Springer Series in Computational Mathematics 18

Wolfgang Hackbusch

Elliptic Differential Equations

Theory and Numerical Treatment

Second Edition

Springer

Springer Series in Computational Mathematics

18

More information about this series at http://www.springer.com/series/797

Wolfgang Hackbusch

Elliptic Differential Equations

Theory and Numerical Treatment

Second Edition

Springer

Wolfgang Hackbusch
MPI für Mathematik
in den Naturwissenschaften
Leipzig, Germany

Original first edition W. Hackbusch, Theorie und Numerik elliptischer Differentialgleichungen, B.G. Teubner, Stuttgart 1987.

ISSN 0179-3632 ISSN 2198-3712 (electronic)
Springer Series in Computational Mathematics
ISBN 978-3-662-54960-5 ISBN 978-3-662-54961-2 (eBook)
DOI 10.1007/978-3-662-54961-2

Library of Congress Control Number: 2017942725

Mathematics Subject Classification (2010): 35J05, 35J08, 35J15, 35J20, 35J25, 35J30, 35J40, 35J50, 65N06, 65N12, 65N15, 65N22, 65N25, 65N30, 65N50, 65N80, 65M50, 35S15

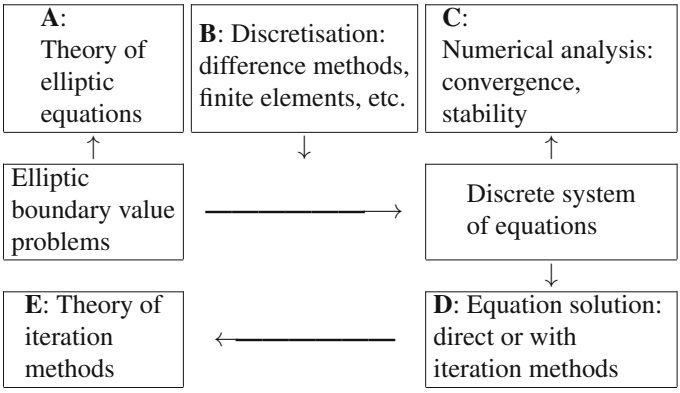
Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer-Verlag GmbH Germany
The registered company address is: Heidelberger Platz 3, 14197 Berlin, Germany

For my granddaughter Finja

Preface

This book has developed from lectures that the author gave for mathematics students at the Ruhr-Universität Bochum and the Christian-Albrechts-Universität Kiel. The present work is restricted to the theory of partial differential equations of elliptic type, which otherwise tends to be given a treatment which is either too superficial or too extensive. The following sketch shows what the problems are for elliptic differential equations.



The *theory* of elliptic differential equations (**A**) is concerned with questions of existence, uniqueness, and properties of solutions. The first problem of numerical treatment is the description of the *discretisation procedures* (**B**), which give finite-dimensional equations for approximations to the solutions. The subsequent second part of the numerical treatment is *numerical analysis* (**C**) of the procedure in question. In particular it is necessary to find out if, and how fast, the approximation converges to the exact solution.

The solution of the finite-dimensional equations (**D**, **E**) is in general no simple problem, since more than 10^6 unknowns can occur. The discussion of this third area of numerical problems is skipped (it is the subject of the author's monographs [142] and [137]).

The descriptions of discretisation procedures and their analyses are closely connected with corresponding chapters of the theory of elliptic equations. In addition, it is not possible to undertake a well-founded numerical analysis without a basic knowledge of elliptic differential equations. Since the latter cannot, in general, be assumed of a reader, it seems to me necessary to present the numerical study along with the theory of elliptic equations.

The book is conceived in the first place as an introduction to the treatment of elliptic boundary-value problems. It should, however, serve to lead the reader to further literature on special topics and applications. It is intentional that certain topics, which are often handled rather summarily, (e.g., eigenvalue problems, regularity properties) are treated here in greater detail.

The exposition is strictly limited to linear elliptic equations. Thus a discussion of the Navier-Stokes equations, which are important for fluid mechanics, is excluded; however, one can approach these matters via the Stokes equation, which is thoroughly treated as an example of an elliptic system.

The exercises that are presented are an integral part of the exposition. Their solution is given in the appendix. The reader should test his understanding of the subject on the exercises.

The first and second editions of the German book are published by Teubner Stuttgart in 1986 and 1996. In 1992, the first English edition [134] appeared in the Springer Series Computational Mathematics, translated by R. Fadiman and P. D. F. Ion. Recently, an extended fourth edition of the German version [143] has appeared. The extensions, which are now available in the present second English edition, concern additional sections (e.g., about finite elements) and the complete proofs of the exercises. Furthermore, the number of references is more than trebled.

The author wishes to thank Springer Verlag for their cordial collaboration.

Kiel, April 2017 *Wolfgang Hackbusch*

Contents

Chapter 1
Partial Differential Equations and Their Classification Into Types

Abstract This chapter introduces the partial differential equations and their distinction into three types.

1.1 Examples

An *ordinary* differential equation describes a function which depends on only *one* variable. However, most of the problems require two or more variables. Almost all physical quantities depend on the spatial variables x, y, and z, and possibly on time t. The time dependence might be omitted for stationary processes, and one might perhaps save one spatial dimension by special geometric assumptions, but even then there would still remain at least two independent variables. Differential equations in two or more variables are called *partial differential equations*. They may contain the first partial derivatives

$$u_{x_i} = u_{x_i}(x_1, x_2, \ldots, x_n) = \partial u(x_1, x_2, \ldots, x_n)/\partial x_i \qquad (1 \le i \le n)$$

with respect to x_i or even higher partial derivatives $u_{x_i x_j}$, etc.

Unlike ordinary differential equations, partial differential equations cannot be analysed all together. Rather, one distinguishes between three[1] types of equations which have different properties and also require different numerical methods.

Before the characteristics for the types are defined, let us introduce some examples of partial differential equations. All of the following examples will contain only two independent variables x, y.[2] The first two examples are partial differential equations of *first order*, since only first partial derivatives occur.

[1] There are even differential equations belonging to none of these three types. More details will follow after Definition 1.14.

[2] In the general case, the independent variables are denoted by the n-tuple $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. In the two- and three-dimensional case, we prefer x, y or respectively x, y, z, to avoid indices. As soon as summations occur, the notation by x_i is more convenient.

W. Hackbusch, *Elliptic Differential Equations*, Springer Series
in Computational Mathematics 18, DOI 10.1007/978-3-662-54961-2_1

Example 1.1. Find a solution $u(x, y)$ of

$$u_y(x, y) = 0. \tag{1.1}$$

It is obvious that $u(x, y)$ must be independent of y, i.e., the solution has the form $u(x, y) = \varphi(x)$. On the other hand, $u(x, y) = \varphi(x)$ with arbitrary φ is a solution of (1.1).

Equation (1.1) is a special case of the next example.

Example 1.2. Find a solution $u(x, y)$ of

$$c\, u_x(x, y) - u_y(x, y) = 0 \qquad (c \text{ constant}). \tag{1.2}$$

Let u be a solution. Introduce new coordinates $\xi := x + cy$, $\eta := y$ and define

$$v(\xi, \eta) := u(x(\xi, \eta), y(\xi, \eta))$$

with the aid of $x(\xi, \eta) = \xi - c\,\eta$ and $y(\xi, \eta) = \eta$. Applying the chain rule, we obtain $v_\eta = u_x x_\eta + u_y y_\eta$ with $x_\eta = -c$ and $y_\eta = 1$. Hence $v_\eta(\xi, \eta) = 0$ follows from (1.2). This equation is analogous to (1.1), and Example 1.1 shows that $v(\xi, \eta) = \varphi(\xi)$. Replacing ξ, η by x, y, we obtain

$$u(x, y) = \varphi(x + cy). \tag{1.3}$$

Conversely, through (1.3) we obviously obtain a solution of (1.2), as long as φ is continuously differentiable.

In order to determine uniquely the solution of an ordinary differential equation $u' - f(u) = 0$ one needs an initial value $u(x_0) = u_0$. The partial differential equation (1.2) can be augmented by the *initial-value* function

$$u(x, y_0) = u_0(x) \qquad \text{for } x \in \mathbb{R} \tag{1.4}$$

on the line $\{(x, y_0) : x \in \mathbb{R}\}$ with y_0 a constant. The comparison of (1.3) and (1.4) shows that $\varphi(x + cy_0) = u_0(x)$. Thus φ is determined by $\varphi(x) = u_0(x - cy_0)$. The unique solution of the initial value problem (1.2) and (1.4) reads

$$u(x, y) = u_0(x - c\,(y_0 - y)).$$

The following three examples involve differential equations of *second order*.

Example 1.3 (potential equation). Let Ω be an open subset of $\mathbb{R}^2$. Find a solution of

$$u_{xx} + u_{yy} = 0 \qquad \text{in } \Omega. \tag{1.5}$$

If one identifies $(x, y) \in \mathbb{R}^2$ with the complex number $z = x + \mathrm{i}y \in \mathbb{C}$, the solutions can be given immediately. The real and imaginary parts of any function

$f(z)$ holomorphic in Ω are solutions of (1.5). Examples are the powers

$$\Re\mathfrak{e}\, z^0 = 1, \qquad \Re\mathfrak{e}\, z^2 = x^2 - y^2$$

as well as

$$\Re\mathfrak{e}\, \log(z - z_0) = \log \sqrt{(x - x_0)^2 + (y - y_0)^2}, \text{ if } z_0 \notin \Omega.$$

To determine the solution uniquely one needs the boundary values $u(x, y) = \varphi(x, y)$ for all (x, y) on the boundary $\Gamma = \partial\Omega$ of Ω.

Another name of the potential equation (1.5) is *Laplace equation*.

Example 1.4 (wave equation). All solutions of

$$u_{xx} - u_{yy} = 0 \qquad \text{in } \Omega \tag{1.6}$$

are given by

$$u(x, y) = \varphi(x + y) + \psi(x - y), \tag{1.7}$$

where φ and ψ are arbitrary twice continuously differentiable functions. Suitable initial values are, for example,

$$u(x, 0) = u_0(x), \quad u_y(x, 0) = u_1(x) \qquad \text{for } x \in \mathbb{R}, \tag{1.8}$$

where u_0 and u_1 are given functions. Inserting (1.7) into (1.8), one finds

$$u_0 = \varphi + \psi, \quad u_1 = \varphi' - \psi' \qquad (\varphi', \psi' : \text{ derivatives of } \varphi \text{ and } \psi)$$

and infers that

$$\varphi' = (u_0' + u_1)/2, \quad \psi' = (u_0' - u_1)/2.$$

From this one can determine φ and ψ up to constants of integration. One constant can be chosen arbitrarily, for example, by $\varphi(0) = 0$, and the other is determined by $u(0, 0) = u_0(0) = \varphi(0) + \psi(0)$.

Exercise 1.5. Prove that every solution of the wave equation (1.6) has the form (1.7). *Hint:* Use $\xi = x + y$ and $\eta = x - y$ as new variables.

The next equation describes the heat conduction of an infinite wire located from $-\infty$ to $+\infty$, where u is the temperature, while y is interpreted as time.

Example 1.6 (heat equation). Find the solution of

$$u_{xx} - u_y = 0 \qquad \text{for } x \in \mathbb{R},\ y \geq 0. \tag{1.9}$$

The separation of variables $u(x, y) = v(x)w(y)$ gives a solution for every $c \in \mathbb{R}$:

$$u(x, y) = \sin(cx) \exp(-c^2 y).$$

Another solution of (1.9) for $y > 0$ is

$$u(x,y) = \frac{1}{\sqrt{4\pi y}} \int_{-\infty}^{\infty} u_0(\xi) \exp\left(\frac{-(x-\xi)^2}{4y}\right) \mathrm{d}\xi, \tag{1.10}$$

where $u_0(\cdot)$ is an arbitrary continuous and bounded function. The initial condition matching equation (1.9), in contrast to (1.8), contains only one function:

$$u(x,0) = u_0(x) \qquad \text{for } x \in \mathbb{R}. \tag{1.11}$$

The solution (1.10), which initially is defined only for $y > 0$, can be extended continuously to $y = 0$ and there satisfies the initial value requirement (1.11).

Exercise 1.7. Let u_0 be bounded in $\mathbb{R}$ and continuous at x. Then prove that the right-hand side of (1.10) converges to $u_0(x)$ for $y \searrow 0$. *Hint:* First show that

$$u(x,y) = u_0(x) + \frac{1}{\sqrt{4\pi y}} \int_{-\infty}^{\infty} [u_0(\xi) - u_0(x)] \exp\left(\frac{-(x-\xi)^2}{4y}\right) \mathrm{d}\xi$$

and then decompose the integral into subintegrals over $[x-\delta, x+\delta]$ and $(-\infty, x-\delta) \cup (x+\delta, \infty)$.

As with ordinary differential equations, equations of higher order can be described by systems of first-order equations. In the following we give some examples.

Example 1.8. Let the pair (u,v) be the solution of the system

$$u_x + v_y = 0, \qquad v_x + u_y = 0. \tag{1.12}$$

If u and v are twice differentiable, the differentiation of (1.12) yields the equations $u_{xx} + v_{xy} = 0$ and $v_{xy} + u_{yy} = 0$, which together imply that $u_{xx} - u_{yy} = 0$. Thus u is a solution of the wave equation (1.6). The same can be shown for v.

Example 1.9 (Cauchy–Riemann differential equations). If u and v satisfy the system

$$u_x + v_y = 0, \qquad v_x - u_y = 0 \qquad \text{in } \Omega \subset \mathbb{R}^2, \tag{1.13}$$

then the same consideration as in Example 1.8 yields that both u and v satisfy the potential equation (1.5).

Example 1.10. If the differentiable functions u and v satisfy the system

$$u_x + v_y = 0, \qquad v_x + u = 0, \tag{1.14}$$

then v solves the heat equation (1.9).

Proof. If u is differentiable, then also v_x because of the second equation. Differentiation in x yields $v_{xx} = -u_x$. Insertion into the first equation proves (1.9). ∎

In fluid mechanics, the following system describes a viscous liquid.

Example 1.11 (Stokes equations). In the system

$$\begin{aligned} u_{xx} + u_{yy} - w_x &= 0, \\ v_{xx} + v_{yy} - w_y &= 0, \\ u_x + v_y &= 0 \end{aligned}$$

u and v denote the flow velocities in x and y directions, while w denotes the pressure. Note that the system is of second order with respect to u and v, whereas no second derivative of w occurs.

1.2 Classification of Second-Order Equations into Types

The general linear differential equation of second order in two variables reads

$$\begin{aligned} a(x,y)u_{xx} + 2b(x,y)u_{xy} + c(x,y)u_{yy} \\ + d(x,y)u_x + e(x,y)u_y + f(x,y)u + g(x,y) = 0. \end{aligned} \tag{1.15}$$

Definition 1.12. (a) Equation (1.15) is said to be *elliptic* at (x,y) if

$$a(x,y)c(x,y) - b^2(x,y) > 0.$$

(b) Equation (1.15) is said to be *hyperbolic* at (x,y) if

$$a(x,y)c(x,y) - b^2(x,y) < 0.$$

(c) Equation (1.15) is said to be *parabolic* at (x,y) if[3]

$$ac - b^2 = 0 \quad \text{and} \quad \operatorname{rank}\begin{bmatrix} a & b & d \\ b & c & e \end{bmatrix} = 2 \quad \text{in } (x,y).$$

(d) Equation (1.15) is said to be elliptic (hyperbolic, parabolic) in $\Omega \in \mathbb{R}^2$ if it is elliptic (hyperbolic, parabolic) at all $(x,y) \in \Omega$.

If different types occur at different $(x,y) \in \Omega$, the differential equation is of *mixed type*.

[3] Usually, the parabolic type is defined by $ac - b^2 = 0$. However, the equation $u_{xx}(x,y) + u_x(x,y) = 0$ or even the purely algebraic equation $u(x,y) = 0$ should not be called parabolic.

Example 1.13. The potential equation (1.5) is elliptic, the wave equation (1.6) is of hyperbolic type, while the heat equation (1.9) is parabolic.

The definition of types can easily be generalised to the case, where more than two independent variables occur. The general *linear differential equation of second order* in n variables $\mathbf{x} = (x_1, \ldots, x_n)$ reads

$$\sum_{i,j=1}^{n} a_{ij}(\mathbf{x})\, u_{x_i x_j} + \sum_{i=1}^{n} a_i(\mathbf{x})\, u_{x_i} + a(\mathbf{x})\, u = f(\mathbf{x}). \tag{1.16}$$

Since $u_{x_i x_j} = u_{x_j x_i}$ holds for twice continuously differentiable functions, one can assume in (1.16) that, without loss of generality,

$$a_{ij}(\mathbf{x}) = a_{ji}(\mathbf{x}) \qquad (1 \le i, j \le n).$$

Thus, the coefficients $a_{ij}(\mathbf{x})$ define a symmetric $n \times n$ matrix

$$A(\mathbf{x}) = \left(a_{ij}(\mathbf{x})\right)_{i,j=1,\ldots,n} \tag{1.17}$$

which therefore has only real eigenvalues (cf. [142, Theorem A.41]).

Translating (1.15) into the notation (1.16) yields

$$A = \begin{bmatrix} a & b \\ b & c \end{bmatrix}, \quad a_1 = d, \quad a_2 = e, \quad a = f, \quad f = -g.$$

Definition 1.14. (a) Equation (1.16) is said to be *elliptic* at $\mathbf{x}$ if all n eigenvalues of the matrix $A(\mathbf{x})$ have the same sign (± 1) (i.e., if $A(\mathbf{x})$ is positive or negative definite).

(b) Equation (1.16) is said to be *hyperbolic* at $\mathbf{x}$ if $n-1$ eigenvalues of $A(\mathbf{x})$ have the same sign (± 1) and one eigenvalue has the opposite sign.

(c) Equation (1.16) is said to be *parabolic* at $\mathbf{x}$ if one eigenvalue vanishes, the remaining $n-2$ eigenvalues have the same sign, and $\operatorname{rank}(A(\mathbf{x}), \mathbf{a}(\mathbf{x})) = n$, where $\mathbf{a}(\mathbf{x}) = (a_1(\mathbf{x}), \ldots, a_n(\mathbf{x}))^{\mathsf{T}}$ are the coefficients of the first derivatives in (1.16).

(d) Equation (1.16) is said to be elliptic in $\Omega \in \mathbb{R}^n$ if it is elliptic at all $\mathbf{x} \in \Omega$.

Definition 1.14 makes it clear that the three types mentioned by no means cover all cases. An unclassified equation occurs, for example, if $A(\mathbf{x})$ has two positive and two negative eigenvalues.

In place of (1.16) we also write

$$Lu = f,$$

where

$$L = \sum_{i,j=1}^{n} a_{ij}(\mathbf{x}) \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^{n} a_i(\mathbf{x}) \frac{\partial}{\partial x_i} + a(\mathbf{x}) \tag{1.18}$$

is a *linear differential operator* of second order. The operator

$$L_0 = \sum_{i,j=1}^{n} a_{ij}(\mathbf{x}) \frac{\partial^2}{\partial x_i \partial x_j},$$

which contains only the highest derivatives of L, is called the *principal part* of L.

Remark 1.15. The ellipticity or hyperbolicity of equation (1.16) depends only on the principal part of the differential operator.

Exercise 1.16 (invariance of the type under coordinate transformations). Let (1.16) be defined for $\mathbf{x} \in \Omega$. The transformation

$$\Phi : \Omega \subset \mathbb{R}^n \to \Omega' \subset \mathbb{R}^n$$

is assumed to have a nonsingular Jacobian matrix $S = \partial\Phi/\partial\mathbf{x} \in C^1(\Omega)$ at $\mathbf{x} \in \Omega$. Prove that equation (1.16) does not change its type at $\mathbf{x}$ if it is written in the new coordinates $\boldsymbol{\xi} = \Phi(\mathbf{x})$. *Hint:* The matrix $A = (a_{ij})$ becomes SAS^{T} after the transformation. Use Remark 1.15 and Sylvester's inertia theorem (cf. Sylvester [280], Gantmacher [108, §X.2], or Liesen–Mehrmann [193, Theorem 18.23]).

1.3 Type Classification for Systems of First Order

The Examples 1.8–1.10 are special cases of the general linear system of first order in two variables:

$$u_x(x,y) - A(x,y)\, u_y(x,y) + B(x,y)\, u(x,y) = f(x,y). \tag{1.19}$$

Here $u = (u_1, \ldots, u_m)^{\mathsf{T}}$ is a vector function, and A, B are $m \times m$ matrices. In contrast to Section 1.2, A need not be symmetric and can have complex eigenvalues If the eigenvalues $\lambda_1, \ldots, \lambda_m$ are real, and if there exists a decomposition $A = S^{-1}DS$ with $D = \operatorname{diag}\{\lambda_1, \ldots, \lambda_m\}$, A is called *real-diagonalisable*.

Definition 1.17. (a) System (1.19) is said to be *hyperbolic* at (x,y) if $A(x,y)$ is real-diagonalisable.

(b) System (1.19) is said to be *elliptic* at (x,y) if all the eigenvalues of $A(x,y)$ are not real.

If A is real or possesses m distinct real eigenvalues the system is hyperbolic since those conditions are sufficient for real diagonalisability. A single real equation, in particular, is always hyperbolic.

Examples 1.1 and 1.2 are hyperbolic according to the preceding remark. System (1.12) from Example 1.8 has the form (1.19) with

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

It is hyperbolic since A is real-diagonalisable:

$$A = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}.$$

The Cauchy–Riemann system (1.13), which is closely connected with the potential equation (1.5), is elliptic since it has the form (1.19) with

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix},$$

and A has the eigenvalues $\pm \mathrm{i}$.

The system (1.14) corresponding to the (parabolic) heat equation can be described as system (1.19) with

$$A = \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix}.$$

The eigenvalues $\lambda_1 = \lambda_2 = 0$ may be real but A is not diagonalisable. Hence, system (1.14) is neither hyperbolic nor elliptic.

A more general system than (1.19) is

$$A_1 u_x + A_2 u_y + Bu = f. \tag{1.20}$$

If A_1 is invertible then multiplication by A_1^{-1} gives the form (1.19) with $A = -A_1^{-1} A_2$. Otherwise one has to investigate the generalised eigenvalue problem $\det(\lambda A_1 + A_2) = 0$. However, system (1.20) with singular A_1 cannot be elliptic, as can be seen from the following (use (1.22) with $\xi_1 = 1$ and $\xi_2 = 0$).

A generalisation of (1.20) to n independent variables is exhibited in the system

$$A_1 u_{x_1} + A_2 u_{x_2} + \ldots + A_n u_{x_n} + Bu = f \tag{1.21}$$

with $m \times m$ matrices $A_i = A_i(\mathbf{x}) = A_i(x_1, \ldots, x_n)$ and $B = B(\mathbf{x})$. As a special case of a later definition (cf. §12.1) we obtain the following definition.

Definition 1.18. The system (1.21) is said to be *elliptic* at $\mathbf{x}$ if

$$\det\left(\sum_{i=1}^{n} \xi_i A_i(\mathbf{x})\right) \neq 0 \qquad \text{for all } 0 \neq (\xi_1, \ldots, \xi_n) \in \mathbb{R}^n. \tag{1.22}$$

1.4 Characteristic Properties of the Different Types

The distinguishing of different types of partial differential equations would be pointless if each type did not have fundamentally different properties. When discussing the examples in Section 1.1 we already mentioned that the solution is uniquely determined if initial values and boundary values are prescribed.

In Example 1.2 the hyperbolic differential equation (1.2) is augmented by the specification (1.4) of u on the line $y = \text{const}$ (see Figure 1.1a). In the case of the hyperbolic wave equation (1.6), u_y must also be prescribed (cf. (1.8)) since the equation is of second order.

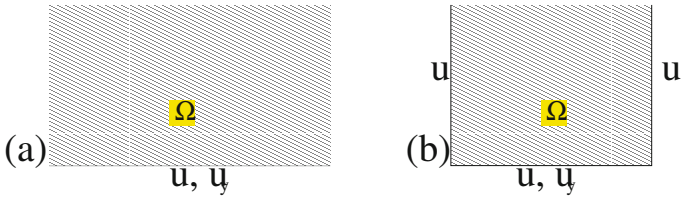


Fig. 1.1 Specification of (a) initial values and (b) initial-boundary values for hyperbolic problems.

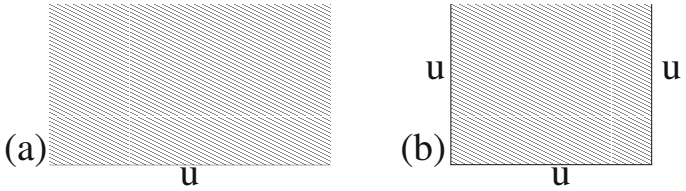


Fig. 1.2 Specification of (a) initial values and (b) initial-boundary values for parabolic problems.

It is also sufficient to give the values u and u_y on a *finite* interval $[x_1, x_2]$ if u is additionally prescribed on the lateral boundaries of the domain Ω of Figure 1.1b. This prescription of *initial-boundary values* occurs, for example, in the following physical problem. A vibrating string is described by the lateral deflection $u(x,t)$ at the point $x \in [x_1, x_2]$ at time t. The function u satisfies the wave equation (1.6) with the coordinate y corresponding to time t. At the initial instant in time, $t = t_0$, the deflection $u(x,0)$ and velocity $u_t(x,0)$ are given for $x_1 < x < x_2$. Under the assumption that the string is firmly clamped at the boundary points x_1 and x_2, one obtains the additional boundary data $u(x_1,t) = u(x_2,t) = 0$ for all t.

For parabolic equations of second order one can also formulate initial-value and initial-boundary value problems (cf. Figure 1.2). However, as initial value only the function $u(x,y_0) = u_0(x)$ may be prescribed. An additional specification of $u_y(x,y_0)$ is not possible, since $u_y(x,y_0) = u_{xx}(x,y_0) = u_0''(x)$ is already determined by the differential equation (1.9) and by u_0.

The heat equation (1.9) with the initial and boundary values

$$u(x,t_0) = u_0(x) \qquad \text{in } [x_1,x_2],$$
$$u(x_1,t) = \varphi_1(t),\ u(x_2,t) = \varphi_2(t) \qquad \text{for } t > t_0$$

(cf. Figure 1.2b) describes the temperature $u(x,t)$ of a wire whose ends at $x = x_1$ and $x = x_2$ have the temperatures $\varphi_1(t)$ and $\varphi_2(t)$. The initial temperature distribution at time t_0 is given by $u_0(x)$.

Aside from the different number of initial data functions in Figures 1.1 and 1.2, there also is the following difference between hyperbolic and parabolic equations.

