\|_V = 0.$$

2.

$$\lim_{n\to\infty} u_n = v \quad \text{\textit{then}} \quad \{u_n\}_{n=1}^\infty \text{ \textit{is bounded}}.$$

3.

$$\lim_{n\to\infty} u_n = v \quad \text{\textit{then}} \quad \lim_{n\to\infty} \|u_{n+1} - u_n\|_V = 0.$$

Proof: Left to the reader as an easy exercise. ■

Let us present a few examples illustrating the concept of convergence in normed spaces. We begin with showing that the third assertion of Lemma A.20 is not an "if and only if":

■ EXAMPLE A.21

Let $V = \mathbb{R}$. The sequence

$$\{u_n\}_{n=1}^\infty = \left\{ \sum_{i=1}^n \frac{1}{n} \right\}_{n=1}^\infty \subset \mathbb{R}$$

satisfies

$$\lim_{n\to\infty} \|u_{n+1} - u_n\|_V = \lim_{n\to\infty} \frac{1}{n+1} = 0,$$

but it is well known that

$$\infty = \sum_{i=1}^\infty \frac{1}{n} = \lim_{n\to\infty} u_n$$

■ EXAMPLE A.22 (Convergence and limit)

1. Consider the space $V = \mathbb{R}^3$ equipped with the discrete maximum norm (A.10) and a sequence of vectors $\{u_n\}_{n=1}^{\infty} \subset V$,

$$u_n = \left(1 - \frac{1}{n}, e^{-n}, \frac{\sin(n)}{n^2}\right)^T.$$

   The only candidate for a limit is $v = (1, 0, 0)^T$. The sequence converges to v since

$$\|u_n - v\|_\infty \leq \frac{1}{n} \quad \text{for all } n \geq 1.$$

2. Let $V = C(0, 1)$ equipped with the maximum norm (A.17),

$$\|v\|_\infty = \sup_{x \in (0,1)} |v(x)|,$$

   and consider a sequence of functions $\{u_n\}_{n=1}^{\infty} \subset V$,

$$u_n(x) = x^n(1 - x) + x^3 + 1.$$

   Since

$$\max_{x \in (0,1)} |u_n - (x^3 + 1)| = \left(\frac{n}{n+1}\right)^n \left(\frac{1}{n+1}\right) \leq \frac{1}{n+1},$$

   the only candidate for a limit is $v(x) = x^3 + 1$. Since

$$\|u_n - v\|_\infty \leq \frac{1}{n+1},$$

   the sequence converges to v.

■ EXAMPLE A.23 (Nonconvergent sequences)

It is easy to show, using Lemma A.20, that the following sequences do not converge.

1. Consider the space $V = P^2([0, 1])$ equipped with the maximum norm (A.17), and the sequence $\{u_n\}_{n=1}^{\infty} \subset V$, $u_n(x) = nx(1 - x)$. The sequence is not bounded ($\|u_n\|_\infty = n/4$ for all n).

2. In the same space V let $\{u_n\}_{n=1}^{\infty} \subset V$, $u_n(x) = (-1)^n x(1 - x)$. The sequence is bounded ($\|u_n\|_\infty = 1/4$ for all n), but $\|u_{n+1} - u_n\|_\infty = 1/2$ for all n.

3. Let $V = C([0, 1])$ be equipped with the integral norm (A.18), and consider the sequence $\{u_n\}_{n=1}^{\infty} \subset V$, $u_n(x) = x^n$. It is $\|u_n\|_1 = 1/(n + 1)$ for all n, but the only function v that could be its limit is defined by $v(x) = 0$ for $x \in [0, 1)$ and $v(1) = 1$. However, this function does not lie in the space V. More about this situation will be said in Paragraph A.2.7.

A.2.3 Open and closed sets

In this paragraph let us continue with the definition of bounded sets, open balls, and open and closed sets:

Definition A.28 (Bounded set) *Let V be a normed space and $S \subset V$ a subset of V. We say that S is bounded in V if there exists a positive constant $C > 0$ such that $\|x\|_V \le C$ for all $x \in S$.*

Definition A.29 (Open ball $B(g, r)$) *Let V be a normed space with the norm $\|\cdot\|_V$, $g \in V$ and $0 < r \in \mathbb{R}$. By the* open ball *with the center g and radius r we mean the set $B(g, r) = \{v \in V;\ \|v - g\|_V < r\}$.*

■ **EXAMPLE A.24 (Open balls in $\mathbb{R}^2$)**

1. Consider the linear space $V = \mathbb{R}^2$ and the norms

$$\|u\|_1 = |u_1| + |u_2|, \quad \|u\|_\infty = \max(|u_1|, |u_2|), \quad \|u\|_2 = \sqrt{u_1^2 + u_2^2},$$

where $u \in V$, $u = (u_1, u_2)^T$. The unit open balls $B(\mathbf{0}, 1)$ in these norms are depicted in Figure A.10.

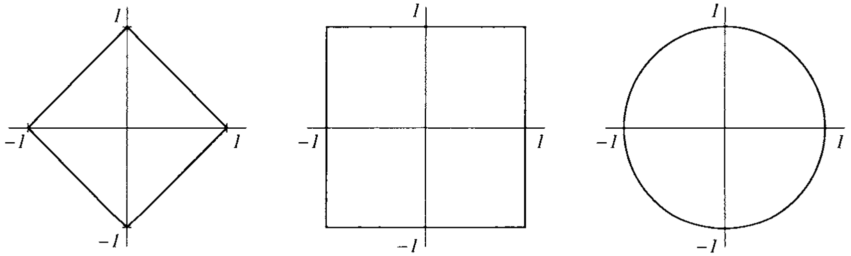


Figure A.10 Examples of unit open balls $B(\mathbf{0}, 1)$ in $V = \mathbb{R}^2$.

2. Let $V = P^5([a, b])$ be equipped with the maximum norm (A.17). Consider a real number $r > 0$ and a function $u \in V$. The open ball $B(u, r)$ comprises all fifth-degree polynomials that lie inside of a "belt" of width $2r$ around u, as shown in Figure A.11.

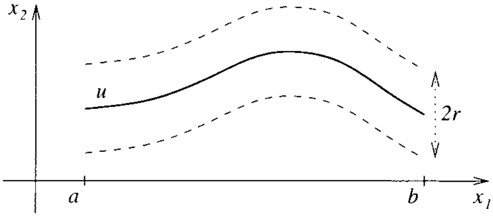


Figure A.11 Open ball $B(u, r)$ in the space $V = P^5([a, b])$ with the maximum norm (A.17).

3. In general, sets of functions are too abstract to be visualized like $B(u, r)$ in Figure A.11. It is sufficient to replace in the previous case the maximum norm (A.17) with the integral norm (A.18), and the visual interpretation of the open ball $B(u, r)$ is lost. In such cases it is good idea to forget about concrete shape of the functions and imagine them as points (see Figure A.12).

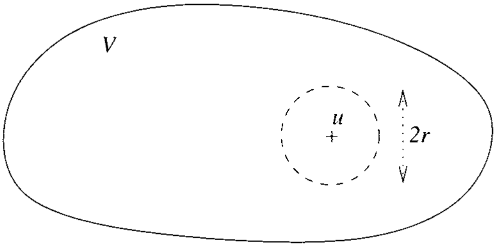


Figure A.12 Open ball $B(u, r)$ in $V = P^5([a, b])$ equipped with the integral norm (A.18).

Definition A.30 (Open and closed set) *Let V be a normed space and $S \subset V$. The set S is* open *in V if for every $g \in S$ there exists $r > 0$ such that $B(g, r) \subset S$. The set S is* closed *if its complement $V \setminus S$ is open.*

Example A.25 shows an open set, a closed set and a set that neither is open nor closed.

■ **EXAMPLE A.25 (Open and closed sets)**

Let $V = C([a, b])$ be equipped with the maximum norm (A.17).

1. Let $u \in V$ and $C > 0$. The set

$$S_1 = \{v \in V; \|v - u\|_\infty \le C\}$$

is closed.

2. Let $C_1 < C_2$ be real numbers. The set

$$S_2 = \{v \in V; \ C_1 < v(x) < C_2 \text{ for all } x \in [a, b]\}$$

is open.

3. The set

$$S_3 = \{v \in V; \ 0 \le v(x) < 1 \text{ for all } x \in [a, b]\}$$

is neither open nor closed.

A.2.4 Continuity of operators

Definition A.31 (Continuous operator) *Let U and V be normed spaces. An operator $f : U \to V$ is* continuous at $g \in U$ *if for every $\epsilon > 0$ there exists a $\delta > 0$ such that*

$$u \in B(g, \delta) \Rightarrow f(u) \in B(f(g), \epsilon). \tag{A.22}$$

We say that f is continuous in U *if it is continuous at every $g \in U$.*

Usually, integral operators are continuous. Example A.26 shows an integral operator that is continuous in the space $C([a,b])$.

■ EXAMPLE A.26 (Integral operator)

Let $U = C([a,b])$ be equipped with the maximum norm (A.17) and $V = \mathbb{R}$. Consider a linear operator $A : U \to V$,

$$A(u) = \int_a^b u(x)\,\mathrm{d}x.$$

Let $g \in U$ be arbitrary. Given a $\delta > 0$, the open ball $B(g, \delta)$ has the form

$$B(g,\delta) = \{v \in U;\ \max_{x \in [a,b]} |v(x) - g(x)| < \delta\}.$$

Every $v \in B(g, \delta)$ satisfies

$$\|A(v) - A(g)\|_V = \left| \int_a^b v(x)\,\mathrm{d}x - \int_a^b g(x)\,\mathrm{d}x \right| \le \int_a^b \underbrace{|v(x) - g(x)|}_{<\delta}\,\mathrm{d}x < \delta(b-a).$$

Hence, for every $\epsilon > 0$ we can find a $\delta := \epsilon/(b-a)$ so that the implication (A.22) holds. Since $g \in U$ was arbitrary, the operator A is continuous in the whole space U.

Differential operators are more tricky. Let us devote Examples A.27, A.28, and A.29 to their study.

■ EXAMPLE A.27 (Derivative operator I)

Let $U = P^k([a,b])$ and $V = P^{k-1}([a,b])$ be equipped with the maximum norm (A.17), and consider the operator $\varphi : U \to V$,

$$\varphi(u) = \frac{du}{dx} \tag{A.23}$$

(here $k \ge 1$ is a natural number). Let us show that the operator φ is continuous.

Define $k+1$ points $a = x_0 < x_1 < x_2 < \ldots < x_k = b$, and consider the Lagrange interpolation polynomials $L_i \in P^k(a,b)$, $L_i(x_j) = \delta_{ij}$, $0 \le i,j \le k$. Recall that the functions $L_0, L_1, \ldots, L_k$ constitute a basis in U. The derivative of each polynomial L_i is bounded by a positive real constant $C_i > 0$,

$$\sup_{x\in(a,b)} |L_i'(x)| \le C_i, \quad 0 \le i \le k. \tag{A.24}$$

Let $g \in U$ and $v \in B(g,\delta)$, where $\delta > 0$ is arbitrary. The functions g and v can be written uniquely as

$$g(x) = \sum_{i=0}^{k} \alpha_i L_i(x), \qquad v(x) = \sum_{i=0}^{k} \beta_i L_i(x).$$

The open ball $B(g,\delta)$ has the form

$$B(g,\delta) = \{v \in U;\ \max_{x\in[a,b]} |v(x) - g(x)| < \delta\},$$

and therefore each pair of coefficients α_m and β_m are related via the inequality

$$\delta > |v(x_m) - g(x_m)| = \left|\sum_{i=0}^{k} \beta_i L_i(x_m) - \sum_{i=0}^{k} \alpha_i L_i(x_m)\right| = |\beta_m - \alpha_m|.$$

From here it follows that

$$\|\varphi(v) - \varphi(g)\|_V = \sup_{x\in(a,b)} |v'(x) - g'(x)| = \sup_{x\in(a,b)} \left|\sum_{i=0}^{k} (\beta_i - \alpha_i) L_i'(x)\right|$$

$$\le \sup_{x\in[a,b]} \sum_{i=0}^{k} \underbrace{|\beta_i - \alpha_i|}_{<\delta}\, \underbrace{|L_i'(x)|}_{\le C_i} \le \delta C,$$

where $C = \sum_{i=0}^{k} C_i$. Hence, for every $\epsilon > 0$ we can find a $\delta := \epsilon/C$ so that the implication (A.22) holds. Since $g \in U$ was arbitrary, the operator φ is continuous in the whole space U.

However, all polynomial spaces are finite-dimensional and as we shall see later, in finite-dimensional spaces all linear operators are continuous. Therefore it is more interesting to look for the largest space where the derivative operator is continuous. We have to be careful, however, not to make the space too big. In Example A.28 we consider the space $U = C^1(a,b) \cap C([a,b])$ equipped with the maximum norm (A.17).

■ EXAMPLE A.28 (Derivative operator II)

Consider the space $V = C([a,b])$ equipped with the maximum norm (A.17) and its subspace $U = C^1(a,b) \cap C([a,b])$. The derivative operator (A.23) is now considered as $\varphi : U \to V$. The function

$$g(x) = \sqrt{x-a}$$

is continuous in $[a, b]$ and therefore $\|g\|_U < \infty$. However, its derivative $g'(x) = 1/(2\sqrt{x-a})$ is unbounded, as shown in Figure A.13. Thus $\varphi(g) \notin V$ and φ is not continuous according to Definition A.31.

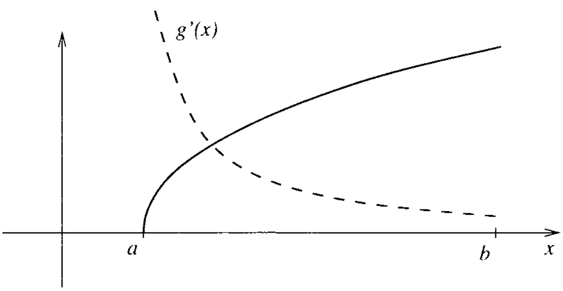


Figure A.13 The function $g(x) = \sqrt{x-a}$ and its derivative $g'(x) = 1/(2\sqrt{x-a})$.

Finally let us show the continuity of the derivative operator in its natural setting of the space of smooth functions with bounded derivatives in Example A.29.

■ EXAMPLE A.29 (Derivative operator III)

Consider the derivative operator (A.23) as $\varphi : U \to V$, where the space

$$U = \{v \in C^1(a,b) \cap C([a,b]);\ \sup_{x\in(a.b)} |v'(x)| < \infty\}$$

is endowed with the norm

$$\|u\|_U = \max_{x\in[a,b]} |u(x)| + \sup_{x\in(a,b)} |u'(x)|,$$

and the space $V = C([a,b])$ is equipped with the maximum norm (A.17). In this case the situation is simple, since

$$\begin{aligned}\delta > \|u-g\|_U &= \max_{x\in[a,b]} |u(x) - g(x)| + \sup_{x\in(a,b)} |u'(x) - g'(x)| \\ &\geq \sup_{x\in(a,b)} |u'(x) - g'(x)| = \|\varphi(u) - \varphi(g)\|_V\end{aligned}$$

for all $g \in U$ and $\delta > 0$. Thus for every $\epsilon > 0$ it is sufficient to define $\delta := \epsilon$ to satisfy the condition of continuity (A.22).

In numerous situations an equivalent definition of continuity, based on sequences, is practical.

Definition A.32 (Heine definition of continuity) *Let U, V be normed spaces. A function $f : U \to V$ is* continuous at $g \in U$ *if for every sequence $\{u_n\}_{n=1}^{\infty} \subset U$ such that*

$$\lim_{n\to\infty} u_n = g,$$

it holds that

$$\lim_{n\to\infty} f(u_n) = f(g).$$

We say that f is continuous in U *if it is continuous at every $g \in U$.*

Theorem A.4 *Definitions A.31 and A.32 are equivalent.*

Proof: The proof for real-valued functions can be found, e.g., in [100]. It can be generalized to functions in normed spaces easily. ■

Another important example of a continuous function is the norm $\|\cdot\|_V$ itself.

Lemma A.21 (Continuity of norm) *Let V be a normed space. The norm $\|\cdot\|_V : V \to \mathbb{R}$ is continuous in V.*

Proof: It is advantageous to use the Heine definition of continuity in this case. Let $v \in V$ and let $u_n \to v$ be an arbitrary sequence in V. It is our aim to show that for every $\epsilon > 0$ there exists an index n_ϵ such that

$$|\|u_n\|_V - \|v\|_V| < \epsilon \quad \text{for all } n > n_\epsilon.$$

Recall the well known "backward triangular inequality" for real numbers,

$$||a| - |b|| \le |a - b| \qquad \text{for all } a, b \in \mathbb{R}. \tag{A.25}$$

Using the triangular inequality for a general norm $\|\cdot\|_V$, this inequality can easily be extended to $|\|u_a\|_V - \|u_b\|_V| \le \|u_a - u_b\|_V$ for all $u_a, u_b \in V$. In particular,

$$|\|u_n\|_V - \|v\|_V| \le \|u_n - v\|_V.$$

Hence, for given ϵ it is sufficient to take n_ϵ from the definition of convergence $u_n \to v$. ■

Next let us introduce an important result for linear operators, which states that the continuity at one element is equivalent to the continuity in the entire space:

Lemma A.22 *Let U, V be normed spaces and $L \in \mathcal{L}(U, V)$. The operator L is continuous in the entire space U if and only if it is continuous at least at one element $u \in U$.*

Proof: The first implication is obvious. Next, without loss of generality, we show that the continuity of L at $0 \in U$ implies its continuity at an arbitrary $u \in U$. Let $\{u_n\}_{n=1}^{\infty} \subset U$ be an arbitrary sequence with the limit u. The following holds:

$$\lim_{n\to\infty} u_n = u \qquad \Leftrightarrow \qquad \lim_{n\to\infty} v_n = 0,$$

where $v_n = u_n - u$ for all n. The continuity of L at zero (Heine definition) implies that

$$0 = \lim_{n\to\infty} Lv_n = \lim_{n\to\infty} L(u_n - u) = \lim_{n\to\infty} (Lu_n - Lu) = \lim_{n\to\infty} Lu_n - Lu.$$

■

An interesting consequence of Lemma A.22 is that if there is some $u \in U$ where a linear operator L is not continuous, then L cannot be continuous at any other element $v \in U$.

Lemma A.22 further easily implies the continuity of linear operators in finite-dimensional spaces:

Proposition A.3 *Let U, V be finite-dimensional normed spaces. Then every linear operator $L \in \mathcal{L}(U, V)$ is continuous.*

Proof: Let $\dim(U) = m$, $\dim(V) = n$ and B_U, B_V be some bases in the spaces U and V, respectively. By Lemma A.11 every $L \in \mathcal{L}(U, V)$ can be represented by an $n \times m$ matrix M_L so that

$$\langle L(g) \rangle_{B_V} = M_L \langle g \rangle_{B_U} \qquad \text{for all } g \in U.$$

An arbitrary sequence $\{g_k\}_{k=1}^{\infty} \subset U$ satisfies

$$g_k \to 0 \quad \Rightarrow \quad \langle g_k \rangle_{B_U} \to \underbrace{(0, 0, \ldots, 0)}_{m \text{ times}}^T,$$

and therefore

$$M_L \langle g_k \rangle_{B_U} \to \underbrace{(0, 0, \ldots, 0)}_{n \text{ times}}^T.$$

Thus the operator L is continuous at $0 \in U$, and Lemma A.22 yields the continuity in the entire space U. ■

Proposition A.3 will be used to prove that all norms in a finite-dimensional space are equivalent (Theorem A.5).

A.2.5 Operator norm and $\mathcal{L}(U, V)$ as a normed space

Let U and V be normed spaces. In this paragraph we show that the space $\mathcal{L}(U, V)$, containing all linear operators from U to V, is a normed space. This allows us to consider the convergence and limit for sequences of linear operators, and define closed and open sets of linear operators. This methodology finds important application in the numerical solution of operator equations (including integral equations, PDEs and integro-differential equations), where typically some complicated operator L is approximated by means of a sequence of simpler operators L_n that converge to L in some appropriate sense.

Definition A.33 (Operator norm, bounded operator) *Let U, V be normed spaces. The norm in $L \in \mathcal{L}(U, V)$, called* operator norm, *is defined by*

$$\|L\| = \sup_{0 \neq u \in U} \frac{\|Lu\|_V}{\|u\|_U}. \tag{A.26}$$

An operator $L \in \mathcal{L}(U, V)$ is said to be bounded *if $\|L\| < \infty$.*

An equivalent definition of the operator norm,

$$\|L\| = \sup_{\|u\|_U = 1} \|Lu\|_V, \tag{A.27}$$

can easily be obtained from (A.26).

Indeed (A.27) is a norm in $\mathcal{L}(U, V)$: If L is nonzero then there exists at least one $0 \neq u \in U$ such that $Lu \neq 0$, and since $\|\cdot\|_V$ is a norm, it is $\|Lu\|_V > 0$ then. The second property in Definition A.24 also follows from the fact that $\|\cdot\|_V$ is a norm. The triangular inequality reads

$$\|L_1 + L_2\| = \sup_{\|u\|_U=1} \|L_1 u + L_2 u\|_V \leq \sup_{\|u\|_U=1} (\|L_1 u\|_V + \|L_2 u\|_V)$$

$$\leq \sup_{\|u\|_U=1} \|L_1 u\|_V + \sup_{\|u\|_U=1} \|L_2 u\|_V = \|L_1\| + \|L_2\|$$

for all $L_1, L_2 \in \mathcal{L}(U, V)$. Also the following proposition is trivial, but frequently used.

Proposition A.4 *Let U, V be normed spaces and $L \in \mathcal{L}(U, V)$. The following holds:*

$$\|Lu\|_V \leq \|L\| \|u\|_U \tag{A.28}$$

for all $u \in U$.

Proof: This follows immediately from the definition of the supremum in (A.26). ■

Lemma A.23 (Composition of linear operators) *Let U, V, and W be normed spaces and $F \in \mathcal{L}(U, V)$, $G \in \mathcal{L}(V, W)$ bounded linear operators. Then the composition $G \circ F$ also is a bounded linear operator, and $\|G \circ F\| \leq \|G\| \|F\|$.*

Proof: By (A.28) it is $\|(G \circ F)u\| = \|G(Fu)\| \leq \|G\| \|Fu\| \leq \|G\| \|F\| \|u\|$ for all $u \in V$. The conclusion follows. ■

Lemma A.24 (Equivalence of continuity and boundedness) *Let U, V be normed spaces and $L \in \mathcal{L}(U, V)$. Then L is continuous if and only if it is bounded.*

Proof: First assume that L is not bounded. Thus there exists a bounded sequence $\{u_n\}_{n=1}^{\infty} \subset U$ such that $\|Lu_n\| \to \infty$. Without loss of generality, we can assume that $Lu_n \neq 0$ for all n. Define a new sequence

$$v_n = \frac{u_n}{\|Lu_n\|},$$

which satisfies

$$v_n \to 0 \quad \& \quad \|Lv_n\| = 1 \text{ for all } n.$$

This means that L is not continuous at 0, and therefore it is not continuous anywhere in U.

Conversely, assume that the operator L is bounded. Then there exists a positive constant C such that

$$\|Lu\|_V \leq C \|u\|_U$$

for all $u \in U$. Thus L is continuous at $0 \in U$ and, by Lemma A.22, in the whole space U.■

A.2.6 Equivalence of norms

Since a linear space V may be endowed with many different norms, it is natural to ask the following questions:

- Does a sequence $\{u_n\}_{n=1}^{\infty} \subset V$, which converges in some norm, remain convergent in another norm?
- Does a set $S \subset V$, which is open in some norm, remain open in another norm?

The answer is related to the notion of equivalent norms.

Definition A.34 (Equivalent norms) *Let V be a normed space and $\|\cdot\|_{V,1}$ and $\|\cdot\|_{V,2}$ norms in V. We say that these norms are* equivalent *if there exist positive constants C_1 and C_2 such that*

$$C_1\|\cdot\|_{V,1} \le \|\cdot\|_{V,2} \le C_2\|\cdot\|_{V,1}$$

for all $v \in V$.

Proposition A.5 *Let V be a normed space and $\|\cdot\|_{V,1}$ and $\|\cdot\|_{V,2}$ norms which are equivalent in V.*

1. *Let $v \in V$. The sequence $\{u_n\}_{n=1}^{\infty} \subset V$ converges to v in the norm $\|\cdot\|_{V,1}$ if and only if it converges to v in the norm $\|\cdot\|_{V,2}$.*
2. *Any subset $S \subset V$ is open in the norm $\|\cdot\|_{V,1}$ if and only if it is open in the norm $\|\cdot\|_{V,2}$.*

Proof: Left to the reader as an easy exercise. ■

Couple of equivalent norms is shown in Examples A.30 and A.31.

■ EXAMPLE A.30 (Equivalent norms in Euclidean spaces)

Let $U = \mathbb{R}^n$. The discrete maximum norm (A.10) and the discrete integral norm (A.11) are equivalent, since

$$\|u\|_\infty \le \|u\|_1 \le n\|u\|_\infty \tag{A.29}$$

for all $u \in U$. The Euclidean norm (A.13) is equivalent to the discrete maximum norm since

$$\|u\|_\infty \le \|u\|_2 \le \sqrt{n}\|u\|_\infty \tag{A.30}$$

for all $v \in U$. This, obviously, makes the norms $\|\cdot\|_1$ and $\|\cdot\|_e$ equivalent.

The situation is even more interesting in polynomial spaces, where one can practise elementary estimates:

■ EXAMPLE A.31 (Equivalent norms in polynomial spaces)

Consider, for example, the space $U = P^k([-1,1])$ with the maximum norm (A.17) and the integral norm (A.18). We will show that these two norms are equivalent in U. The first inequality is easy and it does not even require U to be finite-dimensional:

$$\|f\|_1 = \int_{-1}^{1} |f(x)|\,\mathrm{d}x \le \int_{-1}^{1} \max_{-1\le x\le 1} |f(x)|\,\mathrm{d}x = 2\max_{-1\le x\le 1} |f(x)| = 2\|f\|_\infty. \tag{A.31}$$

When $f(x) = 1$ or $f(x) = -1$, (A.31) becomes an equality. Hence the constant 2 cannot be improved (the estimate is sharp).

The other inequality requires to go deeper into the structure of the polynomial space U, and we shall use the Legendre polynomials for this purpose. These polynomials will be discussed in more detail in Example A.44. Now we need their following properties:

1. The $k+1$ Legendre polynomials $L_0, L_1, \ldots, L_k$ form a basis of $P^k([-1,1])$.
2. It is $-1 \le L_m(x) \le 1$ in $[-1,1]$ for all $m = 0, 1, 2, \ldots$.
3. The following holds:

$$\int_{-1}^{1} L_i(x)L_j(x)\,\mathrm{d}x = \delta_{ij} \qquad \text{for all } i, j \ge 0,$$

   where δ_{ij} is the Kronecker delta.

By the first property, every $f \in U$ can be written as

$$f(x) = \sum_{j=0}^{k} \alpha_j L_j(x),$$

and thus for any $0 \le m \le k$ we can estimate

$$\begin{aligned} \|f\|_1 &= \int_{-1}^{1} |f(x)|\,\mathrm{d}x \ge \int_{-1}^{1} |f(x)|\underbrace{|L_m(x)|}_{\le 1}\,\mathrm{d}x \ge \left|\int_{-1}^{1} f(x)L_m(x)\,\mathrm{d}x\right| \\ &= \left|\int_{-1}^{1}\sum_{j=0}^{k}\alpha_j L_j(x)L_m(x)\,\mathrm{d}x\right| = \left|\sum_{j=0}^{k}\alpha_j \underbrace{\int_{-1}^{1} L_j(x)L_m(x)\,\mathrm{d}x}_{\delta_{jm}}\right| = |\alpha_m|. \end{aligned} \tag{A.32}$$

Summing the inequalities (A.32) over all $m = 0, 1, \ldots, k$, we obtain

$$\begin{aligned} (k+1)\|f\|_1 &\ge \sum_{m=0}^{k} |\alpha_m| \ge \max_{-1\le x\le 1} \sum_{m=0}^{k} |\alpha_m| \underbrace{|L_m(x)|}_{\le 1} \\ &\ge \max_{-1\le x\le 1}\left|\sum_{m=0}^{k}\alpha_m L_m(x)\right| = \max_{-1\le x\le 1}|f(x)| = \|f\|_\infty. \end{aligned}$$

Thus we conclude that

$$\frac{1}{k+1}\|f\|_\infty \le \|f\|_1 \le 2\|f\|_\infty,$$

which means that the norms (A.17) and (A.18) are equivalent in U.

The following theorem confirms the intuition that we gained in the previous two examples:

Theorem A.5 (Finite-dimensional case) *Let U be a finite-dimensional normed space. Then all norms in U are equivalent.*

Proof: Consider a linear space U of dimension n, and two arbitrary norms $\|\cdot\|_{U,1}$ and $\|\cdot\|_{U,1}$. For clarity, by U_1 and U_2 denote the normed spaces obtained when U is equipped with the norms $\|\cdot\|_{U,1}$ and $\|\cdot\|_{U,2}$, respectively. Consider the identity operator $\mathcal{I}_1 : U_1 \to U_2$, $\mathcal{I}_1 u = u$ for all $u \in U$. This indeed is a linear operator, which by Proposition A.3 is continuous. Thus by Lemma A.24 it is bounded, and there exists some $C_1 > 0$ such that

$$\|\mathcal{I}_1\| = \sup_{0 \neq u \in U} \frac{\|u\|_{U,2}}{\|u\|_{U,1}} \leq C_1.$$

This means that we have

$$\|u\|_{U,2} \leq C_1 \|u\|_{U,1} \qquad \text{for all } u \in U.$$

Analogously there is a positive constant $C_2 > 0$,

$$\|\mathcal{I}_2\| = \sup_{0 \neq u \in U} \frac{\|u\|_{U,1}}{\|u\|_{U,2}} \leq C_2,$$

for the identity operator $\mathcal{I}_2 : U_2 \to U_1$. Thus we have

$$\frac{1}{C_2}\|u\|_{U,1} \leq \|u\|_{U,2} \leq C_1 \|u\|_{U,1} \qquad \text{for all } u \in U,$$

which concludes the proof. ■

However, this result does not extend to infinitely-dimensional normed spaces, as the following example shows:

■ EXAMPLE A.32 (Nonequivalent norms and the convergence of sequences)

Let V be the space of bounded integrable functions defined in the interval $[0, 1]$. Consider the maximum norm (A.17) and the p-norm (A.19),

$$\|f\|_\infty = \max_{x \in [0,1]} |f(x)|, \qquad \|f\|_p = \left(\int_0^1 |f(x)|^p \, \mathrm{d}x \right)^{\frac{1}{p}},$$

where $1 \leq p < \infty$. Define a sequence of functions $\{f_n\}_{n=1}^\infty \subset V$ as

$$f_n(x) = \begin{cases} 1, & x \in \left(0, \dfrac{1}{n}\right), \\ 0 & \text{elsewhere.} \end{cases}$$

Clearly the only candidate for a limit is the zero function. However, it is

$$\lim_{n \to \infty} \|f_n\|_\infty = \lim_{n \to \infty} 1 = 1.$$

At the same time,

$$\lim_{n\to\infty} \|f_n\|_p = \lim_{n\to\infty} n^{-\frac{1}{p}} = 0$$

for all $1 \leq p < \infty$. It is easy to show the nonequivalence of the norms using the sequence $\{f_n\}_{n=1}^{\infty}$ and Definition A.34.

Also the following example is related to nonequivalent norms:

■ **EXAMPLE A.33 (Nonequivalent norms and open and closed sets)**

Let $V = C([a,b])$, where $(a,b) \subset \mathbb{R}$. Consider now the maximum norm (A.17) and the integral norm (A.18). The set

$$S = \{f \in V;\ 0 \leq |f(x)| \leq 1;\ f(a) = 0;\ f(b) = 1\}$$

is closed in the maximum norm, since $0 \leq \|f\|_\infty \leq 1$ for all $f \in S$. However, it is open in the integral norm, since $0 < \|f\|_p < (b-a)^{1/p}$ for all $f \in S$. The situation is illustrated in Figure A.14.

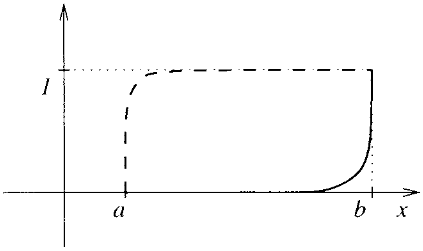


Figure A.14 The set S does not contain the functions $f_1(x) = 0$ and $f_2(x) = 1$. Therefore it is $0 < \|f\|_p < (b-a)^{1/p}$ for all $f \in S$ and S is open in the integral norm.

A.2.7 Banach spaces

In normed spaces one finds sequences which exhibit all signs of convergence except that they miss a limit in V (as it was the case, e.g., in Example A.23). Let us look closer at this behavior. The following definition explains what we mean by "all signs of convergence":

Definition A.35 (Cauchy sequence) *Let V be a normed space. A sequence $\{u_n\}_{n=1}^{\infty} \subset V$ is said to be a* Cauchy sequence *if for every $\epsilon > 0$ there exists an index n_0 such that*

$$\|u_n - u_m\|_V \leq \epsilon \quad \text{for all } n,m \geq n_0. \tag{A.33}$$

Here is a trivial observation on convergent sequences:

Proposition A.6 *Let V be a normed space. Every sequence $\{u_n\}_{n=1}^{\infty} \subset V$ that converges to some element $v \in V$ is a Cauchy sequence.*

Proof: This is an easy exercise using Definition A.27. ■

A nonconvergent Cauchy sequence is shown in Example A.34.

■ EXAMPLE A.34 (Nonconvergent Cauchy sequence)

Consider the linear space $C([0,2])$ endowed with the integral norm (A.18), and a sequence of functions $\{f_n\}_{n=1}^{\infty} \subset V$ defined as

$$f_n(x) = \begin{cases} 1, & x \in (0,1), \\ 1+n-nx, & x \in \left[1, 1+\dfrac{1}{n}\right], \\ 0, & x \in \left(1+\dfrac{1}{n}, 2\right). \end{cases}$$

The sequence f_n is depicted in Figure A.15.

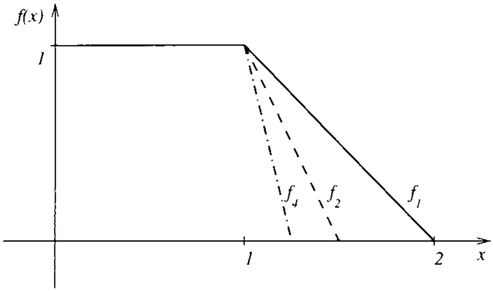


Figure A.15 Nonconvergent Cauchy sequence in the space $C([0,2])$.

It is easy to calculate

$$\|f_n - f_m\| = \frac{1}{2}\left|\frac{1}{n} - \frac{1}{m}\right|$$

and to verify the Cauchy property

$$\lim_{m\to\infty}\left(\lim_{n\to\infty}\|f_n - f_m\|\right) = 0.$$

However, the sequence $\{f_n\}_{n=1}^{\infty}$ does not have a limit in the space $C([0,2])$.

The class of linear spaces where this cannot happen was first defined in the dissertation of a Polish mathematician Stefan Banach in 1920. S. Banach is assumed to be one of the founders of modern functional analysis. He made major contributions to the theory of topological spaces, measure and integration theory, set theory, and the analysis of orthogonal series.

Figure A.16 Stefan Banach (1892–1945).

Definition A.36 (Banach space) *A normed space V is said to be* Banach space *if for every Cauchy sequence $\{v_n\}_{n=1}^{\infty} \subset V$ there exists an element $v \in V$ such that* $\lim_{n\to\infty} v_n = v$.

Here are a few examples of Banach spaces:

1. the real or complex n-dimensional Euclidean space $\mathbb{R}^n$ with the discrete maximum norm (A.10) or the discrete p-norm (A.12), whose special cases are the discrete integral norm (A.11) and the Euclidean norm (A.13),
2. the space $\mathbb{R}^{n\times n}$ of real or complex matrices with the norms analogous to the previous case, for example with the Euclidean norm (A.16).
3. the space $V = l^p$ of infinite real sequences with the discrete p-norm

$$\|\{u_n\}_{n=1}^{\infty}\|_V = \left(\sum_{n=1}^{\infty} |u_n|^p\right)^{\frac{1}{p}}.$$

   where $1 \le p < \infty$,
4. the space $P^k([a,b])$ with the maximum norm (A.17),
5. the space $C([a,b])$ with the maximum norm (A.17).

A sufficient condition for a normed linear space to be Banach space is mentioned in Lemma A.25.

Definition A.37 (Reflexivity) *A normed space V is said to be* reflexive *if the dual space to its dual is V itself, i.e., if $(V')' = V$.*

Lemma A.25 *Every reflexive normed space is a Banach space.*

Proof: This proof can be found in most textbooks on functional analysis, e.g., in [65]. ∎

Let us remark that there exist Banach spaces which are not reflexive. We already know that with U, V being normed space, the space $\mathcal{L}(U,V)$ is a normed space as well. The following theorem moreover gives a sufficient condition for $\mathcal{L}(U,V)$ to be Banach space.

Theorem A.6 (Conditions for $\mathcal{L}(U,V)$ to be a Banach space) *Let U be a normed space and V be a Banach space. Then $\mathcal{L}(U,V)$ is a Banach space.*

Proof: Let $\{L_n\}_{n=1}^{\infty}$ be a Cauchy sequence in $\mathcal{L}(U,V)$ and $u \in U$ arbitrary. We show that $\{L_n u\}_{n=1}^{\infty}$ is a Cauchy sequence in V:

This is clear if $u = 0$. If $u \neq 0$ consider an arbitrary $\epsilon > 0$. There exists an index n_0 so that $\|L_m - L_n\| \leq \epsilon/\|u\|_U$ for all $m, n \geq n_0$. Then

$$\|L_m u - L_n u\|_V \leq \|L_m - L_n\|\|u\|_U \leq \epsilon \qquad \text{for all } m, n \geq n_0.$$

and indeed $\{L_n u\}_{n=1}^{\infty}$ is a Cauchy sequence. Since V is a Banach space, this sequence converges to some element in V. Thus we can define the limit L of the sequence $\{L_n\}_{n=1}^{\infty}$ by

$$Lu = \lim_{n\to\infty} L_n u \qquad \text{for all } u \in U.$$

It is easy to see that L is linear, and thus it remains to be shown that $\|L\| \leq \infty$:

Let $\epsilon > 0$ be given. There exists some index n_1 such that $\|L_m - L_n\| \leq \epsilon/2$ for all $m, n \geq n_1$. Therefore for every $u \in U$ we have

$$\|L_m u - L_n u\|_V \leq \|L_m - L_n\|\|u\|_U \leq \frac{\epsilon\|u\|_U}{2} \qquad \text{for all } m, n \geq n_1.$$

and

$$\|Lu - L_n u\|_V = \left\|\lim_{m\to\infty} L_m u - L_n u\right\|_V \leq \lim_{m\to\infty} \|L_m u - L_n u\|_V \leq \frac{\epsilon\|u\|_U}{2}$$

for all $n \geq n_1$. Finally,

$$\|Lu\|_V = \|L_n u + Lu - L_n u\|_V \leq \|L_n u\|_V + \|Lu - L_n u\|_V \leq \left(\|L_n\| + \frac{\epsilon}{2}\right)\|u\|_U.$$

Thus L is a bounded linear operator. Since $\|L - L_n\| \leq \frac{\epsilon}{2}$ for all $n \geq n_1$, we have that $L_n \to L$ in $\mathcal{L}(U,V)$ as $n \to \infty$. ■

Completion of normed spaces Next let us mention the completion of normed spaces to Banach spaces.

Definition A.38 (Dense subset) *Let V be a normed space and S a subset of V. The set S is said to be* dense *in V if for every $v \in V$ there exists a sequence $\{s_n\}_{n=1}^{\infty} \subset S$ such that*

$$\lim_{n\to\infty} \|s_n - v\|_V = 0.$$

The main result is formulated in the following abstract theorem, which is followed by an illustrating example.

Theorem A.7 (Completion of normed spaces) *To every normed space U there exists a Banach space W such that:*

1. *There is a subspace $V \subset W$ and a linear bijection $\mathcal{I} : U \to V$ satisfying*

$$\|\mathcal{I}u\|_W = \|u\|_U \quad \text{for all } u \in U.$$

The operator $\mathcal{I}$ is called isometric isomorphism between the spaces U and V.

2. *The space V is dense in W.*

The space W is called completion of U, and it is unique up to an isometric isomorphism $\mathcal{I}$.

Proof: See, e.g., [99]. ■

For example, the Banach space $W = \mathbb{R}$ is defined to be the set of equivalence classes of all Cauchy sequences in the space of rational numbers $\mathbb{Q}$. The space V is then identified with Cauchy sequences in $\mathbb{Q}$ whose limit lies in $\mathbb{Q}$. The incompleteness of the space $\mathbb{Q}$ is illustrated in the following example.

■ EXAMPLE A.35 (Completion of rational numbers)

Consider the normed linear space $\mathbb{Q}$ of rational numbers endowed with the standard norm

$$\|q\|_{\mathbb{Q}} = |q|.$$

In order to see that $\mathbb{Q}$ is incomplete, let us describe the way the ancient Babylonians calculated square roots. To find a rational approximation of the square root of an integer $a > 0$, let $0 < x_0 \in \mathbb{Q}$ be such that $x_0^2 < a$. Then $x_0 \leq \sqrt{a} \leq a/x_0 \in \mathbb{Q}$. The average of these two values gives an even closer estimate,

$$x_{k+1} = \frac{x_k + a/x_k}{2} \in \mathbb{Q}, \tag{A.34}$$

as depicted in Figure A.17.

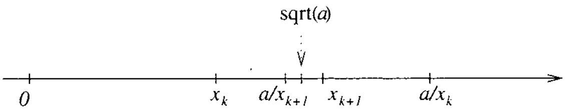


Figure A.17 Approximate calculation of a square root.

Iterating this formula, we obtain a convergent sequence $\{x_k\}_{k=0}^{\infty} \subset \mathbb{Q}$ such that

$$\lim_{k\to\infty} x_k = \sqrt{a}.$$

It is left to the reader as an exercise to prove that the sequence is convergent and that its limit is $\sqrt{a}$. There are several ways to do this, one of them using the fact that

$$\left| x_{k+1} - \frac{a}{x_{k+1}} \right| \leq \frac{1}{2} \left| x_k - \frac{a}{x_k} \right| \leq \ldots \leq \frac{1}{2^{k+1}} \left| x_0 - \frac{a}{x_0} \right| \quad \text{for all } k \geq 1. \tag{A.35}$$

The formula (A.34) is still used in modern digital calculators due to its high efficiency. Its origin sometimes is attributed to Heron of Alexandria who described it in his famous "Metric", but as a matter of fact the formula was already known to the old Babylonians.

A.2.8 Banach fixed point theorem

The concept of contractive operators and the Banach fixed point theorem play an important role in the analysis and numerical solution of nonlinear problems. Let us mention the basic results and show their applications:

Definition A.39 (Contractive operator) *Consider a Banach space V and a (not necessarily linear) operator $L : V \to V$. Then L is said to be* contraction *if there exists a real number $0 \le q < 1$ such that*

$$\|Lu - Lv\|_V \le q\|u - v\|_V \quad \textit{for all } u, v \in V. \tag{A.36}$$

It is important that the number q is strictly less than one. Moreover, if a contractive operator L is linear, we have $\|Lu\|_V \le q\|u\|_V$ for all $u \in V$, which means that

$$\|L\| \le q < 1.$$

The following theorem holds for both linear and nonlinear operators:

Theorem A.8 (Banach fixed point theorem) *Let V be a Banach space and $L : V \to V$ a contraction with a constant $0 \le q < 1$. Then the equation $Lx = x$ has a unique solution $x \in V$. Moreover, the sequence $\{x_n\}_{n=0}^{\infty} \subset V$ defined by $x_{n+1} = Lx_n$ for all n, converges to x for every $x_0 \in V$.*

The element $x \in V$ such that $Lx = x$ is called fixed point of L.

Proof: Let us choose an arbitrary element $x_0 \in V$ and define a sequence $\{x_n\}_{n=0}^{\infty}$, $x_{n+1} = Lx_n$. We begin with showing that this is a Cauchy sequence:

$$\begin{aligned}
\|x_{n+k} - x_n\| &= \|x_{n+1} - x_n + x_{n+2} - x_{n+1} + \ldots + x_{n+k} - x_{n+k-1}\| \\
&\le \sum_{r=1}^{k} \|x_{n+r} - x_{n+r-1}\| \\
&\le \left(\sum_{r=1}^{k} q^{n+r-1}\right) \|x_1 - x_0\| \\
&= q^n \frac{1 - q^k}{1 - q} \|x_1 - x_0\| \\
&\le \frac{q^n}{1 - q} \|x_1 - x_0\|.
\end{aligned}$$

Hence, for an arbitrarily small $\epsilon > 0$ we always can find an index n_0 such that

$$\frac{q^{n_0}}{1 - q} \|x_1 - x_0\| \le \epsilon,$$

and we see that $\|x_m - x_n\| \le \epsilon$ for all $m, n \ge n_0$. Thus the sequence $\{x_n\}_{n=0}^{\infty}$ has a limit in V that can be denoted by x. It remains to be shown that $Lx = x$:

$$\begin{aligned} \|Lx - x\| &= \|Lx - x_{n+1} + x_{n+1} - x\| \\ &\le \|Lx - x_{n+1}\| + \|x_{n+1} - x\| \\ &= \|Lx - Lx_n\| + \|x_{n+1} - x\| \\ &\le q\|x - x_n\| + \|x_{n+1} - x\|. \end{aligned}$$

Since both $\|x - x_n\|$ and $\|x_{n+1} - x\|$ converge to zero as $n \to \infty$, necessarily it is $Lx = x$. Suppose that there exist two different elements x and y in V such that $Lx = x$ and $Ly = y$. Subtracting these relations and taking the norm, we obtain

$$\|x - y\| = \|Lx - Ly\| \le q\|x - y\| < \|x - y\|,$$

which is a contradiction. Thus we conclude that the element $x \in V$ is unique. ■

The procedure for finding the fixed point $x \in V$ of a contractive operator L, which was used in the proof of Theorem A.8, is called fixed point iteration. The next lemma says that under special circumstances the operator L does not even have to be contractive in the whole space V to have a fixed point:

Theorem A.9 (Local fixed point theorem) *Let V be a Banach space and S its closed subset such that $L(S) \subset S$ (i.e., $Ls \in S$ for all $s \in S$). Further, let $L : V \to V$ be an operator which satisfies locally in S the contraction condition*

$$\|Lu - Lv\|_V \le q\|u - v\|_V \qquad \text{for all } u, v \in S \tag{A.37}$$

with some $0 \le q < 1$. Then the equation $Lx = x$ has a unique solution $x \in S$. Moreover, the sequence $\{x_n\}_{n=0}^{\infty}$, $x_{n+1} = Lx_n$, converges to x for an arbitrary $x_0 \in S$.

Proof: Let x_0 be arbitrary element of S. By induction, the sequence $\{x_n\}_{n=0}^{\infty}$ lies in S, and so does its limit x by the closedness of S. For the rest see the proof of Theorem A.8. ■

Many applications of the fixed point theorem are related to the solution of nonlinear problems. Let us present a few examples for illustration:

■ EXAMPLE A.36 (Fixed point iteration)

1. Let $V = \mathbb{R}$. Consider an arbitrary real function $g : V \to V$ which is Lipschitz-continuous with a constant $0 \le q < 1$, i.e., such that

$$|g(x) - g(y)| \le q|x - y| \tag{A.38}$$

   for all $x, y \in \mathbb{R}$. This requirement obviously is satisfied, e.g., by all smooth functions whose derivative satisfies $|g'(x)| \le q$ for all $x \in \mathbb{R}$. Theorem A.8 guarantees the existence of a unique solution to the equation $g(x) = x$. The solution can be found via the fixed point iteration $x_{n+1} = g(x_n)$, starting from an arbitrary $x_0 \in \mathbb{R}$.

2. Now let us apply the fixed point iteration procedure to find all real solutions of the equation

$$x^3 + x - 1 = 0.$$

The easiest way to transform this equation into the form $g(x) = x$ is to define $g(x) = 1 - x^3$. However, then there exists no finite $q > 0$ for condition (A.38) to hold, since $|g'(x)| \to \infty$ as $x \to \infty$. Another attempt,

$$g(x) = \frac{1}{1 + x^2} = x,$$

is successful since $|g'(x)| \le 0.7$ for all $x \in V = \mathbb{R}$. Now Theorem A.8 yields the existence of a unique solution $x \in \mathbb{R}$. Again, the solution can be found via the fixed point iteration $x_{n+1} = g(x_n)$, starting from an arbitrary $x_0 \in \mathbb{R}$. The situation is shown in Figure A.18.

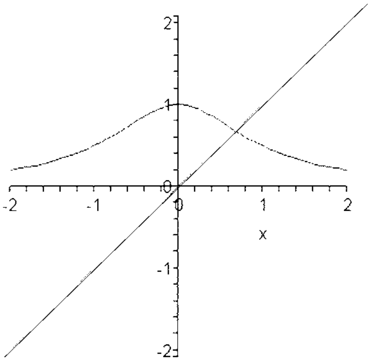


Figure A.18 Solution of the equation $x^3 + x - 1 = 0$ via fixed point iteration in $\mathbb{R}$ (Theorem A.8).

3. Next let us solve the nonlinear equation

$$x - \cos(x) = 0$$

in $V = \mathbb{R}$. The original Banach Theorem A.8 cannot be applied since there exist points $x_k = \pi/2 + k\pi$ such that $|\cos'(x_k)| = 1$, and thus condition (A.38) does not hold for any $0 \le q < 1$. This can be shown easily using some arbitrary sequence $\{y_m\}_{m=1}^{\infty}$ converging to some of the points x_k. Fortunately there is a remedy in the form of Theorem A.9. We can define a closed set $S = [-1, 1]$ and use the fact that $\cos(S) \subset S$. Since $|\cos'(x)| \le 0.9$ for all $x \in S$, Theorem A.9 guarantees the existence of a unique solution $x \in S$. Again, this solution can be found iteratively using an arbitrary $x_0 \in S$. The graphs of the functions $g(x)$ and x are shown in Figure A.19.

4. In the last example we visit numerical linear algebra. Consider a nonsingular real $n \times n$ matrix A and a real vector $b \in V = \mathbb{R}^n$. The Jacobi method for the solution of the system

$$Ax = b$$

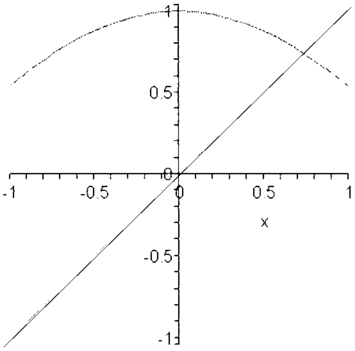


Figure A.19 Solution of the equation $x - \cos(x) = 0$ via fixed point iteration in $[-1, 1] \subset V$ (Theorem A.9).

is based on the decomposition

$$A = L + D + U$$

of the matrix A into the sum of a lower-diagonal matrix L (whose entries on and above the diagonal are zero), diagonal matrix D and upper-diagonal matrix U (whose entries on and below the diagonal are zero). The equation $(L + D + U)x = b$ can be transformed into

$$x = D^{-1}[b - (L + U)x],$$

(without loss of generality, we can assume that the diagonal entries are not zero – otherwise we perform a permutation of rows in the linear system). This leads to the fixed point iteration scheme

$$x_{n+1} = D^{-1}[b - (L + U)x_n],$$

where $x_0 \in V$ is an arbitrary initial guess. We can now define an operator $F : V \longrightarrow V$,

$$Fx = D^{-1}[b - (L + U)x].$$

F is not linear (more precisely, it is nonlinear for all $b \neq 0$), but nevertheless we can use the Banach Theorem A.8 to analyze convergence of the Jacobi method:

Let $\|\cdot\|_\infty$ be the discrete maximum norm (A.10) in V. For any matrix $M \in \mathbb{R}^{n \times n}$ and $w \in V$ we can estimate

$$\begin{aligned} \|Mw\|_\infty &= \max_{1 \leq i \leq n} \left| \sum_{j=1}^{n} m_{ij} w_j \right| \\ &\leq \max_{1 \leq i \leq n} \sum_{j=1}^{n} |m_{ij}||w_j| \\ &\leq \left(\max_{1 \leq i \leq n} \sum_{j=1}^{n} |m_{ij}| \right) \left(\max_{1 \leq j \leq n} |w_j| \right) \\ &\leq \|M\| \|w\|_\infty, \end{aligned} \tag{A.39}$$

where $\|M\|$ is the matrix norm (A.15),

$$\|M\| = \max_{1 \le i \le n} \sum_{j=1}^{n} |m_{ij}|.$$

Let $u, v \in V$. Substituting $D^{-1}(L+U)$ for M and $u-v$ for w in (A.39), we obtain

$$\|Fu - Fv\|_\infty = \|D^{-1}(L+U)(u-v)\|_\infty \le \|D^{-1}(L+U)\| \|u-v\|_\infty.$$

Hence the operator F is a contraction if $\|D^{-1}(L+U)\| < 1$. Looking at the structure of the matrix $D^{-1}(L+U)$, we have

$$\|D^{-1}(L+U)\| = \max_{1 \le i \le n} \frac{1}{|d_{ii}|} \sum_{j=1}^{n} |l_{ij} + u_{ij}| = \max_{1 \le i \le n} \frac{1}{|a_{ii}|} \sum_{j=1, j \ne i}^{n} |a_{ij}|.$$

Thus a sufficient condition for the operator F to be contraction is

$$|a_{ii}| > \sum_{j=1, j \ne i}^{n} |a_{ij}| \quad \text{for all } 1 \le i \le n. \tag{A.40}$$

Every matrix A with this property is called strictly diagonally dominant (SDD). We have shown using Theorem A.8 that the Jacobi method applied to any SDD matrix converges to the solution x of the linear system $Ax = b$ for any right-hand side $b \in \mathbb{R}^n$.

A.2.9 Lebesgue integral and L^p-spaces

The Lebesgue L^p-spaces have a prominent position within the class of Banach spaces because of their importance for the study of partial differential equations. Henri Léon Lebesgue was a French mathematician who generalized the Riemann integration and established the basis of modern measure and integration theory.

Figure A.20 Henri Léon Lebesgue (1875–1941).

Because of space constraints, we only can summarize the basic ideas of the Lebesgue integration theory in the next paragraph. For a systematic introduction we refer the reader to [28, 98] and [99].

A few loose words about the Lebesgue integration theory The measure of a set is a rigorous definition of its volume that remains exact even for very complicated sets whose volume in the traditional sense is difficult to imagine. The definition of the measure of sets precedes the definition of the integral. The Lebesgue measure, which is used to define the Lebesgue integral, is an alternative to the Jordan measure upon which the Riemann integral was built. There exist rare sets whose Lebesgue measure is undefined, but the reader does not have to worry about encountering them, since this is is almost impossible in practice. In what follows, all our considerations are based on the Lebesgue measure, Lebesgue-measurable sets and Lebesgue integrals of functions defined in such sets. We shall not repeat the name of Lebesgue or the measurability assumption anymore.

We shall say that a set $\Omega_0 \subset \mathbb{R}^d$ has zero measure in $\mathbb{R}^d$ if its d-dimensional measure is zero. For example, the d-dimensional measure of a set Ω_0 consisting of a finite number of points or even of a countable infinite set of points, such as the set of rational numbers, is zero if $d \geq 1$. The one-dimensional measure of the interval (a, b) equals to $b - a$, but its two-dimensional measure as an edge of the square $(a, b)^2 \subset \mathbb{R}^2$, or its three-dimensional measure as an edge of the cube $(a, b)^3 \subset \mathbb{R}^3$, is zero. Analogously the measure of any curve $\varphi : (a, b) \to \mathbb{R}^d$ is zero in $\mathbb{R}^d$ if $d \geq 2$. The three-dimensional measure of sets consisting of up to countable infinite number of points, one-dimensional curves or two-dimensional surfaces is zero.

Let $f : \Omega \to \mathbb{R}$, where $\Omega \subset \mathbb{R}^d$ is an open measurable set. The Lebesgue integral of f over Ω is invariant with respect to the values of the function in zero-measure subsets of Ω. This can be written as follows,

$$\int_{\Omega} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{\Omega\setminus\Omega_0} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \qquad \text{for all } \Omega_0 \subset \Omega,\ |\Omega_0| = 0. \tag{A.41}$$

So, for example, it does not matter whether the integral is performed over an open set $\Omega \subset \mathbb{R}^d$ or over its closure $\overline{\Omega}$. Or, when integrating a function g in an interval $(a, b) \subset \mathbb{R}$, it does not matter what values it attains on rational numbers. Because of the above-described properties the Lebesgue integral is defined for functions where the Riemann integration fails, as shown in the following example.

■ EXAMPLE A.37 (Riemann vs. Lebesgue integral)

Let $\Omega = (a, b) \subset \mathbb{R}$ be a nonempty bounded interval. Define $\Omega_0 = \Omega \cap \mathbb{Q}$ and consider the function

$$g(x) = \begin{cases} 1, & x \in \Omega_0, \\ -1, & x \in \Omega \setminus \Omega_0, \end{cases} \tag{A.42}$$

where $\mathbb{Q}$ is the set of rational numbers. The graph of this function is shown in Figure A.21.

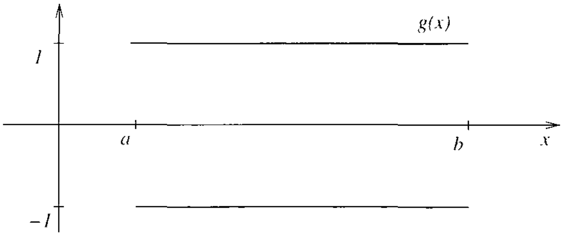


Figure A.21 Graph of the function (A.42).

To define the Riemann integral, cover the interval (a, b) with a partition $a = x_0 < x_1 < \ldots < x_n = b$ such that $x_i - x_{i-1} \leq (b-a)/n$ for all $1 \leq i \leq n$. The Riemann integral is defined as

$$(R)\int_{\Omega} g(x)\,\mathrm{d}x = \lim_{n \to \infty} \sum_{i=1}^{n} g(\xi_i)(x_i - x_{i-1}), \tag{A.43}$$

where ξ_i is an arbitrary point in the subinterval (x_{i-1}, x_i). Clearly the limit (A.43) does not exist, because ξ_i always can either be minus one or one (and thus the result of the Riemann integration can be anything between $-(b-a)$ and $b-a$). The Lebesgue integral, according to (A.41), yields a unique result,

$$(L)\int_{\Omega} g(x)\,\mathrm{d}x = \int_{\Omega\setminus\Omega_0} \underbrace{g(x)}_{=-1}\,\mathrm{d}x + \underbrace{\int_{\Omega_0} g(x)\,\mathrm{d}x}_{=0 \text{ since } |\Omega_0|=0} = -(b-a).$$

We shall say "almost everywhere (a.e.) in Ω" meaning "everywhere in Ω up to a subset $\Omega_0 \subset \Omega$, where $|\Omega_0| = 0$". Two functions f and $\tilde{f}$ defined almost everywhere in Ω are said to be equivalent in the Lebesgue sense if $f = \tilde{f}$ in $\Omega \setminus \Omega_0$, where $|\Omega_0| = 0$. In the above example, the function $g(x)$ was equivalent to $\tilde{g}(x) = -1$.

Definition A.40 (**L^p-norms and L^p-spaces**) *Let $\Omega \subset \mathbb{R}^d$ be an open set. Consider the linear space V of measurable functions defined in Ω. For every $1 \leq p < \infty$ we define the L^p-norm in V as*

$$\|f\|_p = \left(\int_{\Omega} |f(\boldsymbol{x})|^p d\boldsymbol{x} \right)^{\frac{1}{p}}. \tag{A.44}$$

The L^∞ norm in V is defined as

$$\|f\|_\infty = \operatorname{ess}\sup_{\boldsymbol{x}\in\Omega} |f(\boldsymbol{x})|, \tag{A.45}$$

where the essential supremum of a function is defined as

$$\operatorname{ess}\sup_{\boldsymbol{x}\in\Omega} g(\boldsymbol{x}) = \inf_{Z\subset\Omega, |Z|=0}\ \sup_{\Omega\setminus Z} g(\boldsymbol{x}).$$

The spaces $L^p(\Omega)$ are defined as

$$L^p(\Omega) = \{f \in V;\ \|f\|_p < \infty\} \qquad \textit{for all } 1 \le p < \infty$$

and

$$L^\infty(\Omega) = \{f \in V;\ \operatorname*{ess\ sup}_{\boldsymbol{x} \in \Omega} |f| < \infty\}.$$

Despite its rather complicated definition, the essential supremum is just the "supremum that disregards extrema in zero-measure subsets". If the essential supremum of a function is finite, this function is said to be essentially bounded.

Of course the reader has the right to ask if the relations (A.44) and (A.45) define norms: We are going to prove the corresponding triangular inequalities to Paragraph A.2.10. Prior to manipulating with L^p-functions, however, let us get some feeling for their shapes.

Shape of L^p functions in $\Omega \subset \mathbb{R}$

1. Let $\Omega = (a, b) \subset \mathbb{R}$ be bounded. Then every essentially bounded function $f : \Omega \to \mathbb{R}$, $|f(x)| \le C$ a.e. in Ω, belongs to $L^p(\Omega)$ for all $1 \le p \le \infty$. This is obvious for $p = \infty$, and for $p \in [1, \infty)$ we can estimate

$$\|f\|_p = \left(\int_\Omega \underbrace{|f(x)|^p \, \mathrm{d}x}_{\le C^p} \right)^{\frac{1}{p}} \le C|\Omega|^{\frac{1}{p}} < \infty. \tag{A.46}$$

2. In unbounded sets $\Omega \subset \mathbb{R}$, essentially bounded functions still lie in $L^\infty(\Omega)$ (that is the definition). But generally they do not lie in $L^p(\Omega)$ for $p \in [1, \infty)$. This is obvious when taking the function $f(x) = 1$ and rewriting (A.46) for a set $\Omega \subset \mathbb{R}$, $|\Omega| = \infty$.

3. One of the main purposes for L^p-spaces is to control the strength of singularities. Consider, for example, the interval $\Omega = (0, 1)$ and the function $f(x) = 1/x^\alpha$. Then the L^p-norm of $f(x)$ is

$$\|f\|_p = \left(\int_\Omega \left| \frac{1}{x^\alpha} \right|^p \mathrm{d}x \right)^{\frac{1}{p}} = \begin{cases} \dfrac{1}{(1 - \alpha p)^{\frac{1}{p}}} & \alpha < \dfrac{1}{p} \\ \infty & \alpha \ge \dfrac{1}{p}. \end{cases}$$

Hence, real functions defined in $\Omega \subset \mathbb{R}$ lie in the space $L^p(\Omega)$ if either they are essentially bounded or if their singularities are weaker than the singularity of $x^{-1/p}$ (at singular points they go to infinity slower than $x^{-1/p}$).

4. The other purpose of L^p-spaces is to control the rate of decay at infinity on unbounded sets. Consider, for example, the interval $\Omega = (1, \infty)$ and the same function as above, $f(x) = 1/x^\alpha$. Now the L^p-norm of $f(x)$ is

$$\|f\|_p = \left(\int_\Omega \left| \frac{1}{x^\alpha} \right|^p \mathrm{d}x \right)^{\frac{1}{p}} = \begin{cases} \dfrac{1}{(\alpha p - 1)^{\frac{1}{p}}} & \alpha > \dfrac{1}{p} \\ \infty & \alpha \le \dfrac{1}{p}. \end{cases}$$

We conclude that on unbounded sets $\Omega \subset \mathbb{R}$, real functions lie in the space $L^p(\Omega)$ if they decay faster than $1/x^{1/p}$ at infinity.

Shape of L^p functions in $\Omega \subset \mathbb{R}^d$

1. Let $\Omega \subset \mathbb{R}^d$ be an open bounded set. Then every essentially bounded function $f : \Omega \to \mathbb{R}$, $|f(x)| \le C$ a.e. in Ω, belongs to $L^p(\Omega)$ for all $1 \le p \le \infty$ (proof is analogous to the 1D case above).

2. Let $\Omega \subset \mathbb{R}^d$ be an unbounded open set. For the same reason as in the 1D case, essentially bounded functions lie in $L^\infty(\Omega)$, but generally they do not lie in $L^p(\Omega)$, $1 \le p < \infty$. For this they need a sufficient rate of decay at infinity, analogously to the 1D case. We will discuss this in a moment.

3. In order to analyze the behavior of functions with singularities in bounded open sets, it is enough to use the open ball $\Omega = B(\mathbf{0}, R) \subset \mathbb{R}^d$ with a finite radius $R > 0$. Consider the function $f(\boldsymbol{x}) = 1/r^\alpha$, $r(\boldsymbol{x}) = \sqrt{x_1^2 + \ldots + x_d^2}$. Using the integration in polar coordinates (which is left to the reader as an exercise), we obtain that

$$\|f\|_p = \left(\int_\Omega \left|\frac{1}{r^\alpha}\right|^p \mathrm{d}\boldsymbol{x}\right)^{\frac{1}{p}} \begin{cases} < \infty & \alpha < \dfrac{d}{p}, \\ = \infty & \alpha \ge \dfrac{d}{p}. \end{cases}$$

Thus a function defined in an open bounded set $\Omega \subset \mathbb{R}^d$ belongs to $L^p(\Omega)$ if and only if either it is essentially bounded or its singularities are weaker than the singularity of $r^{-d/p}$.

4. On the other hand, for functions on unbounded open sets we can restrict ourselves to the open set $\Omega = \mathbb{R}^d \setminus B(\mathbf{0}, R)$, where $R > 0$. For the function $f(\boldsymbol{x}) = 1/r^\alpha$, $r(\boldsymbol{x}) = \sqrt{x_1^2 + \ldots + x_d^2}$, using the integration in polar coordinates again, we obtain

$$\|f\|_p = \left(\int_\Omega \left|\frac{1}{r^\alpha}\right|^p \mathrm{d}\boldsymbol{x}\right)^{\frac{1}{p}} \begin{cases} < \infty & \alpha > \dfrac{d}{p}, \\ = \infty & \alpha \le \dfrac{d}{p}. \end{cases}$$

We conclude that functions defined in open unbounded sets $\Omega \subset \mathbb{R}^d$ belong to $L^p(\Omega)$ if and only if they decay faster than $r^{-d/p}$ at infinity.

Let $\Omega \subset \mathbb{R}^d$ be an open set. Then the space $L^p(\Omega)$ is infinite-dimensional and its basis, obviously, consists of an infinite number of functions. To give an example, the Legendre polynomials $L_0, L_1, L_2, \ldots$ form a basis in the space $L^2(-1, 1)$. We will discuss them in more detail in Paragraph A.3.3. Another important example is the basis of the space $L^2(0, 2\pi)$ consisting of the functions $\{\cos(nx)\}_{n=0}^\infty$ and $\{\sin(nx)\}_{n=1}^\infty$, which is used to expand L^2-functions into Fourier series. The Fourier series will also be discussed in more detail in Paragraph A.3.3. Before we present the most important inequalities in L^p-spaces in Paragraph A.2.10, let us say a few words about discrete L^p-spaces.

Discrete L^p-spaces Although in this text we focus primarily on infinite-dimensional L^p-spaces defined in open subsets Ω of $\mathbb{R}^d$ (case most relevant for the study of partial differential equations and finite element methods), the finite-dimensional spaces $\mathbb{R}^n$ equipped with the discrete p-norm (A.12) are also worth mentioning.

Discrete L^p-spaces are defined on the so-called counting measure (see, e.g., [99] for details), where the integration is equivalent to summation. Then the integral p-norm (A.44) comes over to the discrete p-norm (A.12), and the L^∞-norm (A.45) naturally is replaced with the discrete maximum norm (A.10). The finite-dimensional case of $\mathbb{R}^n$ can be generalized to the space of infinite real (complex) sequences. Used is an analogy of the discrete p-norm (A.12), with the sum going from one to infinity, and the analogy of the discrete maximum norm (A.10), where the maximum is replaced with the supremum over absolute values of all entries of the sequence. Sometimes the discrete L^p-space in $\mathbb{R}^n$ is denoted by the symbol $l^p(\mathbb{R}^n)$.

A.2.10 Basic inequalities in L^p-spaces

In this paragraph we prove that the relations (A.44) and (A.45) indeed define norms, and we introduce several important inequalities in L^p-spaces: The triangular inequality for the L^p-norms is called Minkowski inequality. The proof of the Minkowski inequality requires the Hölder inequality, which in turn is based on the Young inequality.

The following proofs contain numerous algebraic manipulations involving a pair of real numbers $1 < p, q < \infty$ such that

$$\frac{1}{p} + \frac{1}{q} = 1. \tag{A.47}$$

To get more familiar with this relation, check that

$$pq = p + q$$

and

$$(p-1)(q-1) = 1$$

are equivalent to (A.47).

Lemma A.26 (Young inequality) *Let $a, b \geq 0$ and $1 < p, q < \infty$ such that*

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Then

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}. \tag{A.48}$$

Proof: For an arbitrary $0 \leq b \in \mathbb{R}$ define a function $f : [0, \infty) \to \mathbb{R}$ by

$$f(x) = \frac{x^p}{p} + \frac{b^q}{q} - xb.$$

It is easy to verify that $f'(x_0) = 0$ if and only if $x_0 = b^{\frac{1}{p-1}}$, that $f'(x) < 0$ for all $x \in (0, x_0)$ and that $f'(x) > 0$ for all $x \in (x_0, \infty)$. Evaluating the function $f(x)$ at $x_0 = b^{\frac{1}{p-1}}$, we discover that $f(x_0) = 0$. Therefore it is

$$0 \leq \frac{x^p}{p} + \frac{b^q}{q} - xb$$

for all $x \geq 0$, and thus (A.48) holds for all $a \geq 0$. ■

The following modification of the Young inequality is equally useful:

Lemma A.27 (Modified Young inequality) *Let $a, b \geq 0$, $\epsilon > 0$ and $1 < p, q < \infty$ such that*

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Then

$$ab \leq \epsilon \frac{a^p}{p} + \epsilon^{1-q} \frac{b^q}{q}. \tag{A.49}$$

Proof: Relation (A.49) follows immediately when (A.48) is applied to properly changed values $\tilde{a}, \tilde{b}$ instead of the original values a, b. This is left to the reader as an exercise. ■

The Hölder inequality was first proved by a German mathematician Otto Ludwig Hölder in 1884, in the context of convergence analysis of Fourier series. O.L. Hölder contributed significantly to mathematical analysis and group theory.

Figure A.22 Otto Ludwig Hölder (1859–1937).

Theorem A.10 (Hölder inequality) *Let $\Omega \subset \mathbb{R}^d$ be an open set and $1 \leq p, q \leq \infty$ such that*

$$\frac{1}{p} + \frac{1}{q} = 1$$

(it is understood that $1/\infty = 0$). Let $u \in L^p(\Omega)$ and $v \in L^q(\Omega)$. Then

$$\int_\Omega |u(\boldsymbol{x})v(\boldsymbol{x})| \, d\boldsymbol{x} \leq \|u\|_p \|v\|_q. \tag{A.50}$$

Proof: If $p = 1, p = \infty$ or $\|u\|_p = 0$ then the inequality obviously is satisfied. Otherwise use the modified Young inequality to obtain

$$\int_\Omega \underbrace{|u(\boldsymbol{x})|}_{a} \underbrace{|v(\boldsymbol{x})|}_{b} \, \mathrm{d}\boldsymbol{x} \leq \int_\Omega \epsilon \frac{|u(\boldsymbol{x})|^p}{p} + \epsilon^{1-q} \frac{|v(\boldsymbol{x})|^q}{q} = \frac{\epsilon}{p} \|u\|_p^p + \frac{\epsilon^{1-q}}{q} \|v\|_q^q \tag{A.51}$$

for all $\epsilon > 0$. The function

$$f(\epsilon) = \frac{\epsilon}{p}\|u\|_p^p + \frac{\epsilon^{1-q}}{q}\|v\|_q^q$$

attains its minimum at $\epsilon_0 = \frac{\|v\|_q}{\|u\|_p^{p-1}}$. Using the value ϵ_0 in (A.51), we obtain

$$\int_\Omega |u(\boldsymbol{x})||v(\boldsymbol{x})|\,\mathrm{d}\boldsymbol{x} \le \frac{\epsilon_0}{p}\|u\|_p^p + \frac{\epsilon_0^{1-q}}{q}\|v\|_q^q = \frac{1}{p}\|u\|_p\|v\|_q + \frac{1}{q}\|u\|_p\|v\|_q = \|u\|_p\|v\|_q,$$

which concludes the proof. ■

Let us remark that in the space $l^p(\mathbb{R}^n)$, equipped with the discrete p-norm (A.12) or the discrete maximum norm (A.10), the Hölder inequality (A.50) attains the following form: Let $u, v \in \mathbb{R}^n$, and $1 \le p, q \le \infty$ such that $1/p + 1/q = 1$. Then

$$\sum_{i=1}^{n} |u_i v_i| \le \|u\|_p\|v\|_q = \left(\sum_{i=1}^{n} |u_i|^p\right)^{\frac{1}{p}} \left(\sum_{i=1}^{n} |v_i|^q\right)^{\frac{1}{q}}. \tag{A.52}$$

This inequality sometimes is called discrete Hölder inequality.

The Minkowski inequality carries the name of a German mathematician Hermann Minkowski. Although he was mainly interested in topics of pure mathematics such as quadratic forms and continued fractions, it is commonly assumed that the greatest contribution of H. Minkowski was the coupling of the space and time into a four-dimensional continuum, that provided the foundation for all later work in relativity. Albert Einstein attended several of his courses in Zürich.

Figure A.23 Hermann Minkowski (1864–1909).

Lemma A.28 (Minkowski inequality) *Let $\Omega \subset \mathbb{R}^d$ be an open set and $1 \le p \le \infty$. Then*

$$\|u+v\|_p \le \|u\|_p + \|v\|_p \tag{A.53}$$

for all $u, v \in L^p(\Omega)$.

Proof: The inequality obviously is satisfied for $p = 1$ and $p = \infty$. Applying the Hölder inequality, for $p \in (1, \infty)$ we obtain

$$\begin{aligned} \|u+v\|_p^p &= \int_\Omega |u(\boldsymbol{x}) + v(\boldsymbol{x})|^p \,\mathrm{d}\boldsymbol{x} = \int_\Omega |u(\boldsymbol{x}) + v(\boldsymbol{x})|^{p-1} |u(\boldsymbol{x}) + v(\boldsymbol{x})| \,\mathrm{d}\boldsymbol{x} \\ &\le \int_\Omega |u(\boldsymbol{x}) + v(\boldsymbol{x})|^{p-1} |u(\boldsymbol{x})| \,\mathrm{d}\boldsymbol{x} + \int_\Omega |u(\boldsymbol{x}) + v(\boldsymbol{x})|^{p-1} |v(\boldsymbol{x})| \,\mathrm{d}\boldsymbol{x} \\ &\le \left(\int_\Omega |u(\boldsymbol{x}) + v(\boldsymbol{x})|^{(p-1)q} \,\mathrm{d}\boldsymbol{x} \right)^{\frac{1}{q}} (\|u\|_p + \|v\|_p) \\ &= \left(\int_\Omega |u(\boldsymbol{x}) + v(\boldsymbol{x})|^p \,\mathrm{d}\boldsymbol{x} \right)^{\frac{p-1}{p}} (\|u\|_p + \|v\|_p) \\ &= \|u+v\|_p^{p-1} (\|u\|_p + \|v\|_p), \end{aligned}$$

which concludes the proof. ■

Herewith the triangular inequality for the L^p-spaces is verified and we can be sure that (A.44) and (A.45) are norms. In the space $l^p(\mathbb{R}^n)$, equipped with the discrete p-norm (A.12), the discrete Minkowski inequality has the form

$$\left(\sum_{i=1}^n |u_i + v_i|^p \right)^{\frac{1}{p}} = \|u+v\|_p \le \|u\|_p + \|v\|_p = \left(\sum_{i=1}^n |u_i|^p \right)^{\frac{1}{p}} + \left(\sum_{i=1}^n |v_i|^p \right)^{\frac{1}{p}}. \tag{A.54}$$

The Young, Hölder, and Minkowski inequalities are encountered frequently in the analysis of functions in L^p-spaces. The next lemma summarizes several other important properties of these spaces:

Lemma A.29 *Let $\Omega \subset \mathbb{R}^d$ be an open set.*

1. *For every $1 \le p \le \infty$, $L^p(\Omega)$ is a Banach space.*

2. *For every $1 \le p \le \infty$, every Cauchy sequence in $L^p(\Omega)$ has a subsequence that converges pointwise almost everywhere in Ω.*

3. *Let Ω be bounded. Then for every $1 \le p \le q \le \infty$, $L^q(\Omega) \subset L^p(\Omega)$, and we have*

$$\|v\|_p \le |\Omega|^{\frac{1}{p} - \frac{1}{q}} \|v\|_q$$

for all $v \in L^q(\Omega)$. Moreover, it is

$$\|v\|_\infty = \lim_{p \to \infty} \|v\|_p$$

for all $v \in L^\infty(\Omega)$.

4. *For every $1 \le p \le r \le q \le \infty$ and $\gamma \in [0, 1]$ satisfying*

$$\frac{1}{r} = \frac{\gamma}{p} + \frac{1-\gamma}{q}$$

the following interpolation property of L^p-spaces holds:

$$\|v\|_r \le \|v\|_p^{\gamma} \|v\|_q^{1-\gamma}.$$

Proof: These results are standard, but their proofs are rather technical. See, e.g., [34, 65] and [99]. ■

The assumption of boundedness of Ω in the third assertion of Lemma A.29 is important. The reason why on bounded sets in $\mathbb{R}^d$ the L^p-spaces get smaller as the exponent p rises is that the maximum admissible strength of singularities decreases (Figure A.24).