Remark 1.19. The shaded area Ω in Figures 1.1 and 1.2 corresponds to $t > t_0$ and $y > y_0$, respectively. For hyperbolic equations one can solve in the same way initial-value and initial-boundary value problems in the domain $t \le t_0$, whilst parabolic problems in $t < t_0$ generally do not have a solution.

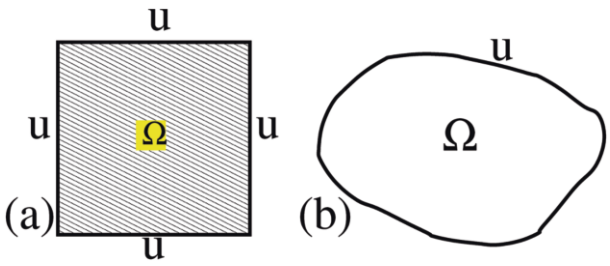


Fig. 1.3 Boundary values for an elliptic problem.

If one changes the parabolic equation $u_t - u_{xx} = 0$ to $u_t + u_{xx} = 0$, the orientation is reversed: solutions exist in general only for $t \le t_0$.

For the solution of an elliptic equation, boundary values are prescribed (cf. Example 1.3, Figure 1.3). A specification such as that in Figure 1.2b would not uniquely determine the solution of an elliptic problem, while the solution of a parabolic problem would be overdetermined by the boundary values of Figure 1.3a.

An elliptic problem with specifications such as in Figure 1.1b in general has no solution. Let us, e.g., impose the conditions $u(x,0) = u(0,y) = u(1,y) = 0$ and $u_y(x,0) = u_1(x)$ on the solution of the potential equation (1.5), where u_1 is not infinitely often differentiable. If a continuous solution u existed in $\overline{\Omega} = [0,1]\times[0,1]$ one could develop $u(x,1)$ into a sine series and the following exercise shows that u_1 would have to be infinitely often differentiable, in contradiction to the assumption.

Exercise 1.20. Let $\varphi \in C^0[0,1]$ have the absolute convergent Fourier expansion

$$\varphi(x) = \sum\nolimits_{\nu=1}^{\infty} \alpha_\nu \sin(\nu\pi x).$$

Show that: (a) the solution of the potential equation (1.5) in the square $\Omega = (0,1)\times(0,1)$ with boundary values $u(0,y) = u(x,0) = u(1,y) = 0$ and

$u(x,1) = \varphi(x)$ is given by

$$u(x,y) = \sum_{\nu=1}^{\infty} \frac{\alpha_\nu}{\sinh(\nu\pi)} \sin(\nu\pi x) \sinh(\nu\pi y).$$

(b) For $0 \le x \le 1$ and $0 \le y < 1$, $u(x,y)$ is differentiable infinitely often. *Hint:*

$$f(x) = \sum_{\nu} \beta_\nu \sin(\nu\pi x) \in C^\infty[0,1] \quad \text{if } \lim_{\nu\to\infty} \beta_\nu \nu^k = 0 \text{ for all } k \in \mathbb{N}\,.$$

Conversely it does not make sense to put boundary value constraints as in Figure 1.3a on a hyperbolic problem. Consider as an example the wave equation (1.6) in $\overline{\Omega} = [0,1]$ with the boundary values

$$u(x,0) = u(0,y) = u(1,y) = 0 \text{ and } u\big(x, \frac{1}{\pi}\big) = \sin(\nu\pi x) \text{ for } \nu \in \mathbb{N}.$$

The solution reads $u(x,y) = \sin(\nu\pi x)\sin(\nu\pi y)/\sin\nu$. Although the boundary values, for all $\nu \in \mathbb{N}$ are bounded by l, the solution in $\overline{\Omega}$ may become arbitrarily large since $\sup\{1/\sin\nu : \nu \in \mathbb{N}\} = \infty$ (cf. Exercise 1.21). Such a boundary-value problem is called *not well-posed* or *ill-posed* (cf. Definition 2.25).

Exercise 1.21. Prove that the set $\{\sin\nu : \nu \in \mathbb{N}\}$ is dense in $[-1,1]$.

Another distinguishing characteristic is the regularity (smoothness) of the solution. Let u be the solution of the potential equation (1.5) in $\Omega \subset \mathbb{R}^2$. As stated in Example 1.3, u is the real part of a function holomorphic in Ω. Since holomorphic functions are infinitely differentiable, this property also holds for u.

In the case of the parabolic heat equation (1.9) with initial values $u(x,0) = u_0$ the solution u is given by (1.10). For $y > 0$, u is infinitely differentiable. The smoothness of u_0 is of no concern here, nor is the smoothness of the boundary values in the case of the potential equation.

One finds a completely different result for the hyperbolic wave equation (1.6). The solution reads $u(x,y) = \phi(x+y) + \psi(x-y)$, where ϕ and ψ result directly from the initial data (1.8). Check that u is k-times differentiable if u_0 is k-times and u_1 is $(k-1)$-times differentiable.

As already mentioned in this section, in the hyperbolic and parabolic equations (1.1), (1.2), (1.6), and (1.9) the variable y often plays the role of time. Therefore one calls processes described by hyperbolic and parabolic equations *nonstationary*. Elliptic equations which only contain space coordinates as variables are called *stationary*. More clearly than in Definitions 1.12b,c the Definitions 1.14b,c distinguish the role of a single variable (time) corresponding to the eigenvalue $\lambda = 0$ in parabolic equations, and to the eigenvalue with opposite sign in hyperbolic equations.

The connection between the different types becomes more comprehensible if one relates the elliptic equations in the variables $x_1, \ldots, x_n$ to the parabolic and hyperbolic equations in the variables $x_1, \ldots, x_n, t$.

Remark 1.22. Let L be a differential operator in the variables $\mathbf{x} = (x_1, \ldots, x_n)$ and let it be of elliptic type (cf. (1.18)). Let L be scaled such that the matrix $A(\mathbf{x})$ in (1.17) has only negative eigenvalues. Then

$$u_t + Lu = 0 \tag{1.23}$$

is a parabolic equation for $u(\mathbf{x}, t) = u(x_1, \ldots, x_n, t)$. In contrast

$$u_{tt} + Lu = 0 \tag{1.24}$$

is of hyperbolic type.

Conversely, the nonstationary problems (1.23) or (1.24) lead to the elliptic equation $Lu = 0$ if one seeks solutions of (1.23) or (1.24) that are independent of time t. One also obtains elliptic equations if one looks for solutions of (1.23) or (1.24) with the aid of a *separation of variables* $u(\mathbf{x}, t) = \varphi(t)v(\mathbf{x})$. The results are

$$\begin{aligned} u(\mathbf{x}, t) &= \mathrm{e}^{-\lambda t} v(\mathbf{x}) && \text{in case (1.23),} \\ u(\mathbf{x}, t) &= \mathrm{e}^{\pm \mathrm{i}\sqrt{\lambda} t} v(\mathbf{x}) && \text{in case (1.24),} \end{aligned}$$

where $v(\mathbf{x})$ is the solution of the elliptic *eigenvalue problem*

$$L\, v = \lambda\, v\,.$$

These eigenvalue problems will be treated in Chapter 11.

1.5 Literature

Finally, we mention some monographs on partial differential equations of the different types.[4] Here the emphasis can be the theory as well as the numerical treatment. Hyperbolic equations are discussed by Alinhac [5], Meister–Struckmeier [203], and Trangenstein [290], parabolic equations are treated by Friedman [105].

A common characteristic of parabolic and hyperbolic types are initial-boundary value problems. On the theoretical side, this fact leads to semigroups. Both types are discussed by Lions–Magenes [195, 196], Pazy [218], and Richtmyer–Morton [237].

Elliptic and parabolic equation are the subject of Trangenstein [291]. These two types have higher regularity properties in common.

Elliptic problems are the subject of Bartels [29], Boccardo–Croce [43], Braess [45], Dupaigne [92], Gilbarg–Trudinger [115], Grivard [123], Ladyženskaja–Ural'ceva [180], Lions–Magenes [194], Miranda [205], and Wienholtz et al. [306]. In the next chapters, Hellwig [150] will often be cited.

All three types can be found in the monographs Dziuk [93], John [160], Jost [162], Knabner–Angermann [172], Larsson–Thomée [183], Mizohata [206], Petrovsky [225], and Wloka [308].

[4] This list is rather incomplete, since there exists an extensive literature on this subject. Further references to particular elliptic topics will follow in the text.

Chapter 2
The Potential Equation

Abstract In **Section 2.1** the simplest but prototypical elliptic differential equation of second order is presented. The solutions of this equation are called harmonic. Together with a boundary condition, one obtains a boundary-value problem. An important tool is the singularity function, which is defined in **Section 2.2**. The Green formulae allow a representation of the solution in Theorem 2.8. In **Section 2.3** functions with mean-value property are introduced. It is shown that these functions coincide with harmonic functions. The mean-value property implies the maximum-minimum principle: non-constant functions have no local extrema. An important conclusion is the uniqueness of the solution (Theorem 2.18). Finally, in **Section 2.4**, it is shown that the solution depends continuously on the boundary data.

2.1 Posing the Problem

The potential equation from Example 1.3 reads[1]

$$-\Delta u = 0 \qquad \text{in } \Omega \subset \mathbb{R}^n, \tag{2.1a}$$

where $\Delta = \partial^2/\partial x_1^2 + \ldots + \partial^2/\partial x_n^2$ is the *Laplace operator*. In physics, equation (2.1a) describes the potentials; for example the electric potential when Ω contains no electric charges, the magnetic potential for vanishing current density, the velocity potential, etc. In these cases the gradient ∇u has the more direct physical meaning.

Equation (2.1a) is also called *Laplace's equation* since it was described by Pierre-Simon Laplace in his five-volume work 'Traité de Mécanique Céleste' (written in the years 1799–1825). However, it was Leonhard Euler [97] who first mentioned the potential equation in 1751.

[1] At this place, the minus sign in $-\Delta u = 0$ only has a symbolic meaning. The background is that $-\Delta$ has 'nicer' properties than Δ, since, in a multiple sense, $-\Delta$ is positive. For instance, the singularity function in (2.4a) will have a positive singularity.

W. Hackbusch, *Elliptic Differential Equations*, Springer Series
in Computational Mathematics 18, DOI 10.1007/978-3-662-54961-2_2

The connection between the potential equation for $n = 2$ and function theory has already been pointed out in Example 1.3. Not only is the Laplace operator an example of an elliptic differential operator, but it actually is the prototype. By using a transformation of variables, any elliptic differential operator of second order can be transformed so that its principal part becomes the Laplace operator (cf. Hellwig [150, Part 2, §1.5]).

Definition 2.1 (domain). The region $\Omega \subset \mathbb{R}^n$ is called a *domain* if Ω is open and connected.[2]

In the following Ω will always be a domain. Its boundary is denoted by

$$\Gamma = \partial\Omega\,.$$

The existence of a second derivative of u is required only in Ω, not on the boundary. For a prescription of *boundary values*

$$u = \varphi \quad \text{on } \Gamma \tag{2.1b}$$

to be meaningful, one has to assume continuity of u on $\overline{\Omega} = \Omega \cup \Gamma$. The combination of an (elliptic) differential equation (here (2.1a)) with a boundary condition (here (2.1b)) is called a *boundary-value problem*.

Definition 2.2 (harmonic). The function u is said to be *harmonic* in Ω if u belongs to $C^2(\Omega) \cap C^0(\overline{\Omega})$ and satisfies the potential equation (2.1a).

Here $C^0(D)$ $[C^k(D),\ C^\infty(D)]$ denotes the set of continuous [k-fold continuously differentiable, infinitely often differentiable] functions on D.

In general one should not expect that the solution of (2.1a,b) lies in $C^2(\overline{\Omega})$, as shown in the following example.

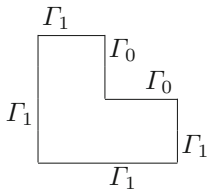


Fig. 2.1 L-shaped domain.

Example 2.3. Let $\Omega = (0,1) \times (0,1)$ (cf. Figure 1.3a). Let the boundary values be given by $\varphi(x,y) = x^2$ for $(x,y) \in \Gamma$. The solution of the boundary-value problem exists but does not belong to $C^2(\overline{\Omega})$.

Proof. The existence of a solution u will be discussed in Theorem 7.21. If $u \in C^2(\overline{\Omega})$, then it follows that $u_{xx}(x,0) = \varphi_{xx}(x,0) = 2$ for $x \in [0,1]$ in particular $u_{xx}(0,0) = 2$. From the analogous result $u_{yy}(0,0) = \varphi_{yy}(0,0) = 0$ one may conclude $\Delta u(0,0) = 2$ in contradiction to $\Delta u = 0$ in Ω. ∎

In the case under discussion one can also show $u \in C^1(\overline{\Omega})$. That this statement is generally false, is shown by the next example with the *L-shaped domain* in Figure 2.1.

[2] Ω is called *connected* if for any $x, y \in \Omega$ there is a continuous curve within Ω connecting x and y, that is, a $\gamma : s \in [0,1] \mapsto \gamma(s) \in \Omega$ continuously with $\gamma(0) = x$, $\gamma(1) = y$.

Example 2.4. In the L-shaped domain $\Omega = (-\frac{1}{2}, \frac{1}{2}) \times (-\frac{1}{2}, \frac{1}{2}) \setminus [0, \frac{1}{2}) \times [0, \frac{1}{2})$ (cf. Figure 2.1) introduce the polar coordinates

$$x = r\cos\varphi, \qquad y = r\sin\varphi. \tag{2.2}$$

The function

$$u(r, \varphi) = r^{2/3}\sin((2\varphi - \pi)/3) \qquad (\pi/2 < \varphi < 2\pi)$$

is the solution of the potential equation (2.1a) and has smooth boundary values on Γ (in particular $u = 0$ holds on $\Gamma_0 \subset \Gamma$) although the first derivatives in $r = 0$ are unbounded, i.e., $u \notin C^1(\overline{\Omega})$.

Proof. This follows from the fact that, along with u_x and u_y, the radial derivative $u_r = u_x\cos\varphi + u_y\sin\varphi$ also has to be bounded. However, $u_r = \mathcal{O}(r^{-1/3})$ holds for $r \to 0$. In order to check that $-\Delta u = 0$, use (2.3a). ∎

Exercise 2.5. Prove the following:
(a) In terms of the polar coordinates (2.2) in $\mathbb{R}^2$, the Laplace operator takes the form

$$\Delta = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \varphi^2}. \tag{2.3a}$$

(b) In terms of the three-dimensional polar coordinates

$$x = r\,\cos\varphi\,\sin\psi, \qquad y = r\,\sin\varphi\,\sin\psi, \qquad z = r\,\cos\psi,$$

the Laplace operator is given by

$$\Delta = \frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\left[\frac{1}{\sin^2\psi}\frac{\partial^2}{\partial\varphi^2} + \cot\psi\frac{\partial}{\partial\psi} + \frac{\partial^2}{\partial\psi^2}\right].$$

In the general n-dimensional case the transformation to polar coordinates leads to

$$\Delta = \frac{\partial^2}{\partial r^2} + \frac{n-1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}B, \tag{2.3b}$$

where the *Beltrami operator* B contains only derivatives with respect to the angle variables.

The Laplace operator is invariant with respect to translation, reflection and rotation.

Exercise 2.6. Let $f \in C^2(\Omega)$ for a domain $\Omega \subset \mathbb{R}^n$, $\mathbf{z} \in \mathbb{R}^n$, and $U \in \mathbb{R}^{n\times n}$ a unitary matrix (i.e., $UU^{\mathsf{T}} = I$). Then $\Phi(\mathbf{x}) := \mathbf{z} + U\mathbf{x}$ describes a combination of translation, reflection, and rotation. Φ maps Ω into $\Omega' := \{\mathbf{y} = \mathbf{z} + U\mathbf{x} : \mathbf{x} \in \Omega\}$. The inverse function is $\Phi^{-1}(\mathbf{y}) = U^{\mathsf{T}}(\mathbf{y} - \mathbf{z})$. The function $F(\mathbf{y}) := f(\Phi^{-1}(\mathbf{y}))$ is defined in Ω'. Prove that $\Delta F(\mathbf{y}) = (\Delta f)(\Phi^{-1}(\mathbf{y}))$.

In particular, it follows that harmonic functions remain harmonic after translation, reflection and rotation.

2.2 Singularity Function

The *singularity function* is defined by

$$s(\mathbf{x},\mathbf{y}) = \begin{cases} -\dfrac{1}{\omega_2}\log|\mathbf{x}-\mathbf{y}| & \text{for } n=2, \\ \dfrac{1}{(n-2)\,\omega_n}|\mathbf{x}-\mathbf{y}|^{2-n} & \text{for } n>2, \end{cases} \tag{2.4a}$$

for $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$, where

$$\begin{aligned} \omega_n &= 2\,\Gamma(\tfrac{1}{2})^n/\,\Gamma(\tfrac{n}{2}), \\ \text{in particular, } \omega_2 &= 2\pi,\ \ \omega_3 = 4\pi, \end{aligned} \tag{2.4b}$$

with Γ the Gamma function, is the surface of the n-dimensional unit sphere. The Euclidean norm of $\mathbf{x}$ in $\mathbb{R}^n$ is denoted by

$$|\mathbf{x}| = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Lemma 2.7. *For fixed* $\mathbf{y}\in\mathbb{R}^n$ *the potential equation in* $\mathbb{R}^n\backslash\{\mathbf{y}\}$ *is solved by* $s(\cdot,\mathbf{y})$.

Proof. The proof can be carried out directly. However, it is simplest to introduce polar coordinates with $\mathbf{y}$ as origin and to use (2.3b), since $s(\mathbf{x},\mathbf{y})$ depends only on $r=|\mathbf{x}-\mathbf{y}|$. ∎

We remark that, in the sense of distributions, $-\Delta_x s(\mathbf{x},\mathbf{y})$ is equal to the delta distribution $\delta(\mathbf{x},\mathbf{y})$. This equation fixes the strength of the singularity and the scaling constant.

For the next theorem we need to introduce the *normal derivative* $\partial/\partial n$. Let Ω be a domain with smooth boundary Γ. Let $\mathbf{n}(\mathbf{x})\in\mathbb{R}^n$ denote the outer normal direction at $\mathbf{x}\in\Gamma$, i.e., $\mathbf{n}$ is a unit vector perpendicular to the tangential hyperplane at $\mathbf{x}$ and points outwards (cf. Figure 2.2). The normal derivative of u at $\mathbf{x}\in\Gamma$ is defined as

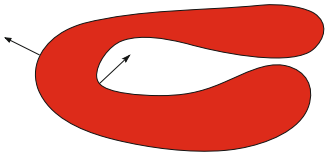


Fig. 2.2 Normal directions.

$$\frac{\partial u(\mathbf{x})}{\partial n} = \langle \mathbf{n}, \nabla u\rangle\,, \qquad \text{where } \nabla u = \operatorname{grad} u = (u_{x_1},\ldots,u_{x_n})^{\mathsf{T}}$$

is the *gradient* of u and

$$\langle \mathbf{x},\mathbf{y}\rangle = \sum_{i=1}^{n} x_i y_i$$

is the scalar product in $\mathbb{R}^n$. In the case of the ball $K_R(\mathbf{y})$ (cf. (2.8)) the normal direction is radial, and $\partial u/\partial n$ becomes $\partial u/\partial r$ with respect to $r=|\mathbf{x}-\mathbf{y}|$, if one

uses polar coordinates with the origin at $\mathbf{y}$. It follows from

$$\partial s(\mathbf{x},\mathbf{y})/\partial r = -|\mathbf{x}-\mathbf{y}|^{1-n}/\omega_n$$

that

$$\frac{\partial s(\mathbf{x},\mathbf{y})}{\partial n} = -\frac{R^{1-n}}{\omega_n} \qquad \text{for all } \mathbf{x} \in \partial K_R(\mathbf{y}). \tag{2.5}$$

The *first Green formula* reads (cf. Green [121])

$$\int_\Omega u(\mathbf{x})\,\Delta v(\mathbf{x})\,\mathrm{d}\mathbf{x} = -\int_\Omega \langle \nabla u(\mathbf{x}), \nabla v(\mathbf{x})\rangle\,\mathrm{d}\mathbf{x} + \int_{\partial\Omega} u(\mathbf{x})\frac{\partial v(\mathbf{x})}{\partial n}\mathrm{d}\Gamma \tag{2.6a}$$

and holds for $u \in C^1(\overline{\Omega})$, $v \in C^2(\overline{\Omega})$ if the domain Ω satisfies suitable conditions. Here $\int_{\partial\Omega} \ldots \mathrm{d}\Gamma$ denotes the surface integral.

Domains for which equation (2.6a) holds are called *normal domains*. To see sufficient conditions for this refer to Kellogg [169, §IV] and Hellwig [150, Part 1, §1.2].

Functions $u, v \in C^2(\overline{\Omega})$ in a normal domain Ω satisfy the *second Green formula*

$$\int_\Omega u(\mathbf{x})\,\Delta v(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_\Omega v(\mathbf{x})\,\Delta u(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int_{\partial\Omega}\left[u(\mathbf{x})\frac{\partial v(\mathbf{x})}{\partial n} - v(\mathbf{x})\frac{\partial u(\mathbf{x})}{\partial n}\right]\mathrm{d}\Gamma. \tag{2.6b}$$

Theorem 2.8. *Let Ω be a normal domain, and let $u \in C^2(\overline{\Omega})$ be harmonic there. Then*

$$u(\mathbf{y}) = \int_{\partial\Omega}\left[s(\mathbf{x},\mathbf{y})\frac{\partial u(\mathbf{x})}{\partial n} - u(\mathbf{x})\frac{\partial s(\mathbf{x},\mathbf{y})}{\partial n_x}\right]\mathrm{d}\Gamma_x \qquad \textit{for all } \mathbf{y} \in \Omega. \tag{2.7}$$

Here $\frac{\partial}{\partial n_x}$ and $\mathrm{d}\Gamma_x$ refer to the variable $\mathbf{x}$.

Proof. By

$$K_r(\mathbf{y}) = \{\mathbf{x} \in \mathbb{R}^n : |\mathbf{x}-\mathbf{y}| < r\} \tag{2.8}$$

we denote the ball with centre $\mathbf{y}$ and radius r. Since the singularity function $s(\cdot,\mathbf{y})$ is not differentiable on $\mathbf{x} = \mathbf{y}$, Green's formula (2.6b) is not directly applicable. Let

$$\Omega_\varepsilon := \Omega \setminus \overline{K_\varepsilon(\mathbf{y})},$$

with ε be so small that $K_\varepsilon(\mathbf{y}) \subset \Omega$. Since Ω_ε is again a normal domain, it follows from $-\Delta u = -\Delta s = 0$ in Ω_ε (cf. Lemma 2.7) and (2.6b) with $v = s(\cdot,\mathbf{y})$ that

$$\int_{\partial\Omega_\varepsilon}\left[u(\mathbf{x})\frac{\partial s(\mathbf{x},\mathbf{y})}{\partial n_x} - s(\mathbf{x},\mathbf{y})\frac{\partial u(\mathbf{x})}{\partial n}\right]\mathrm{d}\Gamma_x = 0. \tag{2.9a}$$

We have $\partial\Omega_\varepsilon = \partial\Omega \cup \partial K_\varepsilon(\mathbf{y})$. However, at $\mathbf{x} \in \partial K_\varepsilon(\mathbf{y})$, the normal directions of $\partial\Omega_\varepsilon$ and of $\partial K_\varepsilon(\mathbf{y})$ differ in their signs. The same holds for the normal derivatives, so that the integral in (2.9a) can be decomposed into

$$\int_{\partial\Omega_\varepsilon} \ldots = \int_{\partial\Omega} \ldots - \int_{\partial K_\varepsilon(\mathbf{y})} \ldots \,.$$

The assertion of the theorem would be proved if we could show that $\int_{\partial K_\varepsilon(\mathbf{y})} \ldots \to -u(\mathbf{y})$ for $\varepsilon \to 0$. The normal derivative $\partial u/\partial n$ is bounded on $\partial K_\varepsilon(\mathbf{y})$, and $\int_{\partial K_\varepsilon(\mathbf{y})} s(\mathbf{x},\mathbf{y})\mathrm{d}\Gamma$ converges like $\mathcal{O}(\varepsilon|\log\varepsilon|))$ or $\mathcal{O}(\varepsilon)$ towards zero, as can be seen from (2.4) and

$$\int_{\partial K_\varepsilon(\mathbf{y})} \mathrm{d}\Gamma = \varepsilon^{n-1}\omega_n. \tag{2.9b}$$

Thus, we obtain

$$\int_{\partial K_\varepsilon(\mathbf{y})} s(\mathbf{x},\mathbf{y})\frac{\partial u(\mathbf{x})}{\partial n}\mathrm{d}\Gamma_x \to 0 \qquad (\varepsilon \to 0). \tag{2.9c}$$

From (2.9b) and (2.5) one infers

$$\int_{\partial K_\varepsilon(\mathbf{y})} u(\mathbf{y})\frac{\partial s(\mathbf{x},\mathbf{y})}{\partial n_x}\,\mathrm{d}\Gamma_x = u(\mathbf{y})\int_{\partial K_\varepsilon(\mathbf{y})} \frac{\partial s(\mathbf{x},\mathbf{y})}{\partial n_x}\,\mathrm{d}\Gamma_x = -u(\mathbf{y}). \tag{2.9d}$$

The continuity of u in $\mathbf{y}$ yields

$$\left|\int_{\partial K_\varepsilon(\mathbf{y})} [u(\mathbf{x}) - u(\mathbf{y})]\,\frac{\partial s(\mathbf{x},\mathbf{y})}{\partial n_x}\mathrm{d}\Gamma_x\right| \le \max_{\mathbf{x}\in\partial K_\varepsilon(\mathbf{y})} |u(\mathbf{x}) - u(\mathbf{y})| \to 0 \tag{2.9e}$$

as $\varepsilon \to 0$. Equations (2.9c–e) show that $\int_{\partial K_\varepsilon(\mathbf{y})}[\,u\frac{\partial}{\partial n}s - s\frac{\partial}{\partial n}u\,]\mathrm{d}\Gamma \to -u(\mathbf{y})$ $(\varepsilon \to 0)$, so that (2.9a) proves the theorem. ∎

Any function of the form

$$\gamma(\mathbf{x},\mathbf{y}) = s(\mathbf{x},\mathbf{y}) + \Phi(\mathbf{x},\mathbf{y}) \tag{2.10}$$

is called a *fundamental solution* of the potential equation in Ω if for fixed $\mathbf{y} \in \Omega$ the function $\Phi(\cdot,\mathbf{y})$ is harmonic in Ω and belongs to $C^2(\overline{\Omega})$.