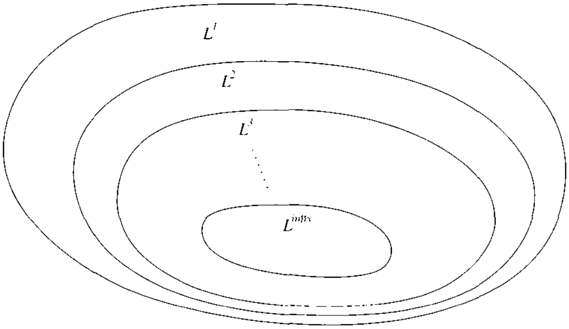


Figure A.24 Structure of L^p-spaces on an open bounded set $\Omega \subset \mathbb{R}^d$. ($L^{infty}$ stands for L^∞.)

On the other hand, if Ω is unbounded, then the implication does not hold since evidently not all bounded functions are integrable.

A.2.11 Density of smooth functions in L^p-spaces

Let $\Omega \subset \mathbb{R}^d$ be an open set. The ability of functions from the space $C^\infty(\Omega)$ to approximate the L^p-functions with an arbitrary accuracy is of great practical importance in the analysis of partial differential equations. This result is formulated in Lemma A.30. At the end of this paragraph we briefly discuss the duality of the L^p-spaces.

Lemma A.30 *Let $\Omega \subset \mathbb{R}^d$ be an open set. Then for any $1 \le p < \infty$ and any $v \in L^p(\Omega)$ there exists a sequence $\{u_n\}_{n=1}^{\infty} \subset C^\infty(\Omega)$ such that*

$$\lim_{n \to \infty} \|u_n - v\|_p = 0.$$

Proof: See, e.g., [34, 65] and [99]. ■

This means that no function $v \in L^p(\Omega)$ is isolated from C^∞-functions in the sense that for arbitrarily small $\epsilon > 0$ there always is some C^∞-function in the open ball $B(v, \epsilon) \subset L^p(\Omega)$. If Ω is bounded, then the result of Lemma A.30 holds even for the space $C_0^\infty(\Omega)$ of infinitely smooth functions that vanish on the boundary of Ω. The result also holds for every space $C^m(\Omega)$, $m \geq 0$, of m-times continuously differentiable functions (including the space $C(\Omega) = C^0(\Omega)$ of continuous functions).

Density argument A frequently used technique for proving various properties of functions in L^p-spaces, called density argument, works as follows:

- Let $v \in L^p(\Omega)$ be a function whose property (P) is to be shown.
- Take some sequence in $C^\infty(\Omega)$ converging to v in the $\|\cdot\|_p$-norm (the existence of such sequence is guaranteed by Lemma A.30).
- Prove that starting with some index n_0, the elements in the sequence have the property (P). This step usually is much easier for infinitely smooth functions than for the original function v.
- Show that also the limit of the sequence has the property (P).

This technique will be used, for example, to prove the Poincaré–Friedrichs' inequality in Paragraph A.4.5. Let us give an example of such sequence:

■ **EXAMPLE A.38 (Sequences of C^∞-functions converging to an L^p-function)**

1. In the interval $\Omega = (-1, 1)$ consider the function

$$v(x) = \begin{cases} 1 & x = 1, \\ 0 & \text{elsewhere in } \Omega. \end{cases}$$

This function belongs to the space $L^p(\Omega)$ for all $1 \leq p \leq \infty$. The sequence of $C^\infty(\Omega)$-functions $\{u_n\}_{n=1}^\infty$,

$$u_n = \left(1 - x^2\right)^n$$

converges to v in the p-norm for all $1 \leq p < \infty$. The functions $u_1, u_{10}, u_{100}, u_{1000}$, and u_{10000} are shown in Figure A.25.

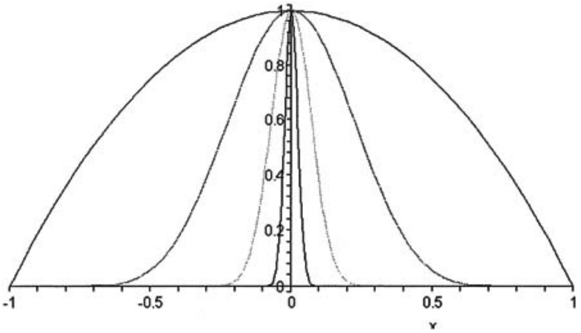


Figure A.25 Example of a sequence converging out of $C(-1, 1)$.

Since the function v is equivalent to the zero function in the Lebesgue sense, one can say that the sequence $\{u_n\}_{n=1}^\infty$ converges to zero in all spaces $L^p(\Omega)$ for all $1 \leq p < \infty$. The sequence does not converge in the L^∞-norm.

2. Next, in the interval $\Omega = (0, 2\pi)$ consider the discontinuous function

$$v(x) = \begin{cases} x & x \in (0, \pi), \\ x - 2\pi & x \in [\pi, 2\pi). \end{cases}$$

   This function again belongs to the space $L^\infty(\Omega)$ on a bounded set and therefore it lies in all spaces $L^p(\Omega)$, $1 \le p \le \infty$. The sequence of $C^\infty(\Omega)$-functions $\{u_n\}_{n=1}^\infty$,

$$u_n = \sum_{k=1}^{n} (-1)^{k+1} \frac{2}{k} \sin(kx),$$

   converges to v in the L^2-norm (actually it converges in the p-norm for all $1 \le p < \infty$). For the explanation, however, the reader will have to wait until Paragraph A.3.3 where we arrive at the Fourier series.

The last topic in the theory of L^p-spaces that we would like to discuss in this section is their duality:

Lemma A.31 (Duality of L^p-spaces) *Let $\Omega \subset \mathbb{R}^d$ be an open set and $1 < p < \infty$. Then any dual space V' to the space $V = L^p(\Omega)$ is isomorphic with the space $L^q(\Omega)$ where*

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Proof: Consider the subset $S \subset V'$ consisting of all linear forms that can be written as

$$f(u) = \int_\Omega v_f(\boldsymbol{x}) u(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x},$$

where v_f are arbitrary functions. Since $u \in V$ and $\|f\|_{V'}$ has to be finite, the Hölder inequality restricts the functions v_f to lie in the space $L^q(\Omega)$. The next step of the proof is to show that $S = V'$. This is more technical and we refer, e.g., to [99]. ■

Remark A.2 (Dual space for $L^\infty(\Omega)$) *Generally, it is not true that the dual space to $L^\infty(\Omega)$ is $L^1(\Omega)$. This only holds if the measure of the set Ω is σ-finite. More details can be found in [99].*

A.2.12 Exercises

Exercise A.23 *Let $V = \mathbb{R}^n$, $n > 0$. Show that the functions*

$$\|v\|_\infty = \max_{i=1,2,\dots,n} |v_i|$$

and

$$\|v\|_1 = \sum_{i=1}^{n} |v_i|$$

are norms on V. These norms are called discrete maximum norm (l^∞-norm), and discrete integral norm (l^1-norm), respectively, because of their relation to Lebesgue spaces of sequences.

Exercise A.24 *Let $V = C([0,1])$. Show that the function*

$$\|f\|_{\max} = \max_{x\in[0,1]} |f(x)|$$

is a norm in V.

Exercise A.25 *Consider the space $V = C^1([0,1])$ and decide which of the following is a norm and which only is a seminorm:*

1. $\max_{0\le x\le 1} |u(x)|$,
2. $\max_{0\le x\le 1} [|u(x)| + |u'(x)|]$,
3. $\max_{0\le x\le 1} |u'(x)|$,
4. $|u(0)| + \max_{0\le x\le 1} |u'(x)|$,
5. $\max_{0\le x\le 1} |u'(x)| + \int_a^b |u(x)|\,dx, \quad a,b\in(0,1), \quad a<b.$

Exercise A.26 *Use Definition A.27 to prove the equivalent characterization of limit in Lemma A.20.*

Exercise A.27 *Use Definition A.27 to show that the sequence in the third item of Example A.22 does not converge in V.*

Exercise A.28 *Prove the "backward triangular inequality" (A.25) and the corresponding result for normed spaces: $|\|u_a\|_V - \|u_b\|_V| \le \|u_a - u_b\|_V$ for all $u_a, u_b \in V$.*

Exercise A.29 *Prove Proposition A.5.*

Exercise A.30 *Prove inequalities (A.29) and (A.30) in Example A.30 (equivalence of the discrete maximum norm, discrete integral norm, and the Euclidean norm in $\mathbb{R}^n$).*

Exercise A.31 *Consider the space $C^1(0,1)\cap C([0,1])$. Prove that the following norms are equivalent:*

$$\|f\|_a = |f(0)| + \int_0^1 |f'(x)|\,dx$$

and

$$\|f\|_b = \int_0^1 |f(x)|\,dx + \int_0^1 |f'(x)|\,dx.$$

Hint: Use the main theorem of calculus or the integral mean value theorem. The latter says that for every $g\in C([0,1])$ there is a $\xi\in[0,1]$ such that

$$\int_0^1 g(x)\,dx = g(\xi).$$

Exercise A.32 *Adjust the procedure from Example A.31 to prove the equivalence of the maximum norm (A.17) and the p-norm (A.19) with $p = 2$ in the space $P^k([-1, 1])$, where k is an arbitrary natural number.*

Exercise A.33 *Let $(a, b) \subset \mathbb{R}$ be a nonempty bounded interval.*

1. *Construct an infinite sequence of functions in the space $V = C([a, b])$ that converges in the p-norm for all $1 \le p < \infty$ but which does not converge in the maximum norm.*
2. *Use this sequence and Definition A.34 to show that the maximum and p-norms are not equivalent in V.*
3. *Is it possible to construct a sequence in V which converges in the maximum norm but does not converge in the p-norm? Present a proof.*
4. *Find a subset $S \subset V$ that is open in the p-norm but is not open in the maximum norm. Can you do this vice versa as well?*

Exercise A.34 *Show that the definitions of the operator norm (A.26) and (A.27) are equivalent.*

Exercise A.35 *Prove Proposition A.6 (every convergent sequence in a normed space is a Cauchy sequence).*

Exercise A.36 *Let V be a normed space and $\{u_n\}_{n=1}^{\infty} \subset V$ a Cauchy sequence. Suppose that there is a subsequence $\{u_{n_k}\}_{k=1}^{\infty} \subset \{u_n\}_{n=1}^{\infty}$ and some element $u \in V$ such that*

$$\lim_{k\to\infty} u_{n_k} = u.$$

Show that

$$\lim_{n\to\infty} u_n = u.$$

Exercise A.37 *Show that the sequence (A.34) is convergent and that the limit is $\sqrt{a}$. Hint: Use, for example, (A.35).*

Exercise A.38 *Consider the open ball $B(\mathbf{0}, R) \subset \mathbb{R}^d$, $d = 3$, with a finite radius $R > 0$, and the function $f(\boldsymbol{x}) = 1/r^{\alpha}$, $r(\boldsymbol{x}) = \sqrt{x_1^2 + \ldots + x_d^2}$. Show that the following statements hold:*

1. *Let $\Omega = B(\mathbf{0}, R)$. Then $f(\boldsymbol{x}) \in L^p(\Omega)$ if and only if $\alpha < d/p$.*
2. *Let $\Omega = \mathbb{R}^d \setminus \overline{B(\mathbf{0}, R)}$. Then $f(\boldsymbol{x}) \in L^p(\Omega)$ if and only if $\alpha > d/p$.*

Describe in detail the application of the Substitution Theorem for integration in spherical coordinates, and write the corresponding resulting finite integral.

Exercise A.39 *Prove the modified Young inequality (A.49) using the Young inequality (A.48).*

Exercise A.40 *Let $\Omega \subset \mathbb{R}^d$ be an open set and $1 \le p_1, p_2, \ldots, p_m$ such that*

$$\frac{1}{p_1} + \frac{1}{p_2} + \ldots + \frac{1}{p_m} = 1.$$

Use the Hölder inequality (A.50) to prove the generalized Hölder inequality

$$\int_\Omega |u_1(\boldsymbol{x})u_2(\boldsymbol{x})\ldots u_m(\boldsymbol{x})|\, d\boldsymbol{x} \le \|u_1\|_{p_1}\|u_2\|_{p_2}\cdots\|u_m\|_{p_m}. \tag{A.55}$$

Hint: Proceed by induction.

Exercise A.41 *Consider the space $V = L^1(-1,1)$ and the step function $v(x) \in V$, $v(x) = 0$ for all $x < 0$ and $v(x) = 1$ for all $x \ge 0$. Construct some concrete sequence $\{u_n\}_{n=1}^{\infty} \subset C^\infty(-1,1)$ such that*

$$\lim_{n\to\infty} \|u_n - v\|_1 = 0.$$

Hint: The norms $\|u_n - v\|_1$ do not have to be calculated exactly if you can estimate them by some values that converge to zero.

A.3 INNER PRODUCT SPACES

Some linear spaces can be endowed with inner product, which is a binary operation similar to the "dot-product"

$$(u,v)_{\mathbb{R}^n} = u \cdot v = \sum_{i=1}^{n} u_i v_i \tag{A.56}$$

of vectors in $\mathbb{R}^n$. Such spaces are called inner product spaces. Orthogonality in a general inner product space V is defined analogously to the orthogonality of vectors in $\mathbb{R}^n$,

$$u \perp v \quad \Leftrightarrow \quad (u,v)_V = 0. \tag{A.57}$$

The notion of orthogonality and orthogonal projection makes inner product spaces extremely convenient for the study of partial differential equations and finite element methods. After discussing the most important concepts and techniques available in inner product spaces, at the end of this section we also mention compactness and weak convergence.

A.3.1 Inner product

Definition A.41 (Inner product) *Let V be a real or complex linear space. An* inner product *in V is any function $(\cdot,\cdot)_V : V \times V \to \mathbb{R}$ (or $\mathbb{C}$) with the following properties:*

1. *For any $u \in V$, $(u,u)_V \ge 0$ and moreover $(u,u)_V = 0$ if and only if $u = 0$.*
2. *For any $u, v \in V$, $(u,v)_V = \overline{(v,u)}_V$.*
3. *For any $u, v, w \in V$ and all $a, b \in \mathbb{R}$ (or $\mathbb{C}$) we have $(au + bv, w)_V = a(u,w)_V + b(v,w)_V$.*

In real inner product spaces the second axiom reduces to the symmetry assumption:

$$(u, v)_V = (v, u)_V \qquad \text{for all } u, v \in V.$$

The subscript V usually is left out when the space V is clear from the context. We restrict ourselves to real inner product spaces in the following.

■ EXAMPLE A.39 (Inner product spaces $l^2(\mathbb{R}^n)$ and $L^2(\Omega)$)

1. It is a simple exercise to verify that the standard "dot-product" (A.56) in $\mathbb{R}^n$ is an inner product in the sense of Definition A.41. Adding this inner product to $\mathbb{R}^n$, we obtain the inner product space $l^2(\mathbb{R}^n)$ that we first encountered at the end of Paragraph A.2.9.

2. Let $\Omega \subset \mathbb{R}^d$ be an open set. It is another simple exercise to verify that the relation

$$(u, v) = \int_\Omega u(\boldsymbol{x})v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \tag{A.58}$$

defines an inner product in the normed space $L^2(\Omega)$.

Next let us show that every inner product space is a normed space:

Lemma A.32 (Inner product induces norm) *Let V be an inner product space. Then the function*

$$\|u\| = \sqrt{(u, u)}, \qquad u \in V$$

is a norm in V.

Proof: Among all required properties of a norm, only the triangular inequality is not obvious. Hence, let us choose any $u, v \in V$ and write

$$\begin{aligned} \|u + v\|^2 &= (u + v, u + v) = (u, u) + 2(u, v) + (v, v) \\ &\leq \|u\|^2 + |2(u, v)| + \|v\|^2. \end{aligned} \tag{A.59}$$

Now let us verify that

$$|(u, v)| \leq \|u\|\|v\| \tag{A.60}$$

for all $u, v \in V$. If $u = 0$ or $v = 0$, (A.60) holds. If $u \neq 0$ and $v \neq 0$ we define a nonnegative real function

$$0 \leq \varphi(t) = (u + tv, u + tv) = (u, u) + 2(u, v)t + (v, v)t^2.$$

Since $\varphi(t)$ is a parabola, its discriminant must be nonpositive,

$$D = [2(u, v)]^2 - 4(u, u)(v, v) \leq 0,$$

which proves (A.60). Returning to (A.59), finally we obtain

$$\|u+v\|^2 \leq \|u\|^2 + |2(u,v)| + \|v\|^2 \leq \|u\|^2 + 2\|u\|\|v\| + \|v\|^2 \leq (\|u\| + \|v\|)^2,$$

which concludes the proof. ■

Inequality (A.60) is of essential importance in inner product spaces, and therefore let us formulate it once more in a separate theorem:

Theorem A.11 (Cauchy–Schwarz inequality) *Let V be an inner product space and $\|\cdot\|$ the norm induced by the inner product $(\cdot,\cdot)$. Then*

$$|(u,v)| \leq \|u\|\|v\| \tag{A.61}$$

for all $u, v \in V$.

Proof: See the proof of inequality (A.60). ■

■ EXAMPLE A.40 (Hölder implies Cauchy–Schwarz)

1. The discrete Hölder inequality (A.52) with $p = q = 2$,

$$\sum_{i=1}^{n} |u_i v_i| \leq \left(\sum_{i=1}^{n} u_i^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^{n} v_i^2\right)^{\frac{1}{2}},$$

together with the triangular inequality

$$\left|\sum_{i=1}^{n} u_i v_i\right| \leq \sum_{i=1}^{n} |u_i v_i|,$$

imply the Cauchy–Schwarz inequality in the inner product space $l^2(\mathbb{R}^n)$,

$$|(u,v)| = \left|\sum_{i=1}^{n} u_i v_i\right| \leq \sum_{i=1}^{n} |u_i v_i| \leq \left(\sum_{i=1}^{n} u_i^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^{n} v_i^2\right)^{\frac{1}{2}} = \|u\|_2 \|v\|_2.$$

2. Let Ω be an open subset of $\mathbb{R}^d$. Also in the space $L^2(\Omega)$ the Cauchy–Schwarz inequality,

$$\begin{aligned} |(u,v)| &= \left|\int_\Omega u(x)v(x)\,\mathrm{d}x\right| \leq \int_\Omega |u(x)v(x)|\,\mathrm{d}x \\ &\leq \left(\int_\Omega |u(\boldsymbol{x})|^2\,\mathrm{d}\boldsymbol{x}\right)^{\frac{1}{2}} \left(\int_\Omega |v(\boldsymbol{x})|^2\,\mathrm{d}\boldsymbol{x}\right)^{\frac{1}{2}} = \|u\|_2\|v\|_2, \end{aligned}$$

is a consequence of the Hölder inequality (A.50) with $p = q = 2$ and the triangular inequality.

We know from Lemma A.32 that every inner product induces a norm. Conversely, there are norms which induce an inner product:

Lemma A.33 (Parallelogram rule) *Let V be a real normed space. If the norm $\|\cdot\|$ satisfies the parallelogram rule*

$$\|u+v\|^2 + \|u-v\|^2 = 2\|u\|^2 + 2\|v\|^2 \qquad \textit{for all } u, v \in V, \tag{A.62}$$

then it induces an inner product in V. This inner product is defined by the relation

$$(u, v) = \frac{1}{4}\left(\|u+v\|^2 - \|u-v\|^2\right). \tag{A.63}$$

Proof: It is easy to see that $(u, u) \geq 0$ for all $u \in V$ and that $(u, u) = 0$ if and only if $u = 0$. Second,

$$(v, u) = \frac{1}{4}\left(\|u+v\|^2 - \|u-v\|^2\right) = (u, v)$$

verifies the symmetry. Next the linearity of the relation (A.63) has to be verified, i.e., we are asking if the following two conditions hold:

$$(u+v, w) = (u, w) + (v, w) \quad \text{for all } u, v, w \in V$$

and

$$(au, v) = a(u, v) \quad \text{for all } a \in \mathbb{R},\ u \in V. \tag{A.64}$$

To begin with, it is

$$\begin{aligned}(u+v, w) &= \frac{1}{4}\left(\|u+v+w\|^2 - \|u+v-w\|^2\right),\\ (u, w) + (v, w) &= \frac{1}{4}\left(\|u+w\|^2 - \|u-w\|^2 + \|v+w\|^2 - \|v-w\|^2\right).\end{aligned}$$

It is left to the reader to verify that the right-hand sides are equal, using the parallelogram rule (A.62). Also use the decomposition

$$\begin{aligned}\tilde{u} &= \frac{\tilde{u}+\tilde{v}}{2} + \frac{\tilde{u}-\tilde{v}}{2},\\ \tilde{v} &= \frac{\tilde{u}+\tilde{v}}{2} - \frac{\tilde{u}-\tilde{v}}{2},\end{aligned}$$

which holds for all elements $\tilde{u}, \tilde{v} \in V$. It remains to verify relation (A.64): For arbitrary $u, v \in V$ define a real-valued function

$$f(x) = \|xu+v\|^2 - \|xu-v\|^2 = 4(xu, v).$$

It is sufficient to show that

$$f(a) = af(1) \qquad \text{for all } a \in \mathbb{R}. \tag{A.65}$$

After some calculation, we obtain that

$$f(x) - f(y) = 2\left(\left\|\frac{x-y}{2}u + v\right\|^2 - \left\|\frac{x-y}{2}u - v\right\|^2\right) = 2f\left(\frac{x-y}{2}\right). \tag{A.66}$$

It is $f(0) = 0$. Taking $y = 0$, we obtain

$$f(x) = 2f\left(\frac{x}{2}\right).$$

Relation (A.66) yields $f(x) - f(y) = f(x - y)$. Therefore $f(x)$ is a linear function passing through the origin $[0, 0]$, and (A.65) holds. ■

Remark A.3

1. *In a complex normed space the relation (A.63) only defines the real part of the inner product. The complex part is defined analogously, replacing u with iu.*

2. *If (u, v) is inner product in a real linear space V, then it satisfies (A.63),*

$$\begin{aligned}\frac{1}{4}\left(\|u+v\|^2 - \|u-v\|^2\right) &= \frac{1}{4}[(u+v, u+v) - (u-v, u-v)]\\ &= \frac{1}{4}[2(u, v) + 2(v, u)]\\ &= (u, v).\end{aligned}$$

3. *The norm $\|\cdot\| = \sqrt{(\cdot, \cdot)}$ induced by this inner product satisfies the parallelogram rule (A.62),*

$$\begin{aligned}\|u+v\|^2 + \|u-v\|^2 &= (u+v, u+v) + (u-v, u-v)\\ &= (u, u) + (u, v) + (v, u) + (v, v)\\ &\quad +(u, u) - (u, v) - (v, u) + (v, v)\\ &= 2\|u\|^2 + 2\|v\|^2.\end{aligned}$$

■ **EXAMPLE A.41 (Parallelogram rules in $l^2(\mathbb{R}^n)$ and $L^2(\Omega)$)**

1. Consider the normed space $l^2(\mathbb{R}^n)$. The parallelogram rule (A.62) reads:

$$\begin{aligned}\|u+v\|^2 + \|u-v\|^2 &= \sum_{i=1}^{n}(u_i + v_i)^2 + \sum_{i=1}^{n}(u_i - v_i)^2\\ &= 2\sum_{i=1}^{n} u_i^2 + 2\sum_{i=1}^{n} v_i^2\\ &= 2\|u\|^2 + 2\|v\|^2.\end{aligned}$$

2. Consider the normed space $L^2(\Omega)$ in an open set $\Omega \subset \mathbb{R}^d$. In this case the parallelogram rule (A.62) has the form

$$\begin{aligned}\|u+v\|^2 + \|u-v\|^2 &= \int_\Omega (u(\boldsymbol{x}) + v(\boldsymbol{x}))^2 \,\mathrm{d}\boldsymbol{x} + \int_\Omega (u(\boldsymbol{x}) - v(\boldsymbol{x}))^2 \,\mathrm{d}\boldsymbol{x}\\ &= 2\int_\Omega u^2(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} + 2\int_\Omega v^2(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}\\ &= 2\|u\|^2 + 2\|v\|^2.\end{aligned}$$

Let $\Omega \subset \mathbb{R}^d$ be an open set. The parallelogram rule determines that $L^2(\Omega)$ is the only inner product space among the Lebesgue L^p-spaces:

Remark A.4 **($L^q(\Omega)$, $q \neq 2$, is not inner product space)** *For any $1 \leq q \leq \infty$, $q \neq 2$, the relation (u, v) defined by (A.63) is not linear in u, and therefore it cannot represent an inner product. Analogous conclusion holds for the discrete Lebesgue spaces $l^p(\mathbb{R}^n)$.*

Every inner product is continuous with respect to the norm it induces:

Lemma A.34 *Let V be an inner product space, $u \in V$ and $\{u_n\}_{n=1}^{\infty} \subset V$. If*

$$\lim_{n \to \infty} \|u_n - u\| = 0$$

then for any $v \in V$

$$\lim_{n \to \infty} (u_n, v) = (u, v).$$

Proof: Using the Cauchy–Schwarz inequality we immediately obtain

$$|(u_n, v) - (u, v)| = |(u_n - u, v)| \leq \|u_n - u\| \|v\|.$$

The conclusion follows from the fact that $\|v\|$ is a finite number and $\|u_n - u\| \to 0$ as $n \to \infty$. ■

A.3.2 Hilbert spaces

From the point of view of convergence analysis it is convenient to work in complete inner product spaces. This class of linear spaces carries the name of a German mathematician David Hilbert, who contributed to many branches of mathematics, including invariants, algebraic number fields, functional analysis, integral equations, mathematical physics, and the calculus of variations.

Figure A.26 David Hilbert (1862–1943).

Because of the importance of the Hilbert spaces, we reserved this paragraph for their definition and a few examples. Most of the time we shall stay in real Hilbert spaces.

Definition A.42 *Every complete inner product space is said to be a* Hilbert space.

■ **EXAMPLE A.42 (Hilbert spaces)**

A sufficient condition for an inner product space to be Hilbert space is that the underlying normed space be a Banach space:

1. The space $V = l^2(\mathbb{R}^d)$, i.e., $\mathbb{R}^d$ equipped with the "dot-product" (A.56), is a Hilbert space.

2. The space $V = l^2(\mathbb{R}^{n \times n})$, i.e., $\mathbb{R}^{n \times n}$ equipped with the Frobenius inner product

$$(A, B) = \sum_{i,j=1}^{n} a_{ij} b_{ij}$$

(which induces the Frobenius norm (A.16)) is a Hilbert space.

3. The space $V = l^2$ of infinite real sequences, equipped with the l^2-product

$$(u, v) = \sum_{i=1}^{\infty} u_i v_i, \tag{A.67}$$

is a Hilbert space.

4. Let Ω be an open subset of $\mathbb{R}^d$. The space $L^2(\Omega)$ equipped with the L^2-product (A.58) is a Hilbert space. By Remark A.4 this is the only Hilbert space among the Lebesgue L^p-spaces.

A.3.3 Generalized angle and orthogonality

In this paragraph we define generalized angle and orthogonality of elements in Hilbert spaces.

Definition A.43 (Generalized angle) *Let V be a Hilbert space. The angle of two elements $0 \neq u.v \in V$ is a real number $\alpha \in [0, \pi)$ such that*

$$\cos(\alpha) = \frac{(u, v)_V}{\|u\| \|v\|}. \tag{A.68}$$

■ **EXAMPLE A.43 (Generalized angle)**

Some of the inner products used below can be found in Example A.42.

1. In the space $V = l^2(\mathbb{R}^d)$ the formula (A.68) reduces to the standard relation

$$\cos(\alpha) = \frac{(u, v)_V}{\|u\| \|v\|} = \frac{u \cdot v}{\|u\| \|v\|}.$$

To give a concrete example, the angle of the vectors $u = (1, 1, 0)^T$ and $v = (0, 1, 1)^T$ in $\mathbb{R}^3$ is

$$\alpha = \arccos\left(\frac{1}{\sqrt{2}\sqrt{2}}\right) = \frac{\pi}{3}.$$

2. Consider the matrices

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

in the space $V = l^2(\mathbb{R}^{3\times 3})$. The angle of A and B is

$$\alpha = \arccos\left(\frac{(A, B)_V}{\|A\|\|B\|}\right) = \arccos\left(\frac{0}{\sqrt{5}\sqrt{4}}\right) = \frac{\pi}{2}.$$

3. Next let us calculate the angle of arbitrary geometrical sequences $u = \{u_n\}_{n=1}^{\infty}$ and $v = \{v_n\}_{n=1}^{\infty}$ in the space l^2 equipped with the l^2-product (A.67). Consider geometrical sequences given by the parameters $u_n = u_0 r^{n-1}$, $v_n = v_0 s^{n-1}$, $0 < r, s < 1$, $0 < u_0, v_0 \in \mathbb{R}$. We have

$$\begin{aligned} \alpha &= \arccos\left(\frac{(u, v)}{\|u\|\|v\|}\right) = \arccos\left(\frac{u_0 v_0 \sum_{i=1}^{\infty}(rs)^{i-1}}{u_0\sqrt{\sum_{j=1}^{\infty}(r^2)^{j-1}}\, v_0\sqrt{\sum_{k=1}^{\infty}(s^2)^{k-1}}}\right) \\ &= \arccos\left(\frac{\sqrt{1-r^2}\sqrt{1-s^2}}{1-rs}\right). \end{aligned}$$

4. Last consider an open set $\Omega = (-1, 1) \subset \mathbb{R}$ and the Hilbert space $L^2(\Omega)$. The angle of the functions $f(x) = 1$ and $g(x) = x$ is

$$\begin{aligned} \alpha &= \arccos\left(\frac{(u, v)}{\|u\|\|v\|}\right) = \arccos\left(\frac{\int_\Omega x \, \mathrm{d}x}{\left(\int_\Omega 1 \, \mathrm{d}x\right)^{1/2}\left(\int_\Omega x^2 \, \mathrm{d}x\right)^{1/2}}\right) \\ &= \arccos\left(\frac{0}{\sqrt{2}\sqrt{\frac{2}{3}}}\right) = \frac{\pi}{2} \end{aligned}$$

(they are orthogonal). The same result will be obtained for any two $L^2(-1, 1)$-functions f and g, where f is even and g odd or vice versa.

Orthogonality of elements in Hilbert spaces is defined as the reader expects:

Definition A.44 (Orthogonality and OG complement) *Let V be an Hilbert space. The elements $0 \neq u, v \in V$ are said to be* orthogonal *if $(u, v) = 0$. An element $0 \neq v \in V$ is said to be* orthogonal *to a nonempty subset $S \subset V$ if $(v, s) = 0$ for all $s \in S$. We define* orthogonal complement *of S as*

$$S^{\perp} = \{v \in V;\ (v, s) = 0 \ \textit{for all } s \in S\}.$$

Lemma A.35 (OG complement) *Let S be a nonempty subset of a Hilbert space V. Then $S^{\perp}$ is a closed subspace of V.*

Proof: Immediately from Definition A.2. ■

Definition A.45 (OG and ON basis) *Let $B = \{v_1, v_2, \ldots\}$ be a basis of a Hilbert space V. B is said to be* orthogonal (OG) *if*

$$(v_i, v_j) = 0 \qquad \textit{whenever } i \neq j,$$

The basis B is orthonormal (ON) *if*

$$(v_i, v_j) = \delta_{ij} \qquad \textit{for all } i, j.$$

Every basis of a separable Hilbert space can be transformed into an orthonormal basis. (A linear space is separable if it has a finite or countable infinite basis.) The orthonormalization procedure is called after a German mathematician Erhardt Schmidt (1876–1959), who contributed significantly to the development of the theory of Hilbert spaces. He published this result in 1907.

Theorem A.12 (Gram–Schmidt orthogonalization) *Let $B = \{w_1, w_2, \ldots\}$ be some basis in a Hilbert space V. Then there exists an orthonormal basis $\tilde{B} = \{v_1, v_2, \ldots\}$ such that*

$$V_n = \operatorname{span}\{v_1, v_2, \ldots, v_n\} = \operatorname{span}\{w_1, w_2, \ldots, w_n\} \tag{A.69}$$

for all $n \geq 1$.

Proof: The proof is done inductively. For $n = 1$ define $v_1 = w_1/\|w_1\|$. Next assume the existence of $n-1$ orthonormal functions $v_1, v_2, \ldots, v_{n-1}$ satisfying

$$V_{n-1} = \operatorname{span}\{v_1, v_2, \ldots, v_{n-1}\} = \operatorname{span}\{w_1, w_2, \ldots, w_{n-1}\}.$$

Define an element $\tilde{w}_n \in V_{n-1}$ by

$$\tilde{w}_n := \sum_{i=1}^{n-1} (w_n, v_i) v_i, \tag{A.70}$$

and another element $\tilde{w}_n^{\perp} := w_n - \tilde{w}_n$. For any $1 \leq k \leq n-1$ we have

$$(\tilde{w}_n^{\perp}, v_k) = (w_n, v_k) - \left(\sum_{i=1}^{n-1} (w_n, v_i) v_i, v_k \right) = (w_n, v_k) - (w_n, v_k) = 0,$$

and thus $\tilde{w}_n^{\perp} \in V_{n-1}^{\perp}$. Since $w_n \notin V_{n-1}$, it is $\|\tilde{w}_n^{\perp}\| \neq 0$, and we can define

$$v_n = \frac{\tilde{w}_n^{\perp}}{\|\tilde{w}_n^{\perp}\|},$$

which finishes the proof. ■

For example, the Legendre polynomials can be constructed via the Gram–Schmidt procedure:

■ EXAMPLE A.44 (Legendre polynomials)

Let us use the monomial basis of the space $V = L^2(-1,1)$, $B_{mon} = \{w_1, w_2, w_3, w_4, \ldots\} = \{1, x, x^2, x^3, \ldots\}$ and the Gram–Schmidt process to create an orthonormal basis of the space V. In the first step normalize w_1, $L_0(x) = v_1 = w_1/\|w_1\| = 1/\sqrt{2}$, and define $V_1 = \mathrm{span}\{v_1\}$. Next define the element $\tilde{w}_2 \in V_1$ by

$$\tilde{w}_2 = \sum_{i=1}^{1}(w_2, v_i)v_i = \left(\int_{-1}^{1} x\frac{1}{\sqrt{2}}\,\mathrm{d}x\right)\frac{1}{\sqrt{2}} = 0.$$

Thus $\tilde{w}_2^{\perp} = w_2 - \tilde{w}_2 = x$. Normalizing $\tilde{w}_2^{\perp}$, we obtain

$$L_1(x) = v_2 = \frac{\tilde{w}_2^{\perp}}{\|\tilde{w}_2^{\perp}\|} = \sqrt{\frac{3}{2}}x.$$

Define $V_2 = \mathrm{span}\{v_1, v_2\}$, and the element $\tilde{w}_3 \in V_2$ by

$$\tilde{w}_3 = \sum_{i=1}^{2}(w_3, v_i)v_i = \left(\int_{-1}^{1} x^2\frac{1}{\sqrt{2}}\,\mathrm{d}x\right)\frac{1}{\sqrt{2}} + \underbrace{\left(\int_{-1}^{1} x^2\sqrt{\frac{3}{2}}x\,\mathrm{d}x\right)}_{=0}\sqrt{\frac{3}{2}}x = \frac{1}{3}.$$

Therefore

$$L_2(x) = v_3 = \frac{\tilde{w}_3^{\perp}}{\|\tilde{w}_3^{\perp}\|} = \sqrt{\frac{45}{8}}\left(x^2 - \frac{1}{3}\right).$$

The fourth Legendre polynomial L_3, obtained analogously, has the form

$$L_3(x) = v_4 = \sqrt{\frac{7}{8}}(5x^3 - 3x).$$

The Legendre polynomials of higher degrees are usually defined by means of recurrent formulae (to be found in many books, see, e.g., [111] and [117]). First few Legendre polynomials are depicted in Figure A.27.

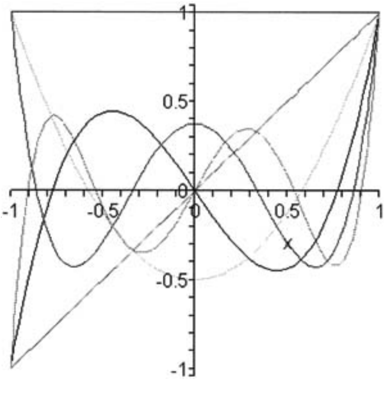


Figure A.27 First five Legendre polynomials $L_0, L_1, \ldots, L_4$.

A.3.4 Generalized Fourier series

The expansion of an element u of a Hilbert space V into an orthonormal basis of V can be viewed as the construction of generalized Fourier series.

Jean Baptiste Joseph Fourier was a French mathematician who made significant contributions to the mathematical theory of propagation of heat in solid bodies. His theory of heat provoked great controversy at his time. The Fourier expansions of real functions into trigonometric series were present already in his famous work "On the Propagation of Heat in Solid Bodies" from 1807.

Figure A.28 Jean Baptiste Joseph Fourier (1768–1830).

After introducing the general theorem, we show an application to the classical Fourier series in Example A.45. Although all results in this paragraph are formulated for infinite-dimensional Hilbert spaces, they obviously hold in the finite-dimensional cases as well.

Theorem A.13 (Generalized Fourier series) *Let V be a Hilbert space and $B = \{v_1, v_2, \dots\}$ an orthonormal basis of V. Then any element $u \in V$ can be written as*

$$u = \sum_{j=1}^{\infty} (u, v_j) v_j. \tag{A.71}$$

Proof: Any element $u \in V$ can be written as

$$u = \sum_{j=1}^{\infty} c_j v_j.$$

From the orthonormality of the basis functions one obtains

$$(u, v_k) = \left(\sum_{j=1}^{\infty} c_j v_j, v_k \right) = \sum_{j=1}^{\infty} c_j \underbrace{(v_j, v_k)}_{\delta_{jk}} = c_k,$$

and (A.71) follows. ■

Important consequences of Theorem A.13 are the generalized Parseval equality and the generalized Bessel inequality:

Lemma A.36 (Generalized Parseval equality) *Let V be a real Hilbert space, $B = \{v_1, v_2, \dots\}$ an orthonormal basis of V and $u \in V$. Then*

$$\|u\| = \sqrt{\sum_{i=0}^{\infty}(u, v_i)^2}. \tag{A.72}$$

Proof: Write $\|u\|^2 = (u, u)$ and apply Theorem A.13. ■

Let us remark that in complex Hilbert spaces, (A.72) holds in the form

$$\|u\| = \sqrt{\sum_{i=0}^{\infty}|(u, v_i)|^2}.$$

Lemma A.37 (Generalized Bessel inequality) *Let V be a real Hilbert space, $B = \{v_1, v_2, \dots\}$ an orthonormal basis of V and $u \in V$. Then*

$$\sum_{i=0}^{n}(u, v_i)^2 \le \|u\|^2 \quad \textit{for any } n \ge 1. \tag{A.73}$$

Proof: Immediately from (A.72). ■

In complex Hilbert spaces (A.73) holds in the form

$$\sum_{i=0}^{n}|(u, v_i)|^2 \le \|u\|^2 \quad \text{for any } n \ge 1.$$

■ EXAMPLE A.45 (Fourier series)

It is well known (see, e.g., [26, 38] and [84]) that the 2π-periodic functions

$$\frac{1}{\sqrt{2\pi}}, \frac{\cos(x)}{\sqrt{\pi}}, \frac{\sin(x)}{\sqrt{\pi}}, \frac{\cos(2x)}{\sqrt{\pi}}, \frac{\sin(2x)}{\sqrt{\pi}}, \dots \tag{A.74}$$

constitute an orthonormal basis in the space $L^2(-\pi, \pi)$. But we are not limited to the interval $(-\pi, \pi)$. Consider, for example, the function $\tilde{g} \in L^2(0, 2\pi)$ defined by

$$\tilde{g}(x) = \begin{cases} x, & x \in (0, \pi), \\ x - 2\pi & x \in [\pi, 2\pi). \end{cases}$$

Using the 2π-periodicity of the basis functions (A.74), equivalently we can consider the function $g(x) = x$ in the interval $(-\pi, \pi)$. The Fourier series, obtained using the procedure from Theorem A.13, has the form

$$g_n(x) = 2\sum_{i=1}^{n}(-1)^{i+1}\frac{\sin(ix)}{i}, \qquad n = 1, 2, \dots$$

This series can be visualized, e.g., in Maple:

```
> g1(x) := ‘if‘(x<3.141593,x,NULL):
> g2(x) := ‘if‘(x>3.141593,x-2*3.141593,NULL):
> n := 100:
> g_n(x) := 2*sum(sin(i*x)*(1./i)*(-1)^(i+1),i=1..n):
> plot([g_n(x),g1(x),g2(x)], x=0..2*3.141593, thickness=1);
```

The functions q_n for $n = 1, 2, 3, 4, 5, 6, 20, 200,$ and 5000 are presented in Figure A.29 (only the period $(0, 2\pi)$ is shown).

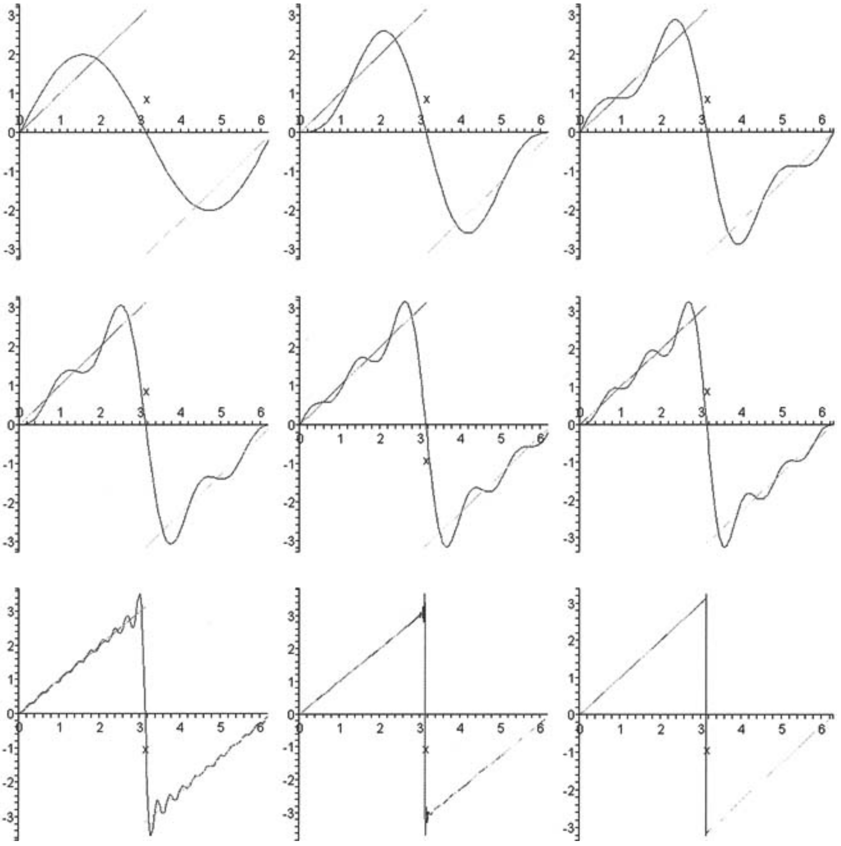


Figure A.29 Fourier series of the discontinuous function $\tilde{g} \in L^2(0, 2\pi)$.

A.3.5 Projections and orthogonal projections

Projections form the basis of many modern numerical methods including the finite element method. As promised at the end of Paragraph A.1.5, let us study in more detail their properties and relations to direct sums.

Definition A.46 (Projection operator) *Let V be a linear space. An operator $P : V \to V$ is said to be a* projection *if it is both linear and idempotent ($P^2 = P$). The range $R(P)$ of a projection operator P is called the* projection subspace.

Sometimes one uses the symbol $P(V)$ for the range $R(P)$. By saying $P^2 = P$ we mean that $P(Pv) = P(v)$ for all $v \in V$. There is a one-to-one relation between projections and direct sums:

Lemma A.38 (Projections and direct sums) *Let V be a linear space. If V is a direct sum $V = V_1 \oplus V_2$ of subspaces $V_1, V_2 \subset V$, then there exists a unique projection operator $P : V \to V$, $P^2 = P$, $P(V) = V_1$, $(I - P)(V) = V_2$. Conversely, every projection operator P determines a decomposition of the space V into the direct sum*

$$V = P(V) \oplus (I - P)(V).$$

Proof: First assume that $V = V_1 \oplus V_2$. Then every element $v \in V$ can be decomposed uniquely into $v = v_1 + v_2$, where $v_1 \in V_1$ and $v_2 \in V_2$. Define the operator $P : V \to V$ by $Pv := v_1$. This operator is unique by its definition and it is easy to verify that P is both linear and idempotent. Moreover, $P(V) = V_1$, and since $v_2 = (I - P)v$ for all $v \in V$, it also holds $(I - P)(V) = V_2$.

Conversely, assume a projection operator $P : V \to V$. Since both P and $I - P$ are linear operators, their ranges $V_1 = \{Pv;\ v \in V\}$ and $V_2 = \{v - Pv;\ v \in V\}$ are subspaces of V (see Lemma A.9). Every element $v \in V$ can be decomposed into $v = v_1 + v_2$, where $v_1 = Pv \in V_1$ and $v_2 = v - Pv \in V_2$. Using the property $P^2 = P$, it is easy to see that $V_1 \cap V_2 = \{0\}$. ■

The interpolation is an example of a projection operator:

■ EXAMPLE A.46 (Lagrange interpolation as a projection operator)

Let $V = C([a, b])$ and $W = P^n([a, b]) \subset V$. Consider a partition $a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b$. Define $P : V \to V$ by

$$Pv = \sum_{i=0}^{n} \left(\prod_{j \neq i} \frac{x - x_j}{x_i - x_j} \right) v(x_i) \qquad \text{for all } v \in V. \tag{A.75}$$

Here Pv is the unique Lagrange interpolant of v, satisfying $Pv \in W$ and $(Pv)(x_i) = v(x_i)$ for all $i = 0, 1, \ldots, n$. It is easy to verify that the operator P is linear and idempotent. The projection subspace $P(V) = W$. According to Lemma A.38 the space V can be written as the direct sum $V = P^n(a, b) \oplus (I - P)(V)$, where the space $(I - P)(V)$ contains continuous functions that vanish at all interpolation points $x_0, x_1, \ldots, x_n$.

Next let us return to Hilbert spaces and introduce orthogonal projections and orthogonal direct sums:

Definition A.47 (Orthogonal projection) *Let V be a Hilbert space. An operator $P : V \to V$ is said to be* orthogonal projection *if it is linear, idempotent ($P^2 = P$) and if*

$$(v - Pv, w)_V = 0 \qquad \textit{for all } v \in V,\ w \in P(V). \tag{A.76}$$

The space $P(V)$ is said to be the projection subspace.

Definition A.48 (Orthogonal direct sum) *Let V be a Hilbert space. A direct sum $V = V_1 \oplus V_2$ is said to be* orthogonal *if $(v_1, v_2) = 0$ for all $v_1 \in V_1$, $v_2 \in V_2$.*

The following theorem summarizes several properties of orthogonal projections and their relation to orthogonal direct sums:

Theorem A.14 (OG projections and OG direct sums) *Let V be a Hilbert space and $V_1 \subset V$ a closed subspace of V endowed with an orthonormal basis $B_{V_1} = \{w_1, w_2, \ldots\}$. Then the relation*

$$P(v) = \sum_{i=1}^{\infty} (v, w_i)_V w_i \qquad \textit{for all } v \in V \tag{A.77}$$

defines a unique orthogonal projection operator $P \in \mathcal{L}(V, V)$, $P^2 = P$, $(v - Pv, w) = 0$ for all $v \in V$ and $w \in P(V)$. Moreover, $\|P\| = 1$ and $V = P(V) \oplus (I - P)(V)$ is an orthogonal direct sum.

The assumption of closedness of the subspace V_1 is essential and we will discuss it in more detail in Remark A.5 and Example A.48. On the contrary, the assumption of the existence of an orthonormal basis B_{V_1} is not necessary, since one always can have such basis by Theorem A.12. Nevertheless, we find it useful to introduce the orthonormal basis explicitly, since this is how the orthogonal projections always are done in practice.

Proof: Given the operator (A.77), let us verify all properties listed in the lemma. First, the linearity of P follows easily from the linearity of the inner product $(\cdot, \cdot)_V$. The operator is idempotent, since

$$\begin{aligned} P(Pv) &= \sum_{j=1}^{\infty} \left(\sum_{i=1}^{\infty} (v, w_i)_V w_i, w_j \right)_V w_j \\ &= \sum_{j=1}^{\infty} \sum_{i=1}^{\infty} (v, w_i)_V \underbrace{(w_i, w_j)_V}_{\delta_{ij}} w_j \\ &= \sum_{j=1}^{\infty} (v, w_j)_V w_j = Pv \qquad \text{for all } v \in V. \end{aligned}$$

Obviously V_1 is the projection subspace. It is sufficient to verify the orthogonality condition $(v - Pv, w_k) = 0$ for all $v \in V$ and all elements of the basis $w_k \in B_{V_1}$:

$$(v - Pv, w_k) = (v, w_k) - \sum_{i=1}^{\infty} (v, w_i)(w_i, w_k) = (v, w_k) - (v, w_k) = 0.$$

It is easy to see that $\|P\| = 1$ since it follows from the Bessel inequality and the Parseval equality immediately that

$$\frac{\|Pv\|_V}{\|v\|_V} \le 1 \quad \text{for all } v \in V \qquad \text{and} \qquad \frac{\|Pv\|_V}{\|v\|_V} = 1 \quad \text{for all } v \in V_1.$$

The rest is straightforward. In particular, $V_2 = (I - P)(V) = V_1^{\perp}$ is a closed subspace of V (Lemma A.35) and $I - P$ is a unique projection operator onto V_2. ■

Remark A.5 (Closedness of projection subspaces) *Every finite-dimensional subspace W of a Hilbert space V obviously is a Hilbert space. However, this generally is no longer true when the subspace W is infinitely-dimensional. The problem is that Cauchy sequences lying in W can "converge out" of W into V. This can happen, for example, when the subspace W is dense in V (see Example A.48). Therefore one has to add to W the limits of all Cauchy sequences lying in W. The completion is possible by Theorem A.7.*

■ EXAMPLE A.47 (An OG projection operator in $V = \mathbb{R}^2$)

Let $V = \mathbb{R}^2$ endowed with the standard Euclidean inner product. Let us define an operator $P : V \to V$ via the relation

$$Pv = (v_1, 0)^T \qquad \text{for all } v \in V, \quad v = (v_1, v_2)^T.$$

(Thus P erases the second component of vectors in V.) Let us see that P is both linear and idempotent, identify the projection subspace $P(V)$, and check whether P is an orthogonal projection. First the linearity:

$$P(\alpha u + \beta v) = (\alpha u_1 + \beta v_1, 0)^T = (\alpha u_1, 0)^T + (\beta v_1, 0)^T = \alpha Pu + \beta Pv.$$

The idempotency follows from the fact that the second vector component only can be erased once,

$$P(Pv) = P((v_1, 0)^T) = (v_1, 0)^T = Pv \qquad \text{for all } v \in V.$$

Hence P is a projection. Clearly the projection subspace $P(V) = [(1,0)^T] = [w_1]$. Since

$$(v - Pv, w_1) = (v_1 - v_1, v_2) \cdot (1, 0)^T = 0 \qquad \text{for all } v \in V,$$

we see that P is an orthogonal projection operator.

Another reason why orthogonal projections are so useful, is that Pv is the closest element to $v \in V$ among all elements in the projection space $P(V)$:

Lemma A.39 *Let V be a Hilbert space and W a closed subspace of V equipped with an orthonormal basis $B_W = \{w_1, w_2, \ldots\}$. Let P be an orthogonal projection operator such that $P(V) = W$. Then for any $v \in V$ we have*

$$\|v - Pv\| = \inf_{w \in W} \|v - w\|.$$

Proof: Write $w = Pv + z$. Since $w \in W$ and $Pv \in W$, necessarily also $z \in W$. We have $\|v - w\|^2 = (v - w, v - w) = (v - Pv - z, v - Pv - z)$. Since $z \in W$ it is $(v - Pv, z) = 0$. Therefore

$$(v - Pv - z, v - Pv - z) = \|v - Pv\|^2 + \|z\|^2 \geq \|v - Pv\|^2,$$

which had to be shown. ■

Let us close this paragraph by showing a subspace of a Hilbert space which is not closed:

■ **EXAMPLE A.48 (Subspaces which cannot be projection subspaces)**

Let V be a Hilbert space and $W \subset V$ a dense subspace of V such that $W \neq V$. In this case an orthogonal projection cannot be defined. Namely, it follows from Lemma A.39 that the projection $Pv \in W$ of any $v \in V \setminus W$ would have to satisfy

$$\|v - Pv\| = \inf_{w \in W} \|v - w\|.$$

However, this minimum is zero by the density of W in V. In turn $v = Pv$ which is in contradiction to $W \neq V$. For illustration take, e.g., the Hilbert space $V = L^2(a,b)$, where $(a,b) \subset \mathbb{R}$ is a bounded interval, and its dense subspace $W = P^{\text{fin}}(a,b)$ of polynomials of finite degrees.

A.3.6 Representation of linear forms (Riesz)

The Riesz representation theorem is a fundamental tool in the solvability analysis of elliptic partial differential equations. It was first proved in 1907 for the Lebesgue L^2-space by Frigyes Riesz, a Hungarian mathematician who is assumed to be one of the founders of functional analysis and operator theory. F. Riesz introduced the concept of weak convergence (to be discussed in Paragraph A.3.8), and he made many contributions to other areas of mathematics including orthonormal series, ergodic theory, and topology.

Figure A.30 Frigyes Riesz (1880–1956).

Theorem A.15 (Riesz) *Let V be a Hilbert space and $\varphi \in V'$ an arbitrary linear form on V. Then there exists a unique element $u \in V$ such that*

$$\varphi(v) = (u, v) \qquad \textit{for all } v \in V.$$

Moreover, $\|\varphi\|_{V'} = \|u\|_V$.

Proof: We restrict ourselves to real Hilbert spaces (see, e.g., [65] for the complex case). First let us prove the uniqueness: If there exist two elements $u, \tilde{u} \in V$ such that

$$\varphi(v) = (v, u) = (v, \tilde{u}) \qquad \text{for all } v \in V,$$

then by the linearity of the inner product it is

$$(v, u - \tilde{u}) = 0 \qquad \text{for all } v \in V.$$

Taking $v = u - \tilde{u}$, we see that $u = \tilde{u}$.

Next let us prove the existence: If the null space $N(\varphi) = V$, then φ is the zero functional and we can define $u = 0$. If $N(\varphi) \neq V$, then there exists an element $v_0 \in V$ such that $\varphi(v_0) \neq 0$. Since $N(\varphi)$ is a closed subspace of V, it is possible to write V as an orthogonal direct sum $V = N(\varphi) \oplus N(\varphi)^{\perp}$. Thus the element $v_0 \in V$ can be decomposed uniquely into the sum $v_0 = v_1 + v_2$, $(v_1, v_2) = 0$. where $v_1 \in N(\varphi)$ and $v_2 \in N(\varphi)^{\perp}$. In particular, it is $\varphi(v_2) \neq 0$. The following holds:

$$\varphi\left(v - \frac{\varphi(v)}{\varphi(v_2)} v_2\right) = \varphi(v) - \frac{\varphi(v)}{\varphi(v_2)} \varphi(v_2) = 0 \qquad \text{for all } v \in V.$$

Thus

$$v - \frac{\varphi(v)}{\varphi(v_2)} v_2 \in N(\varphi) \qquad \text{for all } v \in V.$$

Since $v_2 \in N(\varphi)^{\perp}$, it is

$$\left(v - \frac{\varphi(v)}{\varphi(v_2)} v_2, v_2\right) = 0 \qquad \text{for all } v \in V.$$

From this equation we obtain

$$\varphi(v) = \left(v, \frac{\varphi(v_2)}{\|v_2\|^2} v_2\right) \qquad \text{for all } v \in V,$$

and thus

$$u = \frac{\varphi(v_2)}{\|v_2\|^2} v_2. \tag{A.78}$$

It remains to be shown that $\|\varphi\|_{V'} = \|u\|_V$. The Cauchy–Schwarz inequality yields

$$|\varphi(v)| \leq \|u\|_V \|v\|_V \qquad \text{for all } v \in V,$$

and thus $\|\varphi\|_{V'} \leq \|u\|_V$. Choosing $v = u$, one obtains

$$\varphi(u) = (u, u) = \|u\|_V^2,$$

which establishes the equality $\|\varphi\|_{V'} = \|u\|_V$. ■

The procedure shown in the proof of the Riesz theorem allows us to construct the representants of linear forms over Hilbert spaces explicitly, via the formula (A.78). Another important consequence of the Riesz theorem is the reflexivity of Hilbert spaces:

Lemma A.40 *Every Hilbert space V is reflexive, i.e., $(V')' = V$.*

Proof: See, e.g., [99]. ■

A.3.7 Compactness, compact operators, and the Fredholm alternative

Besides the Lax–Milgram lemma, the Fredholm alternative is another basic tool for proving the existence and uniqueness of solution to certain classes of operator equations. This technique does not assume the V-ellipticity of the underlying operator. Instead, it assumes its compactness.

Definition A.49 (Compact and precompact set) *Let V be a normed space. Then a subset $S \subset V$ is said to be* compact *if every sequence $\{s_n\}_{n=1}^{\infty} \subset S$ contains a subsequence that converges to some element $s \in S$. A subset S of a normed space V is* precompact (relatively compact) *if its closure $\overline{S}$ is compact.*

The following characterization of compactness holds for finite-dimensional spaces:

Theorem A.16 (Heine–Borel) *Let V be a finite-dimensional normed space and let S be a subset of V. Then S is compact if and only if S is both closed and bounded.*

Proof: See, e.g., [99]. ■

The situation is much less trivial in infinite-dimensional spaces, where one can find sets which are both closed and bounded, but not compact. This is illustrated in the following example.

■ **EXAMPLE A.49 (Noncompactness of the closed unit ball in l^2)**

Consider the normed space $V = l^2$ of infinite real sequences with the discrete l^2-norm

$$\|\{u_n\}_{n=1}^{\infty}\|_V = \left(\sum_{n=1}^{\infty} |u_n|^2\right)^{\frac{1}{2}}.$$

The closure of the unit ball

$$\overline{B(0,1)} = \{\{u_n\}_{n=1}^{\infty};\ \|\{u_n\}_{n=1}^{\infty}\|_V \le 1\} \subset V$$

clearly is both closed and bounded. By $v_i \in V$ we denote a sequence which has 1 at the ith position and zeros everywhere else. It is

$$\|v_i\|_V = 1 \qquad \text{for all } i = 1, 2, \ldots \tag{A.79}$$

and therefore $v_i \in \overline{B(0,1)}$ for all $i = 1, 2, \ldots$. The only candidate for the limit of the sequence $\{v_i\}_{i=1}^{\infty} \subset \overline{B(0,1)}$ is the zero sequence, but by (A.79) the sequence $\{v_i\}_{i=1}^{\infty}$ does not contain any convergent subsequence. Therefore, according to Definition A.49, $\overline{B(0,1)}$ is not compact.

Corollary A.2 *An immediate consequence of Theorem A.16 is that in a finite-dimensional normed space every bounded sequence contains a convergent subsequence.*

Next let us introduce the notion of a compact operator:

Definition A.50 (Compact operator) *Let V, W be normed spaces. A linear operator $A : V \longrightarrow W$ is said to be* compact *if the image of any bounded subset of V is relatively compact in W.*

Indeed any compact operator is bounded. A standard way to prove the compactness of an operator is to show that the image $\{Au_n\}_{n=1}^{\infty} \subset W$ of any bounded sequence $\{u_n\}_{n=1}^{\infty} \subset V$ contains a convergent subsequence. The following lemma characterizes the composition of bounded and compact operators:

Lemma A.41 *Let V, W, Z be normed spaces, and let $A : V \longrightarrow W$ and $B : W \longrightarrow Z$ be bounded linear operators. The composition $C = AB$ is compact if at least one of the operators A, B is compact.*

Proof: This is a classical result, see, e.g., [100]. ■

Another basis result characterizes the compactness of the identity operator:

Lemma A.42 *Let V be a normed space. The identity operator $I : V \longrightarrow V$ is compact if and only if the space V is finite-dimensional.*

Proof: The right-left implication is a simple consequence of Theorem A.16. For the other implication see, e.g., [100]. ■

Now we can introduce the Fredholm alternative, or, more precisely, its version that is most suitable for our primary purpose, which is the application to the Maxwell's equations in Chapter 7. This theorem can be found in greater generality, e.g., in [73] and [100].

Theorem A.17 (Fredholm alternative) *Let V be a Hilbert space and $B : V \longrightarrow V$ a bounded linear operator of the form $B = I + A$, where I is the identity operator and A is compact. Then exactly one of the following holds:*

1. *The homogeneous equation $Bu = 0$ has a unique solution $u = 0$. Then the inhomogeneous equation $Bu = f$ has a unique solution for every $f \in V$.*

2. *The homogeneous equation $Bu = 0$ has n linearly independent solutions $u_1, u_2, \ldots, u_n$ in V, where $p > 0$ is an integer number.*

Proof: See, e.g., [73] and [100]. ■

A.3.8 Weak convergence

The concept of weak convergence was introduced by F. Riesz around 1910. It finds important applications in the theory of PDEs and finite element methods by generalizing the standard (strong) convergence in norm:

Definition A.51 (Weak convergence) *Let V be a Hilbert space and V' the dual of V. We say that a sequence $\{u_n\}_{n=1}^{\infty} \subset V$* converges weakly *to an element $u \in V$ if*

$$\lim_{n \to \infty} \varphi(u_n) = \varphi(u) \qquad \textit{for all } \varphi \in V'.$$

The element u is said to be the weak limit *of the sequence.*

It is easy to see that the weak limit is unique and identical to the strong one if they both exist. Moreover, the convergence in norm implies the weak convergence, since for an arbitrary $\varphi \in V'$ we have

$$|\varphi(u_n) - \varphi(u)| \leq \|\varphi\|_{V'} \|u_n - u\|_V.$$

However, the next example shows that the weak convergence does not imply the convergence in norm:

■ EXAMPLE A.50 (Weak convergence $\not\Rightarrow$ convergence in norm)

Consider the sequence $\{u_n\}_{n=1}^{\infty}$, $u_n = \sin(nx)/\sqrt{\pi}$, in the space $V = L^2(-\pi, \pi)$. By the Riesz theorem, for an arbitrary linear form $\varphi \in V'$ we have

$$\varphi(u_n) = (u_\varphi, u_n) \qquad \text{for all } n = 1, 2, \ldots,$$

where $u_\varphi \in V$ is the unique representant of the form φ. Since the elements u_n belong to the orthonormal basis (A.74) of the space V, and the entries $(u_\varphi, u_n)^2$ of the Parseval sum (A.72) of u_φ must converge to zero as $n \to \infty$, we see that

$$\lim_{n\to\infty} \varphi(u_n) = 0 \qquad \text{for all } \varphi \in V'.$$

Thus $\{u_n\}_{n=1}^{\infty}$ converges weakly to zero. However, the sequence cannot converge to zero strongly since $\|u_n\| = 1$ for all n.

From Corollary A.2 we know that every bounded sequence in a finite-dimensional normed space contains a convergent subsequence. This is not true in infinite-dimensional spaces, but there is an important weaker analogy:

Theorem A.18 (Eberlein–Smulyan) *Every bounded sequence in a reflexive Banach space V contains a weakly-convergent subsequence.*

Proof: See, e.g., [34] and [93]. ■

A.3.9 Exercises

Exercise A.42 *Use Definition A.41 to verify in detail that the "dot-product" (A.56) in $\mathbb{R}^n$ and the L^2-inner product (A.58) indeed are inner products.*

Exercise A.43 *In the space $L^2(0,1)$ calculate the angle of the functions $f_m(x) = x^m$ and $g_n(x) = x^n$, where m, n are arbitrary natural numbers.*

Exercise A.44 *In $\mathbb{R}^2$ consider a general parallelogram $ABCD$, as shown in Figure A.31. Use the Theorem of Pythagoras to prove that $|AD|^2 + |BC|^2 = 2|AB|^2 + 2|AC|^2$.*

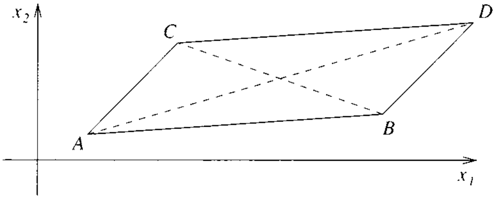


Figure A.31 Parallelogram $ABCD$ in $\mathbb{R}^2$.

Exercise A.45 *Consider a normed space V where the norm $\|\cdot\|$ satisfies the parallelogram rule (A.62),*

$$\|u+v\|^2 + \|u-v\|^2 = 2\|u\|^2 + 2\|v\|^2 \qquad \textit{for all } u, v \in V.$$

Show that the identity

$$\|u+v+w\|^2 - \|u+v-w\|^2 = \|u+w\|^2 - \|u-w\|^2 + \|v+w\|^2 - \|v-w\|^2$$

holds for all $u, v, w \in V$.

Exercise A.46 *Prove Lemma A.35: If S is a nonempty subset of an inner product space V, then $S^{\perp}$ is a subspace of V.*

Exercise A.47 *Consider the Hilbert space $V = \mathbb{R}^3$, endowed with the standard Euclidean inner product, and its subspace W given by a basis $B_W = \{(0,1,0)^T, (1/\sqrt{2}, 0, 1/\sqrt{2})^T\}$.*

1. *Show that B_W is a orthonormal basis of W.*
2. *Use a vector $e_3 = (0,0,1)^T$ and the Gram–Schmidt orthogonalization procedure to extend the basis B_W to an orthogonal basis B_V of V.*

Exercise A.48 *Given $v = (1,2,3)^T \in \mathbb{R}^3$, calculate its projection to the subspace W from Exercise A.47.*

Exercise A.49 *Let $V = L^2(-\pi,\pi)$. Use the relation (A.71) and the Fourier basis of V from Example A.45 to construct the Fourier series of the function $g(x)$, defined as $g(x) = -1$ for all $x \in (-\pi, 0)$ and $g(x) = 1$ in $[0,\pi)$. Present a formula for general n and computer-generated plots of first ten different entries of the series. Hint: Since g is an odd function in $(-\pi,\pi)$, the cosinus part of the series is not present. Additional cancellations occur.*

Exercise A.50 *Repeat Exercise A.49 with the function $\tilde{g}(x) = x$ in the space $L^2(-\pi,\pi)$. Check your result with Figure A.29. Hint: Use the integration-by-parts formula to integrate functions of the form $x\sin(mx)$.*

Exercise A.51 *Use Definition A.46 to verify that that the Lagrange interpolation operator P introduced in Example A.46 is a topological projection. What functions belong to the subspace $V_2 = (I-P)(V) \subset V$?*

Exercise A.52 *Consider the Hilbert space $V = P^5(-1,1)$ endowed with the L^2-inner product, and the subspace $W = P^3(-1,1) \subset V$. Let $B_W = \{L_0, L_1, L_2, L_3\}$, where L_i are the Legendre polynomials derived in Example A.44. Calculate the orthogonal projection $P_W(f_0)$ of the function $f_0(x) = 1 + x^2 - x^3 + x^4 - x^5$ onto W. Calculate the distance* $\mathrm{dist}(f_0, W) = \inf_{w\in W}\|f_0 - w\|$. *Hint: By Lemma A.39 the distance is $\|f_0 - P_W(f_0)\|$.*

Exercise A.53 *Let $V = \mathbb{R}^5$ equipped with the "dot-product" (A.56). Consider the linear operator $f : V \to V$, $f(v) = \boldsymbol{A}v$, where*

$$\boldsymbol{A} = \begin{pmatrix} 1 & -1 & 0 & 2 & 3 \\ 3 & -1 & 3 & 2 & 4 \\ 4 & -2 & 3 & 4 & 7 \\ 0 & 2 & 3 & -4 & -5 \\ 2 & 0 & 3 & 0 & 1 \end{pmatrix}.$$

Construct an orthonormal basis of the nullspace $N(f)$.

Exercise A.54 *Consider the Hilbert space $V = P^5(-1,1)$ equipped with the L^2-inner product.*

- *What is the dimension of the subspace $W = \{w \in V;\ w(0) = 0\}$?*
- *Choose some basis $\mathcal{B}$ of W.*
- *Use the basis $\mathcal{B}$ to construct an orthonormal basis $\mathcal{B}_{ON}$ of W.*
- *Explain why there exists a unique orthogonal projection operator $P : V \to W$.*
- *Assume the function $g(x) = x^3 + x^2 + x + 1 \in V$. Calculate $P(g)$.*
- *Calculate the distance $d = dist(g, W)$, i.e.,*

$$d = \inf_{w \in W} \|g - w\|_V.$$

Exercise A.55 *Consider the Hilbert space $V = l^2$ of infinite real sequences equipped with the inner product*

$$(u, v)_{l^2} = \sum_{i=1}^{\infty} u_i v_i,$$

and the linear form $f \in V'$,

$$f(u) = \sum_{i=1}^{100} u_i.$$

Find the unique Riesz representant of f in the space V.

Exercise A.56 *Consider the Hilbert space $V = L^2(-1,1)$ and the linear form $f \in V'$,*

$$f(u) = \frac{1}{2}\int_{-1/2}^{1/2} u(x)\, dx.$$

Find the unique Riesz representant of f in the space V.

Exercise A.57 *Consider the Hilbert space $V = P^5(-1,1)$ equipped with the L^2-inner product. Let the linear form $f \in V'$ be given by $f(u) = u(0)$. Find the unique Riesz representant of f in the space V. Hint: The results of Exercise A.54 include an orthonormal basis in the nullspace $N(f)$. Choose some suitable $v_0 \in V \setminus N(f)$ and apply the Riesz formula (A.78).*

Exercise A.58 *Let V, W be Hilbert spaces. Use Definition A.50 to show that every compact operator $A : V \to W$ is bounded.*

Exercise A.59 *Consider the Hilbert space l^2 from Exercise A.55. Find a sequence $\{u_n\}_{n=1}^{\infty}$ that converges weakly to the zero sequence but does not converge in norm (prove both statements).*

Exercise A.60 *Let V be a Hilbert space and $\{u_n\}_{n=1}^{\infty} \subset V$ a sequence in V. Show that if $u, v \in V$ are weak limits of the sequence, then $u = v$. Further, suppose that the sequence is convergent in norm to an element $w \in V$. Show that necessarily $u = w$.*

A.4 SOBOLEV SPACES

We know from Paragraph A.2.9 that the Lebesgue L^p-spaces control the regularity of functions: They do not admit functions with singularities whose strength exceeds certain limit or, in unbounded domains, whose decay at infinity is slower than certain rate. The Sobolev spaces $W^{k,p}$ are subspaces of L^p-spaces that, moreover, control the regularity of the derivatives. Their structure and properties make them particularly suitable for the analysis of partial differential equations.

These spaces were introduced in the 1930s by Sergei Lvovich Sobolev, a Russian mathematician who essentially influenced the field of analysis and solution of partial differential equations. To mention at least a few of his results, he derived important inequalities on the norms in the Sobolev spaces, formulated and proved results on their embeddings (some of them to be mentioned in Paragraph A.4.6), introduced the notion of generalized functions (distributions), etc. In the 1950s he turned his attention to the computational mathematics and achieved important results in interpolation of multivariate functions and numerical quadrature in higher spatial dimensions.

Figure A.32 Sergei Lvovich Sobolev (1908–1989).

The Sobolev spaces will be presented in Paragraph A.4.3, after imposing certain regularity to the boundaries of open sets in Paragraph A.4.1 and introducing the concepts of distributions and weak derivatives in Paragraph A.4.2.

A.4.1 Domain boundary and its regularity

Until now we have worked with open bounded sets without paying special attention to their boundaries. This will change in this section, since we will need to use the unit outer normal vector to the boundary and calculate surface integrals. Let us begin with introducing the notion of a domain:

Definition A.52 (Domain in $\mathbb{R}^d$) *A subset $\Omega \subset \mathbb{R}^d$ is said to be a* domain *if it is nonempty, open and connected.*

Set $\Omega \subset \mathbb{R}^d$ is said to be connected if every two points in Ω can be connected by a continuous curve that lies in Ω.

Figure A.33 illustrates what is and what is not a domain:

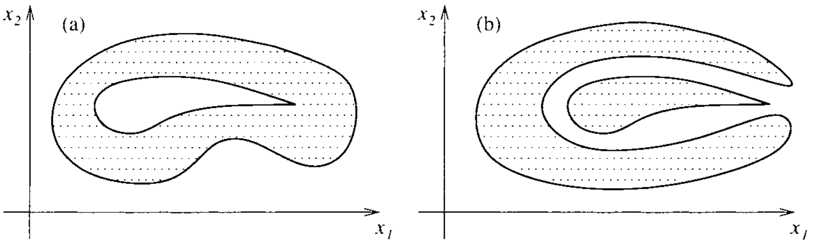


Figure A.33 An open bounded set which (a) is and (b) is not a domain.

The boundary of some domains may be highly irregular, as shown in Example A.51.

■ **EXAMPLE A.51 (Bounded set with infinitely long boundary)**

Consider an infinite sequence of bounded domains $\{\Omega_n\}_{n=0}^{\infty}$, $\Omega_n \subset \Omega_{n+1}$, where the domain Ω_0 is, e.g., a symmetric equilateral hexagon with unit edge length. For every n the domain Ω_{n+1} is obtained from Ω_n as follows: Each edge of Ω_n is split into three equally long parts e_{left}, e_{mid}, and e_{right}. An open equilateral triangle of the edge-length $|e_{mid}|$ is attached from outside to e_{mid}. Points lying in the interior of e_{mid} are added. This is illustrated in Figure A.34.

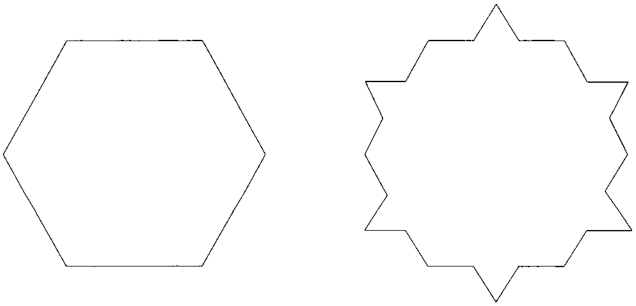


Figure A.34 Construction of a bounded set with infinitely long boundary: domains Ω_0, Ω_1.

It is easy to see that

$$|\partial\Omega_n| = 6\left(\frac{4}{3}\right)^n \qquad \text{for all } n = 0, 1, \dots,$$

and to show that the limit set Ω is bounded. The unit outer normal vector to the boundary $\partial\Omega$ is defined nowhere on $\partial\Omega$.

The above-described situation cannot occur when the boundary $\partial\Omega$ is Lipschitz-continuous:

Lipschitz-continuity of $\partial\Omega$ We assume that the reader knows the definition of Lipschitz-continuity for real-valued functions of one and more variables. The exact definition of the Lipschitz continuity for boundaries of domains in $\mathbb{R}^d$ is rather technical. Roughly speaking, the boundary $\partial\Omega$ is said to be Lipschitz-continuous if there exists a finite covering of $\partial\Omega$ consisting of open d-dimensional rectangles such that in each rectangle, $\partial\Omega$ can be expressed as a Lipschitz-continuous function of $d-1$ variables. See, e.g., [55] for an exact definition. For simply-connected domains (i.e., domains $\Omega \subset \mathbb{R}^d$ such that $\mathbb{R}^d \setminus \Omega$ is connected), the Lipschitz-continuity is equivalent to the cone condition:

Cone condition We say that the boundary $\partial\Omega$ of a d-dimensional bounded domain Ω satisfies the cone condition if and only if there exist constants $\epsilon_1 > 0$ and $\epsilon_2 > 0$ such that for every point $\boldsymbol{x}_0 \in \partial\Omega$ there are two open d-dimensional cones $C_{int}(\boldsymbol{x}_0, \gamma_1, h_1)$ and $C_{ext}(\boldsymbol{x}_0, \gamma_2, h_2)$ sharing the vertex $\boldsymbol{x}_0$, with vertex angles $0 < \epsilon_1 \leq \gamma_1, \gamma_2$ and heights $0 < \epsilon_2 \leq h_1, h_2$, such that $C_{int}(\boldsymbol{x}_0, \gamma_1, h_1) \subset \Omega$ and $C_{ext}(\boldsymbol{x}_0, \gamma_2, h_2) \subset \mathbb{R}^d \setminus \Omega$. Figure A.35 gives examples of domains whose boundary $\partial\Omega$ (a) is, and (b)–(d) is not Lipschitz-continuous. In the case (d) the cone condition is satisfied, but the Lipschitz-continuity is violated at the center of the circle (this situation could not occur if Ω was simply-connected).

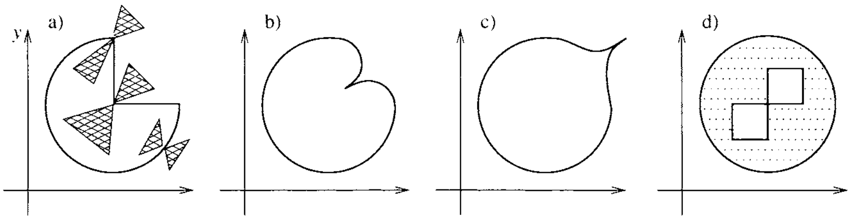


Figure A.35 2D domains whose boundary (a) is, and (b)–(d) is not Lipschitz-continuous.

A unique unit outer normal vector is defined almost everywhere on $\partial\Omega$ when the boundary $\partial\Omega$ is Lipschitz-continuous (see, e.g., [1] and [55]).