Corollary 2.9. Under the conditions of Theorem 2.8 for every fundamental solution in Ω the following holds:

$$u(\mathbf{y}) = \int_{\partial\Omega}\left[\gamma(\mathbf{x},\mathbf{y})\frac{\partial u(\mathbf{x})}{\partial n} - u(\mathbf{x})\frac{\partial\gamma(\mathbf{x},\mathbf{y})}{\partial n_x}\right]\mathrm{d}\Gamma_x \qquad (\mathbf{y}\in\Omega)\,. \tag{2.11}$$

Proof. (2.6b) implies $\int_{\partial\Omega}[\,\Phi\,\partial u/\partial n - u\,\partial\Phi/\partial n\,]\,\mathrm{d}\Gamma = 0\,.$ ∎

For replacing the condition $\Phi = \gamma - s \in C^2(\overline{\Omega})$ by the weaker $\Phi(\cdot,\mathbf{y}) \in C^1(\overline{\Omega}) \cap C^2(\Omega)$ refer to Hellwig [150, Part I, §1.4].

Exercise 2.10 (Green function of a ball). Let $\Omega = K_R(\mathbf{y})$. Define

$$\gamma(\mathbf{x},\boldsymbol{\xi}) = \begin{cases} \frac{1}{(n-2)\omega_n}\left[|\mathbf{x}-\boldsymbol{\xi}|^{2-n} - \left(\frac{|\boldsymbol{\xi}-\mathbf{y}|}{R}\,|\mathbf{x}-\boldsymbol{\xi}'|\right)^{2-n}\right] & \text{for } n \geq 3, \\ -\frac{1}{2\pi}\left[\log|\mathbf{x}-\boldsymbol{\xi}| - \log\left(\frac{|\boldsymbol{\xi}-\mathbf{y}|}{R}\,|\mathbf{x}-\boldsymbol{\xi}'|\right)\right] & \text{for } n = 2, \end{cases} \tag{2.12}$$

with $\mathbf{x}, \boldsymbol{\xi} \in \Omega$, $\boldsymbol{\xi}' = \mathbf{y} + R^2|\boldsymbol{\xi}-\mathbf{y}|^{-2}(\boldsymbol{\xi}-\mathbf{y})$ and show:
(a) $\gamma(\mathbf{x},\boldsymbol{\xi}) = 0$ for $\boldsymbol{\xi} \in \partial\Omega\backslash\{\mathbf{x}\}$,
(b) γ is a fundamental solution in Ω,
(c) $\gamma(\mathbf{x},\boldsymbol{\xi}) = \gamma(\boldsymbol{\xi},\mathbf{x})$,
(d) on the surface $\Gamma = \partial K_R(\mathbf{y})$ the following holds:

$$\frac{\partial}{\partial n_\xi}\gamma(\mathbf{x},\boldsymbol{\xi}) = \frac{\partial}{\partial n_\xi}\gamma(\boldsymbol{\xi},\mathbf{x}) = -\frac{1}{R\omega_n}\frac{R^2-|\mathbf{x}-\mathbf{y}|^2}{|\mathbf{x}-\boldsymbol{\xi}|^n} \qquad (\boldsymbol{\xi} \in \Gamma)\,.$$

2.3 Mean-Value Property and Maximum Principle

Definition 2.11 (mean-value property). A function u has the *mean-value property* in Ω if $u \in C^0(\overline{\Omega})$ and if for all $\mathbf{x} \in \Omega$ and all $R > 0$ with $K_R(\mathbf{x}) \subset \Omega$ the following equation holds:

$$u(\mathbf{x}) = \frac{1}{\omega_n R^{n-1}}\int_{\partial K_R(\mathbf{x})} u(\boldsymbol{\xi})\,\mathrm{d}\Gamma_\xi. \tag{2.13}$$

Since $\int_{\partial K_R(\mathbf{x})}\mathrm{d}\Gamma = \omega_n R^{n-1}$ the right-hand side in (2.13) is the mean value of u taken over the sphere $\partial K_R(\mathbf{x})$. An equivalent characterisation results if one averages over the ball $K_R(\mathbf{x})$.

Exercise 2.12. $u \in C^0(\overline{\Omega})$ has the *second mean-value property* in Ω if

$$u(\mathbf{x}) = \frac{n}{R^n\omega_n}\int_{K_R(\mathbf{x})} u(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} \quad \text{for all } \mathbf{x} \in \Omega,\ R > 0 \qquad \text{with } K_R(\mathbf{x}) \subset \Omega.$$

Show that this mean-value property is equivalent to the mean-value property (2.13).
Hint:

$$\int_{K_R(\mathbf{x})} u(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} = \int_0^R \left(\int_{\partial K_r(\mathbf{x})} u(\boldsymbol{\xi})\,\mathrm{d}\Gamma_\xi\right)\mathrm{d}r\,. \tag{2.14}$$

Functions with the mean-value property satisfy a maximum principle, as is known from the function theory for holomorphic functions.

Theorem 2.13 (maximum-minimum principle). *Let Ω be a domain and let $u \in C^0(\Omega)$ be a nonconstant function which has the mean-value property. Then u takes on neither a maximum nor a minimum in Ω.*

Proof. (i) It suffices to investigate the case of a maximum since a minimum of u is a maximum of $-u$, and $-u$ also has the mean-value property.

(ii) For an indirect proof we assume that there exists a maximum at $\mathbf{y} \in \Omega$:

$$u(\mathbf{y}) = M \geq u(\mathbf{x}) \qquad \text{for all } \mathbf{x} \in \Omega\,.$$

We want to show that $u(\mathbf{y}') = M$ for arbitrary $\mathbf{y}' \in \Omega$, i.e., $u \equiv M$ in contrast to the assumption $u \not\equiv \text{const}$. Let $\mathbf{y}' \in \Omega$. Since Ω is connected, there exists a path connecting $\mathbf{y}$ and $\mathbf{y}'$ running through Ω, i.e., there exists a continuous $\varphi : [0,1] \to \Omega$ with $\varphi(0) = \mathbf{y}$ and $\varphi(1) = \mathbf{y}'$. We set

$$I := \{s \in [0,1] : u(\varphi(t)) = M \text{ for all } 0 \leq t \leq s\}.$$

I contains at least 0, and is closed since u and φ are continuous. Thus there exists $s^\star = \max\{s \in I\}$, and the definition of I shows that $I = [0, s^\star]$. In (iii) it is proved that $s^\star = 1$ so that $\mathbf{y}' = \varphi(1) \in I$ and hence $u(\mathbf{y}') = M$ follows.

(iii) Proof of $s^\star = 1$. The opposite assumption $s^\star < 1$ can be made and shown to be contradicted by proving that $u(\mathbf{x}) = M$ in a neighbourhood of $\mathbf{x}^\star := \varphi(s^\star)$. Since $\mathbf{x}^\star \in \Omega$, there exists $R > 0$ with $K_R(\mathbf{x}^\star) \subset \Omega$. Evidently, it follows that $u = M$ in $K_R(\mathbf{x}^\star)$ if it is shown that $u = M$ on $\partial K_r(\mathbf{x}^\star)$ for all $0 < r \leq R$.

(iv) Proof of $u = M$ on $\partial K_r(\mathbf{x}^\star)$. Equation (2.13) in $\mathbf{x}^\star$ reads

$$M = u(\mathbf{x}^\star) = \frac{1}{\omega_n r^{n-1}} \int_{\partial K_r(\mathbf{x}^*)} u(\boldsymbol{\xi})\, \mathrm{d}\Gamma_\xi\,.$$

In general we have $u(\boldsymbol{\xi}) \leq M$. If one had $u(\boldsymbol{\xi}') < M$ for $\boldsymbol{\xi}' \in \partial K_R(\mathbf{x}^\star)$ and thus also $u < M$ in a neighbourhood of $\boldsymbol{\xi}'$, one would have on the right-hand side a mean value smaller than M. Therefore, $u = M$ on $\partial K_R(\mathbf{x}^\star)$ has been proved. ∎

Simple deductions from Theorem 2.13 are contained in the next statements.

Corollary 2.14. Let Ω be bounded. (a) A function with the mean-value property takes its maximum and its minimum on $\partial\Omega$.

(b) If two functions with the mean-value property coincide on the boundary $\partial\Omega$, they are identical.

Proof. (a) The extrema are assumed on the compact set $\overline{\Omega} = \Omega \cup \partial\Omega$. According to Theorem 2.13, the extremum cannot be in Ω if u is not constant on a connected component of Ω. But in this case the assertion is also obvious.

(b) If u and v with $u = v$ on $\partial\Omega$ satisfy the mean-value property then the latter is also satisfied for $w := u - v$. Since $w = 0$ on $\partial\Omega$, part (a) indicates that $\max w = \min w = 0$. Thus $u = v$ in Ω. ∎

Lemma 2.15. *Harmonic functions have the mean-value property.*

Proof. Let u be harmonic in Ω and $\mathbf{y} \in K_R(\mathbf{y}) \subset \Omega$. We apply the representation (2.7) for $K_R(\mathbf{y})$. The value $s(\mathbf{x}, \mathbf{y})$ is constant on $\partial K_R(\mathbf{y})$: let it be denoted by $\sigma(R)$. Because of (2.5), equation (2.7) becomes

$$u(\mathbf{y}) = \sigma(R) \int_{\partial K_R(\mathbf{y})} \frac{\partial u(\boldsymbol{\xi})}{\partial n} \mathrm{d}\Gamma + \frac{1}{\omega_n R^{n-1}} \int_{\partial K_R(\mathbf{y})} u(\boldsymbol{\xi}) \mathrm{d}\Gamma.$$

The equation agrees with (2.13) if the first integral vanishes. The latter follows from the next lemma. ∎

Lemma 2.16. *Let $u \in C^2(\Omega)$ be harmonic in a normal domain Ω. Then*

$$\int_{\partial\Omega} \frac{\partial u}{\partial n} \mathrm{d}\Gamma = 0.$$

Proof. In Green's formula (2.6a) substitute 1 and u for u and v, respectively. ∎

Lemma 2.15, Theorem 2.13, and Corollary 2.14 together imply Theorems 2.17 and 2.18.

Theorem 2.17 (maximum-minimum principle for harmonic functions). *Let u be harmonic in the domain Ω and nonconstant. There exists no maximum and no minimum in Ω.*

Theorem 2.18 (uniqueness). *Let Ω be bounded. A function harmonic in Ω assumes its maximum and its minimum on $\partial\Omega$ and is uniquely determined by its values on $\partial\Omega$.*

Exercise 2.19. Let Ω be bounded, and let u_1 and u_2 be harmonic in Ω with boundary values φ_1 and φ_2 on $\Gamma = \partial\Omega$. Prove that:
(a) $\varphi_1 \leq \varphi_2$ on Γ implies $u_1 \leq u_2$ in Ω.
(b) If, furthermore, Ω is connected and if $\varphi_1(\mathbf{x}) < \varphi_2(\mathbf{x})$ holds for at least one point $\mathbf{x} \in \Gamma$ then it follows that $u_1 < u_2$ everywhere in Ω.

The representation (2.13) of $u(\mathbf{y})$ by the values on $\partial K_R(\mathbf{y})$ is a special case of the following formula which will be proved on page 22 and which provides equation (2.13) for $\mathbf{x} = \mathbf{y}$.

Theorem 2.20 (Poisson's integral formula). *Assume we have $\varphi \in C^0(\partial K_R(\mathbf{y}))$ and $n \geq 2$. The solution of the boundary-value problem*

$$-\Delta u = 0 \quad \text{in } K_R(\mathbf{y}), \qquad u = \varphi \quad \text{on } \partial K_R(\mathbf{y}),$$

is given by the function

$$u(\mathbf{x}) = \frac{R^2 - |\mathbf{x} - \mathbf{y}|^2}{R\omega_n} \int_{\partial K_R(\mathbf{y})} \frac{\varphi(\boldsymbol{\xi})}{|\mathbf{x} - \boldsymbol{\xi}|^n} \, \mathrm{d}\Gamma_\xi \qquad \text{for } \mathbf{x} \in K_R(\mathbf{y}), \tag{2.15}$$

which belongs to $C^\infty(K_R(\mathbf{y})) \cap C^0(\overline{K_R(\mathbf{y})})$.

The mean-value property only assumes $u \in C^0(\overline{\Omega})$, while harmonic functions belong to $C^2(\Omega)$. This makes the following assertion surprising.

Theorem 2.21. *A function is harmonic in Ω if and only if it has the mean-value property there.*

Proof. Because of Lemma 2.15 it remains to be shown that a function u with the mean-value property is harmonic. Let $\mathbf{x} \in K_R(\mathbf{x}) \subset \Omega$ be given arbitrarily. According to Theorem 2.20 there exists a function u harmonic in $K_R(\mathbf{x})$ with

$$-\Delta u = 0 \quad \text{in } K_R(\mathbf{x}), \qquad u = v \quad \text{on } \partial K_R(\mathbf{x}).$$

According to Lemma 2.15, u has the mean-value property, as does v, and Corollary 2.14b proves that $u = v$ in $K_R(\mathbf{x})$, i.e., v is harmonic in $K_R(\mathbf{x})$. Since $K_R(\mathbf{x}) \subset \Omega$ is arbitrary, v is harmonic in Ω. ∎

An important application of Theorem 2.21 is the following statement of Harnack [149, §20].

Theorem 2.22. *Let $u_1, u_2, \ldots$ be a sequence of functions harmonic in Ω and converging uniformly on Γ. Then $u = \lim_{k\to\infty} u_k$ is harmonic in Ω.*

Proof. Uniform convergence on Γ implies uniform convergence in $\overline{\Omega}$ (cf. Theorem 2.17). Therefore the limit function is continuous: $u \in C^0(\overline{\Omega})$. The limit process, applied to

$$u_k(\mathbf{x}) = \frac{1}{\omega_n R^{n-1}} \int_{\partial K_R(\mathbf{x})} u_k(\boldsymbol{\xi})\, \mathrm{d}\Gamma_\xi$$

yields equation (2.13) for u; i.e., u inherits the mean-value property. According to Theorem 2.21, u is also harmonic in Ω. ∎

In the case of $n = 2$, Example 1.3 shows the connection with holomorphic functions. In fact, u is analytic for all n, i.e., a convergent power series expansion exists in a neighbourhood of any $\mathbf{x} \in \Omega$. The following theorem holds whose proof can be found, for example, in Hellwig [150, Part 3, §1.5].

Theorem 2.23. *A function harmonic in Ω is analytic there.*

The proof of the Poisson formula (2.15) still needs to be carried out.

Proof of Theorem 2.20. (a) First we must show that u in (2.15) is a function harmonic in $K_R(\mathbf{y})$, i.e., it satisfies $-\Delta u = 0$. Since the integrand is twice continuously differentiable and the domain of integration $\Gamma = \partial K_R(\mathbf{y})$ is compact, the Laplace operator commutes with the integral sign:

$$\Delta u(\mathbf{x}) = \frac{1}{R\,\omega_n} \int_\Gamma \varphi(\boldsymbol{\xi})\, \Delta_{\mathbf{x}} \frac{R^2 - |\mathbf{x}-\mathbf{y}|^2}{|\mathbf{x}-\boldsymbol{\xi}|^{-n}}\, \mathrm{d}\Gamma_\xi \qquad \text{for } \mathbf{x} \in K_R(\mathbf{y}). \tag{2.16}$$

According to Exercise 2.10 there exists a fundamental solution $\gamma(\mathbf{x}, \boldsymbol{\xi})$ such that

$$\frac{R^2 - |\mathbf{x}-\mathbf{y}|^2}{|\mathbf{x}-\boldsymbol{\xi}|^n \, R\,\omega_n} = -\frac{\partial \gamma(\mathbf{x}, \boldsymbol{\xi})}{\partial n_\xi} = -\frac{\partial \gamma(\boldsymbol{\xi}, \mathbf{x})}{\partial n_\xi} \qquad \text{for } \boldsymbol{\xi} \in \Gamma,\ \mathbf{x} \in K_R(\mathbf{y}). \quad (2.17)$$

From

$$\Delta_x \frac{\partial \gamma}{\partial n_\xi} = \frac{\partial}{\partial n_\xi} \Delta_x \gamma(\mathbf{x}, \boldsymbol{\xi}) = 0$$

and (2.16) one infers that $-\Delta u = 0$.

(b) The expression (2.15) defines the value $u(\mathbf{x})$ at first only for $\mathbf{x} \in K_R(\mathbf{y})$. It still needs to be shown that u has a continuous extension on $\overline{K_R(\mathbf{y})} = K_R(\mathbf{y}) \cup \Gamma$ (i.e., $u \in C^0(\overline{K_R(\mathbf{y})})$) and that the continuously extended values agree with the boundary values φ:

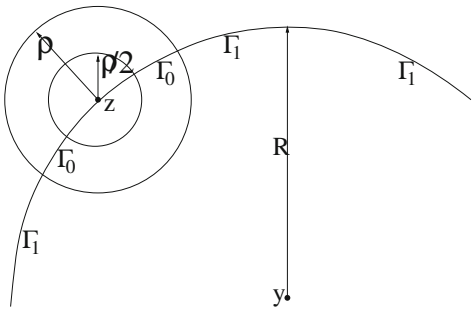


Fig. 2.3 Γ and the balls $K_\rho(z), K_{\rho/2}(z)$.

$$\lim_{K_R(\mathbf{y}) \ni \mathbf{x} \to \mathbf{z}} u(\mathbf{x}) = \varphi(\mathbf{z}) \qquad \text{for } \mathbf{z} \in \Gamma. \quad (2.18)$$

By equation (2.17), putting $u \equiv 1$ in Corollary 2.9 gives the identity

$$\frac{R^2 - |\mathbf{x}-\mathbf{y}|^2}{R\omega_n} \int_\Gamma \frac{\mathrm{d}\Gamma_\xi}{|\mathbf{x}-\boldsymbol{\xi}|^n} = 1 \qquad \text{for } \mathbf{x} \in K_R(\mathbf{y}). \quad (2.19)$$

Let $\mathbf{z} \in \Gamma$ be arbitrary. Due to equation (2.19) one can then write:

$$u(\mathbf{x}) - \varphi(\mathbf{z}) = \frac{R^2 - |\mathbf{x}-\mathbf{y}|^2}{R\omega_n} \int_\Gamma \frac{u(\boldsymbol{\xi}) - \varphi(\mathbf{z})}{|\mathbf{x}-\boldsymbol{\xi}|^n} \, \mathrm{d}\Gamma_\xi. \quad (2.20a)$$

We define $\Gamma_0 = \Gamma \cap K_\rho(\mathbf{z})$, $\Gamma_1 = \Gamma / \Gamma_0$ (see Figure 2.3) and split the expression (2.20a) into $u(\mathbf{x}) - \varphi(\mathbf{z}) = I_0 + I_1$, where

$$I_i = \frac{R^2 - |\mathbf{x}-\mathbf{y}|^2}{R\omega_n} \int_{\Gamma_i} \frac{u(\boldsymbol{\xi}) - \varphi(\mathbf{z})}{|\mathbf{x}-\boldsymbol{\xi}|^n} \, \mathrm{d}\Gamma_\xi \qquad \text{for } i = 0, 1.$$

Since

$$\left| \int_{\Gamma_0} \frac{u(\boldsymbol{\xi}) - \varphi(\mathbf{z})}{|\mathbf{x}-\boldsymbol{\xi}|^n} \, \mathrm{d}\Gamma_\xi \right| \le \max_{\boldsymbol{\xi} \in \Gamma_0} |u(\boldsymbol{\xi}) - \varphi(\mathbf{z})| \int_{\Gamma_0} \frac{\mathrm{d}\Gamma_\xi}{|\mathbf{x}-\boldsymbol{\xi}|^n}$$
$$\le \max_{\boldsymbol{\xi} \in \Gamma_0} |u(\boldsymbol{\xi}) - \varphi(\mathbf{z})| \int_{\Gamma} \frac{\mathrm{d}\Gamma_\xi}{|\mathbf{x}-\boldsymbol{\xi}|^n},$$

it follows from equation (2.19) that

$$I_0 \le \max_{\boldsymbol{\xi} \in \Gamma_0} |u(\boldsymbol{\xi}) - \varphi(\mathbf{z})| \,. \quad (2.20b)$$

Because of the continuity of φ one can choose $\rho > 0$ such that for given $\varepsilon > 0$

$$I_0 \le \varepsilon/2\,. \tag{2.20c}$$

Set $C_\varphi := \max_{\boldsymbol{\xi}\in\Gamma} |\varphi(\boldsymbol{\xi})|$ and choose $\mathbf{x} \in K_R(\mathbf{y})$ sufficiently close to $\mathbf{z}$ such that

$$|\mathbf{x}-\mathbf{z}| \quad \le \quad \delta(\varepsilon) \;:=\; \frac{\varepsilon}{2}\left(\frac{\rho}{2}\right)^n \frac{1}{4C_\varphi R^{n-1}}$$

and $|\mathbf{x}-\mathbf{z}| \le \rho/2$. The last inequality implies

$$|\mathbf{x}-\boldsymbol{\xi}| \ge \frac{\rho}{2} \qquad \text{for } \boldsymbol{\xi} \in \Gamma_1 \qquad \text{(see Figure 2.3)}.$$

Together with

$$\begin{aligned} R^2 - |\mathbf{x}-\mathbf{y}|^2 &= (R+|\mathbf{x}-\mathbf{y}|)(R-|\mathbf{x}-\mathbf{y}|) \\ &\le 2R\,(R-|\mathbf{x}-\mathbf{y}|) \le 2R\,|\mathbf{z}-\mathbf{x}| \le 2R\,\delta(\varepsilon) \end{aligned}$$

one obtains

$$|I_1| \;=\; \left| \frac{R^2-|\mathbf{x}-\mathbf{y}|^2}{R\omega_n} \int_{\Gamma_1} \frac{u(\boldsymbol{\xi})-\varphi(\mathbf{z})}{|\mathbf{x}-\boldsymbol{\xi}|^n} \mathrm{d}\Gamma_\xi \right| \quad \le \quad \frac{2}{\omega_n}\,\delta(\varepsilon)\,\frac{2C_\varphi}{(\rho/2)^n} \int_{\Gamma_1} \mathrm{d}\Gamma.$$

From $\int_{\Gamma_1} \mathrm{d}\Gamma \le \int_\Gamma \mathrm{d}\Gamma = R^{n-1}\omega_n$ and the definition of $\delta(\varepsilon)$ follows

$$|I_1| \;\le\; \varepsilon/2\,. \tag{2.20d}$$

Thus for every $\varepsilon > 0$ there exists a $\delta(\varepsilon) > 0$ such that $|\mathbf{x}-\mathbf{z}| \le \delta(\varepsilon)$ implies the estimate

$$|u(\mathbf{x})-\varphi(\mathbf{z})| = |I_0+I_1| \le |I_0|+|I_1| \le \varepsilon$$

(cf. (2.20c,d)). Hence, (2.18) has been proved, and the continuous extension of u to $\overline{K_R(\mathbf{y})}$ leads to $u \in C^0(\overline{K_R(\mathbf{y})})$. ∎

To ensure that the integral $\int_{\partial K_R(\mathbf{y})} \frac{\varphi(\boldsymbol{\xi})}{|\mathbf{x}-\boldsymbol{\xi}|^n} \mathrm{d}\Gamma_\xi$ in (2.15) is well-defined for $\boldsymbol{\xi} \notin \Gamma := \partial K_R(\mathbf{y})$, φ must belong to the function space $L^1(\Gamma)$. This is the set of all functions which are Lebesgue-integrable in Γ with a finite value $\int_\Gamma |\varphi| \mathrm{d}\Gamma$. By $L^\infty(\Gamma)$ we denote the space of all Lebesgue-integrable functions which are essentially bounded (i.e., bounded in $\Gamma\backslash Z$, where Z has measure zero). The product of $\varphi \in L^1(\Gamma)$ and $\psi \in L^\infty(\Gamma)$ again belongs to $L^1(\Gamma)$. Since $|\cdot - \boldsymbol{\xi}|^{-n}$ for $\boldsymbol{\xi} \notin \Gamma = \partial K_R(\mathbf{y})$ is continuous and in particular bounded, it belongs to $L^\infty(\Gamma)$. This proves that $\int_{\partial K_R(\mathbf{y})} \frac{\varphi(\boldsymbol{\xi})}{|\mathbf{x}-\boldsymbol{\xi}|^n}\,\mathrm{d}\Gamma_\xi$ is well-defined for $\varphi \in L^1(\Gamma)$.

Exercise 2.24. Prove that the function u defined by the Poisson integral formula (2.15) belongs to $C^\infty(K_R(\mathbf{y}))$ and solves $-\Delta u = 0$ in $K_R(\mathbf{y})$ even for boundary values $\varphi \in L^1(\partial K_R(\mathbf{y}))$. For every point of continuity $\mathbf{z} \in \partial K_R(\mathbf{y})$ of φ we have $u(\mathbf{x}) \to \varphi(\mathbf{z})$ $(\mathbf{x} \to \mathbf{z}, \mathbf{x} \in K_R(\mathbf{y}))$.

Finally, we add another derivation of the mean-value property mentioned by Ovall [215]. Let $u \in C^{2p}(\Omega)$ be an *arbitrary* function and let $K_R(\mathbf{x}) \subset \Omega \subset \mathbb{R}^n$. Taylor's formula yields $u(\boldsymbol{\xi}) = \sum_{|\alpha|\le 2p} \frac{1}{\alpha!} D^\alpha u(\mathbf{x}) (\boldsymbol{\xi} - \mathbf{x})^\alpha + o(R^{2p})$ for all $\boldsymbol{\xi} \in K_R(\mathbf{x})$. The integral of $(\boldsymbol{\xi} - \mathbf{x})^\alpha$ over the sphere $\partial K_R(\mathbf{x})$ is explicitly known (cf. Folland [103]) and yields the *Pizzetti series*

$$\frac{1}{\omega_n R^{n-1}} \int_{\partial K_R(\mathbf{x})} u(\boldsymbol{\xi}) \, \mathrm{d}\Gamma_\xi = \sum_{k=0}^{p} \frac{\Delta^k u(\mathbf{x})}{2^k k! \prod_{j=1}^{k} (n+2j)} R^{2k} + o(R^{2p})$$

(cf. Pizzetti [227]). In the case of harmonic functions, Theorem 2.23 allows us to use the infinite Taylor series without remainder. Since $\Delta^k u(\mathbf{x}) = 0$ holds for $k > 0$, the remaining term for $k = 0$ is $u(\mathbf{x})$.

2.4 Continuous Dependence on the Boundary Data

Definition 2.25. An abstract problem of the form

$$A(x) = y, \qquad x \in X,\ y \in Y,$$

is said to be *well-posed* if for all $y \in Y$ it has a unique solution $x \in X$ and if the latter depends continuously on y.

It is important to recognise whether a mathematical problem is well-posed since otherwise essential difficulties may occur in its numerical solution. In the case of the boundary-value problem (2.1a,b), $X \subset C^2(\Omega) \cap C^0(\overline{\Omega})$ is the space of functions harmonic in Ω and $Y = C^0(\Gamma)$ is the set of continuous boundary data on $\Gamma = \partial\Omega$. The topologies of X and Y are given by the supremum norms:

$$\|u\|_\infty := \sup_{\mathbf{x}\in\Omega} |u(\mathbf{x})| \qquad \text{and} \qquad \|\varphi\|_\infty := \sup_{\mathbf{x}\in\Gamma} |\varphi(\mathbf{x})| \,. \tag{2.21}$$

The question of the existence of a solution of (2.1a,b) will have to be postponed (see §7). The uniqueness, however, has been confirmed already in Theorem 2.18, if Ω is bounded. That the boundedness of Ω cannot be dropped without further ado is shown in the next example.

Example 2.26. The functions

$$\begin{aligned} u(x_1, x_2) &= x_1 && \text{in } \Omega = (0,\infty) \times \mathbb{R}, \\ u(x_1, x_2) &= \log\left(x_1^2 + x_2^2\right) && \text{in } \Omega = \mathbb{R}^2 \backslash \overline{K_1(0)}, \\ u(x_1, x_2) &= \sin(x_1)\sinh(x_2) && \text{in } \Omega = (0,\pi) \times (0,\infty) \end{aligned}$$

and also the trivial $u = 0$, are solutions of the boundary-value problem $-\Delta u = 0$ in Ω, $u = 0$ on $\Gamma = \partial\Omega$.

For bounded Ω the harmonic functions (solutions of (2.1a,b)) depend not only continuously but also Lipschitz-continuously on the boundary data.

Theorem 2.27. *Let Ω be bounded. If u^I and u^{II} are solutions of*

$$-\Delta u^I = -\Delta u^{II} = 0 \quad \text{in } \Omega, \qquad u^I = \varphi^I \text{ and } u^{II} = \varphi^{II} \quad \text{on } \Gamma = \partial\Omega$$

then

$$\|u^I - u^{II}\|_\infty \le \|\varphi^I - \varphi^{II}\|_\infty. \tag{2.22}$$

Proof. $v := u^I - u^{II}$ is a solution of $-\Delta v = 0$ in Ω and $v = \varphi^I - \varphi^{II}$ on Γ. According to Theorem 2.18, v takes its maximum and its minimum on Γ:

$$-\|\varphi^I - \varphi^{II}\|_\infty \le v(\mathbf{x}) \le \|\varphi^I - \varphi^{II}\|_\infty \qquad \text{for all } \mathbf{x} \in \overline{\Omega}.$$

The definition (2.21) of $\|\cdot\|_\infty$ implies (2.22). ∎

The continuous dependence of the solution with respect to the boundary values is also shown by Harnack's Theorem 2.22. If $\|\varphi_n - \varphi\|_\infty \to 0$ holds for a sequence of boundary values then the associated solutions satisfy $\|u_n - u\|_\infty \to 0$. Here the existence of a solution of $-\Delta u = 0$ in Ω, $u = \varphi$ on Γ need not be assumed.

Another problem, just as important for numerical mathematics, is rarely discussed in the literature: does the solution also depend continuously on the form of the boundary Γ? Figure 2.4 shows domains Ω' and Ω'' which approximate Ω. A polygonal domain, as, for example, Ω'', occurs in the method of finite elements (see §8.6.3).

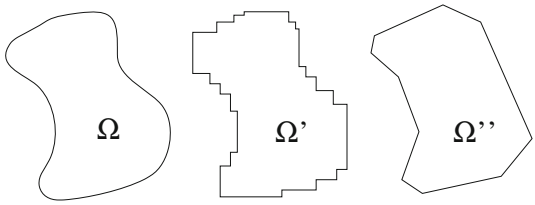


Fig. 2.4 Approximation of Ω by Ω' and Ω''.

Let Ω_n be a sequence of domains with $\Gamma_n = \partial\Omega_n$. We say that $\Gamma_n \to \Gamma$ if $\operatorname{dist}(\Gamma_n, \Gamma) \to 0$. Here, we define

$$\operatorname{dist}(\Gamma_n, \Gamma) := \sup\{\operatorname{dist}(\mathbf{x}, \Gamma) : \mathbf{x} \in \Gamma_n\}, \quad \operatorname{dist}(\mathbf{x}, \Gamma) := \inf\{|\mathbf{x} - \mathbf{y}| : \mathbf{y} \in \Gamma\}.$$

Further, one must specify when $\varphi_n \in C^0(\Gamma_n)$ converges uniformly to $\varphi \in C^0(\Gamma)$. In the following we define the uniform convergence for φ_n and u_n.

$\varphi_n \in C^0(\Gamma_n)$ *converges uniformly* to $\varphi \in C^0(\Gamma)$ if, for each $\varepsilon > 0$, there exist numbers $N(\varepsilon)$ and $\delta(\varepsilon) > 0$ such that the following implication holds:

$$n \ge N(\varepsilon),\ \mathbf{x} \in \Gamma,\ \mathbf{y} \in \Gamma_n,\ |\mathbf{x} - \mathbf{y}| \le \delta(\varepsilon) \implies |\varphi_n(\mathbf{y}) - \varphi(\mathbf{x})| \le \varepsilon. \tag{2.23a}$$

The sequence $u_n \in C^0(\Omega_n)$ *converges uniformly* to $u \in C^0(\Omega)$ if

$$\lim_{n\to\infty} \sup\left\{|u_n(\mathbf{x}) - u(\mathbf{x})| : \mathbf{x} \in \Omega_n \cap \Omega\right\} = 0. \tag{2.23b}$$

Remark 2.28. (a) Let K be a set which is compact (i.e., complete and bounded) with $\Gamma \subset K$ and $\Gamma_n \subset K$ for all n. Let $\phi_n \in C^0(K)$ converge uniformly on K to ϕ. If $\varphi_n = \phi_n$ on Γ_n and $\varphi = \phi$ on Γ then (2.23a) is satisfied.

(b) Let $\Omega_n \subset \Omega$ for all n and let $\bar{u}_n$ be the following (not continuous) continuation of u_n onto $\overline{\Omega}$: $\bar{u}_n = u_n$ on $\overline{\Omega}_n$, $\bar{u}_n = u$ on $\overline{\Omega}\backslash\overline{\Omega}_n$. Then (2.23b) is equivalent to uniform convergence $\bar{u}_n \to u$ on $\overline{\Omega}$ in the usual sense.

Theorem 2.29. *Let $\Omega_n \subset \Omega$ with Ω bounded, and let $\Gamma_n \to \Gamma$. Let the functions u_n which are harmonic in Ω_n, be solutions of*

$$-\Delta u_n = 0 \quad \text{in } \Omega_n\,, \qquad u_n = \varphi_n \quad \text{on } \Gamma_n\,. \tag{2.24a}$$

Let $\varphi_n \in C^0(\Gamma_n)$ converge uniformly in the sense of (2.23a) to $\varphi \in C^0(\Gamma)$. Then the following assertions hold: (a) If there exists a solution $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$ of

$$-\Delta u = 0 \quad \text{in } \Omega\,, \qquad u = \varphi \quad \text{on } \Gamma \tag{2.24b}$$

then $u_n \to u$ holds in the sense of (2.23b).

(b) If conversely $u_n \to u \in C^0(\overline{\Omega})$ is satisfied in the sense of (2.23b), then u is the solution of (2.24b).

Proof. (a) Let the continuation $\bar{u}_n$ be defined as in Remark 2.28b. Since u is uniformly continuous on $\overline{\Omega}$, there exists $\delta_u(\varepsilon) > 0$ for all $\varepsilon > 0$ such that

$$|u(\mathbf{x}) - u(\mathbf{y})| \le \tfrac{\varepsilon}{2} \qquad \text{if } |\mathbf{x} - \mathbf{y}| \le \delta_u(\tfrac{\varepsilon}{2}). \tag{2.25a}$$

Set $\delta^\star(\varepsilon) := \min\{\delta_u(\frac{\varepsilon}{2}), \delta(\frac{\varepsilon}{2})\}$ with δ from (2.23a). Because $\Gamma_n \to \Gamma$ there exists $N_\Gamma(\varepsilon)$, so that $\operatorname{dist}(\Gamma_n, \Gamma) \le \delta^\star(\varepsilon)$ for $n \ge N_\Gamma(\varepsilon)$. For

$$n \ge N^\star(\varepsilon) := \max\{N_\Gamma(\varepsilon), N(\varepsilon/2)\} \qquad (N \text{ from (2.23a)})$$

we want to show that $|\bar{u}_n(\mathbf{x}) - u(\mathbf{x})| \le \varepsilon$ for all $\mathbf{x} \in \overline{\Omega}$. For $\mathbf{x} \in \overline{\Omega}\backslash\overline{\Omega}_n$ the estimate is trivial because $\bar{u}_n(\mathbf{x}) = u(\mathbf{x})$. For all $\mathbf{x} \in \overline{\Omega}_n \subset \overline{\Omega}$, however, there holds

$$|\bar{u}_n(\mathbf{x}) - u(\mathbf{x})| = |u_n(\mathbf{x}) - u(\mathbf{x})| \le \max_{\mathbf{x} \in \Gamma_n} |u_n(\mathbf{x}) - u(\mathbf{x})| \tag{2.25b}$$

(cf. Theorem 2.18), because $u_n - u$ is harmonic in ω_n. It remains to estimate $|u_n(\mathbf{x}) - u(\mathbf{x})|$ for $\mathbf{x} \in \Gamma_n$. For $\mathbf{x} \in \Gamma_n$ with $n \ge N^\star(\varepsilon)$ there exists $\mathbf{y} \in \Gamma$ with $|\mathbf{x} - \mathbf{y}| \le \delta(\varepsilon/2)$. Thus we obtain

$$\begin{aligned} |u_n(\mathbf{x}) - u(\mathbf{x})| &= |\varphi_n(\mathbf{x}) - u(\mathbf{x})| \le |\varphi_n(\mathbf{x}) - \varphi(\mathbf{y})| + |\varphi(\mathbf{y}) - u(\mathbf{x})| \\ &\le |\varphi_n(\mathbf{x}) - \varphi(\mathbf{y})| + |u(\mathbf{y}) - u(\mathbf{x})| \le \tfrac{\varepsilon}{2} + \tfrac{\varepsilon}{2} = \varepsilon \end{aligned}$$

from (2.23a) and (2.25a). Since $\mathbf{x} \in \Gamma_n$ is arbitrary it follows that $|u_n - u| \leq \varepsilon$ on Γ_n, and (2.25b) proves the uniform convergence $\bar{u}_n \to u$ on $\overline{\Omega}$. Hence, from Remark 2.28b it follows that (2.23b) is satisfied.

(b) Let $K \subset \Omega$ be a compact set. Since $\Gamma_n \to \Gamma$ there exists an $N(K)$ such that $K \subset \Omega_n$ for $n \geq N(K)$. Thus, the sequence $\{u_n : n \geq N(K)\}$ converges uniformly in the usual sense on K to u so that one can apply Theorem 2.22: consequently u is harmonic in K. Since $K \subset \Omega$ may be chosen arbitrarily, it follows that $u \in C^2(\Omega)$. By assumption, we already have $u \in C^0(\overline{\Omega})$. That the boundary value $u = \varphi$ is taken on Γ is deduced from $\varphi_n \to \varphi$ and $\Gamma_n \to \Gamma$. ∎

In Theorem 2.22 one was able to derive the existence of a solution u of (2.24b) just from $\varphi_n \to \varphi$. This inference is not possible for the case of $\Omega_n \neq \Omega$ as the following example shows.

Example 2.30. Let

$$\Omega_n := K_1(0) \setminus \overline{K_{1/n}(0)} \subset \Omega := K_1(0) \setminus \{0\} \subset \mathbb{R}^2.$$

The boundaries are $\Gamma_n = \partial K_1(0) \cup \partial K_{1/n}(0)$ and $\Gamma = \partial K_1(0) \cup \{0\}$, and satisfy $\Gamma_n \to \Gamma$. The boundary values

$$\varphi = \varphi_n = 0 \quad \text{on } \partial K_1(0), \qquad \varphi_n = 1 \quad \text{on } \partial K_{1/n}(0), \qquad \varphi(0,0) = 1$$

satisfy the condition $\varphi_n \to \varphi$ (cf. (2.23a) and Remark 2.28a). The solutions u_n of (2.24a) can be given explicitly:

$$u_n(\mathbf{x}) = \log(|\mathbf{x}|) / \log(1/n).$$

Obviously, $u_n(\mathbf{x}) \to u(\mathbf{x}) := 0$ holds pointwise, but $u = 0$ satisfies neither (2.23b) nor the boundary-value problem (2.24b). Conversely, one infers from Theorem 2.29a the following result: In $\Omega = K_1(0)/\{0\} \subset \mathbb{R}^2$ the potential equation has no solution $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$ which assumes the boundary values $u(\mathbf{x}) = 0$ on $\partial K_1(0)$ and $u(\mathbf{x}) = 1$ in $\mathbf{x} = 0$.

Chapter 3
The Poisson Equation

Abstract In **Section 3.1** the Poisson equation $-\Delta u = f$ is introduced, and the uniqueness of the solution is proved. The Green function is defined in **Section 3.2**. It allows the representation (3.6) of the solution, provided it is existing. Concerning the existence, Theorem 3.13 contains a negative statement (cf. **Section 3.3**): The Poisson equation with a continuous right-hand side f may possess no classical solution. A sufficient condition for a classical solution is the Hölder continuity of f as stated in Theorem 3.18. **Section 3.4** introduces Green's function for the ball. In the two-dimensional case, Riemann's mapping theorem allows the construction of the Green function for a large class of domains. In **Section 3.5** we replace the Dirichlet boundary condition by the Neumann condition. The final **Section 3.6** is a short introduction into the integral equation method. The solution of the boundary-value problem can indirectly be obtained by solving an integral equation.

3.1 Posing the Problem

The *Poisson equation* is a slight generalisation of Laplace's equation and reads

$$-\Delta u = f \qquad \text{in } \Omega \tag{3.1a}$$

with given function $f \in C^0(\Omega)$. In the physical interpretation, f is the source term, for example, the charge density in the case of an electrical potential u. In mechanics, f is called the 'load'. In mathematical terminology, f is often called the 'right-hand side'. If L is a linear differential operator, $Lu = f$ is called the *inhomogeneous problem*, whereas $Lu = 0$ is the *homogeneous problem*. In this sense, the Laplace equation is the homogeneous Poisson equation.

To determine the solution uniquely one needs a *boundary-value* specification, for example, the *Dirichlet condition*

$$u = \varphi \qquad \text{on } \Gamma := \partial\Omega\,. \tag{3.1b}$$

W. Hackbusch, *Elliptic Differential Equations*, Springer Series
in Computational Mathematics 18, DOI 10.1007/978-3-662-54961-2_3

Definition 3.1. The function u is called the *classical solution* of the boundary-value problem (3.1a,b) if $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$ satisfies the equations (3.1a,b) pointwise.

Until we introduce weak solutions in §7, 'solution' will always mean 'classical solution'.

The solution of the boundary-value problem (3.1a,b) will in general no longer satisfy the mean-value property and the maximum principle. But these properties still hold for the difference of two solutions u_1 and u_2 of the Poisson equation, since $-\Delta(u_1 - u_2) = f - f = 0$. Thus the uniqueness of the solution of problem (3.1a,b) immediately follows and Theorem 2.27 can be brought over.

Theorem 3.2. *Let Ω be bounded.*

(a) The solution of (3.1a,b) is uniquely determined.

(b) If u^I and u^{II} are solutions of the Poisson equation for boundary values φ^I and φ^{II}, then we have

$$\|u^I - u^{II}\|_\infty \le \|\varphi^I - \varphi^{II}\|_\infty. \tag{3.2}$$

Proof. (b) The proof of Theorem 2.27 can be repeated verbatim here. The inequality (2.22) coincides with (3.2).

(a) If u^I and u^{II} are two solutions of (3.1a,b) then the right-hand side in (3.2) vanishes. Thus $u^I = u^{II}$. ∎

Theorems 2.22 and 2.29 can be transferred likewise.

The boundary-value problem (3.1a,b) can be decomposed into two subproblems corresponding to f and φ, respectively. The following analysis will often refer to one of these subproblems.

Lemma 3.3. *Let u_1 be the solution of the boundary-value problem*

$$-\Delta u_1 = f \quad \text{in } \Omega, \qquad u_1 = 0 \quad \text{on } \Gamma, \tag{3.3a}$$

and u_2 the solution of the boundary-value problem

$$-\Delta u_2 = 0 \quad \text{in } \Omega, \qquad u_2 = \varphi \quad \text{on } \Gamma. \tag{3.3b}$$

Then the sum $u := u_1 + u_2$ solves the boundary-value problem (3.1a,b).

3.2 Representation of the Solution by the Green Function

The following exercise discusses the continuity of improper integrals with respect to parameters in the integrand.

Exercise 3.4. (a) Let $\Omega \subset \mathbb{R}^n$ be bounded, $\mathbf{x}_0 \in \Omega$, $f \in C^0(\overline{\Omega}\backslash\{\mathbf{x}_0\})$ and

$$|f(\mathbf{x})| \leq C\,|\mathbf{x}-\mathbf{x}_0|^{-s} \quad \text{for some } s < n.$$

Show that $\int_\Omega f(\mathbf{x})\mathrm{d}\mathbf{x}$ exists as an improper integral.
(b) Let $\Omega \subset \mathbb{R}^n$ be bounded and let $\mathbf{x}_0(\boldsymbol{\xi}) \in \Omega$ depend continuously on $\boldsymbol{\xi} \in D$, with D compact. Let $f(\mathbf{x},\boldsymbol{\xi})$ be continuous in $(\mathbf{x},\boldsymbol{\xi}) \in \overline{\Omega} \times D$ with $\mathbf{x} \neq \mathbf{x}_0(\boldsymbol{\xi})$ and let $|f(\mathbf{x},\boldsymbol{\xi})| \leq C\,|\mathbf{x}-\mathbf{x}_0(\boldsymbol{\xi})|^{-s}$, $s < n$. Show that

$$F(\boldsymbol{\xi}) := \int_\Omega f(\mathbf{x},\boldsymbol{\xi})\mathrm{d}\mathbf{x}$$

is continuous: $F \in C^0(D)$.

Lemma 3.5. *Let the solution of (3.1a,b) belong to $C^2(\overline{\Omega})$, where Ω is a normal domain. Then u may be represented as*

$$u(\mathbf{x}) = \int_\Omega \gamma(\boldsymbol{\xi},\mathbf{x})\,f(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} + \int_{\partial\Omega}\left[\gamma(\boldsymbol{\xi},\mathbf{x})\frac{\partial}{\partial n}u(\boldsymbol{\xi}) - u(\boldsymbol{\xi})\frac{\partial}{\partial n_\xi}\gamma(\boldsymbol{\xi},\mathbf{x})\right]\mathrm{d}\Gamma_\xi \tag{3.4}$$

for every fundamental solution γ in (2.10).

Proof. The proof is the same as in Theorem 2.8 or in Corollary 2.9. The term $\int_{\Omega_\varepsilon}\gamma(-\Delta u)\mathrm{d}\boldsymbol{\xi}$ with $\Omega_\varepsilon = \Omega\backslash\overline{K_\varepsilon(\mathbf{x})}$ becomes $\int_{\Omega_\varepsilon}\gamma f\mathrm{d}\boldsymbol{\xi}$. Since the singularity of $\gamma(\boldsymbol{\xi},\mathbf{x})$ is integrable at $\boldsymbol{\xi}=\mathbf{x}$ (cf. Exercise 3.4), $\int_{\Omega_\varepsilon}\gamma f\mathrm{d}\boldsymbol{\xi}$ converges to $\int_\Omega\gamma f\mathrm{d}\boldsymbol{\xi}$ as $\varepsilon \to 0$. ∎

In the boundary integral in (3.4) one may replace $u(\boldsymbol{\xi})$ by $\varphi(\boldsymbol{\xi})$ (cf. (3.1b)). The function $\partial u/\partial n$ on Γ, however, is unknown and cannot be specified arbitrarily either, since the boundary values (3.1b) already determine the solution uniquely (cf. Theorem 3.2). To make $\int_\Gamma \gamma\frac{\partial u}{\partial n}\mathrm{d}\boldsymbol{\xi}$ vanish one must select the fundamental solution so that $\gamma(\boldsymbol{\xi},\mathbf{x}) = 0$ for $\boldsymbol{\xi} \in \Gamma$ and $\mathbf{x} \in \Omega$.

Definition 3.6. A fundamental solution g in (2.10) is called a *Green function (of the first kind)* if $g(\boldsymbol{\xi},\mathbf{x}) = 0$ for all $\boldsymbol{\xi} \in \Gamma$, $\mathbf{x} \in \Omega$.

Hence, for all $\mathbf{x} \in \Omega$, the Green function $g(\boldsymbol{\xi},\mathbf{x})$ shares the same singularity at $\mathbf{x}$ as $s(\boldsymbol{\xi},\mathbf{x})$, and solves the following boundary-value problem with respect to the first argument:

$$-\Delta g(\cdot,\mathbf{x}) = 0 \quad \text{in } \Omega\backslash\{\mathbf{x}\}, \qquad g(\cdot,\mathbf{x}) = 0 \quad \text{on } \Gamma. \tag{3.5}$$

The existence of a Green function is closely related to the solvability of the boundary-value problem for the potential equation.

Remark 3.7. The Green function exists if and only if for all $\mathbf{x} \in \Omega$ the boundary-value problem $-\Delta\Phi = 0$ in Ω and $\Phi = -s(\cdot,\mathbf{x})$ on Γ has a solution $\Phi \in C^2(\overline{\Omega})$.

The above consideration results in the following representation theorem.

Theorem 3.8. *Let Ω be a normal domain. Let the boundary-value problem (3.1a,b) have a solution $u \in C^2(\overline{\Omega})$. Assume the existence of a Green function of the first kind. Then one can express u explicitly by*

$$u(\mathbf{x}) = \int_\Omega g(\boldsymbol{\xi}, \mathbf{x})\, f(\boldsymbol{\xi})\, \mathrm{d}\boldsymbol{\xi} - \int_{\partial\Omega} \varphi(\boldsymbol{\xi}) \frac{\partial}{\partial n_\xi} g(\boldsymbol{\xi}, \mathbf{x})\, \mathrm{d}\Gamma_\xi \,. \tag{3.6}$$

In the following we reverse the implication. Let the existence of the Green function be assumed. Then, does function u defined by equation (3.6) represent the classical solution of the boundary-value problem (3.1a,b)? Here it must be proved, in particular, that $u \in C^2(\Omega)$ and $-\Delta u = f$. Firstly, it is not even clear yet whether the function $u(\mathbf{x})$ defined by equation (3.6) depends continuously on $\mathbf{x}$ since the definition of a fundamental solution $\gamma(\boldsymbol{\xi}, \mathbf{x})$ does not require continuity in the second argument $\mathbf{x}$. Despite that, the Green function $g(\boldsymbol{\xi}, \cdot)$ is in $C^2(\Omega \backslash \{\boldsymbol{\xi}\})$, as the following result shows (cf. Leis [189, page 67]).

Exercise 3.9. Let Ω be a normal domain. Let the Green function exist, and for fixed $\mathbf{y} \in \Omega$ let $g(\cdot, \mathbf{y}) \in C^2(\overline{\Omega} \setminus \{\mathbf{y}\})$ (weaker conditions are possible). Now prove that[1]

$$g(\mathbf{x}, \mathbf{y}) = g(\mathbf{y}, \mathbf{x}) \qquad \text{for } \mathbf{x}, \mathbf{y} \in \Omega \,. \tag{3.7}$$

Hint: Apply Green's formula (2.6b) with $\Omega_\varepsilon = \Omega \setminus [K_\varepsilon(\mathbf{x}') \cup K_\varepsilon(\mathbf{x}'')]$, $\mathbf{x}', \mathbf{x}'' \in \Omega$, $u(\mathbf{x}) := g(\mathbf{x}, \mathbf{x}')$, $v(\mathbf{x}) := g(\mathbf{x}, \mathbf{x}'')$ and use (2.11).

Theorem 3.10. *Let $g(\cdot, \cdot)$ be the Green function for a bounded domain Ω. Then*

$$g(\mathbf{x}, \mathbf{y}) > 0 \qquad \textit{for } \mathbf{x}, \mathbf{y} \in \Omega \,. \tag{3.8}$$

Proof. We recall the representation $g = s + \Phi$ in (2.10). Φ is bounded in Ω: $\|\Phi\|_\infty < \infty$. The singularity of s tends to $+\infty$. For sufficiently small $\varepsilon > 0$ we have $K_\varepsilon(\mathbf{y}) \subset \Omega$ as well as $s(\mathbf{x}, \mathbf{y}) > \|\Phi\|_\infty$ for $\mathbf{x} \in \partial K_\varepsilon(\mathbf{y})$. This proves $g > 0$ on $\partial K_\varepsilon(\mathbf{y})$, while by definition $g = 0$ holds on $\partial\Omega$. The union $\partial K_\varepsilon(\mathbf{y}) \cup \partial\Omega$ is the boundary of $\Omega_\varepsilon := \Omega \backslash K_\varepsilon(\mathbf{y})$. The Green function solves $-\Delta g(\cdot, \mathbf{y}) = 0$ in Ω_ε. The maximum principle as formulated in Exercise 2.19b with $u_1 = 0$ and $u_2 = g(\cdot, \mathbf{y})$ implies $g(\cdot, \mathbf{y}) > 0$ and proves (3.8). ∎

Exercise 2.19b concerns the boundary-value problem (3.3b): $-\Delta u = 0$ in Ω and $u = \varphi$ on Γ. It ensures that $u > 0$ in Ω if $\varphi \gneqq 0$. The following theorem refers to the boundary-value problem (3.3a) ($-\Delta u = f$ in Ω and $u = 0$ on Γ) and the representation (3.6) (with $\varphi = 0$).

Theorem 3.11. *Let u be the solution of $-\Delta u = f$ in the domain Ω, and let $u = 0$ on $\Gamma = \partial\Omega$. If $f \in C^0(\overline{\Omega})$ is nonnegative and $f > 0$ for at least one point, then $u > 0$ in Ω.*

Proof. By assumption, there are $\mathbf{x}_0$ and a neighbourhood $K_\varepsilon(\mathbf{x}_0) \subset \Omega$, $\varepsilon > 0$, so that $f > 0$ in $K_\varepsilon(\mathbf{x}_0)$. From (3.6) and (3.8) one concludes that the integrand in $u(\mathbf{x}) = \int_\Omega g(\boldsymbol{\xi}, \mathbf{x}) f(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}$ is nonnegative and positive in $K_\varepsilon(\mathbf{x}_0)$. ∎

[1] Later we shall see that this symmetry holds for all selfadjoint differential operators.