A.4.2 Distributions and weak derivatives

The following compact notation is practical for operations with partial derivatives:

Definition A.53 (Multi-index) *Let d be the spatial dimension.* Multi-index *is a vector $(\alpha_1, \alpha_2, \ldots, \alpha_d)$ consisting of d nonnegative integers. By $|\alpha| = \sum_{i=1}^{d} \alpha_i$ we denote the* length *of the multi-index α. Let f be an m-times continuously differentiable function. We define the αth partial derivative of f by*

$$D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \ldots \partial x_d^{\alpha_d}}.$$

Note that $D^\alpha f = f$ for $\alpha = (0, 0, \ldots, 0)$. To give at least two other examples, we have

$$D^\alpha f = \frac{\partial f}{\partial x_1}$$

for $\alpha = (1, 0, 0, \ldots, 0)$, and for $\alpha = (1, 1, \ldots, 1)$ one obtains

$$D^\alpha f = \frac{\partial^d f}{\partial x_1 \partial x_2 \ldots \partial x_d}.$$

The Lebesgue L^p-spaces contain nonsmooth and discontinuous functions whose derivatives are not defined in the classical sense. However, in many cases the classical derivatives exist almost everywhere. What needs to be done is to generalize the notion of the derivative to be independent of zero-measure subsets. This was done by S.L. Sobolev, who introduced weak derivatives. The basic ingredient for the definition of weak derivatives are distributions:

Definition A.54 (Distributions) *Let $\Omega \subset \mathbb{R}^d$ be an open set. The space of* distributions *(infinitely smooth functions with compact support) is defined by*

$$C_0^\infty(\Omega) = \{\varphi \in C^\infty(\Omega);\ \mathrm{supp}(\varphi) \subset \Omega;\ \mathrm{supp}(\varphi) \text{ is compact}\}.$$

Sometimes one uses the symbols $\mathcal{D}(\Omega)$ or $D(\Omega)$ instead of $C_0^\infty(\Omega)$. Recall that the support

$$\mathrm{supp}(\varphi) = \overline{\{\boldsymbol{x} \in \Omega;\ \varphi(\boldsymbol{x}) \neq 0\}}$$

always is both closed and bounded.

■ **EXAMPLE A.52 (Distributions)**

Consider a bounded domain $\Omega = (-1, 1) \subset \mathbb{R}$ and the functions

$$\varphi(x) = \cos(\pi x) + 1, \qquad \psi(x) = e^{-\frac{1}{1-x^2}},$$

depicted in Figure A.36.

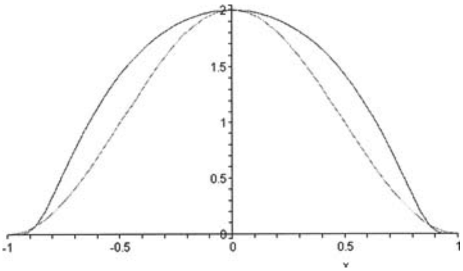


Figure A.36 The functions φ and ψ.

Neither φ nor ψ is a distribution in Ω, since

$$\mathrm{supp}(\varphi) = \mathrm{supp}(\psi) = [-1, 1] \not\subset \Omega.$$

However, the function ψ can be extended by zero to be a distribution in the interval $\tilde{\Omega} = (-1 - \epsilon, 1 + \epsilon)$, where $\epsilon > 0$. This is not possible for the function φ, since already its second derivative would be discontinuous in $\tilde{\Omega}$.

Remark A.6 *Since the support supp(φ) of every distribution $\varphi \in C_0^\infty(\Omega)$ is a closed set, it cannot touch the boundary of the open set Ω. Therefore for every $\varphi \in C_0^\infty(\Omega)$ there is at least a thin belt along the boundary $\partial\Omega$ where φ vanishes entirely.*

Next let us review elementary results related to the integration by parts in higher spatial dimensions, which will be used for the definition of the weak derivatives:

Theorem A.19 (Gauss' theorem) *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz-continuous boundary. For every $u, v \in C^1(\Omega) \cap C(\overline{\Omega})$ we have*

$$\int_\Omega \frac{\partial u}{\partial x_i} v \, dx = -\int_\Omega u \frac{\partial v}{\partial x_i} \, dx + \int_{\partial\Omega} uv\nu_i \, dS. \tag{A.80}$$

Here $\boldsymbol{\nu}(\boldsymbol{x}) = (\nu_1, \nu_2, \ldots, \nu_d)^T(\boldsymbol{x})$ is the unit outer normal vector to the boundary $\partial\Omega$.

Proof: See, e.g., [36]. ■

The formula (A.80) generalizes easily to the divergence of vector fields:

Theorem A.20 (Stokes' theorem) *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz-continuous boundary. Every smooth vector field $\boldsymbol{w} \in [C^1(\Omega) \cap C(\overline{\Omega})]^d$ satisfies*

$$\int_\Omega \nabla \cdot \boldsymbol{w}(\boldsymbol{x}) \, dx = \int_{\partial\Omega} \boldsymbol{w}(\boldsymbol{x}) \cdot \boldsymbol{\nu}(\boldsymbol{x}) \, dS, \tag{A.81}$$

where $\boldsymbol{\nu}(\boldsymbol{x})$ is the unit outer normal to $\partial\Omega$.

Proof: This is an easy exercise using Theorem A.19. ■

By repeated application of Theorem A.19 one easily arrives at the following result:

Theorem A.21 *Let $\Omega \subset \mathbb{R}^d$ be an open set, $f \in C^m(\Omega)$ and α a multi-index such that $|\alpha| \le m$. The following holds:*

$$\int_\Omega D^\alpha f(\boldsymbol{x})\varphi(\boldsymbol{x}) \, dx = (-1)^{|\alpha|} \int_\Omega f(x) D^\alpha \varphi(\boldsymbol{x}) \, dx \quad \text{for all } \varphi \in C_0^\infty(\Omega). \tag{A.82}$$

With this result we are very close to defining the weak derivatives. One last thing we need is the space $L^p_{\text{loc}}(\Omega)$:

Definition A.55 (Space of locally-integrable functions) *Let $\Omega \subset \mathbb{R}^d$ be an open set and $1 \le p < \infty$. A function $f : \Omega \to \mathbb{R}$ is said to be* locally p-integrable in Ω *if $f \in L^p(K)$ for every compact subset $K \subset \Omega$. The space of all locally p-integrable functions in Ω is denoted by $L^p_{\text{loc}}(\Omega)$.*

Remark A.7 *Two remarks to spaces of locally p-integrable functions are in order:*

1. *Obviously it is $L^p(\Omega) \subset L^p_{\text{loc}}(\Omega)$ for every open set $\Omega \subset \mathbb{R}^d$ and every $1 \le p < \infty$. The spaces L^p_{loc} are very large. For example, the function $1/x$ does not belong to the space $L^p(0, \infty)$ for any $p \ge 1$, but it lies in the space $L^p_{\text{loc}}(0, \infty)$ for all $p \ge 1$.*

2. *Let $\Omega \subset \mathbb{R}^d$ be an open (not necessarily bounded) set. The following inclusion holds:*

$$L^p_{\text{loc}}(\Omega) \subset L^q_{\text{loc}}(\Omega) \quad \text{whenever} \quad 1 \le q \le p.$$

Note that for the L^p-spaces, such inclusion was only valid on bounded sets (see Lemma A.29).

Finally, the weak derivatives can be defined:

Definition A.56 (Weak derivative) *Let $\Omega \subset \mathbb{R}^d$ be an open set, $f \in L^1_{\mathrm{loc}}(\Omega)$ and let α be a multi-index. The function $D^\alpha_w f \in L^1_{\mathrm{loc}}(\Omega)$ is said to be the* weak αth derivative *of f if*

$$\int_\Omega D^\alpha_w f\, \varphi\, dx = (-1)^{|\alpha|} \int_\Omega f\, D^\alpha \varphi\, dx \qquad \textit{for all } \varphi \in C^\infty_0(\Omega). \tag{A.83}$$

The following result is needed for the proof of uniqueness of the weak derivative:

Lemma A.43 (Generalized variational lemma) *Let $f \in L^1_{\mathrm{loc}}(\Omega)$ where $\Omega \subset \mathbb{R}^d$ is an open set. If*

$$\int_\Omega f(\boldsymbol{x})\varphi(\boldsymbol{x})\, d\boldsymbol{x} = 0 \qquad \textit{for all } \varphi \in C^\infty_0(\Omega) \tag{A.84}$$

then $f = 0$ almost everywhere in Ω.

Proof: Assume that there exists a nonzero-measure subset $D \subset \Omega$ such that $f \neq 0$ in D. Without loss of generality, we can assume that D is open and $f > 0$ in D. Taking a nonnegative $\varphi \in C^\infty_0(\Omega)$ with a nonempty support supp$(\varphi) \subset D$, we arrive at a positive value of the integral (A.84), which is a contradiction. ■

Lemma A.44 (Uniqueness of weak derivatives) *Let $\Omega \subset \mathbb{R}^d$ be an open set, $f \in L^1_{\mathrm{loc}}(\Omega)$ and let α be a multi-index. The weak αth derivative $D^\alpha_w f \in L^1_{\mathrm{loc}}(\Omega)$ is defined uniquely in Ω up to a zero-measure subset of Ω.*

Proof: Assume that functions $g_1, g_2 \in L^1_{\mathrm{loc}}(\Omega)$ are the weak αth derivatives of f. Then (A.83) implies that

$$\int_\Omega (g_1 - g_2)\varphi\, \mathrm{d}\boldsymbol{x} = 0 \qquad \text{for all } \varphi \in C^\infty_0(\Omega).$$

It follows from Lemma A.43 that $g_1 = g_2$ almost everywhere in Ω. ■

Lemma A.45 (Compatibility of weak and classical derivatives) *Let $\Omega \subset \mathbb{R}^d$ be an open set, $f \in C^m(\Omega)$ and α a multi-index such that $|\alpha| \leq m$. Then the classical αth derivative $D^\alpha f$ is identical to the weak αth derivative $D^\alpha_w f$.*

Proof: This is an immediate consequence of Theorem A.21. ■

■ EXAMPLE A.53 (Weak differentiability in one dimension I)

Continuous, piecewise-smooth functions in 1D are weakly differentiable. This can be illustrated on the function $f(x) = 1 - |x|$ in the interval $\Omega = (-1, 1)$: Since f is smooth in $(-1, 0)$ and in $(0, 1)$, the only candidate for the weak derivative is the function

$$D^1_w f(x) = \begin{cases} 1, & x \in (-1, 0), \\ -1, & x \in (0, 1). \end{cases}$$

It remains to be verified that (A.83) holds, with $D^\alpha\varphi = \varphi'$. For an arbitrary $\varphi \in C_0^\infty(\Omega)$ let us calculate

$$\begin{aligned}
-\int_{-1}^{1} f\varphi' \,\mathrm{d}x &= -\int_{-1}^{0} f\varphi' \,\mathrm{d}x - \int_{0}^{1} f\varphi' \,\mathrm{d}x \\
&= -[f\varphi]_{-1}^{0} + \int_{-1}^{0} (+1)\varphi \,\mathrm{d}x - [f\varphi]_0^1 + \int_0^1 (-1)\varphi \,\mathrm{d}x \\
&= \int_{-1}^{1} D_w^1 f\varphi \,\mathrm{d}x - f(0-)\varphi(0-) + f(0+)\varphi(0+) \\
&= \int_{-1}^{1} D_w^1 f\, \varphi \,\mathrm{d}x.
\end{aligned}$$

Thus (A.83) holds and the above-defined function $D_w^1 f$ is the weak derivative of f.

■ EXAMPLE A.54 (Weak differentiability in one dimension II)

Discontinuous functions in 1D are not weakly differentiable: Consider the function

$$f(x) = \begin{cases} -1, & x \in (-1, 0], \\ 1, & x \in (0, 1). \end{cases}$$

By the same token as in Example A.53, the only candidate for the weak derivative $D_w^1 f$ is the zero function (with an arbitrary value at $x = 0$). If zero is the weak derivative of f, for all $\varphi \in C_0^\infty(-1, 1)$ we have

$$\begin{aligned}
0 = \int_{-1}^{1} D_w^1 f\varphi \,\mathrm{d}x &= -\int_{-1}^{1} f\varphi' \,\mathrm{d}x = -\int_{-1}^{0} f\varphi' \,\mathrm{d}x - \int_0^1 f\varphi' \,\mathrm{d}x \\
&= \int_{-1}^{0} \varphi' \,\mathrm{d}x - \int_0^1 \varphi' \,\mathrm{d}x \\
&= \varphi(0) - \varphi(-1) - (\varphi(1) - \varphi(0)) = 2\varphi(0), \qquad \text{(A.85)}
\end{aligned}$$

which is a contradiction.

A.4.3 Spaces $W^{k,p}$ and H^k

The Sobolev spaces are defined as follows:

Definition A.57 (Sobolev spaces) *Let $\Omega \subset \mathbb{R}^d$ be an open set, $k \geq 1$ an integer number and $p \in [1, \infty]$. We define*

$$W^{k,p}(\Omega) = \{f \in L^p(\Omega);\ D_w^\alpha f \textit{ exists and lies in } L^p(\Omega) \textit{ for all multi-indices } \alpha,\ |\alpha| \leq k\}.$$

For every $1 \leq p < \infty$ the norm $\|\cdot\|_{k,p}$ is defined as

$$\|f\|_{k,p} = \left(\int_\Omega \sum_{|\alpha| \leq k} |D_w^\alpha f|^p \, dx\right)^{\frac{1}{p}} = \left(\sum_{|\alpha| \leq k} \|D_w^\alpha f\|_p^p\right)^{\frac{1}{p}}. \qquad \text{(A.86)}$$

For $p = \infty$ we define

$$\|f\|_{k,\infty} = \max_{|\alpha|\le k} \|D_w^\alpha f\|_\infty. \tag{A.87}$$

In the important special case $p = 2$ we abbreviate $W^{k,p}(\Omega) = H^k(\Omega)$.

In the $W^{k,p}$-spaces we use the following standard seminorms:

$$|f|_{k,p} = \left(\int_\Omega \sum_{|\alpha|=k} |D_w^\alpha f|^p \, \mathrm{d}\boldsymbol{x}\right)^{\frac{1}{p}} = \left(\sum_{|\alpha|=k} \|D_w^\alpha f\|_p^p\right)^{\frac{1}{p}}$$

for $1 \le p < \infty$, and

$$|f|_{k.\infty} = \max_{|\alpha|=k} \|D_w^\alpha f\|_\infty.$$

Classification of Sobolev spaces

We already know that all L^p-spaces are Banach spaces and, moreover, that the space L^2 is a Hilbert space. Let us see about the Sobolev spaces:

Theorem A.22 ($W^{k,p}(\Omega)$ is a Banach space) *Let $\Omega \subset \mathbb{R}^d$ be an open set, $k \ge 1$ an integer number and $p \in [1, \infty]$. The Sobolev space $W^{k,p}(\Omega)$ is a Banach space.*

Proof: We need to show that every Cauchy sequence $\{f_n\}_{n=1}^\infty \subset W^{k,p}(\Omega)$ has a limit $f \in W^{k,p}(\Omega)$. It follows from the Cauchy property of $\{f_n\}_{n=1}^\infty$ in the $W^{k,p}$-norm that for every multi-index $|\alpha| \le k$ the sequence $\{D_w^\alpha f_n\}_{n=1}^\infty \subset L^p(\Omega)$ is a Cauchy sequence in the space L^p. Therefore for every $|\alpha| \le k$ there exists a limit $f_\alpha \in L^p(\Omega)$ such that

$$\lim_{n\to\infty} \|D_w^\alpha f_n - f_\alpha\|_p = 0. \tag{A.88}$$

For $\alpha = (0, 0, \ldots, 0)$ denote $f := f_\alpha$. It remains to be shown that $f_\alpha = D_w^\alpha f$ for every $|\alpha| \le k$. Since $\{f_n\}_{n=1}^\infty \subset W^{k,p}$, we have

$$\int_\Omega D_w^\alpha f_n \varphi \, \mathrm{d}\boldsymbol{x} = (-1)^{|\alpha|} \int_\Omega f_n D^\alpha \varphi \, \mathrm{d}\boldsymbol{x} \qquad \text{for all } \varphi \in C_0^\infty(\Omega)$$

for all n. Passing to the limit for $n \to \infty$, which is justified by (A.88), we obtain

$$\int_\Omega f_\alpha \varphi \, \mathrm{d}\boldsymbol{x} = (-1)^{|\alpha|} \int_\Omega f D^\alpha \varphi \, \mathrm{d}\boldsymbol{x} \qquad \text{for all } \varphi \in C_0^\infty(\Omega).$$

Therefore $f_\alpha = D_w^\alpha f$ for all $|\alpha| \le k$, and thus

$$\lim_{n\to\infty} \|f_n - f\|_{k,p} = 0.$$

■

Lemma A.46 *Let $\Omega \subset \mathbb{R}^d$ be an open set, $k \ge 1$ an integer number. The Sobolev space $W^{k,p}(\Omega)$ is reflexive if and only if $1 < p < \infty$.*

Proof: Follows immediately from the reflexitivy of L^p-spaces. ■

Theorem A.23 ($H^k(\Omega) = W^{k,2}(\Omega)$ **is a Hilbert space)** *Let $\Omega \subset \mathbb{R}^d$ be an open set, $k \geq 1$ an integer number. The Sobolev space $W^{k,2}(\Omega)$, endowed with the inner product*

$$(f, g)_{k,2} = \int_\Omega \sum_{|\alpha| \leq k} D^\alpha f D^\alpha g \, dx = \sum_{|\alpha| \leq k} (D^\alpha f, D^\alpha g)_{L^2(\Omega)}, \qquad \text{(A.89)}$$

is a Hilbert space.

Proof: It is sufficient to show that (A.89) indeed defines an inner product. This follows easily from the fact that (A.89) is a finite sum of L^2-products of the weak derivatives. ■

Density of smooth functions in Sobolev spaces We first explained the density argument in the context of the Lebesgue L^p-spaces in Paragraph A.2.11. The following theorem gives an analogy for the Sobolev spaces.

Theorem A.24 (Density of smooth functions in $W^{k,p}$) *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz-continuous boundary and $v \in W^{k,p}(\Omega)$, $1 \leq p < \infty$. Then there exists a sequence $\{v_n\}_{n=1}^\infty \subset C^\infty(\overline{\Omega})$ such that*

$$\lim_{n \to \infty} \|v - v_n\|_{k,p} = 0.$$

Proof: See, e.g., [1]. ■

Here, $C^\infty(\overline{\Omega})$ is the space of infinitely-smooth functions with all derivatives continuous up to the boundary $\partial\Omega$. Theorem A.24 also holds for unbounded domains.

A.4.4 Discontinuity of H^1-functions in $\mathbb{R}^d$, $d \geq 2$

The density of smooth functions in $W^{k,p}$-spaces, stated in Theorem A.24, does not imply the smoothness, and not even the continuity of $W^{k,p}$-functions. However, there are special cases such as the space H^1 in 1D or H^2 in 2D, whose functions are continuous (this will be explained in more detail in the comments to Theorem A.27). The functions in the frequently used H^1-spaces in 2D and 3D are not continuous in general, but their discontinuity only can have the form of singularities. This is illustrated in the following example:

■ **EXAMPLE A.55 (Discontinuity of H^1-functions in 2D and 3D)**

Let $\Omega \in \mathbb{R}^2$ be an open set. The H^1-functions cannot be discontinuous along lines or curves in Ω, which can be shown using the fact that H^1-functions are continuous in 1D. However, discontinuities can occur in the form of singularities at isolated points in the domain. For example, consider the function

$$f(\boldsymbol{x}) = \log\left(\log\left(\frac{1}{r(\boldsymbol{x})}\right)\right), \qquad r(\boldsymbol{x}) = \sqrt{x_1^2 + x_2^2}, \qquad \text{(A.90)}$$

in a domain $\Omega = B(0, R) \subset \mathbb{R}^2$, $0 < R < 1/e$. It is easy to calculate that

$$\|\nabla f\|_{L^2}^2 = \int_\Omega |\nabla f|^2 \, dx = -\frac{2\pi}{\log R},$$

as well as to verify that $\|f\|_{L^2} < \infty$. Thus $f \in H^1(\Omega)$ despite $f \notin C(\Omega)$.

In 3D, H^1-functions can have singularities both at isolated points and along one-dimensional curves. To illustrate the point-singularities, we can consider the function (A.90) in $B(0,R) \subset \mathbb{R}^3$, $0 < R < 1/e$, with $r(\boldsymbol{x}) = \sqrt{x_1^2 + x_2^2 + x_3^2}$. A line-singularity is obtained by adjusting the function (A.90) to

$$\tilde{f}(x_1, x_2, x_3) = f(x_1, x_2). \tag{A.91}$$

The function (A.91) lies in $H^1(B(0,R))$, and it has a singularity along the x_3-axis.

A.4.5 Poincaré–Friedrichs' inequality

A frequently used subspace of $H^k(\Omega)$ is

$$H_0^k(\Omega) = \{v \in H^k(\Omega);\ D^\alpha v = 0 \text{ on } \partial\Omega \text{ for all } |\alpha| < k\}. \tag{A.92}$$

In the case of $k = 1$ this is the space

$$H_0^1(\Omega) = \{v \in H^1(\Omega);\ v = 0 \text{ on } \partial\Omega\},$$

where, for example, the weak formulation of second-order PDEs with Dirichlet boundary conditions usually takes place. The Poincaré–Friedrichs' inequality says that the H^k-seminorm

$$|u|_{k,2} = \left(\int_\Omega \sum_{|\alpha|=k} |D^\alpha u|^2 \, \mathrm{d}\boldsymbol{x} \right)^{\frac{1}{2}}$$

is a norm in the space $H_0^k(\Omega)$ on every bounded domain $\Omega \subset \mathbb{R}^d$. This norm, moreover, is equivalent to the full H^k-norm

$$\|u\|_{k,2} = \left(\int_\Omega \sum_{|\alpha|\le k} |D^\alpha u|^2 \, \mathrm{d}\boldsymbol{x} \right)^{\frac{1}{2}}.$$

The notion of equivalence of norms was first introduced in Definition A.34. The equivalence of $|\cdot|_{k,2}$ and $\|\cdot\|_{k,2}$ in the space $H_0^k(\Omega)$ finds application in the solvability and uniqueness analysis of partial differential equations as well as in practical computations.

Theorem A.25 (Basic Poincaré–Friedrichs' inequality in $H_0^1(\Omega)$) *Assume a bounded domain $\Omega \subset \mathbb{R}^d$ that is contained in a d-dimensional cube with the edge-length $C > 0$. Then*

$$\|u\|_{L^2} \le C|u|_{1,2} \quad \text{for all } u \in H_0^1(\Omega). \tag{A.93}$$

Proof: Since $C_0^\infty(\Omega)$ is dense in $H_0^1(\Omega)$, it is sufficient to prove the inequality for all $u \in C_0^\infty(\Omega)$. Without loss of generality, let $\Omega \subset S = \{(x_1, x_2, \dots, x_d);\ 0 < x_i < C\}$ and define $u(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in S \setminus \Omega$. Then

$$u(x_1, x_2, \dots, x_d) = u(0, x_2, \dots, x_d) + \int_0^{x_1} \frac{\partial u}{\partial x_1}(t, x_2, \dots, x_d) \, \mathrm{d}t.$$

The boundary term vanishes, and the Cauchy–Schwarz inequality yields

$$\begin{aligned} |u(\boldsymbol{x})|^2 &\le \int_0^{x_1} 1^2\,\mathrm{d}t \int_0^{x_1} \left|\frac{\partial u}{\partial x_1}(t, x_2, \ldots, x_d)\right|^2 \,\mathrm{d}t \\ &\le C \int_0^C \left|\frac{\partial u}{\partial x_1}(t, x_2, \ldots, x_d)\right|^2 \,\mathrm{d}t. \end{aligned}$$

Since the right-hand side is independent of x_1, it follows that

$$\int_0^C |u(\boldsymbol{x})|^2 \,\mathrm{d}x_1 \le C^2 \int_0^C \left|\frac{\partial u}{\partial x_1}(t, x_1, \ldots, x_d)\right|^2 \,\mathrm{d}t = C^2 \int_0^C \left|\frac{\partial u}{\partial x_1}(\boldsymbol{x})\right|^2 \,\mathrm{d}x_1.$$

Now it suffices to integrate over the whole cube S to obtain

$$\|u\|_{L^2}^2 = \int_S |u(\boldsymbol{x})|^2 \,\mathrm{d}\boldsymbol{x} \le C^2 \int_S \left|\frac{\partial u}{\partial x_1}(\boldsymbol{x})\right|^2 \,\mathrm{d}\boldsymbol{x} \le C^2 |u|_{1,2}^2. \tag{A.94}$$

■

Theorem A.26 (General Poincaré–Friedrichs' inequality in $H_0^k(\Omega)$) *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain. Then the seminorm $|\cdot|_{k,2}$ is a norm in the space $H_0^k(\Omega)$, equivalent to the norm $\|\cdot\|_{k,2}$. If Ω is contained in a d-dimensional cube with the side length C, then*

$$|u|_{k,2} \le \|u\|_{k,2} \le (1 + C)^k |u|_{k,2}$$

for all $u \in H_0^k(\Omega)$.

Proof: We use the Poincaré–Friedrichs' inequality (A.94) for the derivatives to obtain

$$\|D^\alpha u\|_{L^2} \le C \left\|\frac{\partial D^\alpha u}{\partial x_1}\right\|_{L^2}$$

for all $|\alpha| \le k - 1$ and $u \in H_0^k(\Omega)$. The rest is shown by induction. ■

Remark A.8 *The proof of the Poincaré–Friedrichs' inequality actually requires weaker assumptions – the space $H_0^k(\Omega)$ can be replaced with the space*

$$V = \{v \in H^k(\Omega);\ D^\alpha v = 0 \ on\ \Gamma \ for\ all\ \ |\alpha| < k\},$$

where Γ is a nonempty open subset of $\partial\Omega$, and the equivalence of norms $|\cdot|_{k,2}$ and $\|\cdot\|_{k,2}$ remains valid.

A.4.6 Embeddings of Sobolev spaces

Sometimes we need to decide whether all functions f that belong to a Banach space U also lie in another Banach space V. Thus we are asking if the following implication holds:

$$\|f\|_U < \infty \Rightarrow \|f\|_V < \infty.$$

This is equivalent to the question whether the identity operator $\mathcal{I} : U \longrightarrow V$ is continuous. The reader already knows the answer in some situations. For example, when $\Omega \subset \mathbb{R}^d$ is a bounded domain, $U = L^p(\Omega)$ and $V = L^q(\Omega)$. In this case the answer is positive if $q \le p$ (see Lemma A.29, Paragraph A.2.10). We also know the answer when U is a Lebesgue L^p-space and V some space of continuous or smooth functions (in this case it is negative). Some more results of this type for Sobolev spaces will be presented in this paragraph. Most of them will be given without proofs, since their difficulty goes beyond the scope of this text. For the following definition recall Definition A.50 of compactness for operators in normed spaces:

Definition A.58 (Embedding of Banach spaces) *Let U, V be Banach spaces such that $U \subset V$. We say that U is* continuously embedded *into V, and write $U \hookrightarrow V$, if there exists a constant $C_{U,V}$ such that for every $u \in U$ it holds*

$$\|u\|_V \le C_{U,V}\|u\|_U. \tag{A.95}$$

We say that the embedding is compact, *and write $U \hookrightarrow\hookrightarrow V$, if the identity operator $\mathcal{I}$ moreover is compact.*

If (A.95) holds, then obviously

$$\|\mathcal{I}\| \le C_{U,V}$$

(see Definition A.33). Since $\mathcal{I}$ is linear, it is continuous if and only if it is bounded (see Lemma A.24). The following definition generalizes the Lipschitz continuity of functions and introduces the Hölder spaces:

Definition A.59 (Hölder continuity, Hölder space $C^{k,\beta}(\Omega)$) *We say that a function $f \in C(\overline{\Omega})$ is* Hölder-continuous *with the exponent $\beta > 0$ if there exists a constant C_f such that*

$$|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)| \le C_f\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^\beta \quad \textit{for all } \boldsymbol{x}_1, \boldsymbol{x}_2 \in \Omega.$$

The space $C^{k,\beta}(\Omega)$ consists of functions whose αth partial derivatives are Hölder-continuous with the exponent $\beta > 0$ for all multi-indices α such that $|\alpha| \le k$.

Theorem A.27 (Embedding theorem) *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz-continuous boundary and $1 \le p < \infty$. We have the following embedding results:*

1. *If $kp < d$, then*

$$W^{k,p}(\Omega) \hookrightarrow\hookrightarrow L^q(\Omega)$$

for all $q < p^*$ such that $1/p^* = 1/p - d/k$.

2. *If $kp = d$, then*

$$W^{k,p}(\Omega) \hookrightarrow\hookrightarrow L^q(\Omega)$$

for all $q < \infty$.

3. *If $kp > d$, then*

$$W^{k,p}(\Omega) \hookrightarrow C^{k-[d/p]-1,\beta}(\Omega)$$

where $\beta = [d/p] + 1 - d/p$ if d/p is not an integer, or $\beta \in (0, 1)$ arbitrary if d/p is an integer.

4. *If $kp > d$, then*

$$W^{k,p}(\Omega) \hookrightarrow\hookrightarrow C^{k-[d/p]-1,\beta}(\Omega)$$

where $\beta \in [0, [d/p] + 1 - d/p)$. Here for $a \in \mathbb{R}$ the symbol $[a]$ stands for the integer part of a.

Proof: The proof can be found, e.g., in [1]. ■

The $W^{k,p}$-functions become smoother as the product kp increases. The critical value is the spatial dimension d. The $W^{k,p}$-functions are continuous (or, more precisely, equivalent to continuous functions) when $kp > d$,

$$W^{k,p} \hookrightarrow C(\overline{\Omega}) \quad \text{if} \ \ k > \frac{p}{d}.$$

By applying this result to the partial derivatives, it is easy to see that

$$W^{k,p} \hookrightarrow C^m(\overline{\Omega}) \quad \text{if} \ \ k - m > \frac{p}{d}.$$

If $kp < d$, then a $W^{k,p}$-function belongs to $L^{p^*}(\Omega)$ for an exponent p^* greater than p. To determine the exponent p^*, one starts from the inequality $kp < d$ written as

$$1/p - d/k > 0.$$

Then $1/p^*$ is defined as $1/p^* = 1/p - d/k$.

Another consequence of Theorem A.27 is the following compact embedding result.

Corollary A.3 (Compact embedding) *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz-continuous boundary. Let k, k' be nonnegative integers such that $k > k'$, and $1 \le p \le \infty$. Then*

$$W^{k,p}(\Omega) \hookrightarrow\hookrightarrow W^{k',p}(\Omega)$$

A.4.7 Traces of $W^{k,p}$-functions

Let $\Omega \subset \mathbb{R}^d$ be a bounded domain. Since the Sobolev space $W^{k,p}(\Omega)$ always is a subset of the corresponding space $L^p(\Omega)$, the $W^{k,p}$-functions are only defined almost everywhere in Ω. Since the boundary $\partial\Omega$ is a zero-measure subset of $\overline{\Omega}$, it might seem that the boundary values (traces) of $W^{k,p}$-functions never can be well defined. However, the notion of trace is associated with the whole class of $W^{k,p}$-equivalent functions, and it is defined using a representant that is continuous up to the boundary:

Definition A.60 (Trace of a $W^{k,p}$-function) *For a function $f \in W^{k,p}(\Omega)$ that is continuous up to the boundary $\partial\Omega$ we define its* trace to the boundary $\partial\Omega$ *as a function $\tilde{f}$ defined on $\partial\Omega$, such that*

$$\tilde{f}(\boldsymbol{x}) = f(\boldsymbol{x}) \quad \textit{for all } \boldsymbol{x} \in \partial\Omega.$$

Theorem A.28 (Traces of $W^{1,p}$-functions) *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz-continuous boundary and $1 \leq p < \infty$. Then there exists a continuous linear operator $\mathcal{T} : W^{1,p}(\Omega) \to L^p(\partial\Omega)$ such that*

1. *$(\mathcal{T}f)(\boldsymbol{x}) = f(\boldsymbol{x})$ for all $\boldsymbol{x} \in \partial\Omega$ if $f \in W^{1,p} \cap C(\overline{\Omega})$.*

2. *There exists a constant $C > 0$ such that*

$$\|\mathcal{T}f\|_{L^p(\partial\Omega)} \leq C\|f\|_{W^{1,p}(\Omega)}$$

for all $f \in W^{1,p}(\Omega)$.

3. *The operator $\mathcal{T} : W^{1,p}(\Omega) \to L^p(\partial\Omega)$ is compact.*

Proof: See, e.g., [1] for the proof of this theorem as well as for more details on traces in general. ■

A.4.8 Generalized integration by parts formulae

In this paragraph let us recall a few standard integral identities that are used frequently in the weak formulation of partial differential equations. Assume a bounded domain $\Omega \subset \mathbb{R}^d$ with Lipschitz-continuous boundary. By

$$\boldsymbol{\nu}(\boldsymbol{x}) = (\nu_1, \nu_2, \ldots, \nu_d)^T(\boldsymbol{x})$$

denote the unit outer normal to $\partial\Omega$ (defined almost everywhere on $\partial\Omega$). The formulae (A.80) and (A.81) are generalized as follows:

Theorem A.29 (Green's theorem for H^1-functions) *For every $u, v \in H^1(\Omega)$ it holds*

$$\int_\Omega \frac{\partial u}{\partial x_i} v \, d\boldsymbol{x} = -\int_\Omega u \frac{\partial v}{\partial x_i} \, d\boldsymbol{x} + \int_{\partial\Omega} uv\nu_i \, d\boldsymbol{S}. \tag{A.96}$$

Proof: See, e.g., [36]. The proof is based on the density of $C^1(\overline{\Omega})$ in $H^1(\Omega)$, and it uses the Trace Theorem (Theorem A.28). ■

Theorem A.29 is the basis for various other useful identities. At least two of them are introduced in the following lemma:

Lemma A.47 *For all $u \in H^1(\Omega)$ and $v \in H^2(\Omega)$ it is*

$$\int_\Omega u\Delta v \, d\boldsymbol{x} = -\int_\Omega \nabla u \nabla v \, d\boldsymbol{x} + \int_{\partial\Omega} u \frac{\partial v}{\partial \boldsymbol{\nu}} \, d\boldsymbol{S}.$$

where $\partial v/\partial\boldsymbol{\nu} = \nabla v(\boldsymbol{x}) \cdot \boldsymbol{\nu}(\boldsymbol{x})$, $\boldsymbol{x} \in \partial\Omega$. For all $u \in [H^1(\Omega)]^d$ and $v \in H^1(\Omega)$ it holds

$$\int_\Omega (\mathrm{div}\boldsymbol{u})v \, d\boldsymbol{x} = -\int_\Omega \boldsymbol{u} \cdot \nabla v \, d\boldsymbol{x} + \int_{\partial\Omega} (\boldsymbol{u} \cdot \boldsymbol{\nu})v \, d\boldsymbol{S}.$$

Proof: Immediately from Theorem A.29. ■

A.4.9 Exercises

Exercise A.61 *Show that the sequence of domains $\{\Omega_n\}_{n=0}^{\infty}$ in Example A.51 is uniformly bounded, i.e., that there exists a bounded domain $\hat{\Omega}$ such that $\Omega_n \subset \hat{\Omega}$ for all n.*

Exercise A.62 *Prove Lemma A.46.*

Exercise A.63 *Show in detail that (A.89) defines an inner product in $H^k(\Omega)$.*

Exercise A.64 *Consider a function $f \in C([-1,1])$ defined as $f(x) = x^3$ for $-1 \le x \le 0$ and $f(x) = x^2$ for $0 \le x \le 1$. Find the largest k for which $f \in H^k(-1,1)$.*

Exercise A.65 *Consider a domain $\Omega = B(0,R) \subset \mathbb{R}^2$, $0 < R < 1/e$. Show in detail that the function*

$$f(\boldsymbol{x}) = \log\left(\log\left(\frac{1}{r(\boldsymbol{x})}\right)\right), \qquad r(\boldsymbol{x}) = \sqrt{x_1^2 + x_2^2},$$

lies in the space $H^1(\Omega)$.

Exercise A.66 *Prove Theorem A.20 using Theorem A.19.*

Exercise A.67 *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz-continuous boundary. Use the Green's Theorem A.29 to see that every divergence-free vector field $\boldsymbol{w} \in [H^1(\Omega)]^d$ satisfies*

$$\int_{\partial\Omega} \boldsymbol{w} \cdot \boldsymbol{\nu} \, dS = 0,$$

where $\boldsymbol{\nu}(\boldsymbol{x})$ is the unit outer normal vector to Ω at the point $\boldsymbol{x} \in \partial\Omega$.

Exercise A.68 *Prove Lemma A.47 using Theorem A.19.*

Exercise A.69 *Consider a domain $\Omega = (-1,1) \times (-1,1) \subset \mathbb{R}^2$ and a function*

$$f(t, x_1, x_2) = \left(2 - e^{\frac{1}{t}} + \frac{1}{t^3}\sin(t^2)\right)(1 - x_1^2)(1 - x_2^2).$$

A sequence of functions $\{f_n\}_{n=1}^{\infty}$ is defined by putting $f_n(\boldsymbol{x}) = f(n, \boldsymbol{x})$.

1. *Verify that this sequence lies in $H^1(\Omega)$.*
2. *Show that it converges in H^1-norm and find the limit $g \in H^1(\Omega)$ of the sequence.*
3. *Use the fact that $\{f_n\}_{n=1}^{\infty} \subset H_0^1(\Omega)$ and a consequence of the Poincaré–Friedrichs' inequality to simplify the convergence analysis.*

APPENDIX B

SOFTWARE AND EXAMPLES

This chapter is devoted to the discussion of selected topics related to finite element software. In Section B.1 we present an efficient way of connecting the packages PETSc, Trilinos and UMFPACK to a finite element solver. Section B.2 gives a brief description of the high-performance modular finite element system HERMES. The full manual posted on our web page offers additional technical details. At the end of Section B.2 we present numerical results obtained with HERMES, where the efficiency of the lowest-order FEM and the *hp*-FEM is compared. Since it was not possible to include color pictures with this book, a color PDF file with the visualizations is available on our web page.

B.1 SPARSE MATRIX SOLVERS

Efficient solvers for sparse systems of linear algebraic equations are key ingredients of finite element programs. Nowadays an engineer or researcher hardly can afford developing matrix solvers on his/her own, and thus public domain software packages play an increasingly important role. Moreover, as the finite element software becomes more complex, the question of efficient simultaneous interfacing to multiple matrix solver packages matters. Every matrix solver comes with its own unique interface. Hardcoding this interface into a FEM solver means an unwanted coupling. If the FEM solver deals with multiple PDEs that produce matrices with substantially different properties (this is the case, e.g., with second-order elliptic PDEs and Maxwell's equations), the application of multiple matrix solvers becomes a need.

For this reason we decided to put an additional interface between the finite element solver and the matrix solvers, that we call the `sMatrix` utility. The basic version of the `sMatrix` utility is described in Paragraph B.1.1. The class `sMatrix` allows the user to operate with sparse matrices in the same way as with full matrices. Internally, the sparse matrices are represented via arrays of pointer chains, so that not even an estimate of the number of nonzero entries per row has to be provided. The library comprises 10 iterative matrix solvers based on ILU-preconditioned CG, BiCG, and other standard methods. These solvers work fine for moderately ill-conditioned problems. In Paragraph B.1.2 we provide an example where the `sMatrix` utility is incorporated into a simple finite element code. Both the data structures and algorithms in the `sMatrix` utility can be replaced easily while maintaining the original interface to the finite element solver. This is demonstrated in Paragraphs B.1.3–B.1.5, where we provide a brief description of the packages PETSc, Trilinos, and UMFPACK, and show how to connect them with a finite element solver through the `sMatrix` interface.

B.1.1 The sMatrix utility

The software package `sMatrix` comprises the files

```
src/s.cpp,
inc/sMatrix.h,
inc/sMatrix_f.cpp,
inc/sMatrix_PETSc.cpp,
inc/sMatrix_Trilinos.cpp,
inc/sMatrix_UMFPACK.cpp,
inc/Solvers.f,
Makefile,
obj/.
```

The file `s.cpp` contains a simple piecewise-affine one-dimensional finite element solver for the model problem from Paragraph 2.2.1, which uses the `sMatrix` utility. The file `inc/sMatrix_f.cpp` is the default version of the `sMatrix` utility which contains several standard ILU-preconditioned matrix solvers for both symmetric and nonsymmetric problems. These solvers are collected in the Fortran file `inc/Solvers.f`. The files `sMatrix_PETSc.cpp`, `sMatrix_Trilinos.cpp`, and `sMatrix_UMFPACK.cpp` employ, under the same interface, iterative and direct matrix solvers provided by the packages PETSc, Trilinos, and UMFPACK. These packages must be downloaded and installed separately. The directory `obj/` is used to store object files.

Including the* `sMatrix` *utility The `sMatrix` utility is included into a C/C++ code via the header file `sMatrix.h` after the standard system includes.

Initialization of a sparse matrix The size of the matrix `Ndof` must be known at the time of initialization. An empty sparse matrix `S` is initialized by the command `sMatrix *S = new sMatrix(Ndof, Nnz)`. The input parameter `Nnz` is ignored unless the packages PETSc, Trilinos, or UMFPACK are employed (this will be discussed later). The parameter `Nnz` either is a single integer number defining the maximum number of nonzeros per row, or it is an integer array of the length `Ndof` whose entries define the maximum number of nonzero entries in each row.

Adding nonzero entries The operation $S_{ij} = S_{ij} + value$, where $value$ is an arbitrary real number, is performed via the command `S->Add(i, j, value)`. The indices start from 0, which is a standard C/C++ convention (as opposed to Fortran, where indices start from 1). This means that the indices `i = 0` and `j = 0` correspond to the upper left corner of the matrix. Nonzero entries can be added to any position and as many times as the user wishes. A new entry in the sparse matrix structure is created when a first contribution to a position `(i,j)` is made, and further contributions to an existing entry only change its value.

Transforming the matrix into the CSR format After the process of filling the sparse matrix is finished, the number of nonzero entries in the matrix can be calculated by calling

```
S->ComputeNonZeros();
```

Next the matrix can be transformed into the Compressed Sparse Row (CSR) format (see Paragraph 2.5.1) by calling the functions

```
S->Alloc_CSR_arrays();
S->Fill_CSR_array_IA();
S->Fill_CSR_array_JA();
S->Fill_CSR_array_A();
```

If for some reason one needs to reset all entries of the matrix to zero while preserving its sparse structure (i.e., the arrays `IA` and `JA`), the command

```
S->SetZero();
```

can be used. After the `IA`, `JA`, and `A` arrays have been created, the matrix is stored twice in the computer memory. Hence the original sparse structure may be deleted via the command

```
S->DeleteSparseStructure();
```

Solving the system of linear algebraic equations With a right-hand side vector F of the length S->Rank, the system SY = F is solved by calling the function

```
SolveSparseSystem(S->Rank, S->NonZeroNum, S->IA, S->JA, S->A,
                  F, Solver, Max_iter_num, Iter_error,
                  Num_of_iter);
```

The solution vector `Y` is returned in the vector `F`. This function is defined outside of the `sMatrix` class so that it can be used independently, with any sparse matrix represented in terms of the three CSR arrays. The input parameters `int Solver`, `int Max_iter_num`, and `double Iter_error`, and the output parameter `int &Num_of_iter` have the following meaning:

`Solver` is a nonzero integer number between -10 and 10 that determines which numerical method from the file `Solver.f` is be used. This parameter is ignored when PETSc, Trilinos, or UMFPACK are employed:

1 . . . pbcg(): BiCG (biconjugate gradient method), nonsymmetric
2 . . . pbcgmr(): BiCGMR2 (BiCGStab2 with full two-dimensional minimization of the symmetric method), nonsymmetric
3 . . . pcgs(): (squared BiCG), nonsymmetric
4 . . . pscgs(): SCGS (smoothed squared BiCG), nonsymmetric
5 . . . pdcgs(): DCGS (twice smoothed squared BiCG), nonsymmetric
6 . . . pqmr(): QMR (quasi minimum residual method), nonsymmetric
7 . . . pstab(): BiCGStab, nonsymmetric

8 . . . pstab2(): BiCGStab2, nonsymmetric
9 . . . pgcgm(): GCGM (generalized conjugate gradient method), symmetric
10 . . . pmr(): (minimum residual method), symmetric

See, e.g., [103] for the description of these methods. The 'p' in the name of the method stands for "ILU-preconditioned". When flipping the sign, the same method is used without preconditioning. Max_iter_num is the maximum allowed number of iterations (to avoid an infinite loop in the case of convergence problems), Iter_error is the accuracy of the solver (more precisely, the upper bound for the residual), and Num_of_iter returns the number of iterations actually performed. After the sparse linear algebraic system is solved, the function

```
S->Delete_CSR_arrays();
```

may be used to delete the CSR arrays from the computer memory. Additional functions can be found in the header file sMatrix.h, and the source code sMatrix.cpp contains the description of additional internal variables that we have not mentioned. The source code sMatrix_f.cpp is too lengthy to be printed here, but the reader finds it included with the sMatrix package.

B.1.2 An example application

Let us return to the model problem stated in Paragraph 2.2.1: Given the real coefficients $a_1 > 0$ and $a_0 \ge 0$, an interval $\Omega = (x_0, x_1)$, and a finite element mesh $\mathcal{T}_{h,p}$ over Ω consisting of $M \ge 1$ equally-spaced affine elements, find a piecewise-affine approximate solution to the equation

$$-(a_1 u')' + a_0 u = 1 \quad \text{in } \Omega, \tag{B.1}$$

equipped with homogeneous Dirichlet boundary conditions.

The corresponding C++ finite element code that employs the sMatrix utility is shown below. The input parameters $a_1 > 0, a_0 \ge 0, x_0 < x_1$ and $M \ge 1$ are hardcoded for simplicity, but the user is free to change them.

```
/*
 This short code illustrates the use
 of the utility sMatrix. Solved is
 a model problem -(a1 u')' + a0 u = 1
 with homogeneous Dirichlet conditions
 in a 1D interval (x0, x1) by piecewise
 -affine equally-spaced elements.
 */

//system includes
# include <stdio.h>
# include <string.h>
# include <math.h>
# include <stdlib.h>
# include <unistd.h>

//the sMatrix utility
# include "../inc/sMatrix.h"

int main(int argv, char **argc) {
```

```
/***   DEFINING PROBLEM PARAMETERS   ***/

double x0 = 0, x1 = 1; //computational domain Omega = (x0, x1)
double a1 = 1;         //coefficient, must be greater than zero
double a0 = 0;         //coefficient, must be greater than
                       //or equal to zero
int M = 10000;         //number of equally-spaced elements in (x0,x1)

/***   PRINTING PROBLEM PARAMETERS   ***/

printf("\n----------------------------------------\n");
printf(" This is a demo for the sMatrix utility\n");
printf("----------------------------------------\n");
printf("Domain: Omega = (%g, %g).\n", x0, x1);
printf("Equation: -(a1 u')' + a0 u = 1\n");
printf("Bdy conditions: u(x0) = u(x1) = 0.\n");
printf("Coeffs: a1 = %g, a0 = %g.\n", a1, a0);
printf("Subdivision into %d elements.\n", M);

/***   DEFINING ELEMENT LENGTH   ***/

double h = (x1 - x0)/M;

/***   DEFINING THE NUMBER OF UNKNOWNS   ***/

int ndof = M - 1;

/***   ALLOCATING THE SPARSE MATRIX   ***/

//the second parameter is ignored unless
//the PETSc or Trilinos packages are used
//(to be explained later)
sMatrix *S = new sMatrix(ndof, 3);

/***   ALLOCATING THE RHS   ***/

double *f = (double*)malloc(sizeof(double)*ndof);
if(f == NULL) Error("Not enough memory for the right-hand side.");

/***   FILLING THE SPARSE MATRIX AND THE RHS  ***/
/***   (THE ELEMENT LOOP)   ***/

//setting the RHS zero
for(int i=0; i<ndof; i++) f[i] = 0;

//first element
S->Add(0,0, a1/h + a0*h/3);
f[0] = h/2;

//loop over internal elements
for(int i=2; i<M; i++) {
  S->Add(i-2, i-2, a1/h + a0*h/3);
  S->Add(i-2, i-1, -a1/h + a0*h/6);
  S->Add(i-1, i-2, -a1/h + a0*h/6);
  S->Add(i-1, i-1, a1/h + a0*h/3);
  f[i-2] += h/2;
  f[i-1] += h/2;
}

//last element
```

```
S->Add(M-2,M-2, a1/h + a0*h/3);
f[M-2] += h/2;

/***   CONTROL OUTPUT OF THE MATRIX IN MATLAB FORMAT     ***/

/*
string name = "matrix";
S->OutputMatrixMatlab(name);
printf("Control output of the matrix in Matlab format.\n");
*/

/***   CONTROL OUTPUT OF THE MATRIX IN ASCI FORMAT     ***/

/*
//string name = "matrix";
S->OutputMatrixASCI(name);
printf("Control output of the matrix in ASCI format.\n");
*/

/***   CONTROL OUTPUT OF THE RHS IN ASCI FORMAT   ***/

/*
printf("RHS = ");
for(int i=0; i<ndof; i++) printf("%g ", f[i]);
printf("\n");
*/

/***   TRANSLATING THE MATRIX INTO THE CSR FORMAT   ***/

S->ComputeNonZeros();
S->Alloc_CSR_arrays();
S->Fill_CSR_array_IA();
S->Fill_CSR_array_JA();
S->Fill_CSR_array_A();

/***   RELEASING MEMORY FOR THE SPARSE MATRIX STRUCTURE   ***/

S->DeleteSparseStructure();

/***   DEFINING SPARSE MATRIX SOLVER PARAMETERS   ***/

//choosing the iterative sparse matrix solver
//(ignored when the PETSc or Trilinos packages are used)
int solver = 9; //9 for ILU-preconditioned Conjugate Gradients

//defining accuracy of the iterative sparse matrix  solver
double iter_error = 1e-10;

//setting a limit to the number of iterations of the solver
int max_iter_num = 1000;

//declaring a variable where the solver returns the number
//of iterations actually performed
int num_of_iter;

/***   SOLVING THE SPARSE LINEAR ALGEBRAIC SYSTEM   ***/

SolveSparseSystem(S->Rank, S->NonZeroNum, S->IA,
                  S->JA, S->A, f, solver,
                  max_iter_num, iter_error, num_of_iter);
```

```
  printf("Matrix solver performed %d iterations.\n", num_of_iter);

  /***   RELEASING MEMORY FOR THE CSR SPARSE MATRIX ARRAYS   ***/

  S->Delete_CSR_arrays();

  /***   OUTPUT OF SOLUTION   ***/

  FILE *g = fopen("solution.gnu", "wb");
  if(g == NULL) Error("Could not open the output file.");
  fprintf(g, "%g  0\n", x0);
  for(int i=0; i<ndof; i++) fprintf(g, "%g %g\n", x0 + (i+1)*h, f[i]);
  fprintf(g, "%g 0\n", x1);
  fclose(g);
  printf("Gnuplot file solution.gnu created.\n");

  delete S;
  free(f);
  printf("Bye.\n");
  return 1;
}
```

B.1.3 Interfacing with PETSc

The PETSc solver package (*Portable, Extensible Toolkit for Scientific Computation*) was developed at Argonne National Labs. It can be downloaded from the web page `http://www-unix.mcs.anl.gov/petsc/petsc-2/index.html` where also installation instructions, documentation and an extensive amount of additional information can be found. Rather than trying to give another description of the package here, let us present the source code sMatrix_PETSc.cpp. This is a PETSc version of the sMatrix utility, with an interface identical to the original sMatrix. Every finite element solver that works with the original sMatrix will work with sMatrix_PETSc as well.

The source code of sMatrix_PETSc.cpp

```
//implementation of PETSc solvers under the sMatrix interface

#include "sMatrix.h"

#ifdef __cplusplus
extern "C" {
#endif

#include <petsc.h>
#include <petscvec.h>
#include <petscmat.h>
#include <petscksp.h>

#ifdef __cplusplus
}
#endif

#include <sstream>
#include <fstream>
#include <iomanip>
#include <cstdio>
#include <cstdlib>
```

```
using namespace std;

void Error(char *msg) {
  fprintf(stderr, "%s\n", msg);
  fflush(stderr);
  exit(0);
}

//class sMatrix
sMatrix::sMatrix(int iRank, int nz = 0)
{
    PetscInitializeNoArguments();
    Rank = iRank;
    A = (double*)malloc(sizeof(Mat));
    MatCreateSeqAIJ(PETSC_COMM_SELF, Rank, Rank,
      nz > 0 ? nz : PETSC_DEFAULT, PETSC_NULL, (Mat*)A);
    IA = NULL;
    JA = NULL;
    NonZeroNum = 0;
}

sMatrix::sMatrix(int iRank, int *nz)
{
    PetscInitializeNoArguments();
    Rank = iRank;
    A = (double*)malloc(sizeof(Mat));
    MatCreateSeqAIJ(PETSC_COMM_SELF, Rank, Rank,
      PETSC_DEFAULT, nz, (Mat*)A);
    IA = NULL;
    JA = NULL;
    NonZeroNum = 0;
}

double sMatrix::GiveEntry(int iRow, int iColumn)
{
  if (iRow < 0 || iColumn < 0 ||
      iRow >= Rank || iColumn >= Rank)
    Error("internal in sMatrix::GiveEntry().");
  PetscScalar val;
  MatGetValues(*(Mat*)A, 1, &iRow, 1, &iColumn, &val);
  return (double)val;
}

void sMatrix::SetZero()
{
  MatZeroEntries(*(Mat*)A);
}

void sMatrix::ComputeNonZeros()
{
  //counting nonzeros
  MatAssemblyBegin(*(Mat*)A, MAT_FINAL_ASSEMBLY);
  MatAssemblyEnd(*(Mat*)A, MAT_FINAL_ASSEMBLY);
  MatInfo info;
  MatGetInfo(*(Mat*)A, MAT_LOCAL, &info);
  NonZeroNum = (int)info.nz_used;
}

void sMatrix::Alloc_CSR_arrays()
{
  MatAssemblyBegin(*(Mat*)A, MAT_FINAL_ASSEMBLY);
  MatAssemblyEnd(*(Mat*)A, MAT_FINAL_ASSEMBLY);
}

void sMatrix::Delete_CSR_arrays()
{
}

void sMatrix::Fill_CSR_array_IA()
{
  MatAssemblyBegin(*(Mat*)A, MAT_FINAL_ASSEMBLY);
  MatAssemblyEnd(*(Mat*)A, MAT_FINAL_ASSEMBLY);
}

void sMatrix::Fill_CSR_array_JA()
{
  MatAssemblyBegin(*(Mat*)A, MAT_FINAL_ASSEMBLY);
  MatAssemblyEnd(*(Mat*)A, MAT_FINAL_ASSEMBLY);
}

void sMatrix::Fill_CSR_array_A()
{
  MatAssemblyBegin(*(Mat*)A, MAT_FINAL_ASSEMBLY);
  MatAssemblyEnd(*(Mat*)A, MAT_FINAL_ASSEMBLY);
}

void sMatrix::Add(int Row, int Column, double Value)
{
  MatSetValue(*(Mat*)A, Row, Column, Value, ADD_VALUES);
}

//this function compares the sparse matrix with
//a given full matrix (for debug purposes)
void sMatrix::TestVsFullMatrix(double **A, double precision)
{
  Error("TestVsFullMatrix() not done in PETSc version.");
}

void sMatrix::OutputMatrixMatlab(string ProjectName)
{
  string MatlabFileName = ProjectName + ".mat";

  PetscViewer w;
  PetscViewerASCIIOpen(PETSC_COMM_SELF,
    MatlabFileName.c_str(), &w);
  PetscViewerSetFormat(w, PETSC_VIEWER_ASCII_MATLAB);
  MatView(*(Mat*)A, w);
  PetscViewerDestroy(w);
}

void sMatrix::OutputMatrixASCI(string ProjectName)
{
  string TxtFileName = ProjectName + ".txt";

  PetscViewer w;
  PetscViewerASCIIOpen(PETSC_COMM_SELF,
    TxtFileName.c_str(), &w);
  MatView(*(Mat*)A, w);
  PetscViewerDestroy(w);
}

void sMatrix::DeleteSparseStructure() {
}

void SolveSparseSystem(int Rank, int NonZeroNum, int *IA,
                       int * JA, double *A, double *X,
                       int Solver, int Max_iter_num,
                       double Iter_error, int &Num_of_iter)
{
    Vec rhs, x;
    VecCreateSeqWithArray(PETSC_COMM_SELF, Rank, X, &rhs);
    VecDuplicate(rhs, &x);

    KSP ksp;
    KSPCreate(PETSC_COMM_SELF, &ksp);
    KSPSetTolerances(ksp, Iter_error, PETSC_DEFAULT,
      PETSC_DEFAULT, Max_iter_num);
    KSPSetFromOptions(ksp);
    KSPSetOperators(ksp, *(Mat*)A, *(Mat*)A, SAME_PRECONDITIONER);
    //VERSION 2.2.0:
    //KSPSetRhs(ksp, rhs);
    //KSPSetSolution(ksp, x);
    //KSPSolve(ksp);
    //VERSION 2.2.1:
    KSPSolve(ksp,rhs,x);

    PetscReal r_norm;
    KSPGetResidualNorm(ksp, &r_norm);
    KSPGetIterationNumber(ksp, &Num_of_iter);
    printf("Matrix solver step %d, residual %g.\n",
      Num_of_iter, r_norm);

    PetscScalar *p;
    VecGetArray(x, &p);
    for(int i=0; i<Rank; i++) {
        X[i] = p[i];
    }
    VecRestoreArray(x, &p);

    KSPDestroy(ksp);
    VecDestroy(rhs);
    VecDestroy(x);
}
```

B.1.4 Interfacing with Trilinos

Trilinos is a large collection of linear and nonlinear algebraic solvers developed at Sandia National Labs. Documentation on the Trilinos project can be found on the web page http://software.sandia.gov/trilinos/. Here we present its rather simple serial application. The following source code sMatrix_Trilinos.cpp is the Trilinos version of the sMatrix utility, that again preserves the original sMatrix interface.

The source code of sMatrix_Trilinos.cpp

```
//implementation of Trilinos solvers under the sMatrix interface

#include "sMatrix.h"

#include <Epetra_ConfigDefs.h>
#include <Epetra_SerialComm.h>
#include <Epetra_CrsMatrix.h>
```

```
#include <Epetra_Map.h>
#include <Epetra_Vector.h>
#include <Epetra_LinearProblem.h>
#include <AztecOO.h>

#include <sstream>
#include <fstream>
#include <iomanip>
#include <cstdio>
#include <cstdlib>

using namespace std;

void Error(char *msg) {
  fprintf(stderr, "%s\n", msg);
  fflush(stderr);
  exit(0);
}

//class sMatrix
sMatrix::sMatrix(int iRank, int nz)
{
    Rank = iRank;
    Epetra_SerialComm *Comm = new Epetra_SerialComm;
    Epetra_Map *Map = new Epetra_Map(Rank, 0, *Comm);
    Epetra_CrsMatrix *Crs = new Epetra_CrsMatrix(Copy, *Map, nz);
    A = (double*)Crs;
    IA = NULL;
    JA = NULL;
    NonZeroNum = 0;
}

sMatrix::sMatrix(int iRank, int *nz)
{
    Rank = iRank;
    Epetra_SerialComm *Comm = new Epetra_SerialComm;
    Epetra_Map *Map = new Epetra_Map(Rank, 0, *Comm);
    Epetra_CrsMatrix *Crs = new Epetra_CrsMatrix(Copy, *Map, nz);
    A = (double*)Crs;
    IA = NULL;
    JA = NULL;
    NonZeroNum = 0;
}

double sMatrix::GiveEntry(int iRow, int iColumn)
{
  if (iRow < 0 || iColumn < 0 ||
      iRow >= Rank || iColumn >= Rank)
    Error("internal in sMatrix::GiveEntry().");
  Error("GiveEntry() not done in Trilinos version.");
  return 0.0;
}

void sMatrix::SetZero()
{
    Epetra_CrsMatrix *Crs = (Epetra_CrsMatrix*)A;
    Crs->PutScalar(0.0);
}

void sMatrix::ComputeNonZeros()
{
  //counting nonzeros
  Epetra_CrsMatrix *Crs = (Epetra_CrsMatrix*)A;
  Crs->FillComplete();
  NonZeroNum = Crs->NumGlobalNonzeros();
}

void sMatrix::Alloc_CSR_arrays()
{
  Epetra_CrsMatrix *Crs = (Epetra_CrsMatrix*)A;
  Crs->FillComplete();
}

void sMatrix::Delete_CSR_arrays()
{
}

void sMatrix::Fill_CSR_array_IA()
{
  Epetra_CrsMatrix *Crs = (Epetra_CrsMatrix*)A;
  Crs->FillComplete();
}

void sMatrix::Fill_CSR_array_JA()
{
  Epetra_CrsMatrix *Crs = (Epetra_CrsMatrix*)A;
  Crs->FillComplete();
}

void sMatrix::Fill_CSR_array_A()
{
  Epetra_CrsMatrix *Crs = (Epetra_CrsMatrix*)A;
  Crs->FillComplete();
}

void sMatrix::Add(int Row, int Column, double Value)
{
  Epetra_CrsMatrix *Crs = (Epetra_CrsMatrix*)A;
  int ret = Crs->SumIntoGlobalValues(Row, 1, &Value, &Column);
  if (ret!=0) {
    Crs->InsertGlobalValues(Row, 1, &Value, &Column);
  }
}

//this function compares the sparse matrix with
//a given full matrix (for debug purposes)
void sMatrix::TestVsFullMatrix(double **A, double precision)
{
  Error("TestVsFullMatrix() not done in Trilinos version.");
}

void sMatrix::OutputMatrixMatlab(string ProjectName)
{
  Error("OutputMatrixMatlab() not done in Trilinos version.");
}

void sMatrix::OutputMatrixASCI(string ProjectName)
{
  Epetra_CrsMatrix *Crs = (Epetra_CrsMatrix*)A;
  cout << *Crs << endl;
}
```

```
void sMatrix::DeleteSparseStructure() {
}

void SolveSparseSystem(int Rank, int NonZeroNum, int *IA,
                       int * JA, double *A, double *X,
                       int Solver, int Max_iter_num,
                       double Iter_error, int &Num_of_iter)
{
    Epetra_CrsMatrix *Crs = (Epetra_CrsMatrix*)A;
    Epetra_Vector rhs(Copy, Crs->RowMap(), X);
    Epetra_Vector x(Crs->RowMap());
    x.Random();
    Epetra_LinearProblem Problem(Crs, &x, &rhs);

    AztecOO KSP(Problem);
    KSP.Iterate(Max_iter_num, Iter_error);
    Num_of_iter = KSP.NumIters();

    x.ExtractCopy(X);
}
```

B.1.5 Interfacing with UMFPACK

UMFPACK is a set of routines for solving nonsymmetric sparse linear systems by means of the Unsymmetric MultiFrontal method. The software was developed at the University of Florida at Gainesville. Being a *direct* solver, UMFPACK is substantially different from the iterative solvers Trilinos and PETSc. We use it successfully for indefinite problems arising in the discretization of the time-harmonic Maxwell's equations, where the iterative solvers do not perform well. Documentation and source codes can be found on the web page `http://www.cise.ufl.edu/research/spar se/umfpack`. Building the UMFPACK functionality into the `sMatrix` utility was slightly more technical because of its specific data structures, but the original `sMatrix` interface could be preserved exactly. The source code `sMatrix_UMFPACK.cpp` is not printed here because of its length.

B.2 THE HIGH-PERFORMANCE MODULAR FINITE ELEMENT SYSTEM HERMES

Nodal elements (such as the Lagrange, Whitney, or Nédélec elements) are naturally suited for meshes where all elements have the same polynomial degree. This is why they are most suitable for problems with "nice" solutions. However, many problems in computational engineering and science exhibit significant local behavior in the form of steep gradients, singularities, boundary and/or internal layers, etc. These phenomena can be most efficiently resolved by means of hierarchic finite element methods (hp-FEM), which are capable of combining elements of variable size and polynomial degree. The impact on the efficiency of the method is tremendous. A few numerical examples at the end of this section give the reader some feeling for the difference. An introduction to hierarchic finite element methods can be found, e.g., in [111], on which the implementation of HERMES is based.

B.2.1 Modular structure of HERMES

HERMES is a modular object-oriented FEM system designed to facilitate the portability of the *hp*-FEM technology to various PDE models in engineering and science. The system consists of two main modules:

- FEM/*hp*-FEM Module containing the finite element discretization technology, such as the mesh processing algorithms, interior mode elimination algorithms, assembling algorithms, a-posteriori error estimation algorithms, etc.
- Algebraic Module with a variety of solvers for systems of linear and nonlinear algebraic equations, such as ILU preconditioned CG and BiCG methods, and solvers provided by the packages PETSc, Trilinos, and UMFPACK. Additional solvers can be added easily.

The FEM/*hp*-FEM and Algebraic Modules communicate through the universal `sMatrix` interface that was described in Paragraph B.1.1.

The FEM/*hp*-FEM module is the most complex part of the system. It comprises:

- FEM/*hp*-FEM kernel containing PDE-independent algorithms,
- smaller modules representing PDE-dependent data, such as various types of finite elements.

In this way the discretization technology is fully separated from the physics of the solved problems, which reduces the development cost and increases the portability of the system to various PDE applications. The modular structure is depicted in Figure B.1.

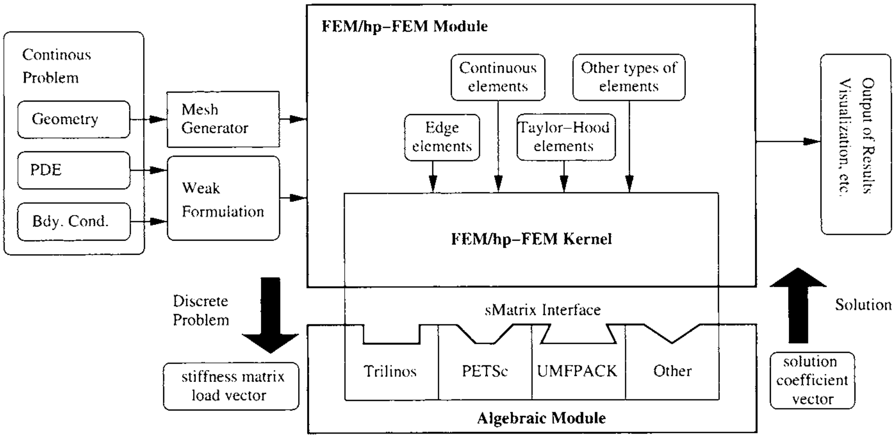


Figure B.1 Structure of the modular FEM system HERMES.

Currently two different PDE modules are implemented in the FEM/*hp*-FEM module:

- Elliptic Module with hierarchic continuous elements for systems of arbitrary number of nonlinear elliptic equations,

- Maxwell's Module containing hierarchic edge elements for time-harmonic Maxwell's equations.

The Stokes Module with hierarchic higher-order Taylor–Hood elements is under construction. Each type of PDE problem can be supplemented with appropriate boundary conditions.

B.2.2 The elliptic module

The system of nonlinear PDEs is entered via the definition of nonzero components of the vectors and matrices in the matrix equation

$$\begin{aligned}
&\frac{\partial}{\partial x_1}\left(P_1(\boldsymbol{x}, u, \nabla u)\frac{\partial u}{\partial x_1}(\boldsymbol{x})\right) + \frac{\partial}{\partial x_1}\left(P_2(\boldsymbol{x}, u, \nabla u)\frac{\partial u}{\partial x_2}(\boldsymbol{x})\right) \\
&+ \frac{\partial}{\partial x_2}\left(P_3(\boldsymbol{x}, u, \nabla u)\frac{\partial u}{\partial x_1}(\boldsymbol{x})\right) + \frac{\partial}{\partial x_2}\left(P_4(\boldsymbol{x}, u, \nabla u)\frac{\partial u}{\partial x_2}(\boldsymbol{x})\right) \\
&+ \frac{\partial}{\partial x_1}\left(P_5(\boldsymbol{x}, u, \nabla u)u(\boldsymbol{x})\right) + \frac{\partial}{\partial x_2}\left(P_6(\boldsymbol{x}, u, \nabla u)u(\boldsymbol{x})\right) \\
&+ P_7(\boldsymbol{x}, u, \nabla u)u(\boldsymbol{x}) = F(\boldsymbol{x}, u, \nabla u).
\end{aligned} \tag{B.2}$$

Here $u(\boldsymbol{x}) = (u_1(\boldsymbol{x}), u_2(\boldsymbol{x}), \dots, u_{N_{eq}}(\boldsymbol{x}))^T$ is the unknown solution with the components $u_i \in H^1(\Omega), i = 1, \dots N_{eq}$. The matrix parameters $P_1, P_2, \dots, P_7$ of the type $N_{eq} \times N_{eq}$ may be arbitrary, some of them can be zero, but each equation in the system must stay a scalar second-order elliptic PDE. All parameters may depend on the spatial variable $\boldsymbol{x}$, on the solution u and/or on the gradient ∇u. The same applies to the right-hand side function $F = (F_1, \dots, F_{N_{eq}})^T(\boldsymbol{x}, u, \nabla u)$. By applying $\partial/\partial x_1$ and $\partial/\partial x_2$ to vectors, we mean that the derivatives are applied to every component.

The variable N_{eq} is set to one when a single scalar equation is solved, in which case obviously the matrix and vector parameters become scalars. In this way the model equation (B.2) covers a large variety of nonlinear, possibly vector-valued second-order elliptic problems.

Boundary conditions Boundary conditions can be prescribed in a general form. For each solution component, the boundary of the domain is split into two parts $\Gamma_{D,i}$ and $\Gamma_{N,i}$, $i = 1, 2, \dots, N_{eq}$. These boundary parts do not have to be connected, and moreover either $\Gamma_{N,i}$ or $\Gamma_{D,i}$ can be empty, provided that the other parameters guarantee unique solvability.

On $\Gamma_{D,i}$ one prescribes the Dirichlet boundary condition

$$u(\boldsymbol{x}) = g_{D,i}(\boldsymbol{x}), \quad \boldsymbol{x} \in \Gamma_{D,i}, \; i = 1, 2, \dots, N_{eq},$$

where $g_{D,i}$ are given functions.

The Neumann boundary conditions, $g_{N,i}$, on $\Gamma_{N,i}$ are defined in the form

$$n_1\left(P_1\frac{\partial u}{\partial x_1} + P_2\frac{\partial u}{\partial x_2} + P_5 u\right)_i + n_2\left(P_3\frac{\partial u}{\partial x_1} + P_4\frac{\partial u}{\partial x_2} + P_6 u\right)_i = g_{N,i},$$

where $g_{N,i}(\boldsymbol{x})$ are the given functions, $\boldsymbol{\nu} = (\nu_1(\boldsymbol{x}), \nu_2(\boldsymbol{x}))^T$ is the unitary outer normal vector to $\partial\Omega$, $\boldsymbol{x} \in \Gamma_{N,i}$, and $i = 1, 2, \dots, N_{eq}$.

Vector formulation With the notation

$$\mathcal{A} = \begin{pmatrix} P_1 & P_2 \\ P_3 & P_4 \end{pmatrix},$$

the problem (B.2) can be rewritten to

$$\begin{aligned} \text{div } (\mathcal{A}\nabla u) + \frac{\partial}{\partial x}(P_5 u) + \frac{\partial}{\partial y}(P_6 u) + P_7 u &= F \quad \text{in } \Omega, \\ u_i &= g_{D,i} \quad \text{on } \Gamma_{D,i}, \\ ([\mathcal{A}\nabla u]\cdot \boldsymbol{\nu})_i + (P_5 u \nu_1 + P_6 u \nu_2)_i &= g_{N,i} \quad \text{on } \Gamma_{N,i}. \end{aligned}$$

Spatial discretization The polynomial degrees $1 \le p \le 10$ on the finite elements can be defined either via a data file or by means of a function in the code that assigns the polynomial degrees to elements based on the coordinates of their vertices. The polynomial degrees may differ from element to element. The solver constructs the corresponding hierarchic basis of the (vector-valued) finite element space $V_{h,p} \subset (H^1(\Omega))^{N_{eq}}$, $\mathcal{B} = \{\varphi_1, \varphi_2, \dots, \varphi_N\}$. The unknown solution is sought in the form

$$u(\boldsymbol{x}) = \sum_{j=1}^{N} y_j \varphi_j(\boldsymbol{x}),$$

where $\boldsymbol{y} = (y_1, y_2, \dots, y_N)^T$ is the vector of unknown coefficients of the length N. Equation (B.2) is formulated in the variational sense and the usual finite element discretization is performed. The result of the discretization is a system of nonlinear algebraic equations in the form

$$A(\boldsymbol{y})\boldsymbol{y} = f(\boldsymbol{y}). \tag{B.3}$$

Here A is a square matrix of the type $N \times N$ depending on the vector $\boldsymbol{y}$. It contains the nonlinearity coming from the coefficients $P_1, P_2, \dots, P_7$. The right-hand side $f(\boldsymbol{y})$ is a vector of the length N that also depends on $\boldsymbol{y}$. The discrete system is solved via a fixed point iteration

$$A(\boldsymbol{y}^k)\boldsymbol{y}^{k+1} = f(\boldsymbol{y}^k), \qquad k = 0, 1, 3, \dots, \tag{B.4}$$

starting from a suitable initial guess $\boldsymbol{y}^0$. Each iteration of this process includes the solution of a system of linear algebraic equations with a given matrix $A^k = A(\boldsymbol{y}^k)$. For this purpose we use the previously mentioned `sMatrix` utility.

B.2.3 The Maxwell's module

The time-harmonic Maxwell's equation (7.62) is considered in the two-dimensional form

$$\bar{\nabla}\left(\mu_r^{-1}\nabla \boldsymbol{E}\right) - \kappa^2 \epsilon_r \boldsymbol{E} = \boldsymbol{F} \qquad \text{in } \Omega, \tag{B.5}$$

where $\boldsymbol{E}$ is the complex phasor of the electric field (i.e., the underlined quantity in (7.61)). In the two-dimensional setting, the relative permeability $\mu_r = \mu_r(\boldsymbol{x})$ is a scalar in 2D, whereas the relative permittivity $\epsilon_r = \epsilon_r(\boldsymbol{x})$ is a 2×2 tensor. By ω and $\kappa = \omega\sqrt{\mu_0\epsilon_0}$ we denote the frequency and wave number, respectively. Here, as in Chapter 7, the symbol $\bar{\nabla} = (\partial/\partial x_2, -\partial/\partial x_1)^T$ stands for the vector-valued curl of a scalar quantity, and $\nabla \boldsymbol{E} =$

$\partial E_2/\partial x_1 - \partial E_1/\partial x_2$ is the surface curl. Equation (B.5) is considered in an open bounded domain $\Omega \subset \mathbb{R}^2$.

The code works with the normalized values (7.67),

$$\sqrt{\epsilon_0}\underline{\boldsymbol{E}} \to \boldsymbol{E}, \quad \sqrt{\mu_0}\underline{\boldsymbol{H}} \to \boldsymbol{H}, \quad \frac{\underline{\boldsymbol{J}_a}}{\sqrt{\epsilon_0}} \to \boldsymbol{J}_a,$$

and (7.68),

$$\begin{aligned} \epsilon_r &= \frac{1}{\epsilon_0}\left(\epsilon + \frac{j\gamma}{\omega}\right), \\ \mu_r &= \frac{\mu}{\mu_0}, \end{aligned} \tag{B.6}$$

where the conductivity γ is a function of the spatial variable.

Boundary conditions The boundary $\partial\Omega$ can be split into two open (not necessarily connected) disjoint parts Γ_P and Γ_I. We consider the perfect conductor boundary condition (7.57),

$$\boldsymbol{E} \cdot \boldsymbol{t} = 0 \qquad \text{on } \Gamma_P,$$

and the impedance boundary condition (7.71),

$$\mu_r^{-1}\nabla \boldsymbol{E} - j\kappa\lambda \boldsymbol{E} \cdot \boldsymbol{t} = \boldsymbol{g} \cdot \boldsymbol{t} \qquad \text{on } \Gamma_I.$$

Here $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x})$ is the positively-oriented unit tangential vector $\boldsymbol{t} = (-\nu_2, \nu_1)^T$, where $\boldsymbol{\nu} = (\nu_1, \nu_2)^T$ is the unit outer normal vector to the boundary $\partial\Omega$. The impedance $\lambda = \lambda(\boldsymbol{x}) > 0$ was defined in (7.72). Only the tangential component of $\boldsymbol{g} = \boldsymbol{g}(\boldsymbol{x})$ is relevant.

Spatial discretization The discretization of the time-harmonic Maxwell's equations is analogous to second-order elliptic problems. The solver constructs a hierarchic basis of the corresponding finite element subspace of $\boldsymbol{H}(\text{curl}, \Omega_h)$, $\mathcal{B} = \{\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_N\}$. Recall that the space $\boldsymbol{H}(\text{curl})$ only requires the continuity of the tangential component of the approximation across element interfaces. The unknown solution $\boldsymbol{E}_{h,p}$ is sought in the form

$$\boldsymbol{E}_{h,p}(\boldsymbol{x}) = \sum_{j=1}^{N} z_j \boldsymbol{\psi}_j(\boldsymbol{x}),$$

where $z = (z_1, z_2, \ldots, z_N)^T \in \mathbb{C}^N$ is the vector of unknown complex coefficients.

The usual finite element discretization yields a system of complex-valued linear algebraic equations of the form

$$Az = f. \tag{B.7}$$

Here A is an $N \times N$ complex matrix and f a complex vector. Only the complex version of UMFPACK can handle the complex arithmetics. However, the code is written in such a way that also real solvers can be employed. This is done by representing the complex system as a $2N \times 2N$ real system.