Exercise 3.12. Let $\Omega,\ \Omega_1, \Omega_2$ be bounded domains.
(a) Let g be the Green function in Ω. Show that

$$g(\mathbf{x},\mathbf{y}) < s(\mathbf{x},\mathbf{y}) \qquad \text{for } \mathbf{x},\mathbf{y} \in \Omega \subset \mathbb{R}^n,\ n \geq 3. \tag{3.9}$$

What is the corresponding statement for $n = 2$?

(b) Let g_1, g_2 be the respective Green functions in $\Omega_1 \subset \Omega_2$. Prove with the aid of Exercise 2.19b that

$$g_1(\mathbf{x},\mathbf{y}) < g_2(\mathbf{x},\mathbf{y}) \qquad \text{for } \mathbf{x},\mathbf{y} \in \Omega_1 \subset \Omega_2\,.$$

3.3 Existence of a Solution

If one tries to reverse the assertion of Theorem 3.8, one encounters the surprising difficulty of having to set precise conditions on the source term f. The natural requirement $f \in C^0(\overline{\Omega})$ is necessary for $u \in C^2(\overline{\Omega})$, but it is not sufficient, as the following theorem shows. The statement underlines that the classical function spaces C^0 and C^2 are not well suited for boundary-value problems.

Theorem 3.13. *Even if the boundary $\Gamma = \partial\Omega$ and the boundary values φ are sufficiently smooth and if the Green function exists, there are functions $f \in C^0(\overline{\Omega})$ to which no solutions $u \in C^2(\overline{\Omega})$ correspond.*

Proof. The counter-example uses the square $(-1,1)\times(-1,1)$ with constant boundary values $\varphi = 1$ and the solution $u(x,y) = |x|\,|y|\,\log(|x|+|y|)$. One verifies that $f := -\Delta u$ is continuous, but the mixed derivative $\partial^2 u/\partial x\partial y$ contains the unbounded term $\log(|x|+|y|)$, proving $u \notin C^2(\overline{\Omega})$.

The latter result is not a consequence of the fact that the boundary of the square is not smooth. For this purpose we modify the example. The function u from above can be extended outside $K_\varepsilon(\mathbf{0})$ to a function in C^∞. Then u has boundary values $\varphi \in C^\infty$ on $\partial K_1(\mathbf{0})$, f is continuous in $\overline{K_1(\mathbf{0})}$, while $\frac{\partial^2 u}{\partial x \partial y}$ still has the same singularity at $\mathbf{x} = \mathbf{0}$. ∎

The statement of Theorem 3.21 will be equivalent to that of Theorem 3.13.

Theorem 3.13 shows that equation (3.6) need not represent a classical solution for $f \in C^0(\overline{\Omega})$. However, a sufficient condition for f to do so is Hölder-continuity.

Definition 3.14 (Hölder-continuity). $f \in C^0(\overline{\Omega})$ is said to be Hölder-continuous in $\overline{\Omega}$ with the exponent $\lambda \in (0,1)$ if there exists a constant $C = C(f)$ such that

$$|f(\mathbf{x}) - f(\mathbf{y})| \leq C\,|\mathbf{x}-\mathbf{y}|^\lambda \qquad \text{for } \mathbf{x},\mathbf{y} \in \overline{\Omega}\,. \tag{3.10a}$$

We write $f \in C^\lambda(\overline{\Omega})$ and define the norm $\|f\|_{C^\lambda(\overline{\Omega})}$ as the smallest constant C which satisfies (3.10a) and $|f(\mathbf{x})| \leq C$:

$$\|f\|_{C^\lambda(\overline{\Omega})} := \max\left\{\sup\left\{\frac{|f(\mathbf{x})-f(\mathbf{y})|}{|\mathbf{x}-\mathbf{y}|^\lambda} : \mathbf{x},\mathbf{y}\in\overline{\Omega},\ \mathbf{x}\neq\mathbf{y}\right\},\ \|f\|_\infty\right\}. \tag{3.10b}$$

The function $f \in C^k(\overline{\Omega})$ is said to be k-fold Hölder-continuously differentiable in $\overline{\Omega}$ (with the exponent λ), if $D^\nu f \in C^\lambda(\overline{\Omega})$ for all $|\nu| \le k$. Here ν is a multi-index of length $|\nu|$,

$$\nu = (\nu_1,\ldots,\nu_n) \qquad \text{with } \nu_i \in \mathbb{N}_0, \quad |\nu| = \nu_1 + \ldots + \nu_n, \tag{3.11a}$$

and D^ν is a $|\nu|$-fold partial differentiation operator:

$$D^\nu = \frac{\partial^{|\nu|}}{\partial x_1^{\nu_1}\partial x_2^{\nu_2}\ldots\partial x_n^{\nu_n}}. \tag{3.11b}$$

The k-fold Hölder-continuously differentiable functions form the Banach space $C^{k+\lambda}(\overline{\Omega})$ with the norm

$$\|f\|_{C^{k+\lambda}(\overline{\Omega})} = \max\left\{\|D^\nu f\|_{C^\lambda(\overline{\Omega})} : |\nu| \le k\right\}.$$

If $s = k+\lambda$ one also writes $C^s(\overline{\Omega})$ for $C^{k+\lambda}(\overline{\Omega})$. The k-fold Lipschitz-continuously differentiable functions $f \in C^{k,1}(\overline{\Omega})$ are the result of the choice $\lambda = 1$ in (3.10a,b). For reasons of completeness let us add that

$$\|f\|_{C^k(\overline{\Omega})} = \max\{\|D^\nu f\|_\infty : |\nu| \le k\} \tag{3.12}$$

is the norm in $C^k(\overline{\Omega})$ for integer $k \ge 0$.

Exercise 3.15. f is said to be *locally Hölder-continuous* in D if for each $\mathbf{x} \in D$ there exists a neighbourhood $K_\varepsilon(\mathbf{x})$ such that $f \in C^\lambda(K_\varepsilon(\mathbf{x}) \cap D)$.
(a) Prove that if D is compact then $f \in C^\lambda(D)$ follows from the local Hölder-continuity in D. Formulate and prove corresponding statements for $C^{k+\lambda}(D)$ and $C^{k,1}(D)$.
(b) Let $s > 0$. Prove $|\mathbf{x}|^s \in C^s(\overline{K_R(0)})$ if $s \notin \mathbb{N}$, otherwise $|\mathbf{x}|^s \in C^{s-1,1}(\overline{K_R(0)})$. *Hint:* $1 - t^s \le (1-t)^s$ for $0 \le t \le 1$, $s \ge 0$.

The function u from equation (3.6) can be decomposed into $u_1 + u_2$, where $u_1 = \int_\Omega g f \,\mathrm{d}\boldsymbol{\xi}$ and $u_2 = -\int_{\partial\Omega} \varphi \partial g/\partial n \,\mathrm{d}\Gamma$. u is the solution of the boundary-value problem (3.1a,b) if we are able to show that u_1 and u_2 are solutions of (3.3a) and (3.3b), respectively.

Theorem 3.16. *If the Green function exists and satisfies suitable conditions then*

$$u(\mathbf{x}) = -\int_{\partial\Omega} \varphi(\boldsymbol{\xi}) \frac{\partial}{\partial n_\xi} g(\boldsymbol{\xi},\mathbf{x}) \,\mathrm{d}\Gamma_\xi$$

is a classical solution of (3.3b): $-\Delta u = 0$ in Ω and $u = \varphi$ on $\Gamma = \partial\Omega$.

The proof goes in principle just as for Theorem 2.20 (cf. Leis [189, page 69]).

In preparation for the next theorem, the following exercise discusses whether differentiation and integration can be interchanged in the presence of singular integrands.

Exercise 3.17. Let $\Omega \subset \mathbb{R}^n$ be bounded and $A := \{(\boldsymbol{\xi}, \mathbf{x}) \in \overline{\Omega} \times \overline{\Omega} : \boldsymbol{\xi} \neq \mathbf{x}\}$. For the derivatives of f with respect to $\mathbf{x}$ assume

$$D_x^\nu f \in C^0(A) \text{ and } |D_x^\nu f(\boldsymbol{\xi}, \mathbf{x})| \leq C\,|\mathbf{x} - \boldsymbol{\xi}|^{-s} \quad \text{with } s < n \text{ for all } |\nu| \leq k.$$

Prove that then $F(\mathbf{x}) := \int\limits_\Omega f(\boldsymbol{\xi}, \mathbf{x}) \mathrm{d}\boldsymbol{\xi} \in C^k(\overline{\Omega})$ and $D^\nu F(\mathbf{x}) = \int\limits_\Omega D_x^\nu f(\boldsymbol{\xi}, \mathbf{x}) \mathrm{d}\boldsymbol{\xi}$ for $|\nu| \leq k$.

Theorem 3.18. *Suppose that the Green function $g(\cdot, \mathbf{x}) \in C^2(\Omega \backslash \{\mathbf{x}\})$ for $\mathbf{x} \in \Omega$ exists, and let $f \in C^\lambda(\overline{\Omega})$. Then*

$$u(\mathbf{x}) = \int_\Omega f(\boldsymbol{\xi})\, g(\boldsymbol{\xi}, \mathbf{x})\, \mathrm{d}\boldsymbol{\xi} \tag{3.13}$$

is a classical solution of (3.3a): $-\Delta u = f$ in Ω and $u = 0$ on Γ.

Proof. The boundary condition $u(\mathbf{x}) = 0$ for $\mathbf{x} \in \Gamma$ follows easily from (3.7) and $g(\mathbf{x}, \boldsymbol{\xi}) = 0$. The property $u \in C^1(\overline{\Omega})$ and the representation of the derivative $u_{x_i}(\mathbf{x}) = \int_\Omega f(\boldsymbol{\xi}) g_{x_i}(\boldsymbol{\xi}, \mathbf{x}) \mathrm{d}\boldsymbol{\xi}$ result from Exercise 3.17. To prove $u \in C^2(\overline{\Omega})$ this step cannot be repeated since $g_{x_i x_j} = \mathcal{O}(|\boldsymbol{\xi} - \mathbf{x}|^{-n})$ has a singularity which is not integrable. We write the derivative in the form

$$u_{x_i}(\mathbf{x}) = \int_\Omega [f(\boldsymbol{\xi}) - f(\hat{\mathbf{x}})]\; g_{x_i}(\boldsymbol{\xi}, \mathbf{x})\, \mathrm{d}\boldsymbol{\xi} + f(\hat{\mathbf{x}}) \int_\Omega g_{x_i}(\boldsymbol{\xi}, \mathbf{x})\, \mathrm{d}\boldsymbol{\xi}\,, \tag{3.14a}$$

where $\hat{\mathbf{x}}$ will be fixed later. Let $\partial_j F(\mathbf{x})$ be the difference quotient $\frac{F(\mathbf{x}^\varepsilon) - F(\mathbf{x})}{\varepsilon}$ with $x_j^\varepsilon = x_j + \varepsilon$ and $x_i^\varepsilon = x_i$ $(i \neq j)$. The difference ∂_j can be taken under the integral sign:

$$\partial_j u_{x_i}(\mathbf{x}) = \int_\Omega [f(\boldsymbol{\xi}) - f(\hat{\mathbf{x}})]\; \partial_j g_{x_i}(\boldsymbol{\xi}, \mathbf{x})\, \mathrm{d}\boldsymbol{\xi} + f(\hat{\mathbf{x}})\, \partial_j \frac{\partial}{\partial x_i} \int_\Omega g(\boldsymbol{\xi}, \mathbf{x})\, \mathrm{d}\boldsymbol{\xi}\,.$$

The values $\boldsymbol{\xi}$ lying on the line $[\mathbf{x}, \mathbf{x}^\varepsilon]$ form a set of measure zero. Otherwise, the intermediate value theorem yields $\partial_j g_{x_i}(\boldsymbol{\xi}, \mathbf{x}) = g_{x_i x_j}(\boldsymbol{\xi}, \mathbf{x}^{\mu\varepsilon})$ with $\mathbf{x}^{\mu\varepsilon} = \mathbf{x} + \mu(\mathbf{x}^\varepsilon - \mathbf{x})$ and a factor $\mu = \mu(\boldsymbol{\xi}, \mathbf{x}) \in [0, 1]$. Now we choose $\hat{\mathbf{x}} := \mathbf{x}^{\mu\varepsilon}$. Since

$$[f(\boldsymbol{\xi}) - f(\mathbf{x}^{\mu\varepsilon})] g_{x_i x_j}(\boldsymbol{\xi}, \mathbf{x}^{\mu\varepsilon}) = \mathcal{O}(|\boldsymbol{\xi} - \mathbf{x}^{\mu\varepsilon}|^{\lambda - n})$$

is integrable, the limit process $\varepsilon \to 0$ and $\mathbf{x}^{\mu\varepsilon} \to \mathbf{x}$ can be carried out in the integrand:

$$u_{x_i x_j}(\mathbf{x}) = \int_\Omega [f(\boldsymbol{\xi}) - f(\mathbf{x})]\; g_{x_i x_j}(\boldsymbol{\xi}, \mathbf{x})\, \mathrm{d}\boldsymbol{\xi} + f(\mathbf{x})\, \frac{\partial^2}{\partial x_i \partial x_j} \int_\Omega g(\boldsymbol{\xi}, \mathbf{x})\, \mathrm{d}\boldsymbol{\xi}\,. \tag{3.14b}$$

Equation (3.14b) implies

$$-\Delta u(\mathbf{x}) = \int_\Omega [f(\boldsymbol{\xi}) - f(\mathbf{x})]\,(-\Delta g)\,\mathrm{d}\boldsymbol{\xi} - f(\mathbf{x})\,\Delta \int_\Omega g \mathrm{d}\boldsymbol{\xi} = -f(\mathbf{x})\,\Delta \int_\Omega g(\boldsymbol{\xi}, \mathbf{x}) \mathrm{d}\boldsymbol{\xi},$$

so that it remains only to show that $-\Delta \int_\Omega g \mathrm{d}\boldsymbol{\xi} = 1$. Choose $K_R(\mathbf{z})$ so that $\mathbf{x} \in K_R(\mathbf{z}) \subset \Omega$. The Green function has the form (2.10): $g = s + \Phi$. The first two terms in

$$\int_\Omega g(\boldsymbol{\xi}, \mathbf{x})\,\mathrm{d}\boldsymbol{\xi} = \int_{\Omega \setminus K_R(\mathbf{z})} g(\boldsymbol{\xi}, \mathbf{x})\,\mathrm{d}\boldsymbol{\xi} + \int_{K_R(\mathbf{z})} \Phi(\boldsymbol{\xi}, \mathbf{x})\,\mathrm{d}\boldsymbol{\xi} + \int_{K_R(\mathbf{z})} s(\boldsymbol{\xi}, \mathbf{x})\,\mathrm{d}\boldsymbol{\xi}$$

are harmonic in $K_R(\mathbf{z})$, so that $-\Delta \int_{K_R(\mathbf{z})} s(\boldsymbol{\xi}, \mathbf{x}) \mathrm{d}\boldsymbol{\xi} = 1$ is what has to be proved.

Let $\sigma(r)$ be defined by $s(\boldsymbol{\zeta}, \mathbf{x}) = \sigma(|\boldsymbol{\zeta} - \mathbf{x}|)$ (cf. (2.4)). For fixed $r > 0$ set

$$v(\mathbf{x}) := \frac{1}{\omega_n r^{n-1}} \int_{\partial K_r(\mathbf{z})} s(\boldsymbol{\xi}, \mathbf{x})\,\mathrm{d}\Gamma_\xi\,. \tag{3.14c}$$

For all $\mathbf{x} \notin \partial K_r(\mathbf{z})$ (i.e., $|\mathbf{x} - \mathbf{z}| \neq r$) v is harmonic, since $s(\boldsymbol{\zeta}, \mathbf{x})$ is nonsingular on $\partial K_r(\mathbf{z})$ and satisfies $-\Delta_x s(\boldsymbol{\zeta}, \mathbf{x}) = 0$. Since $s(\cdot, \mathbf{x})$ is harmonic in $K_r(\mathbf{z})$ for $r < |\mathbf{z} - \mathbf{x}|$, the mean-value property (2.13) holds, which can now be written

$$v(\mathbf{x}) = s(\mathbf{z}, \mathbf{x}) = \sigma(|\mathbf{z} - \mathbf{x}|) \qquad \text{for } |\mathbf{z} - \mathbf{x}| > r. \tag{3.14d}$$

Using Exercise 3.4b, we see that $v(\mathbf{x})$ is continuous in $\mathbb{R}^n$, so that we also have

$$v(\mathbf{x}) = \sigma(r) \qquad \text{for } |\mathbf{z} - \mathbf{x}| = r. \tag{3.14e}$$

Thus v is harmonic in $K_r(\mathbf{x})$ with the constant boundary values (3.14e). The unique solution is therefore

$$v(\mathbf{x}) = \sigma(r) \qquad \text{for } |\mathbf{z} - \mathbf{x}| \leq r\,. \tag{3.14f}$$

The equations (3.14c,d,f) yield

$$\int_{\partial K_r(\mathbf{z})} s(\boldsymbol{\xi}, \mathbf{x})\,\mathrm{d}\Gamma_\xi = \omega_n\, r^{n-1}\, \sigma(\max\{r, |\mathbf{z} - \mathbf{x}|\})$$

and then, since $0 < |\mathbf{z} - \mathbf{x}| < R$,

$$\begin{aligned}
&\int_{K_r(\mathbf{z})} s(\boldsymbol{\xi}, \mathbf{x})\,\mathrm{d}\boldsymbol{\xi} = \int_0^R \int_{\partial K_r(\mathbf{z})} s(\boldsymbol{\xi}, \mathbf{x})\,\mathrm{d}\Gamma_\xi\,\mathrm{d}r \\
&= \omega_n \int_0^{|\mathbf{z}-\mathbf{x}|} r^{n-1} \sigma(|\mathbf{z} - \mathbf{x}|)\,\mathrm{d}r + \omega_n \int_{|\mathbf{z}-\mathbf{x}|}^R r^{n-1} \sigma(r)\,\mathrm{d}r \\
&= \omega_n \frac{|\mathbf{z} - \mathbf{x}|^n}{n} \sigma(|\mathbf{z} - \mathbf{x}|) + \omega_n \frac{r^n}{n} \sigma(r)\Big|_{|\mathbf{z}-\mathbf{x}|}^R - \omega_n \int_{|\mathbf{z}-\mathbf{x}|}^R \frac{r^n}{n}\,\sigma'(r)\,\mathrm{d}r
\end{aligned}$$

$$= \frac{\omega_n}{n} R^n \sigma(R) - \omega_n \int_{|\mathbf{z}-\mathbf{x}|}^{R} \frac{r^n}{n} \left(-\frac{r^{1-n}}{\omega_n}\right) \mathrm{d}r \quad = \quad \frac{\omega_n}{n} R^n \sigma(R) + \int_{|\mathbf{z}-\mathbf{x}|}^{R} \frac{r}{n} \mathrm{d}r$$

$$= \frac{\omega_n}{n} R^n \sigma(R) + \frac{R^2}{2n} - \frac{|\mathbf{z}-\mathbf{x}|^2}{2n}.$$

From this we see that $-\Delta \int_{K_R(\mathbf{z})} s(\boldsymbol{\zeta}, \mathbf{x}) \mathrm{d}\boldsymbol{\zeta} = 1$. ∎

Using to Lemma 3.3, Theorems 3.16 and 3.18 prove the following.

Theorem 3.19. *Under the same assumptions as for Theorems 3.16 and 3.18 equation (3.6) gives a representation for the classical solution of the boundary-value problem (3.1a,b).*

Remark 3.20. If the boundary Γ is sufficiently smooth, the statement of Theorem 3.19 can be strengthened to $u \in C^{2+\lambda}(\overline{\Omega})$. This results from the Schauder theory mentioned in Theorem 9.22.

Finally, we return to the negative result of Theorem 3.13. According to Theorem 6.12, Theorem 3.13 is equivalent to the following one.

Theorem 3.21. *The solution u does not depend continuously on f, if the $C^0(\overline{\Omega})$-norm (2.21) is used for f, and the $C^2(\overline{\Omega})$-norm from (3.12) is used for u.*

Proof. Let $\Omega = K_1(0) \subset \mathbb{R}^2$ and $\varphi = 0$. The disk Ω is a normal region for which the Green function is known (cf. Theorem 3.22). The functions

$$f_n(\mathbf{x}) = \frac{x_2^2 - x_1^2}{r^2} \rho_n(r), \quad r := |\mathbf{x}|, \qquad \rho_n(r) := \min\left\{nr, 1/|\log \tfrac{r}{2}|\right\},$$

are not only continuous, but also Lipschitz-continuous: $f \in C^{0,1}(\overline{\Omega})$. By Theorem 3.18 there exist solutions $u^n \in C^2(\overline{\Omega})$ of $-\Delta u^n = f_n$ in Ω, with $u^n = 0$ on Γ.

The functions $f_n \in C^0(\overline{\Omega})$ are uniformly bounded: $\|f_n\|_\infty = 1/\log 2$ (the maximum is taken on the boundary $r = 1$). By Theorem 3.18 we have

$$u^n(\mathbf{x}) = \int_\Omega g(\boldsymbol{\xi}, \mathbf{x}) f_n(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi}.$$

Since $f_n(\mathbf{0}) = 0$, equation (3.14b) shows that

$$u^n_{x_1 x_1}(\mathbf{0}) = \int_\Omega g_{x_1 x_1}(\boldsymbol{\xi}, \mathbf{0}) \, f_n(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi}$$
$$= \int_\Omega \Phi_{x_1 x_1}(\boldsymbol{\xi}, \mathbf{0}) \, f_n(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi} + \int_\Omega s_{x_1 x_1}(\boldsymbol{\xi}, \mathbf{0}) \, f_n(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi},$$

where $g = \Phi + s$. The first integral is bounded since $\Phi \in C^2(\overline{\Omega})$. The derivative of the singularity function is $s_{x_1 x_1}(\boldsymbol{\xi}, \mathbf{0}) = (\xi_1^2 - \xi_2^2)/|\boldsymbol{\xi}|^4$. For the special choice of f_n, the second integral is equal to

$$I_n := \int_\Omega \frac{\left[\xi_1^2 - \xi_2^2\right]^2}{|\boldsymbol{\xi}|^6} \rho_n(|\boldsymbol{\xi}|)\,\mathrm{d}\boldsymbol{\xi} = \int_0^1 \left[\int_{\partial K_r(0)} \frac{[\xi_1^2 - \xi_2^2]^2}{|\boldsymbol{\xi}|^5}\mathrm{d}\Gamma\right] \frac{\rho_n(r)}{r}\mathrm{d}r$$

The surface integral $K := \int_{\partial K_r(0)} [\xi_1^2 - \xi_2^2]^2\, |\boldsymbol{\xi}|^{-5}\,\mathrm{d}\Gamma > 0$ does not depend on $r \in (0,1]$ so that the second integral takes on the form $I_n := K\int_0^1 r^{-1}\rho_n(r)\mathrm{d}r$. Since $\int_\varepsilon^1 \left[r|\log(\frac{r}{2})|\right]^{-1}\mathrm{d}r$ diverges as $\varepsilon \to 0$, we deduce $I_n \to \infty$ as $n \to \infty$. Since $\|u^n\|_{C^2(\overline{\Omega})} \geq \left|u^n_{x_1x_1}(0)\right|$, it follows that the map $f \mapsto u$ is not bounded, and thus not continuous. ∎

3.4 The Green Function for the Ball

Theorem 3.22. *The Green function for the ball $K_R(\mathbf{y})$ is given by the function in (2.12).*

Proof. According to Exercise 2.10b, the function γ defined in (2.12) is a fundamental solution in $\Omega = K_R(\mathbf{y})$. It remains to show that $\gamma(\mathbf{x},\mathbf{y})$ vanishes for $\mathbf{x} \in \partial\Omega$ and $\mathbf{y} \in \Omega$. This fact follows from combining the results in Exercise 2.10a,c. ∎

In the case $n = 2$ the plane $\mathbb{R}^2$ can be identified with $\mathbb{C}$ by the correspondence $(x,y) \leftrightarrow z = x + \mathrm{i}y$. The following considerations are based on the next exercise.

Exercise 3.23. Let the map $\Phi : z = x + \mathrm{i}y \in \Omega \mapsto \zeta = \xi + \mathrm{i}\eta = \Phi(z) \in \Omega'$ be conformal, i.e., holomorphic with nonvanishing derivative. Show

$$\Delta_z u(\Phi(z)) = |\Phi'|^2\, \Delta_\zeta u(\zeta), \qquad \Phi' = \xi_x + \mathrm{i}\eta_x \tag{3.15}$$

for $u \in C^2(\Omega')$.

Equation (3.15) shows, in particular, that a conformal coordinate transformation maps harmonic functions into harmonic functions. An arbitrary simply connected region with at least two boundary points can, by the Riemann mapping theorem (cf. Hurwitz–Courant [159, §III.5]), be mapped by a conformal mapping $\Phi_{z_0} : z \in \Omega \mapsto \Phi_{z_0}(z) \in K_1(0)$ onto the unit disk such that $\Phi_{z_0}(z_0) = 0$ for any given $z_0 \in \Omega$. Let $g(\zeta,\zeta')$ be the Green function for $K_1(0)$ (cf. Salamon [247, §5]). One may check that $G(z,z_0) := g(\Phi_{z_0}(z),0)$ is again a fundamental solution. Now $z \in \partial\Omega$ implies $\Phi_{z_0}(z) \in \partial K_1(0)$, i.e., $G(z,z_0) = 0$. Thus $G(z,z_0)$ is the Green function in Ω. This proves the next statement.

Theorem 3.24. *Let $\Omega \subset \mathbb{R}^2$ be simply connected with at least two boundary points. Then there exists a Green function of the first kind for Ω.*

The explicit forms of various Green functions can be found, for example, in the book by Wloka [308, Exercises 21.1–21.8]. Of numerical interest might be the fact that with conformal mappings one may remove corners which are disturbing (e.g., re-entrant corners) (cf. Gladwell–Wait [119, page 70]).

Example 3.25. Let Ω be the L-shaped region in Example 2.4. Choose $\Phi(z) = z^{2/3} : \Omega \mapsto \Omega'$. Then Φ is conformal in Ω. The sides of the angle $\Gamma_0 \subset \partial\Omega$ (cf. Fig. 2.1) are mapped into a single line segment, so that Ω' has no reentrant corners. The Poisson equation $\Delta u = f$ in Ω corresponds to the equation $\Delta v(\zeta) = \frac{9}{4}|\zeta| f(\zeta)$ in Ω'.

A generalisation of the latter statements to the case of more than two variables is impossible. For general $n \geq 3$, only the mappings from Exercise 2.6 and the Kelvin transformation – the reflection at a sphere – are invariant with respect to the Laplace equation. More precisely, the following theorem holds. A proof can be found in Walter [302, pages 17f].

Theorem 3.26. *$T : \mathbf{x} \in \mathbb{R}^n \backslash \{\mathbf{0}\} \mapsto \mathbf{x}/|\mathbf{x}|^2 \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ is the reflection at the unit sphere. It satisfies $T = T^{-1}$. Let Ω be a domain with $\mathbf{0} \notin \Omega \subset \mathbb{R}^n$. Let $\hat{\Omega}$ be the image under T. For a function $u \in C^2(\Omega)$, define $v \in C^2(\hat{\Omega})$ by*

$$v(\mathbf{x}) := |\mathbf{x}|^{2-n} u(T\mathbf{x}) \qquad (\mathbf{x} \in \hat{\Omega}).$$

Δv and Δu are connected via

$$\Delta v(\mathbf{x}) = |\mathbf{x}|^{-2-n} \Delta u(T\mathbf{x}) \qquad (\mathbf{x} \in \hat{\Omega}).$$

In particular, $-\Delta u = 0$ implies $-\Delta v = 0$.

3.5 The Neumann Boundary-Value Problem

In (3.1b) and in (2.1b) the boundary values $u = \varphi$ were given on Γ. These so-called Dirichlet conditions or 'boundary conditions of the first kind' are not the only possibility. An alternative is the *Neumann condition*

$$\frac{\partial}{\partial n} u(\mathbf{x}) = \varphi(\mathbf{x}) \qquad \text{on } \Gamma. \tag{3.16}$$

In physics this *second boundary condition*, as it is also called, occurs more frequently than the Dirichlet condition. For example, if u is the velocity potential of a gas, then $\partial u/\partial n = 0$ means that the gas can only move tangentially at the boundary.

Remark 3.27. The Neumann boundary-value problem $Lu = f$ in Ω and $\partial u/\partial n = \varphi$ on Γ is not uniquely solvable if the differential operator L contains no absolute term[2]. If a solution u does exist, then $u + c$, with c any constant, is also a solution.

Proof. Without absolute term L only contains derivatives of first or higher order. Therefore $Lc = 0$ holds for the constant function c. Since also $\frac{\partial}{\partial n} c = 0$, c is a solution of the homogeneous Neumann problem. Let u be the solution of the original inhomogeneous problem. Then $u + c$ is again a solution. ∎

[2] This is the term $a(\mathbf{x})$ in (1.18).

In particular, Remark 3.27 applies to the Laplace operator. In general, non-uniqueness of the solution implies that the problem is not solvable for all right-hand sides. The next theorem describes the requirements for f and φ.

Theorem 3.28. *Let Ω be a normal region. The Poisson equation $-\Delta u = f$ with the Neumann boundary condition (3.16) is only solvable if*

$$\int_\Gamma \varphi(\boldsymbol{\xi}) \mathrm{d}\Gamma_\xi + \int_\Omega f(\mathbf{x}) \mathrm{d}\mathbf{x} = 0. \tag{3.17}$$

If a solution u does exist, then $u + c$, with c any constant, is also a solution.

Proof. Repeat the proof of Lemma 2.16 for $-\Delta u = f$. ∎

Later, in Example 7.30, we shall show that the Neumann boundary-value problem for the Poisson equation has a solution if and only if (3.17) is satisfied, and that two solutions can differ only by a constant.

In the representation (3.4) both the values $u(\boldsymbol{\xi})$, $\boldsymbol{\xi} \in \Gamma$, and the normal derivative $\partial u/\partial n$ occur. The Green function of the first kind was chosen in such a way that $g(\boldsymbol{\xi}, \mathbf{x}) = 0$ for $\boldsymbol{\xi} \in \Gamma$. In the case of the second boundary conditions (3.16) one makes the assumption that $\partial \gamma(\boldsymbol{\xi}, \mathbf{x})/\partial n_\xi = c$ (c: constant), i.e.,

$$\frac{\partial}{\partial n_\xi} \Phi(\boldsymbol{\xi}, \mathbf{x}) = c - \frac{\partial}{\partial n_\xi} s(\boldsymbol{\xi}, \mathbf{x})$$

for $\Phi = \gamma - s$. Then Corollary 2.9 with $u \equiv 1$ and $\gamma = s$ gives

$$\int_\Gamma \frac{\partial}{\partial n_\xi} \Phi(\boldsymbol{\xi}, \mathbf{x}) \mathrm{d}\Gamma_\xi = cL + 1 \qquad \text{with } L := \int_\Gamma \mathrm{d}\Gamma\,.$$

Since Φ must be harmonic (i.e., $f := -\Delta\Phi = 0$), from equation (3.17) we see that $cL + 1 = 0$ is a necessary condition for the existence of Φ. Thus the condition on the *Green function of the second kind* for the potential equation is

$$\frac{\partial}{\partial n_x} \gamma(\boldsymbol{\xi}, \mathbf{x}) = -1 \,/ \int_\Gamma \mathrm{d}\Gamma\,.$$

Thus the term $\int_\Gamma u\, \partial\gamma/\partial n \,\mathrm{d}\Gamma$ in (3.4) becomes $\mathrm{const} \cdot \int_\Gamma u \mathrm{d}\Gamma$. Since u is only determined up to a constant (cf. Theorem 3.28), one can fix this constant with the additional condition $\int_\Gamma u \mathrm{d}\Gamma = 0$. This gives the following result, if we write g for γ:

$$u(\mathbf{x}) = \int_\Omega f(\boldsymbol{\xi})\, g(\boldsymbol{\xi}, \mathbf{x})\, \mathrm{d}\boldsymbol{\xi} + \int_\Gamma \varphi(\boldsymbol{\xi})\, g(\boldsymbol{\xi}, \mathbf{x})\, \mathrm{d}\Gamma_\xi.$$

The Green function of the second kind for the ball $K_R(0) \subset \mathbb{R}^3$ can be found in Leis [189, page 79].

3.6 The Integral Equation Method

In the representation (3.4) of the Poisson solution one may choose the singularity function $\gamma := s$. If in addition one imposes the given Neumann data (3.16), one obtains

$$u(\mathbf{x}) = -\int_{\Gamma} k(\mathbf{x},\boldsymbol{\xi})\, u(\boldsymbol{\xi})\, \mathrm{d}\Gamma_\xi + g(\mathbf{x}) \qquad \text{for } \mathbf{x} \in \Omega \tag{3.18}$$

with the kernel function $k(\mathbf{x},\boldsymbol{\xi}) := \partial s(\boldsymbol{\xi},\mathbf{x})/\partial n_\xi$ and the following function g:

$$g(\mathbf{x}) := g_1(\mathbf{x}) + g_2(\mathbf{x}),\ g_1(\mathbf{x}) := \int_{\Gamma} s(\boldsymbol{\xi},\mathbf{x})\varphi(\boldsymbol{\xi})\mathrm{d}\Gamma_\xi\,,\ g_2(\mathbf{x}) := \int_{\Omega} s(\boldsymbol{\xi},\mathbf{x}) f(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi}\,.$$

The right-hand side in equation (3.18) with the unknown boundary value $u(\boldsymbol{\xi})$, $\boldsymbol{\xi} \in \Gamma$, can be used as an ansatz solution:

$$\Phi(\mathbf{x}) = -\int_{\Gamma} k(\mathbf{x},\boldsymbol{\xi})\, u(\boldsymbol{\xi})\, \mathrm{d}\Gamma_\xi + g(\mathbf{x}). \tag{3.19}$$

The first term on the right-hand side of (3.19) is called the *double-layer potential* (dipole potential); g_1 is the *single-layer potential*, while g_2 is a *volume potential*.

For each $u \in C^0(\Gamma)$ the Φ in (3.19) is a solution of the Poisson equation (3.1a) in Ω. However, Φ is also defined for an argument $\mathbf{x}$ in the *exterior domain* $\Omega_+ := \mathbb{R}^n \backslash \overline{\Omega}$. A closer look at the kernel function $k(\mathbf{x},\zeta)$ shows that it is in fact only weakly singular for the case of smooth boundaries Γ. Thus Φ is also defined for $\mathbf{x} \in \Gamma$. The function Φ which is now defined on all $\mathbb{R}^n$ is not continuous at points of the boundary Γ. At $\mathbf{x}_0 \in \Gamma$ there exists both an interior limit $\Phi_-(\mathbf{x}_0)$ for $\Omega \ni \mathbf{x} \to \mathbf{x}_0$ and an exterior limit $\Phi_+(\mathbf{x}_0)$ for $\Omega_+ \ni \mathbf{x} \to \mathbf{x}_0$. In addition we have the third function value $\Phi(\mathbf{x}_0)$ of (3.19). Their connection is given by the following jump discontinuity relation (cf. Hackbusch [136, Theorem 8.2.8]):

$$\Phi_+(\mathbf{x}_0) - \Phi_-(\mathbf{x}_0) = -u(\mathbf{x}_0) \qquad \text{for } \mathbf{x}_0 \in \Gamma, \tag{3.20a}$$

$$\Phi_+(\mathbf{x}_0) + \Phi_-(\mathbf{x}_0) = 2\,\Phi(\mathbf{x}_0) \qquad \text{for } \mathbf{x}_0 \in \Gamma, \tag{3.20b}$$

In order that the ansatz (3.19) does indeed give the solution u in (3.18), the boundary value Φ_-, continued from the interior, must agree with the function u which is put in the integral: $\Phi_- = u$. Now one can solve equation (3.20a,b) for Φ_-. This gives $\Phi_-(\mathbf{x}_0) = \Phi(\mathbf{x}_0) + \frac{1}{2}u(\mathbf{x}_0)$. The equation $\Phi_- = u$ thus leads to

$$u(\mathbf{x}) = 2\Phi(\mathbf{x}) = -2\int_{\Gamma} k(\mathbf{x},\boldsymbol{\xi}) u(\boldsymbol{\xi})\mathrm{d}\Gamma_\xi + 2g(\mathbf{x}) \qquad \text{for } \mathbf{x} \in \Gamma. \tag{3.21}$$

Equation (3.21) is called a *Fredholm integral equation of the second kind* for the unknown function $u \in C^0(\Gamma)$. The original Neumann boundary-value problem (3.1a), (3.16) and the integral equation (3.21) are equivalent in the following sense:

(a) If u is the solution of the Neumann boundary-value problem, then the boundary values, $u(\boldsymbol{\zeta})$, $\boldsymbol{\zeta} \in \Gamma$, satisfy the integral equation (3.21).

(b) If $u \in C^0(\Gamma)$ is a solution of the integral equation (3.21), then the expression (3.18) gives a solution of (3.1a) and (3.16) in the entire domain Ω.

The integral equation (3.21) is only one example of several possible formulations by integral equations, which differ by the underlying kernel function $k(\mathbf{x}, \boldsymbol{\xi})$. Since the kernel k is derived from the singularity function, this method is applicable to all differential equation with known singularity function.

A further example of an integral equation formulation will be mentioned in §7.5.

The transformation of a boundary-value problem into an integral equation, and the subsequent solution of the integral equation is referred to as the *integral equation method*. It allows, for example, a new approach to existence statements, in that one shows the solvability of (3.21).

The integral equation (3.21) can also be attacked numerically. As in the finite-element method described in Chapter 8, one replaces u by ansatz functions of an n-dimensional space V_n, e.g., by piecewise linear functions defined on a triangulation of the surface Γ (decomposition of Γ into plane or spherical triangles). Requiring equation (3.21) for all $\mathbf{x}_0$ corresponding to the n corners of the triangles, we obtain the collocation method. Integration of equation (3.21) weighted by the basis functions of V_n leads to the Galerkin method, which corresponds directly to the discretisation procedure in §8. Both discretisation methods are called *boundary-element method* (BEM).

Unlike the finite-element method in §8, the boundary-element method leads to cumbersome integrals (surface integration, singular integrands) which nevertheless can be treated by efficient numerical methods. Another difficulty is the fact that the matrices produced by the boundary-element method are fully populated and require special numerical methods (cf. Hackbusch [140, §10] and [142]).

One can find references to the integral equation method and the boundary element method, e.g., in Sauter–Schwab [250], Hsiao–Wendland [157], Kreß [176], Hackbusch [136, §§7-9], and Steinbach [272].

Chapter 4
Difference Methods for the Poisson Equation

Abstract The difference method replaces the derivatives by difference quotients. **Section 4.1** describes the difference method applied to the one-dimensional Poisson equation $-u'' = f$. This simple example is chosen to show the generation of the discrete system of equations. The difference equations are complemented by the Dirichlet boundary condition. The equations of the resulting linear system correspond to the inner grid points, while the boundary data appear in the right–hand side of the system.

The generalisation to two spatial dimensions in **Section 4.2** leads to a regular grid and the five-point discretisation of the Poisson equation. The corresponding linear system is defined. Moreover, the notation of difference stars is explained.

Typical properties of the matrices of the discrete linear systems are described in **Section 4.3**. M-matrices, matrix norms, and positive-definite matrices are introduced. The property $A^{-1} > 0$ of M-matrices is the discrete analogue of the positive sign of Green's function. Another important property is the diagonal dominance. For proving the criteria for M-Matrices we use the Gershgorin circles.

Section 4.4 is devoted to the properties of the matrix L_h corresponding to the five-point discretisation of the Poisson equation. In particular, concrete norm estimates of L_h and L_h^{-1} are given. The maximum-minimum principle of the Laplace equation has a discrete analogue. Also the representation of the Poisson solution in Chapter 3 has discrete counterparts. For this purpose we introduce and analyse the discrete Green function. The logarithmic singularity in the continuous case corresponds to the upper bound $\mathcal{O}(\log h)$ (h: step size) in the discrete case (cf. Lemma 4.42).

The convergence of the discrete solution u_h to the solution u of the Poisson equation is studied in **Section 4.5**. The typical proof uses that consistency and stability implies convergence. The order of consistency can be increased by choosing difference schemes of higher order as discussed in **Section 4.6**.

The discretisation of the Neumann boundary condition is the subject of **Section 4.7**. In this case there are several discretisations. The corresponding matrix properties and the convergence behaviour are analysed. As in the continuous case the solution of the discrete Neumann problem requires a side condition. Since floating-

W. Hackbusch, *Elliptic Differential Equations*, Springer Series
in Computational Mathematics 18, DOI 10.1007/978-3-662-54961-2_4

point errors may lead to a perturbation of the side condition, the influence of this perturbation is studied. In the Neumann case the proof of stability is more cumbersome. For this purpose, estimates of the discrete Green function and the discrete singularity function are derived in §4.7.4.

The previous statements refer to the square $\Omega = (0,1) \times (0,1)$. The generalisation to arbitrary domains is described in **Section 4.8**. Two discretisation schemes are discussed: the Shortley–Weller method and linear interpolation close to the boundary. In both cases convergence is analysed.

4.1 Introduction: The One-Dimensional Case

Before developing difference methods for the partial differential Poisson equation, let us first recall the discretisation of ordinary differential equations. The equation

$$a(x)u''(x) + b(x)u'(x) + c(x)u(x) = f(x)$$

can be supplemented with initial conditions $u(x_1) = u_1$, $u'(x_1) = u'_1$ or with boundary conditions $u(x_1) = u_1$, $u(x_2) = u_2$. The ordinary initial value problems correspond to the hyperbolic and parabolic initial value problems, while an ordinary boundary-value problem may be viewed as an elliptic boundary-value problem in one variable. In particular one can view

$$-u''(x) = f(x) \qquad \text{in } x \in (0,1)\,, \tag{4.1a}$$

$$u(0) = \varphi_0, \qquad u(1) = \varphi_1, \tag{4.1b}$$

as the one-dimensional Poisson equation $-\Delta u = f$ in the domain $\Omega = (0,1)$ with Dirichlet conditions on the boundary $\Gamma = \{0,1\}$.

Difference methods are characterised by the fact that derivatives are replaced by difference quotients (divided differences), in the following called, for short, 'differences'. The first derivative $u'(x)$ can be approximated by several (so-called 'first') differences, for example, by

- the *forward* or *right difference*

$$\left(\partial^+ u\right)(x) := \left[\,u(x+h) - u(x)\right]/h,$$

- the *backward* or *left difference*

$$\left(\partial^- u\right)(x) := \left[\,u(x) - u(x-h)\right]/h$$

  (right and left differences are also called *one-sided differences*),

- the *symmetric difference*

$$\left(\partial^0 u\right)(x) := \left[\,u(x+h) - u(x-h)\right]/\left(2h\right),$$

  where $h > 0$ is called the *step size*.

An obvious second difference for $u''(x)$ is

$$\left(\partial^-\partial^+ u\right)(x) := \left[\, u(x+h) - 2u(x) + u(x-h) \,\right]/h^2. \tag{4.2}$$

One also calls $\partial^+, \partial^-, \partial^0$, and $\partial^-\partial^+$ *difference operators*. $\partial^-\partial^+$ may be viewed as the product $\partial^- \circ \partial^+$ or as $\partial^+ \circ \partial^-$, i.e., $(\partial^+\partial^-)u(x) = \partial^+(\partial^- u(x))$.

Lemma 4.1. *Let* $[x-h, x+h] \subset \overline{\Omega}$. *Then*

$$\begin{aligned}
\left(\partial^+ u\right)(x) &= u'(x) + hR \quad &&\text{with } |R| \le \frac{1}{2}\,\|u\|_{C^2(\overline{\Omega})}\,, \quad \text{if } u \in C^2(\overline{\Omega}),\\
\left(\partial^0 u\right)(x) &= u'(x) + h^2 R \quad &&\text{with } |R| \le \frac{1}{6}\,\|u\|_{C^3(\overline{\Omega})}\,, \quad \text{if } u \in C^3(\overline{\Omega}),\\
\left(\partial^-\partial^+ u\right)(x) &= u''(x) + h^2 R \quad &&\text{with } |R| \le \frac{1}{12}\,\|u\|_{C^4(\overline{\Omega})}\,, \quad \text{if } u \in C^4(\overline{\Omega}).
\end{aligned} \tag{4.3}$$

Proof. We give the proof only for (4.3). Inserting Taylor's formula

$$u(x \pm h) = u(x) \pm hu'(x) + h^2 u''(x)/2 \pm h^3 u'''(x)/6 + h^4 R_4$$

with

$$R_4 = \frac{h^{-4}}{3!}\int_x^{x\pm h} (x \pm h - \xi)^3 u''''(\xi)\,\mathrm{d}\xi = \frac{1}{4!}u''''(\xi \pm \vartheta h) \tag{4.4}$$

and $\vartheta \in (0,1)$ into (4.2), the result is (4.3) because

$$R = \frac{u''''(x+\vartheta_1 h) - u''''(x-\vartheta_2 h)}{24}\,. \qquad \blacksquare$$

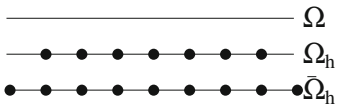


Fig. 4.1 Grid for $h = 1/8$.

We replace $\Omega = (0,1)$ and $\overline{\Omega} = [\,0,1\,]$ by the *grids*

$$\begin{aligned}
\Omega_h &= \{h, 2h, \ldots, (n-1)\,h = 1-h\}\,, \\
\overline{\Omega}_h &= \{0, h, 2h, \ldots, 1-h, 1\}
\end{aligned} \tag{4.5}$$

of *step size* $h = 1/n$ (cf. Figure 4.1). For $x \in \Omega_h$, $\partial^-\partial^+ u(x)$ only contains the values of u at $x, x \pm h \in \overline{\Omega}_h$. Under the assumption that the solution u of (4.1a,b) belongs to $C^4(\overline{\Omega})$, (4.3) yields the equations

$$-\partial^-\partial^+ u(x) = f(x) + \mathcal{O}(h^2) \qquad \text{for } x \in \Omega_h.$$

Neglecting the remainder term $\mathcal{O}(h^2)$, we obtain

$$-\partial^-\partial^+ u_h(x) = \frac{-u_h(x-h) + 2u_h(x) - u_h(x+h)}{h^2} = f(x) \quad (x \in \Omega_h). \tag{4.6a}$$

These are $n-1$ equations in $n+1$ unknowns $\{u_h(x) : x \in \overline{\Omega}_h\}$. The two missing equations are supplied by boundary conditions (4.1b):

$$u_h(0) = \varphi_0, \qquad u_h(1) = \varphi_1. \tag{4.6b}$$

u_h is a *grid function* defined on $\overline{\Omega}_h$. Its restriction to Ω_h yields the vector

$$u_h = \begin{bmatrix} u_h(h) \\ u_h(2h) \\ \vdots \\ u_h(1-h) \end{bmatrix}.$$

If in (4.6a) one eliminates the components $u_h(0)$ and $u_h(1)$ with the aid of equation (4.6b), one gets the system of equations

$$L_h\, u_h = q_h \tag{4.7a}$$

with

$$L_h = h^{-2}\begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix}, \quad q_h = \begin{bmatrix} f(h) + h^{-2}\varphi_0 \\ f(2h) \\ f(3h) \\ \vdots \\ f(1-2h) \\ f(1-h) + h^{-2}\varphi_1 \end{bmatrix}. \tag{4.7b}$$

Remark 4.2. Note that the vector q_h contains the boundary values φ_0 and φ_1 together with values of f. u_h can also be interpreted as the solution of the difference equations $-\partial^-\partial^+ u_h(x) = q_h$ with homogeneous boundary values $u_h(0) = u_h(1) = 0$.

4.2 The Five-Point Formula

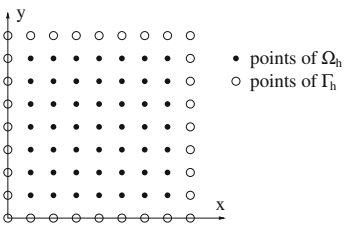


Fig. 4.2 Two-dimensional grid.

First we select the unit square

$$\Omega = \{(x, y) : 0 < x < 1,\ 0 < y < 1\}$$

as the fundamental domain. More general domains will be discussed in Section 4.8. In the discretisation process Ω is replaced by the grid

$$\Omega_h = \left\{(x, y) \in \Omega : \tfrac{x}{h}, \tfrac{y}{h} \in \mathbb{Z}\right\} \tag{4.8a}$$

for the equidistant step size $h = 1/n$ ($n \in \mathbb{N}$, cf. Figure 4.2). The discrete boundary points form the set

$$\Gamma_h = \{(x,y) \in \Gamma : x/h, y/h \in \mathbb{Z}\}\,. \tag{4.8b}$$

As in (4.5) we set

$$\overline{\Omega}_h = \Omega_h \cup \Gamma_h = \left\{(x,y) \in \overline{\Omega} : x/h, y/h \in \mathbb{Z}\right\}. \tag{4.8c}$$

In the Poisson equation

$$-\Delta u = -u_{xx} - u_{yy} = f \qquad \text{in } \Omega, \tag{4.9a}$$
$$u = \varphi \qquad \text{on } \Gamma \tag{4.9b}$$

the second derivatives u_{xx} and u_{yy} can each be replaced by the respective differences (4.3) in the x and y directions:

$$\begin{aligned} -\left(\Delta_h u\right)(x,y) &:= -\left(\partial_x^- \partial_x^+ + \partial_y^- \partial_y^+\right) u(x,y) \\ &= h^{-2}\left[\,4u(x,y) - u(x-h,y) - u(x+h,y) - u(x,y-h) - u(x,y+h)\,\right]. \end{aligned} \tag{4.10}$$

Since on the right-hand side of (4.10) the function u is evaluated at five points, Δ_h is also called the *five-point formula*. The discretisation of the boundary-value problem (4.9a,b) using Δ_h leads to the difference equations

$$-\Delta_h u_h(\mathbf{x}) = f(\mathbf{x}) \qquad \text{for } \mathbf{x} \in \Omega_h, \tag{4.11a}$$
$$u_h(\mathbf{x}) = \varphi(\mathbf{x}) \qquad \text{for } \mathbf{x} \in \Gamma_h. \tag{4.11b}$$

Through (4.11a,b) one obtains one equation per grid point $\mathbf{x} = (x,y) \in \overline{\Omega}_h$, and hence one equation per component of the grid function $u_h = (u_h(\mathbf{x}))_{\mathbf{x}\in\overline{\Omega}_h}$. Except in the one-dimensional case, there exists no natural arrangement of grid points, thus one cannot immediately obtain a matrix representation as in (4.7b). The only natural indexing of u_n is that through $\mathbf{x} \in \Omega_h$ or the pair $(i,j) \in \mathbb{N}^2$ with $\mathbf{x} = (x,y) = (ih, jh)$. Let the matrix elements be given by

$$L_{h,\mathbf{x}\boldsymbol{\xi}} := L_{\mathbf{x}\boldsymbol{\xi}} := \begin{cases} -h^{-2} \text{ if } \mathbf{x} \in \Omega_h,\ \boldsymbol{\xi} \in \overline{\Omega}_h,\ \mathbf{x} - \boldsymbol{\xi} = \left(\begin{smallmatrix} 0 \\ \pm h \end{smallmatrix}\right),\ \text{or } \mathbf{x} - \boldsymbol{\xi} = \left[\begin{smallmatrix} \pm h \\ 0 \end{smallmatrix}\right], \\ 4\,h^{-2} \text{ if } \mathbf{x} = \boldsymbol{\xi} \in \Omega_h, \\ 0 \qquad \text{otherwise.} \end{cases} \tag{4.12}$$

For $\mathbf{x} = \boldsymbol{\xi}$, $L_{\mathbf{x}\mathbf{x}}$ is a diagonal element; in the first case of (4.12) we say that $\mathbf{x}$ and $\boldsymbol{\xi}$ are *neighbours*. If one eliminates the components $u_h(\mathbf{x})$, $\mathbf{x} \in \Gamma_h$, with the aid of equation (4.11b), then equation (4.11a) assumes the following form:

$$\sum_{\boldsymbol{\xi}\in\Omega_h} L_{\mathbf{x}\boldsymbol{\xi}}\, u_h(\boldsymbol{\xi}) = q_h(\mathbf{x}) \qquad \text{for } \mathbf{x} \in \Omega_h, \tag{4.13a}$$

where

$$q_h := f_h + \varphi_h \text{ with } \begin{cases} f_h(\mathbf{x}) := f(\mathbf{x}), \\ \varphi_h(\mathbf{x}) := -\sum\limits_{\boldsymbol{\xi}\in\Gamma_h} L_{\mathbf{x}\boldsymbol{\xi}}\, u_h(\boldsymbol{\xi}) = -\sum\limits_{\boldsymbol{\xi}\in\Gamma_h} L_{\mathbf{x}\boldsymbol{\xi}}\, \varphi(\boldsymbol{\xi}). \end{cases} \tag{4.13b}$$

For the proof split the sum

$$-\Delta_h u_h(\mathbf{x}) = \sum_{\boldsymbol{\xi}\in\overline{\Omega_h}} L_{\mathbf{x}\boldsymbol{\xi}} u_h(\boldsymbol{\xi})$$

into $\sum_{\boldsymbol{\xi}\in\Omega_h} \dots$ and $\sum_{\boldsymbol{\xi}\in\Gamma_h} \dots$. The second partial sum is $-\varphi_h$ in (4.13b) and is moved to the right-hand side of the equation.