All matrix solvers available in HERMES (ILU preconditioned CG and BiCG methods, PETSc, Trilinos, UMFPACK, and Gaussian elimination) can be employed to solve the linear system (B.7). However, the convergence of the iterative solvers in this case may be unsatisfactory due to the indefinite nature of the matrix A. As we said earlier, UMFPACK seems to be most appropriate for the discretized time-harmonic Maxwell's equations.

B.2.4 Example 1: L-shape domain problem

The first numerical example deals with a problem whose exact solution u is known, and thus the error function $e_{h,p} = u - u_{h,p}$ can be calculated exactly. We consider an L-shape domain $\Omega \subset \mathbb{R}^2$ with a reentrant corner, shown in Figure B.2.

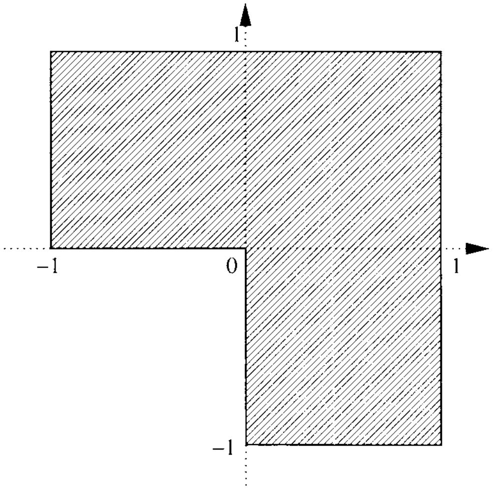


Figure B.2 Geometry of the L-shape domain.

Considered is the equation $-\Delta u = 0$ in Ω with the Dirichlet boundary conditions

$$u(\boldsymbol{x}) = R(\boldsymbol{x})^{2/3} \sin(2\theta(\boldsymbol{x})/3 + \pi/3) \qquad \text{for all } \boldsymbol{x} \in \partial\Omega.$$

Here $R(\boldsymbol{x})$ and $\theta(\boldsymbol{x})$ are the standard spherical coordinates in the plane. The exact solution has the form

$$u(\boldsymbol{x}) = R(\boldsymbol{x})^{2/3} \sin(2\theta(\boldsymbol{x})/3 + \pi/3) \qquad \text{for all } \boldsymbol{x} \in \Omega.$$

The magnitude of the gradient $|\nabla u|$ of the exact solution (whose calculation is left to the reader as an exercise) exhibits a singularity at the reentrant corner. Singularities are typical for second-order elliptic problems in domains with reentrant corners, and they make their numerical solution challenging. Despite being very local in space, they are a significant source of error. The error can be measured in a variety of different ways. The H^1-norm

$$\|e_{h,p}\|_{H^1(\Omega)} = \left(\int_\Omega |u - u_{h,p}|^2 + |\nabla u - \nabla u_{h,p}|^2 \, \mathrm{d}\boldsymbol{x} \right)^{\frac{1}{2}}$$

is a natural choice from the point of view of the weak formulation of the problem. The L^∞-norm

$$\|e_{h,p}\|_{L^\infty(\Omega)} = \sup_{\boldsymbol{x} \in \Omega} |u(\boldsymbol{x}) - u_{h,p}(\boldsymbol{x})|,$$

on the other hand, gives the maximum difference of u and $u_{h,p}$. We use the H^1-norm in what follows. The problem was solved twice, using the piecewise-affine FEM and the hp-FEM. In both cases it was our goal to attain the best possible accuracy using as few degrees of freedom as possible. The approximate solution, its gradient, finite element meshes, and a-posteriori error estimate $e_{h,p} \approx u_{ref} - u_{h,p}$ based on a very accurate reference solution u_{ref}, are shown in Figures B.3–B.7. The efficiency of the piecewise-affine FEM and the hp-FEM is compared in Table B.1.

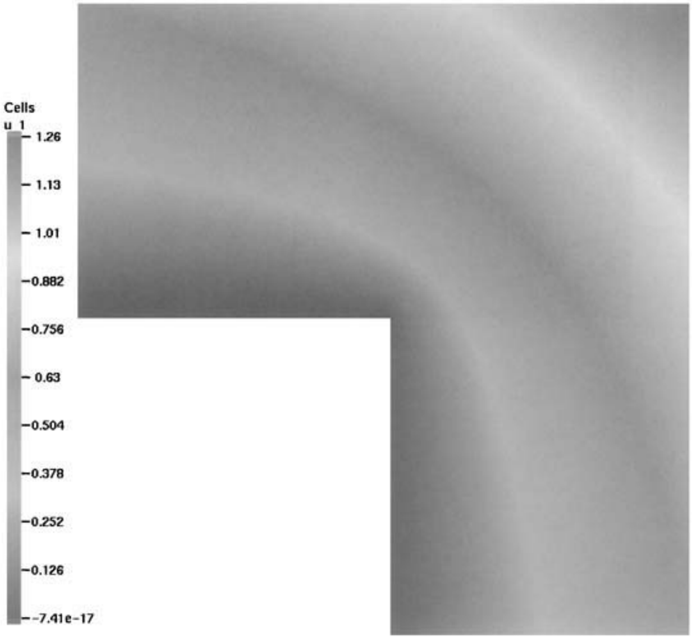


Figure B.3 Approximate solution $u_{h,p}$ of the L-shape domain problem.

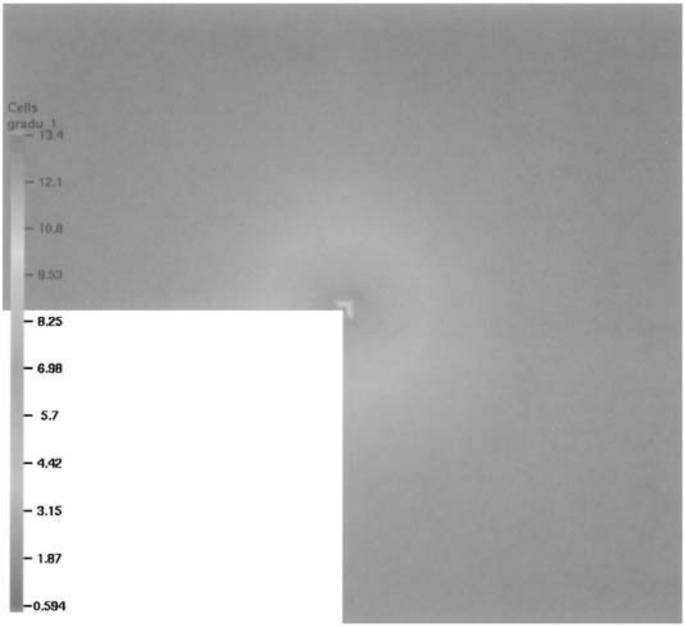


Figure B.4 Detailed view of $|\nabla u_{h,p}|$ at the reentrant corner (zoom = 70).

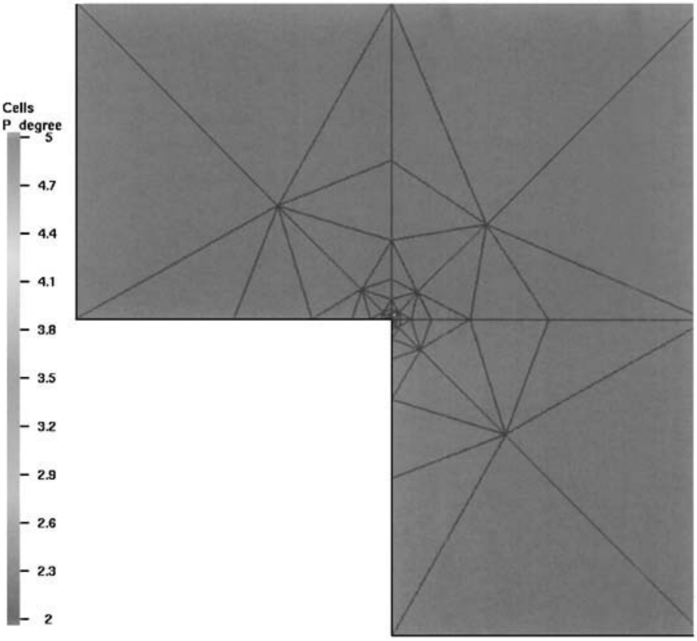


Figure B.5 The *hp*-mesh. Large fifth-order elements are used far from the singularity, and small quadratic elements cover the vicinity of the reentrant corner.

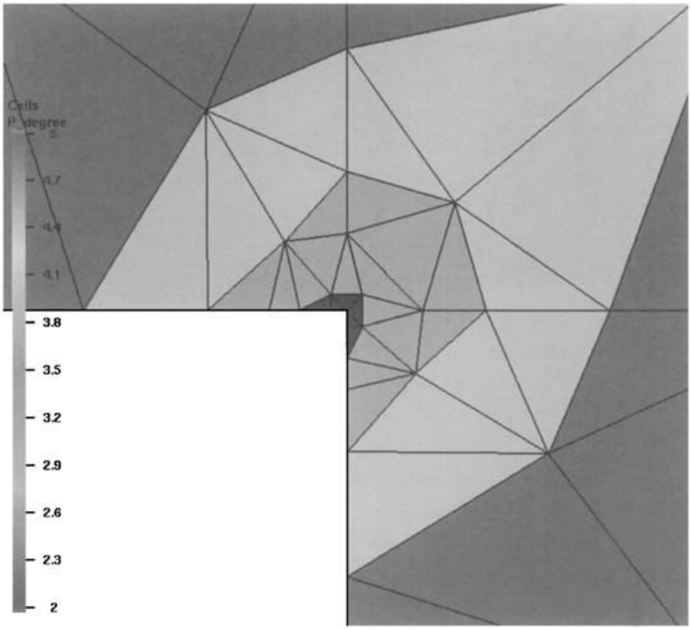


Figure B.6 The *hp*-mesh, details of the reentrant corner (zoom = 70).

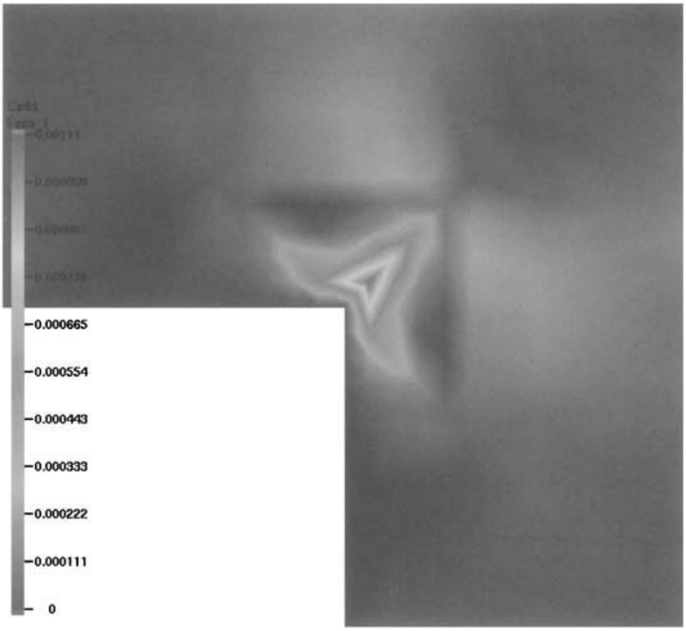


Figure B.7 A-posteriori error estimate for $u_{h,p}$, details of the reentrant corner (zoom = 70).

The geometry of the piecewise-affine mesh was identical to the hp-mesh, but the piecewise-affine mesh was moreover uniformly subdivided to reach the required accuracy (each edge was split into 60 equally long parts).

An efficiency comparison of the piecewise-affine FEM and hp-FEM is shown in Table B.1. Both computations, as well as all other computations shown in the following, were performed using our modular FEM system HERMES under identical conditions on a desktop Linux PC with a 3 GHz Pentium 4 processor and 2 GB of memory.

Table B.1 Comparison of the number of DOF, relative error in the H^1-norm, number of iterations of the matrix solver, and the CPU-time.

	Affine elements	hp elements
DOF	143161	839
Error	0.1876%	0.1603%
Iterations	421	30
CPU time	**2.1 min**	**0.35 sec**

Acknowledgment The numerical results presented in this section were obtained with the help of the triangular mesh generator Triangle [107] by Richard Shewchuk (see `http://www-2.cs.cmu.` and the visualization tool General Mesh Viewer (GMV) by Frank Ortega (see `http://www-xdiv.lanl.` `GMVHome.html`).

B.2.5 Example 2: Insulator problem

This time it is our goal to calculate the distribution of the electric field induced by an insulated conductor in the vicinity of a point where the conductor leaves the wall. The computational domain $\Omega \subset \mathbb{R}^2$ corresponding to this axisymmetric problem is depicted in Figure B.8.

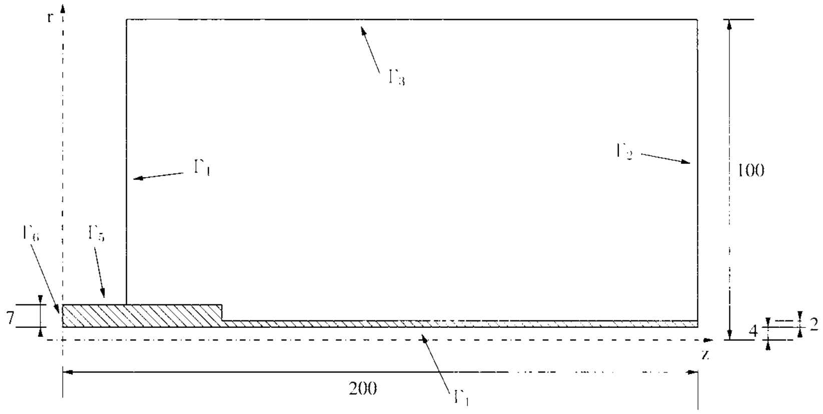


Figure B.8 Computational domain (all measures are in millimeters).

The wall itself, where we are not interested in the solution, is not included in the domain Ω. The same holds for the conductor along the horizontal axis of symmetry. Both the wall and the conductor are handled via suitable boundary conditions (to be defined below). The hatched subdomain $\Omega_2 \subset \Omega$ represents the insulator with the relative permittivity $\epsilon_r = 10$. The relative permittivity in the rest of the domain is $\epsilon_r = 1$. This problem is more difficult compared to the previous one, because in addition to a reentrant corner there is a material interface in the domain along which the electric field $\boldsymbol{E}$ is discontinuous (i.e., across which the scalar potential φ has a significant jump in the derivative). Solved is the standard potential equation of electrostatics (7.25) in cylindrical coordinates, equipped with the following boundary conditions:

$$\varphi = 220\,\mathrm{V} \qquad \text{on } \Gamma_1.$$

$$\varphi = 0\,\mathrm{V} \qquad \text{on } \Gamma_4 \cup \Gamma_5,$$

and

$$\frac{\partial \varphi}{\partial \nu} = 0 \qquad \text{on } \Gamma_2 \cup \Gamma_3 \cup \Gamma_6.$$

Again we compare the results obtained by means of the piecewise-affine FEM and *hp*-FEM. The approximate solution, its gradient, finite element meshes, and an a-posteriori error estimate are shown in Figures B.9–B.13. The efficiency of the piecewise-affine FEM and the *hp*-FEM is compared in Table B.2.

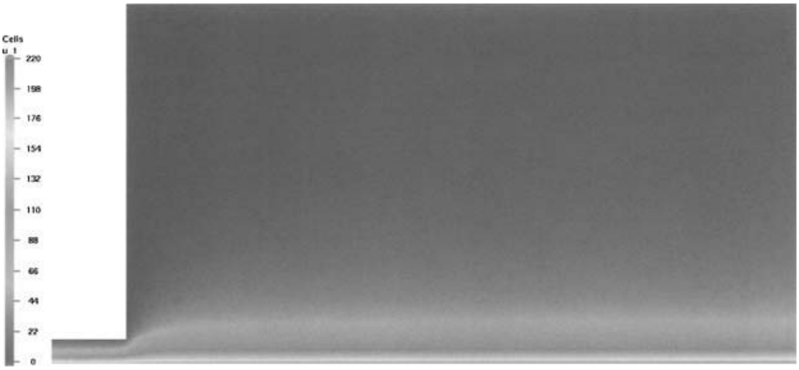


Figure B.9 Approximate solution $\varphi_{h,p}$ of the insulator problem.

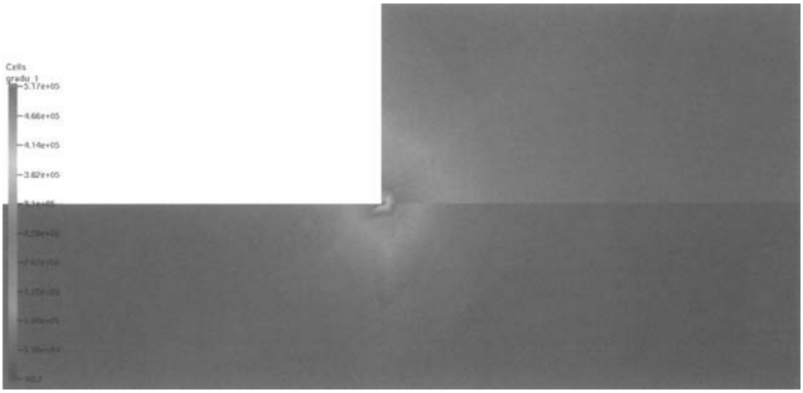


Figure B.10 Details of the singularity of $|\boldsymbol{E}_{h,p}| = |-\nabla\varphi_{h,p}|$ at the reentrant corner, and the discontinuity along the material interface (zoom = 1000).

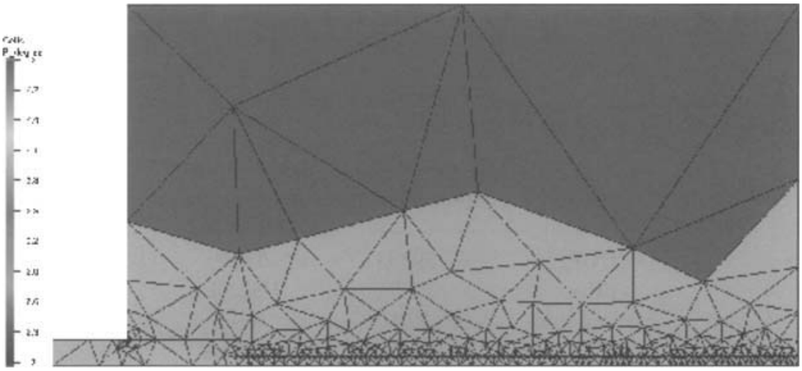


Figure B.11 The hp-mesh, global view. Large fifth-order elements are used far from the singularity and material interface, small quadratic elements are placed close to the reentrant corner and the material interface

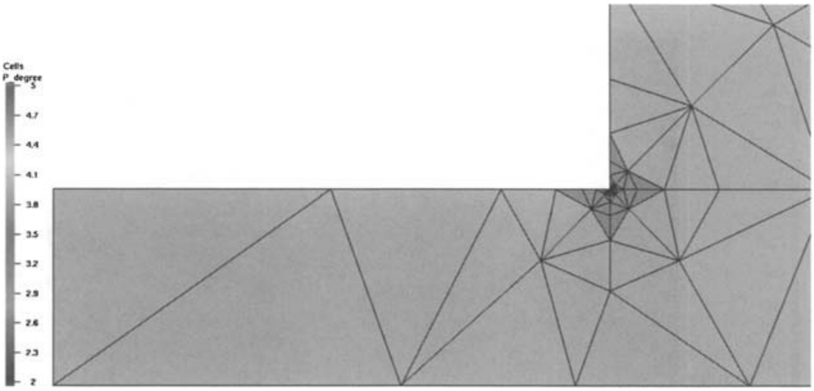


Figure B.12 The hp-mesh, details of the reentrant corner (zoom = 1000).

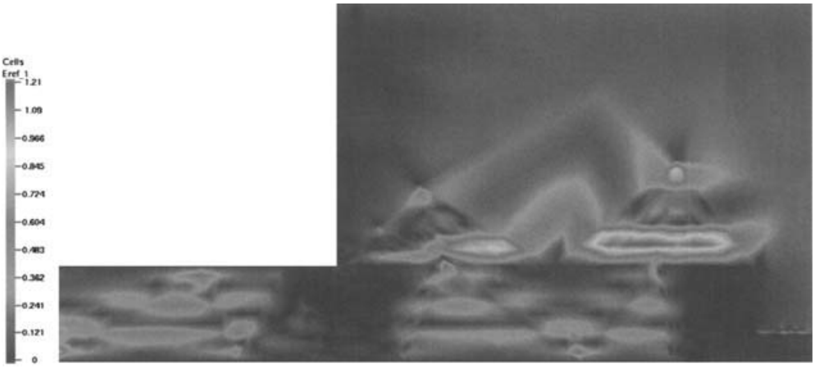


Figure B.13 A-posteriori error estimate for $\varphi_{h,p}$, details of the reentrant corner (zoom = 4).

The piecewise-affine mesh had geometry identical to the hp-mesh, but it was uniformly subdivided so that an accuracy similar to the hp-FEM could be reached (each edge was split into 23 equally long parts). An efficiency comparison is shown in Table B.2.

Table B.2 Comparison of the number of DOF, relative error in the H^1-norm, number of iterations of the matrix solver, and the CPU-time.

	Affine elements	hp elements
DOF	259393	6331
Error	1.617%	1.521%
Iterations	228	60
CPU time	**34 min**	**11.58 sec**

B.2.6 Example 3: Sphere-cone problem

The next problem also deals with electrostatics. A metallic sphere of the radius 200 mm carries an electric potential $\varphi_s = 100$ kV. The distance of the sphere to the ground is 1000 mm. There is a metallic cone 100 mm above the sphere with zero electric potential. The cone is 500 mm high and its bottom has the radius 100 mm. The axisymmetric computational domain Ω is depicted in Figure B.14 (notice that the figure describes the boundary conditions used). We solve equation (7.25) in cylindrical coordinates and compare the performance of the piecewise-affine and hp-FEM. The approximate solution, its gradient, the finite element meshes, and an a-posteriori error estimate are shown in Figures B.15–B.19. The efficiency of the piecewise-affine FEM and the hp-FEM is compared in Table B.3.

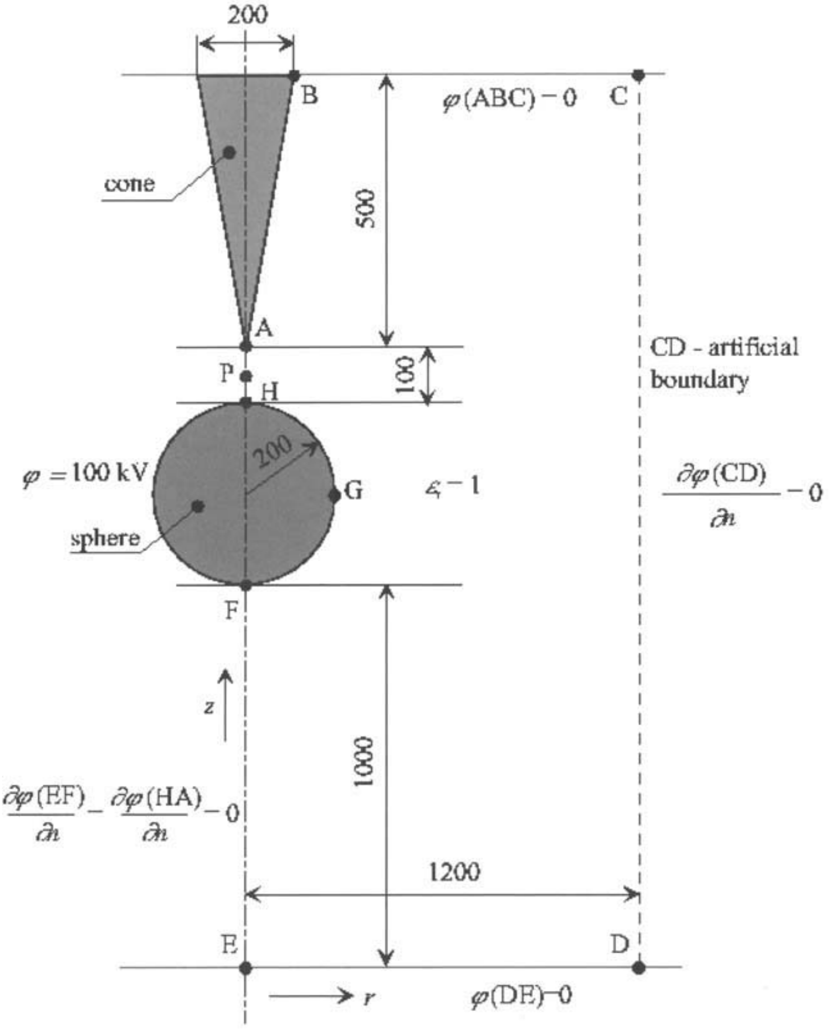


Figure B.14 Computational domain of the cone-sphere problem.

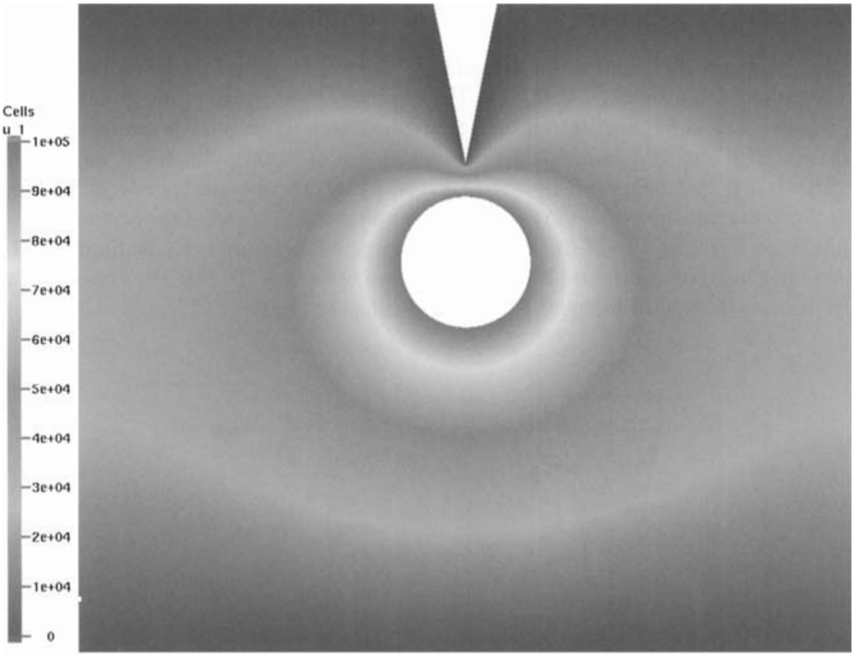


Figure B.15 Approximate solution $\varphi_{h,p}$ of the cone-sphere problem.

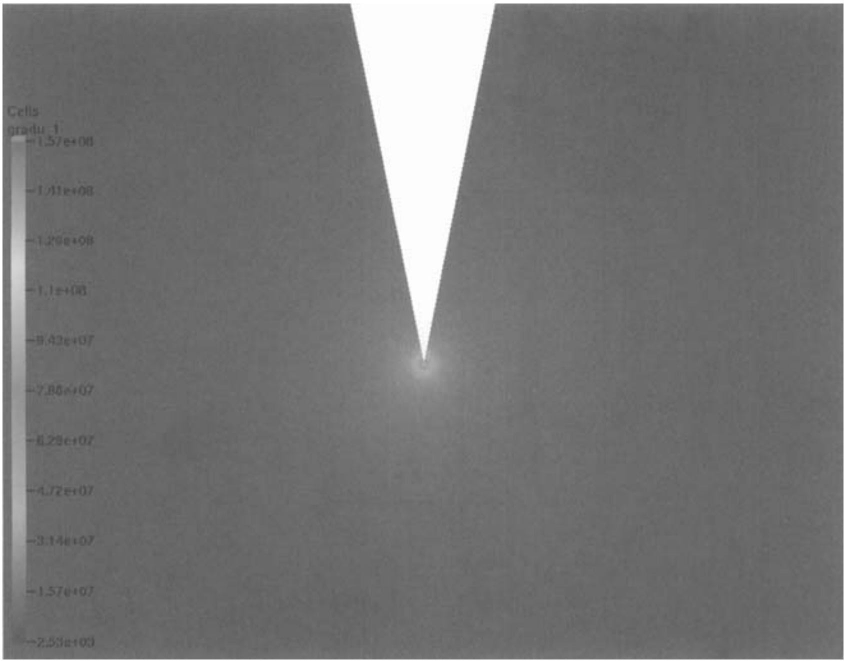


Figure B.16 Details of the singularity of $|E_{h,p}| = |-\nabla\varphi_{h,p}|$ at the tip of the cone (zoom = 50,000).

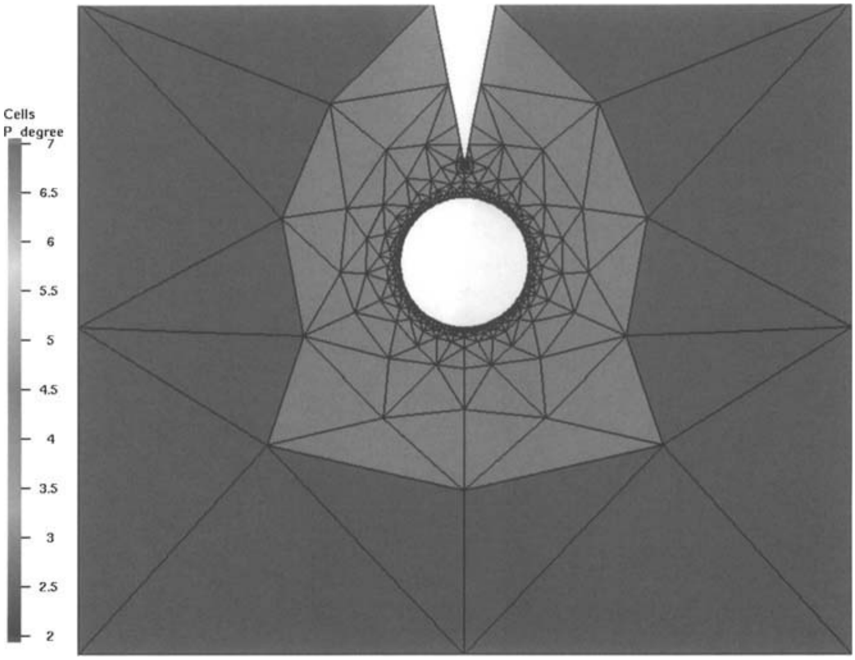


Figure B.17 The *hp*-mesh, global view. Large seventh-order elements are used far from the singularity and small quadratic elements at the tip of the cone.

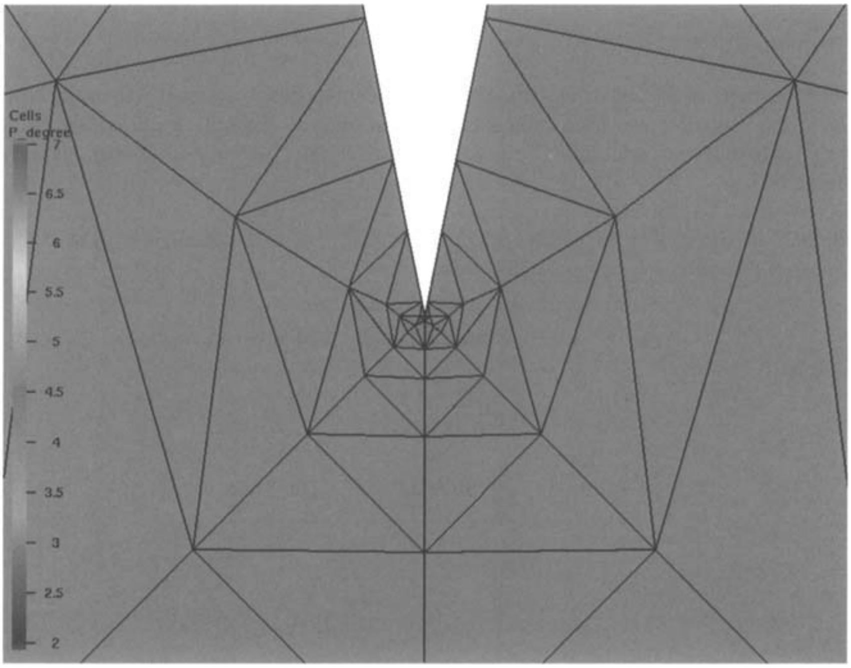


Figure B.18 The *hp*-mesh, details of the tip of the cone (zoom = 50,000).

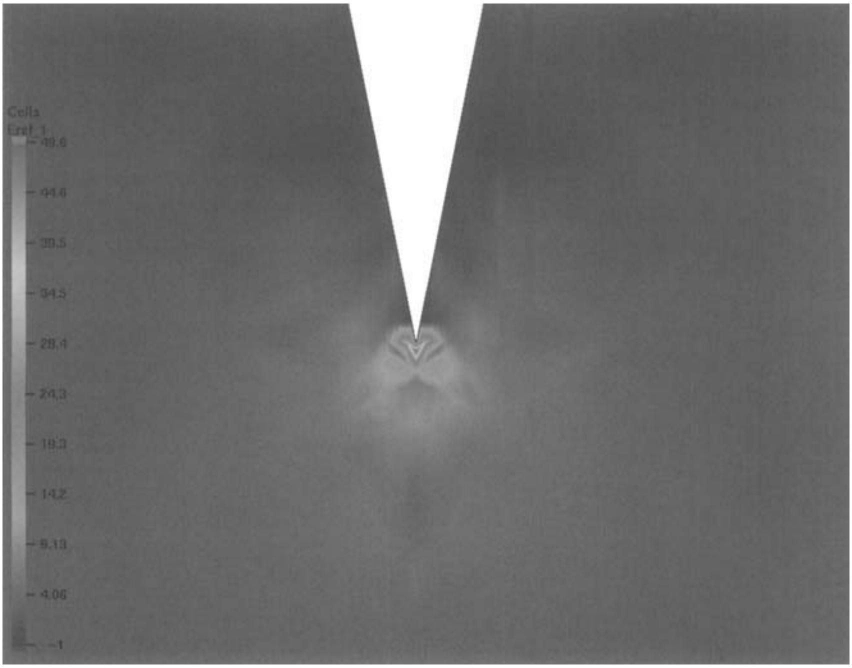


Figure B.19 A-posteriori error estimate for $\varphi_{h,p}$, details of the reentrant corner (zoom = 200,000).

The geometry of the piecewise-affine mesh was identical to the hp-mesh, but the piecewise-affine mesh was moreover uniformly subdivided in order to reach the required level of accuracy (each edge was split into 48 equally long parts). An efficiency comparison is shown in Table B.3.

Table B.3 Comparison of the number of DOF, relative error in the H^1-norm, number of iterations of the matrix solver, and the CPU-time.

	Affine elements	hp elements
DOF	488542	3317
Error	0.5858%	0.2804%
Iterations	859	44
CPU time	**30 min**	**10.53 sec**

B.2.7 Example 4: Electrostatic micromotor problem

This computation is rooted in the construction of electrostatic micromotors. These devices, which are capable of transforming the electric energy into motion analogously to standard electromotors, do not contain any coils or electric circuits that could be destroyed by strong electromagnetic waves. The goal of this computation is a highly-accurate approximation of the distribution of the electric field in a domain containing two electrodes and a thin object placed between them. The problem is plane-symmetric, and Figure B.20 shows one-half of the domain Ω.

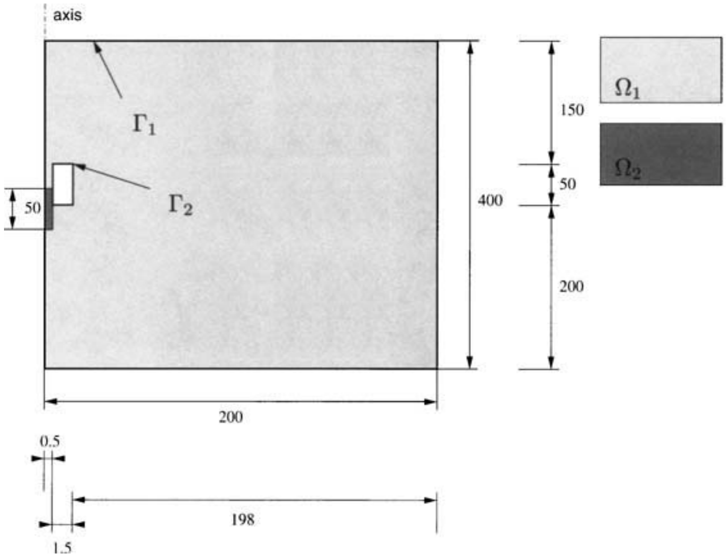


Figure B.20 Computational domain (the scaling was adjusted, but true measures in millimeters are provided). The electrode is modeled via a Dirichlet boundary condition.

The gray subdomain Ω_2 represents the moving part of the device, while the white subdomain Ω_2 represents the electrodes that are fixed. The distribution of the electric potential φ is governed equation (7.25),

$$-\nabla \cdot (\epsilon_r(\boldsymbol{x})\nabla\varphi(\boldsymbol{x})) = 0 \qquad \text{in } \Omega,$$

equipped with the Dirichlet boundary conditions

$$\varphi = 0\,\text{V} \qquad \text{on } \Gamma_1,$$

and

$$\varphi = 50\,\text{V} \qquad \text{on } \Gamma_2.$$

The relative permittivity ϵ_r is piecewise-constant, $\epsilon = 1$ in Ω_1 and $\epsilon = 10$ in Ω_2. We solve the problem twice, using the piecewise-affine and hp-FEM. The solution, gradient of the solution, a-posteriori error estimate, and the meshes are shown in Figures B.21–B.22. The efficiency of the piecewise-affine FEM and the hp-FEM is compared in Table B.4.