Remark 4.3. f_h is the restriction of f to the grid Ω_h. For all far-boundary points we have $\varphi_h(\mathbf{x}) = 0$; here $\mathbf{x} \in \Omega_h$ is said to be a *far-boundary point* if all neighbours $\mathbf{x}\pm(0,h)$ and $\mathbf{x}\pm(h,0)$ belong to Ω_h. The other $\mathbf{x} \in \Omega_h$ are called *near-boundary points*. In the case of homogeneous boundary values $\varphi = 0$ we have $q_h = f_h$.

The system of equations (4.13a) can be expressed in the form (4.7a):

$$L_h u_h = q_h,$$

where the matrix

$$L_h = (L_{\mathbf{x}\boldsymbol{\xi}})_{\mathbf{x},\boldsymbol{\xi}\in\Omega_h} \tag{4.14}$$

and the grid functions $u_h = (u_h(\mathbf{x}))_{\mathbf{x}\in\Omega_h}$ and $q_h = (q_h(\mathbf{x}))_{\mathbf{x}\in\Omega_h}$ are described by their components (cf. (4.12), (4.13b)). Since the grid points of Ω_h are not (or not uniquely) ordered, we refer to the following notation.

Notation 4.4. Let I be a finite index set (not necessarily ordered). The vector space $\mathbb{R}^I$ consists of all I-tuples $(u_\alpha)_{\alpha\in I}$ with $u_\alpha \in \mathbb{R}$. Correspondingly $\mathbb{R}^{I\times I}$ is the vector space of all matrices $M = (M_{\alpha\beta})_{\alpha,\beta\in I}$. Note that most of the matrix properties (symmetry, positive definiteness, etc.) do not depend of the ordering of the indices.

A possible linear enumeration of indices $\mathbf{x} \in \Omega_h$ is lexicographical ordering

$$\begin{array}{llll} (h,h), & (2h,h), & (3h,h), & \dots (1-h,h), \\ (h,2h), & (2h,2h), & (3h,2h), & \dots (1-h,2h), \\ & & \vdots & \\ (h,1-h), & (2h,1-h), & (3h,1-h), & \dots (1-h,1-h). \end{array} \tag{4.15}$$

Generally, the point $\mathbf{x} = (x_1, \dots, x_d)$ precedes the point $\mathbf{y} = (y_1, \dots, y_d)$ in lexicographical order, if for some $j \in \{1, \dots, d\}$ the conditions $x_i = y_i$ (for all $i > j$) and $x_j < y_j$ hold. Each line in (4.15) corresponds to a so-called x-row in the grid Ω_h (cf. Figure 4.2). A vector u_h whose $(n-1)^2$ components are enumerated in the series (4.15) thus separates into $n-1$ blocks (so-called x-blocks). The block decomposition of the vectors generates a block decomposition of the matrix L_h, which is given in (4.16).

Exercise 4.5. (a) With the lexicographical numbering of grid points the matrix L_h has the form

$$L_h = h^{-2} \begin{bmatrix} T & -I & & \\ -I & T & -I & \\ & \ddots & \ddots & \ddots \\ & & -I & T & -I \\ & & & -I & T \end{bmatrix}, \quad T = \begin{bmatrix} 4 & -1 & & \\ -1 & 4 & -1 & \\ & \ddots & \ddots & \ddots \\ & & -1 & 4 & -1 \\ & & & -1 & 4 \end{bmatrix}, \tag{4.16}$$

where T is an $(n-1) \times (n-1)$ matrix and L_h contains $(n-1)^2$ blocks. I is the $(n-1) \times (n-1)$ identity matrix.

(b) Let Ω be the rectangle $\Omega = (0,a) \times (0,b) = \{(x,y) : 0 < x < a, 0 < y < b\}$. Let the step size h satisfy the conditions $a = nh$ and $b = mh$. Show that the discretisation (4.11a,b) in the corresponding grid Ω_h leads to a matrix which also has the form (4.16). But here L_h contains $(m-1)^2$ blocks of the size $(n-1)\times(n-1)$.

Another frequently used arrangement is the *chequer-board ordering* (or *red-black ordering*). The latter name corresponds to dividing the board pattern into 'red' and 'black' fields:

$$\begin{aligned} \Omega_h^r &:= \{(x,y) : (x+y)/h \ \text{ odd}\}, \\ \Omega_h^b &:= \{(x,y) : (x+y)/h \ \text{ even}\}. \end{aligned} \tag{4.17}$$

First one numbers the red squares $(x,y) \in \Omega_h^r$ lexicographically, and then those of Ω_h^b. The partition (4.17) induces a partition of vectors into 2 blocks and a partition of the matrix L_h into $2 \cdot 2 = 4$ blocks.

Exercise 4.6. With respect to the chequer-board ordering, the matrix L_h assumes the form

$$L_h = h^{-2} \left[\begin{array}{ccc|ccc} 4 & & & & & \\ & \ddots & & & A & \\ & & 4 & & & \\ \hline & & & 4 & & \\ & A^{\mathsf{T}} & & & \ddots & \\ & & & & & 4 \end{array}\right], \tag{4.18}$$

where, in general, A is a rectangular block matrix because for n even, Ω_h^r and Ω_h^b contain a different number of points.

The complete $(n-1)^2 \times (n-1)^2$ matrix L_h in (4.16) or (4.18) is needed neither for the theoretical investigation of the system of equations $L_h u_h = q_h$ nor for its numerical solution. All properties of L_h considered in the following are invariant with respect to re-numbering of the grid points. Even though numerical methods for the solution of $L_h u_h = q_h$ implicitly use an arrangement of grid points (with the exception of special algorithms for parallel computers), they never employ the complete $(n-1)^2 \times (n-1)^2$ matrix L_h. Every usable algorithm must take into account that L_h is *sparse*, i.e., it has substantially more zero than nonzero elements.

In the following we again return to indexing by $\mathbf{x} \in \Omega_h$.

The *difference operator* Δ_h in (4.10) is also described by the 'five-point star'

$$\Delta_h = h^{-2} \begin{bmatrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{bmatrix}. \tag{4.19}$$

The general definition of a difference star (with variable coefficients) reads

$$\begin{bmatrix} & \vdots & \\ & c_{-1,1}(x,y) \quad c_{0,1}(x,y) \quad c_{1,1}(x,y) & \\ \dots & c_{-1,0}(x,y) \quad c_{0,0}(x,y) \quad c_{1,0}(x,y) & \dots \\ & c_{-1,-1}(x,y) \quad c_{0,-1}(x,y) \quad c_{1,-1}(x,y) & \\ & \vdots & \end{bmatrix} u_h(x,y) \tag{4.20}$$

$$:= \sum\nolimits_{i,j} c_{ij}(x,y) u_h(x+ih, y+jh),$$

in which the zero coefficients have not been written out.

Attention. The star (4.19) does not represent a submatrix of L_h! The coefficients of the star appear in each row of L_h (coefficients do not appear if $(x+ih, y+jh)$ does not belong to Ω_h).

Remark 4.7. Note that the difference operator $D_h := -\Delta_h$ cannot be equated with the matrix L_h since Δ_h does not contain information on the type or place of the boundary conditions. The equations in $D_h u = f_h$ belong to $\mathbf{x} \in \Omega_h$, but use values of u on $\overline{\Omega}_h$. After elimination of the Dirichlet boundary values $u = \varphi$, we obtain the system $L_h u_h = q_h$ with q_h as in (4.13b).

4.3 M-matrices, Matrix Norms, Positive-Definite Matrices

The elements of the matrix A are denoted by $a_{\alpha\beta}$ $(\alpha, \beta \in I)$. Here A and the index set I assume the places of L_h and Ω_h. We denote the componentwise inequality by

$$A \geq B, \qquad \text{if} \quad a_{\alpha\beta} \geq b_{\alpha\beta} \quad \text{for all } \alpha, \beta \in I,$$

and define analogously $A \leq B$, $A > B$, $A < B$. The zero matrix is denoted by O.

Definition 4.8 (M-matrix). A is called an M-matrix if

$$a_{\alpha\alpha} > 0 \quad \text{for all } \alpha, \beta \in I, \qquad a_{\alpha\beta} \leq 0 \quad \text{for all } \alpha \neq \beta, \tag{4.21a}$$

$$A \text{ nonsingular and } A^{-1} \geq O. \tag{4.21b}$$

The inequalities (4.21a) can immediately be proved for L_h (cf. (4.12)). However we still need criteria and auxiliary results to prove (4.21b).

Definition 4.9 (matrix graph). Let A be an $I \times I$ matrix. Its corresponding *matrix graph* is the following subset of $I \times I$:

$$G(A) := \{(\alpha, \beta) \in I \times I : a_{\alpha\beta} \neq 0\}.$$

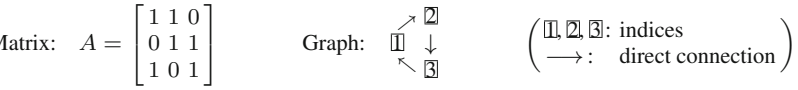


Fig. 4.3 Irreducible matrix A and corresponding matrix graph $G(A)$.

The set $G(A)$ represents a *directed graph*. The index $\alpha \in I$ is said to be *directly connected* with $\beta \in I$ if $(\alpha, \beta) \in G(A)$, i.e., $a_{\alpha\beta} \neq 0$ (cf. Figure 4.3). We say that $\alpha \in I$ is connected with $\beta \in I$ if there exists a *connection* (chain of direct connections)

$$\alpha = \alpha_0, \alpha_1, \ldots, \alpha_k = \beta \quad \text{with} \quad (\alpha_{i-1}, \alpha_i) \in G(A) \qquad (1 \leq i \leq k). \tag{4.22}$$

The pairs (α, β) in (4.22) form the set

$$\overline{G(A)} := \{(\alpha, \beta) \in I \times I : \ (\alpha, \beta) \ \text{satisfies (4.22)}\}.$$

Definition 4.10 (irreduzibel). A matrix A is said to be *irreducible* if every $\alpha \in I$ is connected with every $\beta \in I$, i.e., $\overline{G(A)} = I \times I$.

In the case of the matrix $A = L_h$ in (4.14) two indices $\mathbf{x}, \mathbf{y} \in \Omega_h$ are directly connected if and only if $\mathbf{y} = \mathbf{x}$ or if $\mathbf{y}$ is a neighbour of $\mathbf{x}$. Arbitrary $\mathbf{x}, \mathbf{y} \in \Omega_h$ can evidently be connected by a chain $\mathbf{x} = \mathbf{x}(0), \mathbf{x}(1), \ldots, \mathbf{x}(k) = \mathbf{y}$ of neighbouring points. Thus L_h is irreducible.

Exercise 4.11. Prove the following: (a) The irreducibility of an $I \times I$ matrix does not depend on the ordering of the index set I.
(b) Let $\#I \geq 2$. Prove that A is irreducible if and only if there is no ordering of the indices such that the resulting matrix has the form $A = \begin{bmatrix} A_{11} & A_{12} \\ O & A_{22} \end{bmatrix}$, where A_{11} and A_{22} are respectively square $n_1 \times n_1$ and $n_2 \times n_2$ matrices ($n_1 \geq 1, n_2 \geq 1$), and A_{12} is an $n_1 \times n_2$ submatrix.

The following considerations concern the enclosure of the eigenvalues of A, i.e., supersets of the *spectrum*

$$\sigma(A) := \{\lambda \in \mathbb{C} : \ \lambda \text{ is an eigenvalue of } A\}.$$

The important question as to whether $A = L_h$ is nonsingular (i.e., $0 \notin \sigma(A)$) can be treated as a special case of Criterion 4.12. For a detailed discussion of the following Gershgorin circles refer to Varga [295]. We use the usual notation $K_r(z)$ for the open circle $\{\zeta \in \mathbb{C} : |z - \zeta| < r\}$ around $z \in \mathbb{C}$ and $\overline{K_r(z)} := \{\zeta \in \mathbb{C} : |z - \zeta| \leq r\}$ for the closed circle. A special definition is used for $r = 0$:

$$\overline{K_0(z)} = K_0(z) := \{z\}.$$

Criterion 4.12 (Gershgorin [113]). *(a) All eigenvalues of A lie in*

$$\bigcup_{\alpha\in I} \overline{K_{r_\alpha}(a_{\alpha\alpha})} \qquad \text{with} \quad r_\alpha = \sum_{\beta\in I\setminus\{\alpha\}} |a_{\alpha\beta}| \,.$$

(b) If A is irreducible, the eigenvalues even lie in

$$\Big(\bigcup_{\alpha\in I} K_{r_\alpha}(a_{\alpha\alpha}) \Big) \;\cup\; \Big(\bigcap_{\alpha\in I} \partial K_{r_\alpha}(a_{\alpha\alpha}) \Big) \,.$$

Proof. (a) Let λ be an eigenvalue of A and u a corresponding eigenvector which, without loss of generality, satisfies $\|u\|_\infty = 1$, where

$$\|u\|_\infty := \max\{|u_\alpha| : \alpha \in I\} \tag{4.23}$$

is the *maximum norm*. There exists (at least) one index $\gamma \in I$ with $|u_\gamma| = 1$.

Assertion 1. $|u_\gamma| = 1$ implies

$$|\lambda - a_{\gamma\gamma}| \le \sum_{\beta\in I\setminus\{\gamma\}} |a_{\gamma\beta}|\,|u_\beta| \le \sum_{\beta\in I\setminus\{\gamma\}} |a_{\gamma\beta}| = r_\beta. \tag{4.24}$$

From (4.24) follows $\lambda \in \overline{K_{r_\gamma}(a_{\gamma\gamma})}$ and hence the statement (a). To prove the assertion use the equation from $Au = \lambda u$ associated to the index γ:

$$\lambda u_\gamma = \sum_{\beta\in I} a_{\gamma\beta}\, u_\beta, \qquad \text{i.e.,}\;\; (\lambda - a_{\gamma\gamma})\, u_\gamma = \sum_{\beta\in I\setminus\{\gamma\}} a_{\gamma\beta}\, u_\beta.$$

From $|u_\gamma| = 1$ follows

$$|\lambda - a_{\gamma\gamma}| = |(\lambda - a_{\gamma\gamma})u_\gamma| \le \Big|\sum\nolimits_{\beta\neq\gamma} a_{\gamma\beta}u_\beta\Big| \,. \tag{4.25}$$

By taking the modulus into the sum and by using $|u_\beta| \le \|u\|_\infty = 1$, (4.24) follows.

(b) Let the matrix A be irreducible and let λ be an arbitrary eigenvalue of A with associated eigenvector u which in turn is again normalised by $\|u\|_\infty = 1$. The case $\lambda \in \bigcup_{\alpha\in I} K_{r_\alpha}(a_{\alpha\alpha})$ immediately leads to the statement. Therefore let $\lambda \notin \bigcup_{\alpha\in I} K_{r_\alpha}(a_{\alpha\alpha})$ be assumed.

Assertion 2. Let $a_{\gamma\beta} \neq 0$, i.e., γ is directly connected with β; then $|u_\gamma| = 1$ and $|\lambda - a_{\gamma\gamma}| = r_\gamma$ implies $|u_\beta| = 1$ and $|\lambda - a_{\beta\beta}| = r_\beta$.

Part (a) proves the existence of a $\gamma \in I$ with $|u_\gamma| = 1$ and $|\lambda - a_{\gamma\gamma}| \le r_\gamma$. According to the assumption, $|\lambda - a_{\gamma\gamma}| = r_\gamma$ must hold so that Assertion 2 is applicable to γ. Since A is irreducible, for any $\beta \in I$ there exists a connection (4.22) of γ with $\beta:\ \alpha_0 = \gamma,\ \alpha_1, \ldots, \alpha_k = \beta,\ a_{\alpha_{i-1}\alpha_i} \neq 0$. Assertion 2 shows

$$|u_{\alpha_i}| = 1 \qquad \text{and} \qquad |\lambda - a_{\alpha_i\alpha_i}| = r_{\alpha_i} \quad \text{for all } i = 0, \ldots, k\,;$$

in particular, $\lambda \in \partial K_{r_\beta}(a_{\beta\beta})$ for $\beta = \alpha_k$. Since β was chosen arbitrarily, it follows that $\lambda \in \bigcap \partial K_{r_\alpha}(a_{\alpha\alpha})$, and the statement is proved.

Proof of Assertion 2. Besides the inequality chain (4.25) there also holds $|\lambda - a_{\gamma\gamma}| = r_\gamma$, so that all the inequalities in (4.25) become equations. In particular

$$\sum_{\beta \in I\setminus\{\gamma\}} |a_{\gamma\beta}|\,|u_\beta| = \sum_{\beta \in I\setminus\{\gamma\}} |a_{\gamma\beta}|$$

must hold. Since $|u_\beta| \le \|u\|_\infty = 1$, the identity $|a_{\gamma\beta}|\,|u_\beta| = |a_{\gamma\beta}|$ must be satisfied for each term. Hence $a_{\gamma\beta} \neq 0$ implies $|u_\beta| = 1$. Applying Assertion 1 to β yields $|\lambda - a_{\beta\beta}| \le r_\beta$. The assumption $\lambda \notin \bigcup_{\alpha\in I} K_{r_\alpha}(a_{\alpha\alpha})$ proves $|\lambda - a_{\beta\beta}| = r_\beta$. ■

Even if A is not irreducible, the statement of Criterion 4.12a may hold.

Exercise 4.13. Let $I_\gamma := \{\beta \in I : (\gamma,\beta) \in \overline{G(A)}\}$. Prove that

$$\sigma(A) \subset \bigcup_{\gamma\in I} \left\{ K_{r_\gamma}(a_{\gamma\gamma}) \cup \bigcap_{\beta\in I_\gamma} \partial K_{r_\beta}(a_{\beta\beta}) \right\}.$$

Definition 4.14 (diagonal dominance). (a) A is said to be (strongly) *diagonally dominant* if (4.26a) holds for all $\alpha \in I$:

$$\sum_{\beta \in I\setminus\{\alpha\}} |a_{\alpha\beta}| < |a_{\alpha\alpha}|\,. \tag{4.26a}$$

(b) A is said to be *weakly diagonally dominant* if

$$\sum_{\beta \in I\setminus\{\alpha\}} |a_{\alpha\beta}| \le |a_{\alpha\alpha}| \qquad \text{for all } \alpha \in I. \tag{4.26b}$$

(c) A is said to be irreducibly diagonally dominant if A is irreducible and weakly diagonally dominant, and the inequality (4.26a) holds for at least one index $\alpha \in I$.

Note that while an irreducible *and* diagonally dominant matrix is irreducibly diagonally dominant, the reverse need not hold.

The matrix L_h from §4.2, while not diagonally dominant, is irreducibly diagonally dominant, for L_h is irreducible and satisfies (4.26b). At all near-boundary points—i.e., those $\mathbf{x} \in \Omega_h$ which have a boundary point $\mathbf{y} \in \Gamma_h$ as a neighbour—however, (4.26a) holds: $\sum_{\beta\neq\alpha} |a_{\alpha\beta}| \le 3h^{-2} < 4h^{-2} = a_{\alpha\alpha}$.

The *spectral radius* $\rho(A)$ of a matrix A is characterised by the eigenvalue that is largest in modulus:

$$\rho(A) := \max\{|\lambda| : \lambda \text{ eigenvalue of } A\}. \tag{4.27}$$

In the following we split A into

$$A = D - B, \qquad D := \operatorname{diag}\{a_{\alpha\alpha} : \alpha \in I\}, \tag{4.28a}$$

where D is the diagonal part of A:

$$d_{\alpha\alpha} = a_{\alpha\alpha}, \qquad d_{\alpha\beta} = 0 \quad \text{for } \alpha \neq \beta. \tag{4.28b}$$

$B := D - A$ is the off-diagonal part:

$$b_{\alpha\alpha} = 0, \qquad b_{\alpha\beta} = -a_{\alpha\beta} \quad \text{for } \alpha \neq \beta. \tag{4.28c}$$

Criterion 4.15. *Let (4.28a–c) hold. Sufficient conditions for*

$$\rho(D^{-1}B) < 1 \tag{4.29}$$

are the diagonal dominance or the irreducible diagonal dominance of A.

Proof. (a) The coefficients of $C := D^{-1}B$ read

$$c_{\alpha\beta} = -a_{\alpha\beta}/a_{\alpha\alpha} \quad (\alpha \neq \beta), \qquad c_{\alpha\alpha} = 0.$$

From the diagonal dominance (4.26a) follows that $r_\alpha := \sum_{\beta\neq\alpha} |c_{\alpha\beta}| < 1$ for all $\alpha \in I$. By Gershgorin's Criterion 4.12a all eigenvalues λ of C lie in

$$\bigcup_{\alpha\in I} \overline{K_{r_\alpha}(c_{\alpha\alpha})} = \bigcup_{\alpha\in I} \overline{K_{r_\alpha}(0)},$$

so that $|\lambda| \leq \max r_\alpha < 1$ and hence $\rho(C) = \rho(D^{-1}B) < 1$ follows.

(b) If A is irreducibly diagonally dominant then $r_\beta \leq 1$ for all $\beta \in I$ and $r_\alpha < 1$ for at least one α. According to Criterion 4.12b all eigenvalues of C lie in $\bigcup_{\beta\in I} K_{r_\beta}(0) \cup (\bigcap_{\beta\in I} \partial K_{r_\beta}(0))$. This set lies in $K_1(0)$ if

$$\bigcap_{\beta\in I} \partial K_{r_\beta}(0)) \subset K_1(0).$$

At first let us assume that all r_β agree: $r_\beta = r$ for all β. Since $r_\alpha < 1$ for some $\alpha \in I$, it follows that $r < 1$ and $\bigcap_{\beta\in I} \partial K_{r_\beta}(0)) = \partial K_r(0) \subset K_1(0)$. But if all r_β are not equal then $\bigcap_{\beta\in I} \partial K_{r_\beta}(0))$ is empty. Thus in both cases $\lambda \in K_1(0)$ holds and (4.29) is proved. ∎

Exercise 4.16. (a) Weaken irreducible diagonal dominance as follows: Let A satisfy the inequalities (4.26b) and for all $\gamma \in I$ let a connection exist for an index $\alpha \in I$ (i.e., $(\gamma, \alpha) \in \overline{G(A)}$) for which the strict inequality (4.26a) holds. Prove that even under this assumption $\rho(D^{-1}B) < 1$ holds. *Hint.* Use Exercise 4.13.

(b) Show: the geometric series $S = \sum_{\nu=0}^{\infty} C^\nu$ converges if and only if $\rho(C) < 1$. Then the following holds: $S = (I - C)^{-1}$. *Hint.* Represent C in the form QRQ^{T} (Q a unitary, and R an upper triangular matrix) and show $\|C^\nu\|_\infty \leq \text{const} \cdot [\rho(C)]^\nu$.

(c) Let u be a vector. We define $|u|$ as the vector (sic!) with the entries $|u|_\alpha := |u_\alpha|$. For two vectors one writes $v \leq w$ if $v_\alpha \leq w_\alpha$ ($\alpha \in I$). Show that:

1) $AB \geq O$ if $A \geq O$ and $B \geq O$; $\quad AB > O$ if $A > O$ and $B > O$;

2) $AD > O$ if $A > O$ and $D \geq O$ is a nonsingular diagonal matrix;

3) $Av \leq Aw$ if $A \geq O$ and $v \leq w$; $\quad \|v\|_\infty \leq \|w\|_\infty$ if $0 \leq v \leq w$;

4) $Au \leq |Au| \leq A\,|u|$ if $A \geq O$.

The importance of inequality (4.29) results from the next statement.

Lemma 4.17. *Let A satisfy (4.21a). Let D and B be defined by (4.28a–c). A is an M-matrix if and only if*

$$\rho(D^{-1}B) < 1.$$

Proof. (a) Let $C := D^{-1}B$ satisfy $\rho(C) < 1$. Then the geometric series

$$S := \sum_{\nu=0}^{\infty} C^{\nu}$$

converges (cf. Exercise 4.16b). From $D^{-1} \geq O$ and $B \geq O$ one infers $C \geq O$, $C^{\nu} \geq O$, and $S \geq O$. Since $I = S(I - C) = SD^{-1}(D - B) = SD^{-1}A$, A has the inverse $A^{-1} = SD^{-1}$. $D^{-1} \geq O$ and $S \geq O$ result in $A^{-1} \geq O$. From this (4.21b) also results, i.e., A is an M-matrix.

(b) Let A be an M-matrix. For an eigenvalue λ of $D^{-1}B$ select an eigenvector $u \neq 0$. According to Exercise 4.16c we have

$$|\lambda|\,|u| = |\lambda u| = \left|D^{-1}Bu\right| \leq D^{-1}B\,|u|\,.$$

Because $A^{-1}D \geq O$ (cf. (4.21a,b)) we obtain $-A^{-1}DD^{-1}B\,|u| \leq -A^{-1}D\,|\lambda||u|$ so that

$$\begin{aligned} |u| &= A^{-1}\,(D - B)\,|u| = A^{-1}D\left(I - D^{-1}B\right)|u| \leq A^{-1}D\,|u| - A^{-1}D\,|\lambda|\,|u| \\ &= (1 - |\lambda|)\,A^{-1}D\,|u| \end{aligned}$$

follows. For $|\lambda| \geq 1$ we would get the inequality $|u| \leq 0$, i.e., $u = 0$ in contradiction to the assumption $u \neq 0$. From this follows $|\lambda| < 1$ for every eigenvalue of $C = D^{-1}B$, thus $\rho(D^{-1}B) < 1$. ∎

Criterion 4.15 and Lemma 4.17 imply the next criterion.

Criterion 4.18. *If a matrix A with the property (4.21a) is diagonally dominant or irreducibly diagonally dominant, then A is an M-matrix.*

Theorem 4.19. *An irreducible M-matrix A has an elementwise positive inverse: $A^{-1} > O$.*

Proof. Let $\alpha, \beta \in I$ be selected arbitrarily. There exists a connection (4.22): $\alpha = \alpha_0, \alpha_1, \ldots, \alpha_k = \beta$. Set $C := D^{-1}B$. Since $c_{\alpha_{i-1}\alpha_i} > 0$, it follows that

$$(C^k)_{\alpha\beta} = \sum_{\gamma_1,\ldots,\gamma_{k-1} \in I} c_{\alpha\gamma_1}\, c_{\gamma_1\gamma_2} \cdots c_{\gamma_{k-1}\beta} \geq c_{\alpha\alpha_1}\, c_{\alpha_1\alpha_2} \cdots c_{\alpha_{k-1}\beta} > 0\,.$$

According to Lemma 4.17, $\rho(C) < 1$ holds, so that $S := \sum_{\nu=0}^{\infty} C^{\nu}$ converges. Since $S_{\alpha\beta} \geq (C^k)_{\alpha\beta} > 0$ and $\alpha, \beta \in I$ are arbitrary, $S > O$ is proved. The assertion results from $A^{-1} = SD^{-1} > O$ (cf. proof of Lemma 4.17). ∎

In the following we derive norm estimates for A^{-1}.