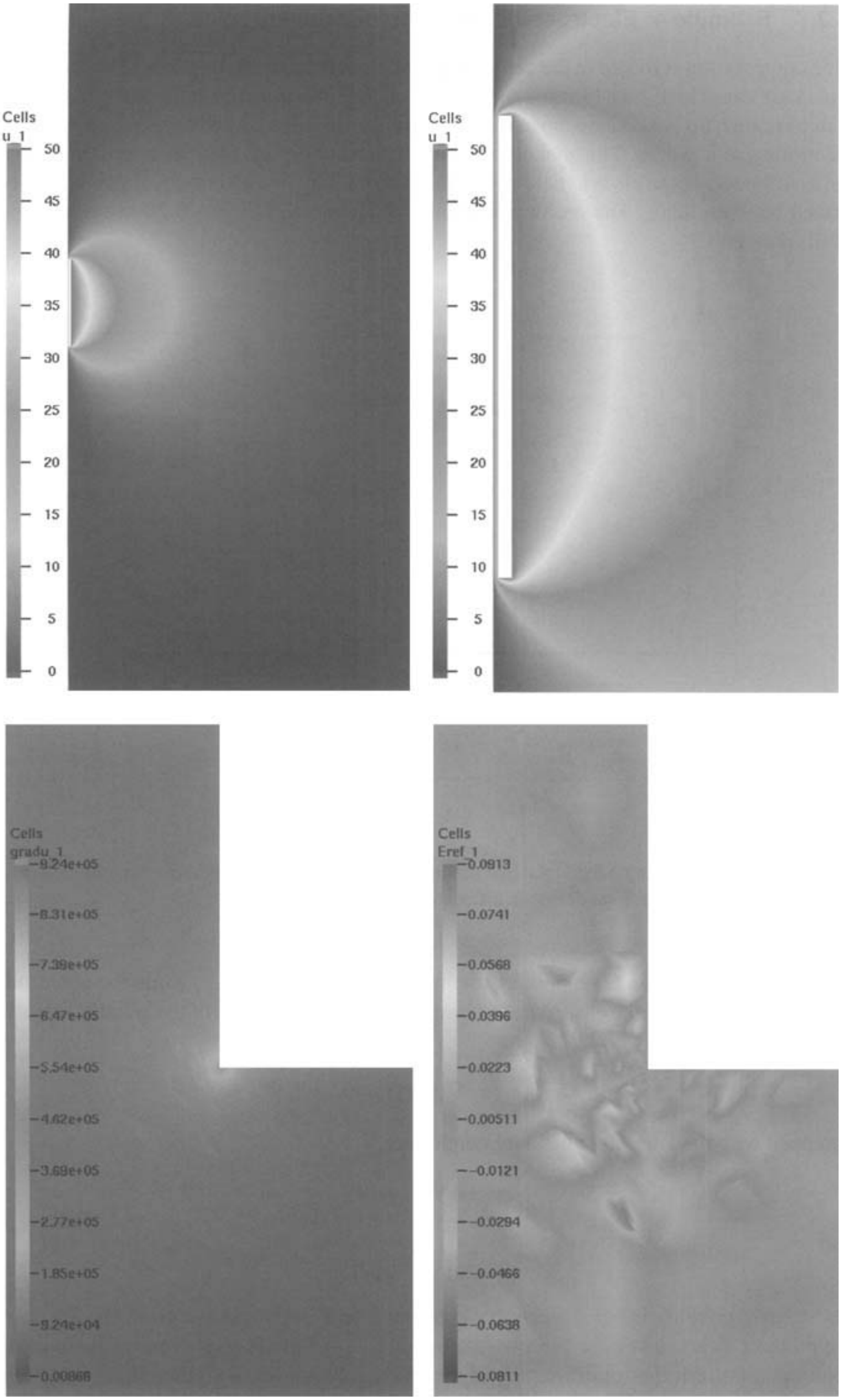


Figure B.21 Approximate solution of the micromotor problem. Top: electric potential $\varphi_{h,p}$ (zoom = 1 and 6). Bottom left: detailed view of the singularity of $|\boldsymbol{E}_{h,p}| = | - \nabla\varphi_{h,p}|$ at a corner of the electrode (zoom = 1000). Bottom right: Error estimate based on a reference solution (zoom = 1000).

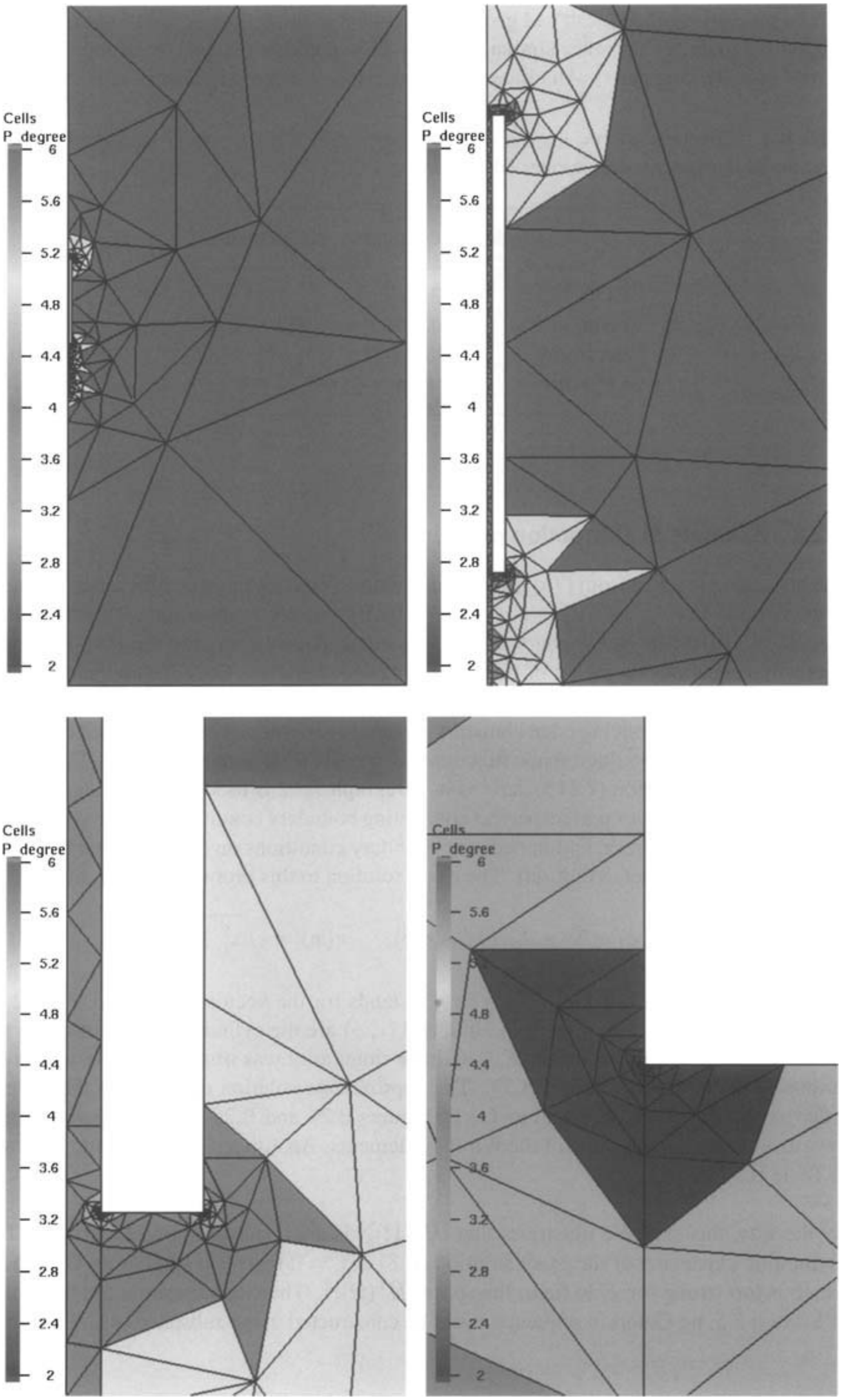


Figure B.22 The *hp*-mesh (zoom = 1, 6, 50, 1000). Large sixth-order elements are used far from the electrodes and small quadratic elements are placed at the reentrant corners.

The piecewise-affine mesh had geometry identical to the hp-mesh, but it was uniformly subdivided so that an accuracy similar to the hp-FEM could be reached (each edge was split into 44 equally long parts). An efficiency comparison is shown in Table B.4.

Table B.4 Comparison of the number of DOF, relative error in the H^1-norm, number of iterations of the matrix solver, and the CPU-time.

	Affine elements	hp elements
DOF	472384	4511
Error	0.2024%	0.173%
Iterations	387	71
CPU time	**32 min**	**17 sec**

B.2.8 Example 5: Diffraction problem

The last example taken from [76] is concerned with an electromagnetic diffraction problem in the domain $\Omega = (-10,10)^2 \setminus (0,10) \times (-10,0)$ with reentrant corner. The Maxwell's module of HERMES (see Paragraph B.2.3) is employed to discretize the time-harmonic Maxwell's equations by means of hierarchic hp edge elements. The edge elements use the same hp-FEM kernel as the elliptic module that was described in Paragraph B.2.2. The technology of the hierarchic edge elements is slightly different from the Nédélec elements. The hierarchic vector-valued shape functions used in HERMES can be found in [111]. The reference transformation (7.113) derived in Paragraph 7.5.2 is used without changes.

The problem involves perfect perfect conducting boundary conditions on the edges meeting at the reentrant corner, and impedance boundary conditions on the rest of the boundary (see [76] for their exact definition). The exact solution to this problem is given by

$$\boldsymbol{E}(\boldsymbol{x}) = \bar{\nabla} \times J_\alpha(r)\cos(\alpha\phi), \qquad r(\boldsymbol{x}) = \sqrt{x_1^2 + x_2^2}, \tag{B.8}$$

where the symbol $\bar{\nabla} = (\partial/\partial x_2, -\partial/\partial x_1)^T$ stands for the vector-valued curl, $\alpha = 2/3$, J_α is the Bessel function of the first kind, and (r, ϕ) are the cylindrical coordinates in the plane. The approximate solution $\boldsymbol{E}_{h,p}$ (whose singularity was truncated for visualization purposes) is depicted in Figure B.23. The approximate solution obtained on the lowest-order mesh is optically identical to $\boldsymbol{E}_{h,p}$. Figures B.24 and B.25 show the hp-mesh and lowest-order mesh consisting of the Whitney elements. An efficiency comparison is shown in Table B.5.

By the way, this example illustrates that $[H^1(\Omega)]^2$ is not a subspace of $\boldsymbol{H}(\mathrm{curl}, \Omega)$: The asymptotic expansion of the exact solution (B.8) at $r = 0$ reveals a singularity $O(r^{-4/3})$, which is too strong for $\boldsymbol{E}$ to lie in the space $[H^1(\Omega)]^2$. Thus, as we said at the beginning of Section 7.5, no Galerkin sequence could be constructed using subspaces of $[H^1(\Omega)]^2$.

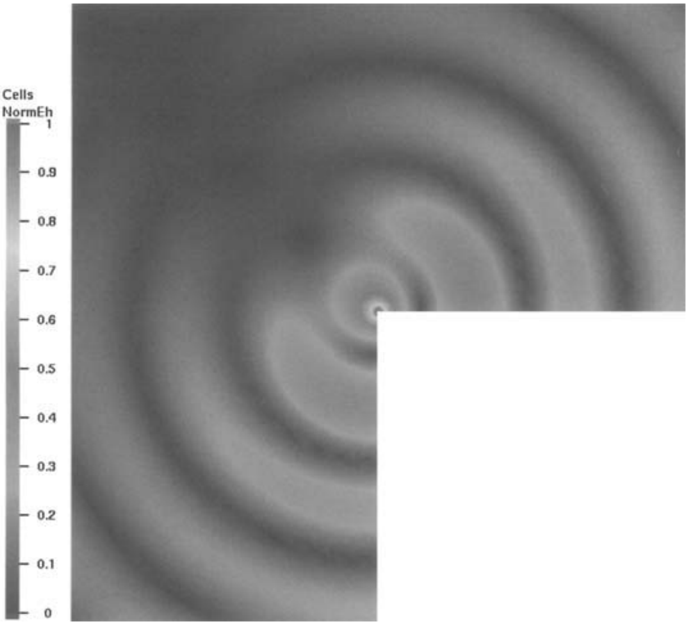


Figure B.23 Approximate solution to the diffraction problem (the magnitude of the phasor of the electromagnetic field $|\boldsymbol{E}_{h,p}|$). The singularity at the reentrant corner was truncated for visualization purposes.

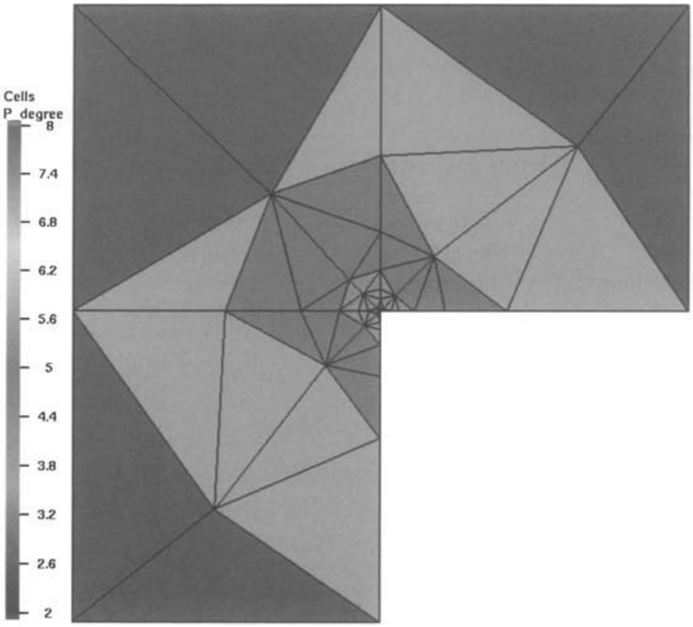


Figure B.24 The hp-mesh consisting of hierarchic edge elements.

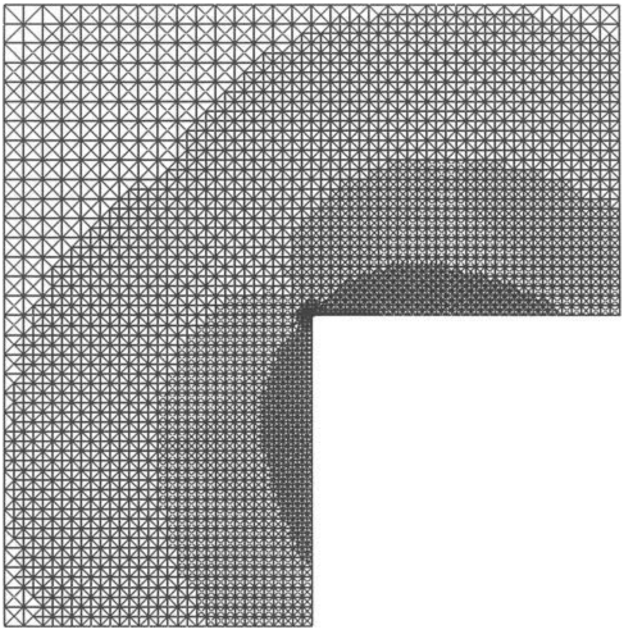


Figure B.25 The mesh consisting of the lowest-order (Whitney) edge elements.

The lowest-order mesh shown in Figure B.25 was uniformly subdivided in order to reach an accuracy comparable to the *hp*-FEM (each edge was split into 10 equally long parts). An efficiency comparison is shown in Table B.5.

Table B.5 Comparison of the number of DOF, relative error in the $H(\mathrm{curl})$-norm, and the CPU-time.

	Whitney edge elements	*hp* edge elements
DOF	2586540	4324
Error	0.6445%	0.6211%
CPU time	**21.2 min**	**2.49 sec**

REFERENCES

1. R.A. Adams, J.J.F. Fournier: *Sobolev Spaces*, 2nd edition, Academic Press/Elsevier, Amsterdam, 2003.

2. A. Ahagon, K. Fujiwara, T. Nakata: Comparison of various kinds of edge elements for electromagnetic field analysis, IEEE Trans. Magn. **32**, 898–901, 1996.

3. M. Ainsworth, J. Coyle: Hierarchic *hp*-edge element families for Maxwell's equations on hybrid quadrilateral/triangular meshes, Comput. Methods Appl. Mech. Engrg. **190**, 6709–6733, 2001.

4. F. Assons, P. Ciarlet (Jr.), P.A. Raviart, E. Sonnendrücker: Characterization of the singular part of the solution of the Maxwell's equations in a polyhedral domain, *Math. Meth. Appl. Sci.* **22**, 485–499, 1999.

5. K. Atkinson: *An Introduction to Numerical Analysis*, 2nd edition, John Wiley & Sons, New York, 1989.

6. K. Atkinson, W. Han: *Theoretical Numerical Analysis*, Springer, New York, 2001.

7. I. Babuška, T. Scapolla: Benchmark computation and performance evaluation for a rhombic plate-bending problem, *Int. J. Numer. Meth. Engrg* **28**, 155–180, 1989.

8. I. Babuška, The stability of domains and the question of formulation of plate problems, *Appl. Math.*, 463–467, 1962.

9. I. Babuška, T. Strouboulis: *Finite Element Method and Its Reliability*, Clarendon Press, Oxford, 2001.

10. M.L. Barton, Z.J. Cendes: New vector finite elements for three-dimensional magnetic computation, J. Appl. Phys **61**, 3919–3921, 1987.

11. E. Bertolazzi: Discrete conservation and discrete maximum principle for elliptic PDEs, Math. Models Methods Appl. Sci. **8**, 685-711, 1998.

12. A.M. Bespalov: Finite element method for the eigenmode problem of a RF cavity, Sov. J. Numer. Anal. Math. Model. **3**, 163–178, 1988.

13. G. Birkhoff, M. H. Schultz, R.S. Varga: Piecewise Hermite interpolation in one and two variables with application to partial differential equations, Numer. Math. **11**, 232–256, 1968.

14. V.S. Borisov: On discrete maximum principles for linear equation systems and monotonicity of difference schemes, SIAM J. Matrix Anal. Appl. **24**, No. 4, 1110–1135, 2004.

15. L. Bos: On certain configurations of points in $\mathbb{R}^n$ which are uniresolvant for polynomial interpolation, J. Approx. Theory **64**, 271–280, 1991.

16. L. Bos: M. A. Taylor, B. A. Wingate: Tensor product Gauss-Lobatto points are Fekete points for the cube, Math. Comp. **70**, 1543–1547, 2001.

17. A. Bossavit, J. Verite: A mixed FEM-BEM method to solve 3D eddy current problems, IEEE Trans. Magn. **18**, 431–435, 1982.

18. D. Braess: *Finite Elements. Theory, Fast Solvers, and Applications in Solid Mechanics*, 2nd Edition, Cambridge University Press, 2001.

19. J.H. Bramble, B.E. Hubbard: New monotone-type approximations of elliptic problems, Math. Comp. **18**, 349–367, 1964.

20. M. Brdička, *Continuum Mechanics*, ČSAV, Prague, 1959.

21. K.E. Brenan, S.L. Campbell, L.R. Petzold: *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, Classics in Applied Mathematics **14**, SIAM, Philadelphia, PA, 1996.

22. S. C. Brenner and L. R. Scott: *The Mathematical Theory of Finite Element Methods*, 2nd edition, Springer, New York, 2002.

23. F. Brezzi, M. Fortin: *Mixed and Hybrid Finite Element Methods*, Springer, Berlin, 1991.

24. L. Bucciarelly, N. Dworsky; *Sophie Germain, An Essay on the History of Elasticity*, Reidel, New York, 1980.

25. J. C. Butcher: *Numerical Methods for Ordinary Differential Equations*, John Wiley & Sons, New York, 2003.

26. H.S. Carslaw: *Introduction to the Theory of Fourier's Series and Integrals*, 3rd edition, Dover, New York, 1950.

27. J. Céa: Approximation variationelle des problémes aux limites, Ann. Inst. Fourier (Grenoble) **14**, 345–444, 1964.

28. S.B. Chae: *Lebesgue Integration*, 2nd edition, Springer, New York, 1995.

29. D. Chapelle, K. J. Bathe: *The Finite Element Analysis of Shells – the Fundamentals*, Springer, Berlin, 2003.

30. P.G. Ciarlet: *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1979.

31. P.G. Ciarlet, P.A. Raviart: Maximum principle and uniform convergence for the finite element method, Comput. Methods Appl. Mech. Engrg. **2**, 17–31, 1973.

32. D. Colton, R. Kress: *Integral Equation Methods in Scattering Problems*, John Wiley & Sons, New York, 1983.

33. D. Colton, R. Kress: *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd edition, Springer, New York, 1998.

34. J. Conway: *A course in Functional Analysis*, 2nd edition, Springer, New York, 1990.

35. R. Cools: Advances in multidimensional integration, J. Comp. Appl. Math. **149**, 1–12, 2002.

36. R. Courant: *Differential and Integral Calculus* (two volume set), John Wiley & Sons, New York, 1992.

37. M. Crouzeix, P.A. Raviart: Conforming and non-conforming finite element methods for solving the stationary Stokes equations, RAIRO R-3, 77–104, 1973.

38. H.F. Davis: *Fourier Series and Orthogonal Functions*, Dover, New York, 1963.

39. L. Demkowicz: Edge finite elements of variable order for Maxwell's equations. In: Proc. International Workshop on Scientific Computing in Electrical Engineering (SCEE), Warnemünde, August 20-23, 2000.

40. L. Demkowicz, I. Babuška: Optimal p-interpolation error estimates for edge finite elements of variable order in 2D, TICAM Report 01-11, The University of Texas at Austin, April 2001.

41. P. Deuflhard, U. Nowak: Extrapolation integrators for quasilinear implicit ODEs. In: *Large Scale Scientific Computing* (P. Deuflhard, B. Engquist, eds.), Progress in Scientific Computing 7, Birkhäuser, Basel, 37–50, 1987.

42. J. R. Dormand, P. J. Prince: A family of embedded Runge-Kutta formulae, J. Comp. Appl. Math **6**, 19–26, 1980.

43. J. R. Dormand, P. J. Prince: Higher-order embedded Runge-Kutta formulae, J. Comp. Appl. Math **7**, 67–75, 1981.

44. M. Dubiner: Spectral element methods on triangles and other domains, J. Sci. Comput. **6**, 345–390, 1991.

45. F. Dubois: Discrete vector potential representation of a divergence-free vector-field in three-dimensional domains. Numerical analysis of a model problem, SIAM J. Numer. Anal. **27**, 1103–1141, 1990.

46. D.A. Dunavant: Economical symmetrical quadrature rules for complete polynomials over a square domain, Int. J. Numer. Methods Engrg. **21**, 1777–1784, 1985.

47. D.A. Dunavant: High degree efficient symmetrical Gaussian quadrature rules for the triangle, Int. J. Numer. Methods Engrg. **21**, 1129–1148, 1985.

48. D.A. Dunavant: Efficient symmetrical cubature rules for complete polynomials of high degree over the unit cube, Int. J. Numer. Methods Engrg. **23**, 397–407, 1986.

49. H. Engels: *Numerical Quadrature and Cubature*, Academic Press, London, 1980.

50. L. C. Evans: *Partial Differential Equations*, Berkeley Mathematics Lecture Notes, Volume 3, 1994.

51. E. Fehlberg: Low-order classical Runge-Kutta formulas with step-size control and their application to some heat transfer problems. Computing **6**, 61–71, 1970.

52. M. Feistauer, J. Felcman and I. Straškraba: *Mathematical and Computational Methods for Compressible Flow*, Oxford University Press, 2004.

53. L. Fejér: Bestimmung derjenigen Abszissen eines Intervalles für welche die Quadratsumme der Grundfunktionen der Lagrangeschen Interpolation im Intervalle $[-1, 1]$ ein möglichst kleined Maximum besitzt. Ann. Scuola Norm. Sup. Pisa Sci. Fis. Mt. Ser. II, **1**, 263–276, 1932.

54. J. Felcman, P. Šolín: On the construction of the Osher-Solomon scheme for 3D Euler equations, East-West J. Num. Math., 43–64, 1998.

55. S. Fučík, A. Kufner: *Nonlinear Differential Equations*, Elsevier, Amsterdam, 1980.

56. W.C. Gear: The simultaneous numerical solution of differential-algebraic equations, IEEE Trans. Circuit Theory **18**, 89–95, 1971.

57. D. Gilbarg, N.S. Trudinger: *Elliptic Partial Differential Equations of Second Order*, Springer, Berlin, 1977.

58. I.S. Gradshteyn, I.M. Ryzhik: *Tables of Integrals, Series, and Products*, 6th ed., Academic Press, San Diego, 2000.

59. E. Hairer, S.P. Norsett, G. Wanner: *Solving Ordinary Differential Equations I – Nonstiff Problems*, 2nd edition, Springer, New York, 1993.

60. E. Hairer, G. Wanner: *Solving Ordinary Differential Equations II – Stiff and DAE Problems*, 2nd edition, Springer, New York, 1996.

61. M. Hano: Finite element analysis of dielectric-loaded waveguides, IEEE Trans. Microwave Theory Tech. **32**, 1275–1279, 1984.

62. M. T. Heath: *Scientific Computing, An Introductory Survey*, 2nd edition, McGraw-Hill, New York, 2002.

63. J.S. Hesthaven: From electrostatics to almost optimal nodal sets for polynomial interpolation in a simplex, SIAM J. Numer. Anal. **35**, 655–676, 1998.

64. J.S. Hesthaven, D. Gottlieb: Stable spectral methods for conservation laws on triangles with unstructured grids, Comput. Methods Appl. Mech. Engrg. **175**, 361–381, 1999.

65. H.G. Heuser: *Functional Analysis*, John Wiley & Sons, New York, 1982.

66. A.C. Hindmarsh: ODEPACK. A systematized collection of ODE solvers. In: *Scientific Computing* (R.S. Stapleman et al., eds.), IMACS Transactions on Scientific Computation 1, North-Holland, Amsterdam, 55–64, 1983.

67. W. Höhn, H.D. Mittelmann: Some remarks on the discrete maximum principle for finite elements of higher order, Computing **27**, 145–154, 1981.

68. L.O. Jay: Inexact simplified Newton iterations for implicit Runge-Kutta methods. SIAM J. Numer. Anal. **38**, 1369–1388, 2000.

69. G.E. Karniadakis, S.J. Sherwin: *Spectral/hp Element Methods for CFD*, Oxford University Press, 1999.

70. P. Keast: Moderate-degree tetrahedral quadrature formulas, Comput. Methods Appl. Mech. Engrg. **55**, 339–348, 1986.

71. G. Kirchhoff: ¨Uber das Gleichgewicht und die Bewegung einer Elastischen Scheibe, J. Reine und Angewandte Mathematik **40**, 51–88, 1850.

72. A. Kirsch, P. Monk: A finite element/spectral element method for approximation of the time-harmonic Maxwell's system in $\mathbb{R}^3$. *SIAM J. Appl. Math.* **55**, 1324–44, 1995.

73. R. Kress: *Linear Integral Equations*, 2nd edition, Springer, Berlin, 1999.

74. W. Kutta: Beitrag zur näherungsweisen Integration totaler Differentialgleichungen, Zeitschrift f. Math. u. Physik **46**, 435–453, 1901

75. David C. Lay: *Linear Algebra and Its Applications (3rd Edition)*, Pearson Addison Wesley, 2002.

76. R. Leis: *Initial Boundary Value Problems in Mathematical Physics*, John Wiley & Sons, New York, 1988.

77. R.J. LeVeque: *Numerical Methods for Conservation Laws*, Birkhäuser, 1992.

78. R.J. LeVeque: *Finite Volume Methods for Hyperbolic Problems*, Cambridge University Press, 2002.

79. J. L. Lions, E. Magenes: *Problèmes aux Limites non Homogènes et Applications, 2*, Dunod, Paris, 1968.

80. J.N. Lyness, D. Jespersen: Moderate degree symmetric quadrature rules for the triangle, J. Inst. Math. Appl. **15**, 15–32, 1975.

81. R. D. Mindlin: Influence of rotatory inertia and shear and flexural motions of isotropic elastic plates, J. Appl. Mech. **18**, 31–38, 1951.

82. P. Monk: An analysis of Nédélec's method for spatial discretization of Maxwell's equations, J. Comput. Appl. Math., 103–121, 1993.

83. P. Monk: *Finite Element Methods for Maxwell's Equations*, Clarendon Press, Oxford, 2002.

84. N. Morrison: *Introduction to Fourier Analysis*, John Wiley & Sons, New York, 1994.

85. G. Mur: Edge elements, their advantages and their disadvantages, IEEE Trans. Magn. **30**, 3552–3557, 1994.

86. J. Nečas, I. Hlaváček: *Mathematical Theory of Elastic and Elasto-Plastic Bodies: An Introduction*, Elsevier, Amsterdam, 1981.

87. J.C. Nédélec: Mixed finite elements in $\mathbb{R}^3$, Numer. Math. **35**, 315–341, 1980.

88. J.C. Nédélec: A new family of mixed finite elements in $\mathbb{R}^3$, Numer. Math. **50**, 57–81, 1986.

89. S. P. Norsett, G. Wanner: The real-pole sandwich for rational approximations and oscillation equations, BIT **19**, 79–94, 1979.

90. J. E. Pasciak: Spectral and pseudospectral methods for advection equations, Math. Comput. **35**, 1081–1092, 1980.

91. L.R. Petzold: A description of DDASSL: A differential/algebraic system solver, Sandia Report Sand 82-8637, Sandia National Laboratory, Livermore, CA, 1982.

92. M. Práger, J. Taufer, E. Vitásek: Overimplicit multistep methods, Appl. Math. **18**, 399–421, 1973.

93. A. Quarteroni, A. Valli: *Numerical Approximation of Partial Differential Equations*, Springer, New York, 1994.

94. P. A. Raviart, J. M. Thomas: *Introduction à l'Analyse Numérique Des Équations aux Dérivées Partielles*, Masson, Paris, 1983.

95. J. N. Reddy: *An Introduction to the Finite Element Method*, 2nd edition, McGraw-Hill, New York, 1993.

96. E. Reissner: Reflections on the theory of elastic plates, Appl. Mech. Rev. **38**, 1453–1464, 1985.

97. E. Reissner: The effect of transverse shear deformation on the bending of elastic plates, J. Appl. Mech. **12**, 69–76, 1945.

98. H. Royden: *Real Analysis*, 3rd edition, Prentice-Hall, Englewood Cliffs, NJ, 1988.

99. W. Rudin: *Real and Comple Analysis*, McGraw-Hill, New York, 1986.

100. W. Rudin: *Functional Analysis*, McGraw-Hill, New York, 1976.

101. C. Runge: Über die numerische Auflösung von Differentialgleichungen, Math. Ann. **46**, 167-178, 1895.

102. P. Silvester, R. Ferrari: *Finite Elements for Electrical Engineers*, Cambridge University Press, 1996.

103. Y. Saad: *Iterative Methods for Sparse Linear Systems*, PWS Series in Computer Science, PWS Publishing Company, Boston, MA, 1996.

104. H.E. Salzer: Tables for facilitating of Chebyshev's quadrature formulae, J. Math. Phys. **26**, 191–194, 1947.

105. Ch. Schwab: *p- and hp-Finite Element Methods*, Clarendon Press, Oxford, 1998.

106. G. Sewell: *The Numerical Solution of Ordinary and Partial Differential Equations*, Academic Press, New York, 1988.

107. J.R. Shewchuk: Triangle – a two-dimensional quality mesh generator and Delaunay triangulator. Software available at `http://www.cs.cmu.edu/~quake/triangle.html`.

108. S.L. Sobolev, V.L. Vaskevich: *The Theory of Cubature Formulas*, Kluwer Academic Publishers, Dordrecht, 1996.

109. P. Šolín: On a mesh generation technique based on a special smoothing procedure for uniform inner point distribution, Acta Technica ASCR 45, No. 4, 397–417, 2000. Software available at `http://www.math.utep.edu/Faculty/solin`.

110. P. Šolín, K. Segeth: Optimal hierarchic higher-order Hermite elements for H^2-problems in 1D, in print.

111. P. Šolín, K. Segeth, I. Doležel: *Higher-Order Finite Element Methods*, Chapman & Hall/CRC Press, 2003.

112. P. Šolín, T. Vejchodský: On the discrete maximum principle for *hp*-FEM, submitted.

113. G. Strang, G. J. Fix: *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

114. A. Stroud, D. Secrest: *Gaussian Quadrature Formulas*, Prentice-Hall, Englewood Cliffs, NJ, 1966.

115. D. Sun et al.: Spurious modes in finite element methods, IEEE Trans. Antennas and Propagation **37**, 12–24, 1995.

116. B. Szabó, I. Babuška: *Finite Element Analysis*, John Wiley & Sons, New York, 1991.

117. G. Szegö: *Orthogonal Polynomials*, AMS Colloquium Publications **23**, Providence, RI, 1939.

118. M. A. Taylor, B. A. Wingate, R. E. Vincent: An algorithm for computing Fekete points in the triangle, SIAM J. Numer. Anal. **38**, 1707–1720, 2000.

119. S. P. Timoshenko, S. Woinowski-Krieger: *Theory of Plates and Shells*, 2nd edition, McGraw-Hill, New York, 1959.

120. A. Vande Wouwer, P. Saucez, W.E. Schiesser (eds.), *Adaptive Method of Lines*, Chapman & Hall/CRC, Boca Raton, FL, 2001.

121. V. Vogelsang: On the strong unique continuation principle for inequalities of Maxwell type. *Math. Ann.* **289**, 285–295, 1991.

122. J. Webb: Hierarchical vector based functions of arbitrary order for triangular and tetrahedral finite elements, IEEE Trans. Antennas and Propagation **47**, 1244–1253, 1999.

123. H. Whitney: *Geometric Integration Theory*, Princeton University Press, 1957.

124. O. Zienkiewicz, R. L. Taylor: *The Finite Element Method, Vol. 2 - Solid Mechanics*, 5th edition, Butterworth-Heinemann, Woburn, MA, 2002.

INDEX

PURE AND APPLIED MATHEMATICS

A Wiley-Interscience Series of Texts, Monographs, and Tracts

ADÁMEK, HERRLICH, and STRECKER—Abstract and Concrete Catetories
ADAMOWICZ and ZBIERSKI—Logic of Mathematics
AINSWORTH and ODEN—A Posteriori Error Estimation in Finite Element Analysis
AKIVIS and GOLDBERG—Conformal Differential Geometry and Its Generalizations
ALLEN and ISAACSON—Numerical Analysis for Applied Science
*ARTIN—Geometric Algebra
AUBIN—Applied Functional Analysis, Second Edition
AZIZOV and IOKHVIDOV—Linear Operators in Spaces with an Indefinite Metric
BERG—The Fourier-Analytic Proof of Quadratic Reciprocity
BERMAN, NEUMANN, and STERN—Nonnegative Matrices in Dynamic Systems
BERKOVITZ—Convexity and Optimization in $\mathbb{R}^n$
BOYARINTSEV—Methods of Solving Singular Systems of Ordinary Differential Equations
BURK—Lebesgue Measure and Integration: An Introduction
*CARTER—Finite Groups of Lie Type
CASTILLO, COBO, JUBETE, and PRUNEDA—Orthogonal Sets and Polar Methods in Linear Algebra: Applications to Matrix Calculations, Systems of Equations, Inequalities, and Linear Programming
CASTILLO, CONEJO, PEDREGAL, GARCIÁ, and ALGUACIL—Building and Solving Mathematical Programming Models in Engineering and Science
CHATELIN—Eigenvalues of Matrices
CLARK—Mathematical Bioeconomics: The Optimal Management of Renewable Resources, Second Edition
COX—Galois Theory
†COX—Primes of the Form $x^2 + ny^2$: Fermat, Class Field Theory, and Complex Multiplication
*CURTIS and REINER—Representation Theory of Finite Groups and Associative Algebras
*CURTIS and REINER—Methods of Representation Theory: With Applications to Finite Groups and Orders, Volume I
CURTIS and REINER—Methods of Representation Theory: With Applications to Finite Groups and Orders, Volume II
DINCULEANU—Vector Integration and Stochastic Integration in Banach Spaces
*DUNFORD and SCHWARTZ—Linear Operators
Part 1—General Theory
Part 2—Spectral Theory, Self Adjoint Operators in Hilbert Space
Part 3—Spectral Operators
FARINA and RINALDI—Positive Linear Systems: Theory and Applications
FOLLAND—Real Analysis: Modern Techniques and Their Applications
FRÖLICHER and KRIEGL—Linear Spaces and Differentiation Theory
GARDINER—Teichmüller Theory and Quadratic Differentials

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in paperback.

GILBERT and NICHOLSON—Modern Algebra with Applications, Second Edition

*GRIFFITHS and HARRIS—Principles of Algebraic Geometry

GRILLET—Algebra

GROVE—Groups and Characters

GUSTAFSSON, KREISS and OLIGER—Time Dependent Problems and Difference Methods

HANNA and ROWLAND—Fourier Series, Transforms, and Boundary Value Problems, Second Edition

*HENRICI—Applied and Computational Complex Analysis

Volume 1, Power Series—Integration—Conformal Mapping—Location of Zeros

Volume 2, Special Functions—Integral Transforms—Asymptotics—Continued Fractions

Volume 3, Discrete Fourier Analysis, Cauchy Integrals, Construction of Conformal Maps, Univalent Functions

*HILTON and WU—A Course in Modern Algebra

*HOCHSTADT—Integral Equations

JOST—Two-Dimensional Geometric Variational Procedures

KHAMSI and KIRK—An Introduction to Metric Spaces and Fixed Point Theory

*KOBAYASHI and NOMIZU—Foundations of Differential Geometry, Volume I

*KOBAYASHI and NOMIZU—Foundations of Differential Geometry, Volume II

KOSHY—Fibonacci and Lucas Numbers with Applications

LAX—Functional Analysis

LAX—Linear Algebra

LOGAN—An Introduction to Nonlinear Partial Differential Equations

MARKLEY—Principles of Differential Equations

MORRISON—Functional Analysis: An Introduction to Banach Space Theory

NAYFEH—Perturbation Methods

NAYFEH and MOOK—Nonlinear Oscillations

PANDEY—The Hilbert Transform of Schwartz Distributions and Applications

PETKOV—Geometry of Reflecting Rays and Inverse Spectral Problems

*PRENTER—Splines and Variational Methods

RAO—Measure Theory and Integration

RASSIAS and SIMSA—Finite Sums Decompositions in Mathematical Analysis

RENELT—Elliptic Systems and Quasiconformal Mappings

RIVLIN—Chebyshev Polynomials: From Approximation Theory to Algebra and Number Theory, Second Edition

ROCKAFELLAR—Network Flows and Monotropic Optimization

ROITMAN—Introduction to Modern Set Theory

*RUDIN—Fourier Analysis on Groups

SENDOV—The Averaged Moduli of Smoothness: Applications in Numerical Methods and Approximations

SENDOV and POPOV—The Averaged Moduli of Smoothness

SEWELL—The Numerical Solution of Ordinary and Partial Differential Equations, Second Edition

SEWELL—Computational Methods of Linear Algebra, Second Edition

*SIEGEL—Topics in Complex Function Theory

Volume l—Elliptic Functions and Uniformization Theory

Volume 2—Automorphic Functions and Abelian Integrals

Volume 3—Abelian Functions and Modular Functions of Several Variables

SMITH and ROMANOWSKA—Post-Modern Algebra

ŠOLÍN–Partial Differential Equations and the Finite Element Method

STADE—Fourier Analysis

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in paperback.

STAKGOLD—Green's Functions and Boundary Value Problems, Second Editon
STAHL—Introduction to Topology and Geometry
STANOYEVITCH—Introduction to Numerical Ordinary and Partial Differential Equations Using MATLAB^{R}
*STOKER—Differential Geometry
*STOKER—Nonlinear Vibrations in Mechanical and Electrical Systems
*STOKER—Water Waves: The Mathematical Theory with Applications
WATKINS—Fundamentals of Matrix Computations, Second Edition
WESSELING—An Introduction to Multigrid Methods
†WHITHAM—Linear and Nonlinear Waves
†ZAUDERER—Partial Differential Equations of Applied Mathematics, Second Edition

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in paperback.