Definition 4.20. Let V be a linear space (vector space) over the field of real numbers ($\mathbb{K} := \mathbb{R}$) or complex numbers ($\mathbb{K} := \mathbb{C}$). The functional $\|\cdot\| : V \to [0, \infty)$ is called a *norm* in V if

$$\begin{aligned} \|u\| &= 0 && \text{only for } 0 \neq u \in V, \\ \|u + v\| &\leq \|u\| + \|v\| && \text{for all } u, v \in V, \\ \|\lambda u\| &= |\lambda| \, \|u\| && \text{for all } \lambda \in \mathbb{K},\ u \in V. \end{aligned}$$

For instance, for $V = \mathbb{R}^I$ the maximum norm defined in (4.23) satisfies the norm axioms.

If one views the elements $u \in V$ as vectors, $\|\cdot\|$ is called a vector norm. But the matrices also form a linear space. In the latter case one calls $\|\cdot\|$ a *matrix norm*. A special class of matrix norms is the following one.

Definition 4.21 (associated matrix norm). Let V be the vector space with vector norm $\|\cdot\|$. Then one calls

$$|||\, A \,||| := \sup\{\|Au\| \,/\, \|u\| : 0 \neq u \in V\} \tag{4.30}$$

the matrix norm associated with the vector norm $\|\cdot\|$.

Exercise 4.22. Let $|||\cdot|||$ be defined by (4.30). Show that: (a) $|||\cdot|||$ is a norm; (b) the following holds:

$$\begin{aligned} |||\, AB \,||| &\leq |||\, A \,||| \; |||\, B \,||| && \text{(submultiplicativity)}, && \text{(4.31a)} \\ |||\, I \,||| &= 1 && (I\text{: unit matrix}), \\ \|Au\| &\leq |||\, A \,||| \; \|u\| && \text{for all } u \in V, \\ |||\, A \,||| &\geq \rho(A). && && \text{(4.31b)} \end{aligned}$$

As usual, the associated matrix norm is denoted by the same symbol, i.e., we again write $\|\cdot\|$ instead of $|||\cdot|||$. If the vector norm $\|\cdot\|_{\mathrm{xyz}}$ carries some subscript "xyz", $\|\cdot\|_{\mathrm{xyz}}$ is also used for the associated matrix norm.

Example. The matrix norm associated with the maximum norm $\|\cdot\|_\infty$ (cf. (4.23)) is called the *row-sum norm* and is also denoted by $\|\cdot\|_\infty$. It has the explicit representation

$$\|A\|_\infty = \max_{\alpha \in I} \Big\{ \sum\nolimits_{\beta \in I} |a_{\alpha\beta}| \Big\}. \tag{4.32}$$

Exercise 4.23. (a) Prove (4.32) and (b) $\|B\|_\infty \leq \|C\|_\infty$ for matrices $O \leq B \leq C$.

In the next theorem we denote by $\mathbf{1}$ the vector having only ones as components:

$$\mathbf{1}_\alpha = 1 \qquad \text{for all } \alpha \in I.$$

For the notation $v \leq w$ see Exercise 4.16c.

Theorem 4.24. *Let A be an M-matrix and let a vector w exist with $Aw \geq \mathbf{1}$. Then $\|A^{-1}\|_\infty \leq \|w\|_\infty$.*

Proof. As in the proof of Lemma 4.17, let $|u|$ be the vector with the components $|u_\alpha|$. For each u we have $|u| \le \|u\|_\infty \mathbf{1} \le \|u\|_\infty Aw$. Since $A^{-1} \ge O$, we

$$|A^{-1}u| \le A^{-1}|u| \le \|u\|_\infty A^{-1}Aw = \|u\|_\infty w$$

(cf. Exercise 4.16c) and $\|A^{-1}u\|_\infty / \|u\|_\infty \le \|w\|_\infty$. Definition 4.21 implies that $\|A^{-1}\|_\infty \le \|w\|_\infty$. ∎

How to estimate with the aid of a majorising matrix is shown next.

Theorem 4.25. *Let A and A' be M-matrices with $A' \ge A$. Then the following holds:*

$$O \le A'^{-1} \le A^{-1} \qquad \textit{and} \qquad \|A'^{-1}\|_\infty \le \|A^{-1}\|_\infty. \tag{4.33}$$

Proof. $A'^{-1} \le A^{-1}$ follows from $A^{-1} - A'^{-1} = A^{-1}(A' - A)A'^{-1}$ and $A^{-1} \ge O$, $A' - A \ge O$, $A'^{-1} \ge O$. The remainder follows from Exercise 4.23b. ∎

Exercise 4.26. Prove (4.33) under the following weaker assumptions: A is an M-matrix, A' satisfies (4.21a) and $A' \ge A$. *Hint*: Repeat the considerations from the first part of the proof of Lemma 4.17 with the matrices D' and B' associated to A'.

Exercise 4.27. Let B be a principal submatrix of A, i.e., there exists a subset $I' \subset I$ such that B is given by the entries $b_{\alpha\beta} = a_{\alpha\beta}$ $(\alpha, \beta \in I')$. Prove that if A is an M-matrix, then so is B and

$$O \le (B^{-1})_{\alpha\beta} \le (A^{-1})_{\alpha\beta} \quad \text{for all } \alpha, \beta \in I'.$$

Hint: Apply Exercise 4.26 to the following matrix A': $a'_{\alpha\beta} = a_{\alpha\beta}$ for $\alpha, \beta \in I'$, $a'_{\alpha\alpha} = a_{\alpha\alpha}$ for $\alpha \in I \backslash I'$, and $a'_{\alpha\beta} = 0$ otherwise.

Another well-known vector norm is the Euclidean norm

$$\|u\|_2 := \sqrt{c \sum\nolimits_{\alpha \in I} |u_\alpha|^2} \tag{4.34}$$

with fixed scaling constant $c > 0$. For example, the choice $c = h^2$ in connection with the grid functions from §4.2 results in the fact that $c \sum_{\alpha \in I}$ represents an approximation to the integration $\int_\Omega$. The matrix norm associated to $\|\cdot\|_2$ is independent of the factor c. It is called the *spectral norm* and is also denoted by $\|\cdot\|_2$. The name derives from the following characterisation.

Exercise 4.28. Prove:
(a) For symmetric matrices there holds $\|A\|_2 = \rho(A)$ (cf. (4.27)).
(b) $\|A\|_2 = \|A^\mathsf{T}\|_2$.
(c) For each real matrix holds

$$\|A\|_2 = \sqrt{\rho(A^\mathsf{T}A)} = [\,\text{maximal eigenvalue of } A^\mathsf{T}A\,]^{1/2}.$$

(d) For each matrix holds $\|A\|_2^2 \le \|A\|_\infty \|A^\mathsf{T}\|_\infty$. *Hint:* (c) and (4.31b).

For the proof in the exercise use the *scalar product*

$$\langle u, v\rangle = c \sum_{\alpha\in I} u_\alpha v_\alpha \tag{4.35a}$$

(c as in (4.34)) and its properties

$$\langle u, u\rangle = \|u\|_2^2\,, \quad \langle Au, v\rangle = \left\langle u, A^{\mathsf{T}} v\right\rangle, \quad |\langle u, v\rangle| \le \|u\|_2 \|v\|_2\,. \tag{4.35b}$$

Here $\mathbb{K} = \mathbb{R}$ is always used as field, i.e., all matrices and vectors are real. In the case of $\mathbb{K} = \mathbb{C}$, v_α in (4.35a) must be replaced by $\bar{v}_\alpha$, and A^{T} in (4.35b) becomes A^{H}. In Exercise 4.28 'symmetric' must be replaced by 'Hermitian'.

Definition 4.29 (positive definite). A matrix A is said to be positive definite if it is symmetric and

$$\langle Au, u\rangle > 0 \qquad \text{for all } u \neq 0.$$

Exercise 4.30. Prove: (a) A symmetric matrix is positive definite if and only if all eigenvalues are positive.

(b) All principal submatrices of a positive-definite matrix are positive definite (note the similarity to Exercise 4.27).

(c) The diagonal elements $a_{\alpha\alpha}$ of a positive-definite matrix are positive.

(d) A positive-definite matrix A has a unique positive-definite square root $B = A^{1/2}$, which has the property $B^2 = A$.

(e) A is called *positive semidefinite* if $\langle Au, u\rangle \ge 0$ for all u. Then all principal submatrices of A are positive semidefinite, $a_{\alpha\alpha} \ge 0$, and $A^{1/2}$ is positive semidefinite.

A corollary to Exercise 4.30a is Lemma 4.31.

Lemma 4.31. *A positive-definite matrix is nonsingular and has a positive-definite inverse.*

The property 'A^{-1} is positive definite' is neither necessary nor sufficient to ensure the property '$A^{-1} \ge O$' of an M-matrix. In both cases, however, (irreducible) diagonal dominance is a sufficient criterion (cf. Criterion 4.18).

Criterion 4.32. *If a symmetric matrix with positive diagonal entries is diagonally dominant or irreducibly diagonally dominant then it is positive definite.*

Proof. Since $r_\alpha < a_{\alpha\alpha}$, resp. $r_\alpha \le a_{\alpha\alpha}$, the Gershgorin circles which occur in Criterion 4.12 do not intersect the semi-axis $(-\infty, 0]$, so that all the eigenvalues must be positive. By Exercise 4.30a then A is positive definite. ∎

Lemma 4.33. *Let $\lambda_{\min}$ and $\lambda_{\max}$ be respectively the smallest and largest eigenvalues of a positive-definite matrix A. Then there holds*

$$\|A\|_2 = \lambda_{\max}\,, \qquad \|A^{-1}\|_2 = 1/\lambda_{\min}\,.$$

Proof. Exercise 4.28a shows that $\|A\|_2 = \rho(A)$ and $\|A^{-1}\| = \rho(A^{-1})$. From (4.27) then result $\rho(A) = \lambda_{\max}$ and $\rho(A^{-1}) = 1/\lambda_{\min}$, since $\lambda_{\min} > 0$. ∎

4.4 Properties of the Matrix L_h

Theorem 4.34. *The matrix L_h (five-point formula) defined in (4.12) has the following properties:*

$$L_h \text{ is an M-matrix,} \tag{4.36a}$$

$$L_h \text{ is positive definite,} \tag{4.36b}$$

$$\|L_h\|_\infty \le 8h^{-2}, \qquad \|L_h^{-1}\|_\infty \le 1/8, \tag{4.36c}$$

$$\|L_h\|_2 \;\le\; 8h^{-2}\cos^2\big(\frac{\pi h}{2}\big) \;<\; 8h^{-2}, \tag{4.36d}$$

$$\|L_h^{-1}\|_2 \;\le\; \frac{1}{8}h^2\sin^{-2}\big(\frac{\pi h}{2}\big) \;=\; \frac{1}{2\pi^2}+\mathcal{O}(h^2) \;\le\; \frac{1}{16}. \tag{4.36e}$$

Proof. (a) In §4.3 we already noticed that L_h is irreducibly diagonally dominant and satisfies the inequality (4.21a). By Criterion 4.18 then L_h is an M-matrix.

(b) Since L_h is symmetric and irreducibly diagonally dominant, (4.36b) follows from Criterion 4.32.

(c) That $\|L_h\| \le 8h^{-2}$ can be read from (4.12) and (4.32). To estimate L_h^{-1} one uses Theorem 4.24 with $w(x,y) = x(1-x)/2$. Then we have $L_h w \ge \mathbf{1}$ (even that $(L_h w)(x,y) = 1$ unless $y = h$ and $y = 1-h$) and $\|w\|_\infty \le w(1/2, y) = 1/8$.

(d) The inequalities (4.36d,e) result from Lemma 4.33 and the next lemma. ∎

Lemma 4.35. *The $(n-1)^2$ eigenvectors of L_h are $u^{\nu\mu}$ $(1 < \nu, \mu < n-1)$:*

$$u^{\nu\mu}(x,y) = \sin(\nu\pi x)\sin(\mu\pi y), \qquad (x,y) \in \Omega_h. \tag{4.37a}$$

The corresponding eigenvalues are

$$\lambda_{\nu\mu} = 4h^{-2}\left(\sin^2(\frac{\nu\pi h}{2}) + \sin^2(\frac{\mu\pi h}{2})\right), \qquad 1 \le \nu, \mu \le n-1. \tag{4.37b}$$

Proof. Let $\Omega_h^{1\mathrm{D}}$ be the one-dimensional grid (4.5) and set $u^\nu(x) := \sin(\nu\pi x)$. For each $x \in \Omega_h^{1\mathrm{D}}$ there holds

$$\begin{aligned}\partial^-\partial^+ u^\nu(x) &= h^{-2}\left[\sin(\nu\pi(x-h)) + \sin(\nu\pi(x+h)) - 2\sin(\nu\pi x)\right] \\ &= 2h^{-2}\sin(\nu\pi x)\left[\cos(\nu\pi x) - 1\right]\end{aligned}$$

since $\sin(\nu\pi(x \pm h)) = \sin(\nu\pi x)\cos(\nu\pi h) \pm \cos(\nu\pi x)\sin(\nu\pi h)$. The identity $1 - \cos\xi = 2\sin^2(\xi/2)$ then implies

$$-\partial^-\partial^+ u^\nu(x) = 4h^{-2}\sin^2(\frac{\nu\pi h}{2})\, u^\nu(x), \qquad x \in \Omega_h^{1\mathrm{D}}. \tag{4.37c}$$

Let $L_h^{1\mathrm{D}}$ be the matrix (4.7b). Note that the difference $(\partial^-\partial^+ u)(h)$—in contrast to $(L_h^{1\mathrm{D}}u)(h)$—also involves the boundary value $u(0)$; similarly $(\partial^-\partial^+ u)(1-h)$ depends on $u(1)$. However, since $u(0) = \sin(0) = 0$ and $u(1) = \sin(\nu\pi) = 0$

we have $L_h^{1D}u^\nu = -\partial^-\partial^+u^\nu$, and (4.37c) can be brought over:

$$L_h^{1\mathrm{D}}u^\nu = 4h^{-2}\sin^2\left(\frac{\nu\pi h}{2}\right)u^\nu, \qquad 1 \le \nu \le n-1. \tag{4.37d}$$

The two-dimensional grid function $u^{\nu\mu}$ in (4.37a) can be written as the (tensor) product $u^\nu(x)u^\mu(y)$. Now we have that $(L_h u^{\nu\mu})(x,y)$ is equal to the sum $u^\mu(y)(L_h^{1\mathrm{D}}u^\nu)(x) + u^\nu(x)(L_h^{1\mathrm{D}}u^\mu)(y)$, so that (4.37b) follows from (4.37d). ∎

In the sequel we want to show the analogies between the properties of the Poisson equation (4.9a,b) and the discrete five-point formula (4.11a,b).

The analogue of the mean-value property (2.13) is the equation

$$u_h(x,y) = \frac{1}{4}[u_h(x-h,y) + u_h(x+h,y) + u_h(x,y-h) + u_h(x,y+h)] \tag{4.38}$$

From (4.10) and (4.11a) with $f = 0$ we obtain the following result.

Remark 4.36. The solution u_h of the discrete potential equation (4.11a) with $f = 0$ satisfies equation (4.38) at all grid points $(x,y) \in \Omega_h$.

As in the continuous case the mean-value property (4.38) implies the maximum-minimum principle.

Remark 4.37. Let u_h be a nonconstant solution of the discrete potential equation (4.11a) with $f = 0$. The extrema $\max\{u_h(\mathbf{x}) : \mathbf{x} \in \overline{\Omega_h}\}$ and $\min\{u_h(\mathbf{x}) : \mathbf{x} \in \overline{\Omega_h}\}$ are assumed not on Ω_h but on Γ_h.

Proof. If u_h were maximal in $(x,y) \in \Omega_h$, then because of equation (4.38), all neighbouring points $(x \pm h, y)$ and $(x, y \pm h)$ would have to carry the same values. Since every pair of points can be linked by a chain of neighbouring points, it follows that $u_h = \mathrm{const}$, in contradiction to the assumption. ∎

The last proof indirectly uses the fact that L_h is irreducible. The irreducibility of L_h corresponds to the assumption in Theorem 2.17 that Ω is a domain, i.e., connected.

The result of carrying over Theorems 2.27 and 3.2 reads as follows.

Theorem 4.38. *(a) Let u_h^1 and u_h^2 be two solutions of (4.11a): $-\Delta_h u_h^i = f$ for different boundary values $u_h^i = \varphi^i$ $(i = 1, 2)$. Then the following holds:*

$$\|u_h^1 - u_h^2\|_\infty \le \max_{\mathbf{x}\in\Gamma_h} \left|\varphi^1(\mathbf{x}) - \varphi^2(\mathbf{x})\right|, \tag{4.39a}$$

$$u_h^1 \le u_h^2 \ \text{ in } \Omega_h, \quad \text{if } \varphi^1 \le \varphi^2 \ \text{ on } \Gamma_h\,. \tag{4.39b}$$

(b) A solution u_h of $-\Delta_h u_h = f \ge 0$ with boundary values $u_h = \varphi \ge 0$ satisfies $u_h \ge 0$ everywhere in Ω_h.

Proof. (a) Let $w_h := u_h^2 - u_h^1$.

(a1) In the case that $\varphi^1 \le \varphi^2$ one has $w_h \ge 0$ on Γ_h and $-\Delta_h w_h = 0$. Remark 4.37 proves that $w_h = \text{const} \ge 0$ or $w_h > 0$, and hence (4.39b).

(a2) Let M be the right-hand side of (4.39a). $-M \le w_h \le M$ on Γ_h implies the inequalities $-M \le w_h \le M$ on Ω_h, and hence (4.39a).

(b) $f \ge 0$, $\varphi \ge 0$ imply $q_h \ge 0$ in (4.13b). The M-matrix property (4.36a) yields $u_h = L_h^{-1} q_h \ge 0$. ∎

The discrete analogue of the Green function $g(\mathbf{x}, \boldsymbol{\xi})$ is $h^{-2} L_h^{-1}$. Let $\delta_{\boldsymbol{\xi}}$ be the scaled unit vector

$$\delta_{\boldsymbol{\xi}}(\mathbf{x}) = \begin{cases} h^{-2} & \text{if } \mathbf{x} = \boldsymbol{\xi}, \\ 0 & \text{if } \mathbf{x} \ne \boldsymbol{\xi}, \end{cases} \qquad (\mathbf{x}, \boldsymbol{\xi} \in \Omega_h). \tag{4.40a}$$

The column of the matrix $h^{-2} L_h^{-1}$ with index $\boldsymbol{\xi} \in \Omega_h$ is given by

$$g_h(\cdot, \boldsymbol{\xi}) := L_h^{-1} \delta_{\boldsymbol{\xi}} \qquad (\boldsymbol{\xi} \in \Omega_h). \tag{4.40b}$$

For $\boldsymbol{\xi} \in \Omega_h$ fixed, $g_h(\cdot, \boldsymbol{\xi})$ is a grid function defined on Ω_h. The domain of definition is extended to $\overline{\Omega}_h \times \overline{\Omega}_h$:

$$\bar{g}_h(\mathbf{x}, \boldsymbol{\xi}) = \begin{cases} g_h(\mathbf{x}, \boldsymbol{\xi}) & \text{if } \mathbf{x}, \boldsymbol{\xi} \in \Omega_h, \\ 0 & \text{if } \mathbf{x} \in \Gamma_h \text{ or } \boldsymbol{\xi} \in \Gamma_h \end{cases} \qquad (\mathbf{x}, \boldsymbol{\xi} \in \overline{\Omega}_h).$$

The values $g_h(\mathbf{x}, \boldsymbol{\xi})$ are entries of $h^{-2} L_h^{-1}$: $g_h(\mathbf{x}, \boldsymbol{\xi}) = h^{-2} (L_h^{-1})_{\mathbf{x}\boldsymbol{\xi}}$. The symmetry of L_h implies the following statement.

Remark 4.39. $\bar{g}_h(\mathbf{x}, \boldsymbol{\xi}) = \bar{g}_h(\boldsymbol{\xi}, \mathbf{x})$ for all $\mathbf{x}, \boldsymbol{\xi} \in \overline{\Omega}_h$ (cf. (3.7)).

The representation (3.13) is recalled in the next remark.

Remark 4.40. The solution u_h of the system of equations (4.11a) with boundary values $\varphi = 0$ reads

$$u_h(\mathbf{x}) = h^2 \sum_{\boldsymbol{\xi} \in \Omega_h} \bar{g}_h(\mathbf{x}, \boldsymbol{\xi}) f(\boldsymbol{\xi}) \qquad \text{for } \mathbf{x} \in \overline{\Omega}_h. \tag{4.41}$$

Equation (4.41) is the componentwise representation of the equation $u_h = L_h^{-1} f_h$. The factor h^2 compensates for h^{-2} in (4.40a). It was introduced so that the summation $h^2 \sum$ in (4.41) approximates the integral $\int_\Omega$.

The discrete Green function is positive also (cf. (3.8)).

Remark 4.41. $0 < g_h(\mathbf{x}, \boldsymbol{\xi}) \le h^{-2}/8$ for $\mathbf{x}, \boldsymbol{\xi} \in \Omega_h$.

Proof. The upper bound follows from $g_h(\mathbf{x}, \boldsymbol{\xi}) \le \|g_h(\cdot, \boldsymbol{\xi})\|_\infty \le \|L_h^{-1}\|_\infty \|\delta_{\boldsymbol{\xi}}\|_\infty$, $\|\delta_{\boldsymbol{\xi}}\|_\infty = h^{-2}$, and (4.36c). $g_h > 0$ can be inferred from $L_h^{-1} > 0$ (cf. Theorem 4.19). ∎

The bound $g_h(\mathbf{x},\boldsymbol{\xi}) \le \frac{h^{-2}}{8}$ is too pessimistic and can be improved considerably.

Lemma 4.42. *The discrete Green's function g_h in (4.40b) satisfies the estimate*

$$0 < g_h(\mathbf{x},\boldsymbol{\xi}) \le \frac{\log 2}{4\log 3}\left(1 - \frac{\log\left(|\mathbf{x}-\boldsymbol{\xi}|^2 + h^2/2\right)}{\log 2}\right) \le \frac{\log \frac{2}{h}}{\log\ 9} \tag{4.42}$$

for all $\mathbf{x},\boldsymbol{\xi} \in \overline{\Omega}_h$. The upper bound $\mathcal{O}(|\log h|)$ reflects the logarithmic singularity of the singularity function $s(\mathbf{x},\boldsymbol{\xi}) = -\log(|\mathbf{x}-\boldsymbol{\xi}|^2)/(4\pi)$.

Proof. For the proof of inequality (4.42) define

$$s_h(\mathbf{x},\boldsymbol{\xi}) := \frac{\log 2}{4\log 3}\left(1 - \frac{\log\left(|\mathbf{x}-\boldsymbol{\xi}|^2 + h^2/2\right)}{\log 2}\right). \tag{4.43a}$$

(a) First we want to show $s_h(\mathbf{x},\boldsymbol{\xi}) \ge 0$ for all $\mathbf{x} \in \overline{\Omega}_h,\ \boldsymbol{\xi} \in \Omega_h$. Since $\boldsymbol{\xi} \in \Omega_h$, we have $|x_i - \xi_i| \le 1-h$ $(i=1,2)$ and hence $|\mathbf{x}-\boldsymbol{\xi}|^2 \le 2(1-h)^2$. Because $1/2$ is the coarsest possible step size, h varies in $(0,1/2]$, so that

$$|\mathbf{x}-\boldsymbol{\xi}|^2 + h^2/2 \le 2(1-h)^2 + \frac{1}{2}h^2 < 2.$$

Therefore s_h in (4.43a) satisfies

$$s_h(\mathbf{x},\boldsymbol{\xi}) > \frac{\log 2}{4\log 3}\left(1 - \frac{\log 2}{\log 2}\right) = 0 \qquad \text{for all } \mathbf{x} \in \overline{\Omega}_h,\ \boldsymbol{\xi} \in \Omega_h.$$

(b) Define $u_h(\mathbf{x}) := s_h(\mathbf{x},0)$ for all $\mathbf{x} \in \{(\nu h, \mu h) : \nu,\mu \in \mathbb{Z}\}$. The five-point formula applied to u_h gives

$$-\Delta_h u_h = \frac{1}{4\log 3}\Delta_h \log\left(|\cdot|^2 + h^2/2\right). \tag{4.43b}$$

Concerning the evaluation at the origin, note that $\log(|\mathbf{x}|^2 + \frac{h^2}{2})|_{\mathbf{x}=\mathbf{0}} = \log(\frac{h^2}{2})$ and $\log(|\mathbf{x}'|^2 + \frac{h^2}{2}) = \log(h^2 + \frac{h^2}{2}) = \log(\frac{3}{2}h^2) = \log 3 + \log\frac{h^2}{2}$ for all neighbours $\mathbf{x}'$ of $\mathbf{x} = \mathbf{0}$. Therefore (4.43b) shows that

$$\begin{aligned} -\left(\Delta_h u_h\right)(\mathbf{0}) &= \frac{h^{-2}}{4\log 3}\Big(4\log\left(3h^2/2\right) - 4\log\left(h^2/2\right)\Big) \\ &= h^{-2}\,\frac{\log\left(3h^2/2\right) - \log\left(h^2/2\right)}{\log 3} = h^{-2}. \end{aligned}$$

(c) The evaluation of $-\Delta_h u_h$ in general grid points $\mathbf{x} = (x,y)$ yields

$$-\Delta_h u_h(\mathbf{x}) = \frac{h^{-2}}{4\log 3}\left\{\begin{array}{l}\log\left(x^2+(y+h)^2+\frac{h^2}{2}\right)+\log\left((x-h)^2+y^2+\frac{h^2}{2}\right)\\ +\log\left(x^2+(y-h)^2+\frac{h^2}{2}\right)+\log\left((x+h)^2+y^2+\frac{h^2}{2}\right)\\ -4\log\left(x^2+y^2+\frac{h^2}{2}\right)\end{array}\right\}$$

$$= \frac{h^{-2}}{4\log 3}\log\frac{\left\{\begin{array}{l}\left((x+h)^2+y^2+\frac{h^2}{2}\right)\cdot\left((x-h)^2+y^2+\frac{h^2}{2}\right)\\ \cdot\left(x^2+(y+h)^2+\frac{h^2}{2}\right)\cdot\left(x^2+(y-h)^2+\frac{h^2}{2}\right)\end{array}\right\}}{\left(x^2+y^2+\frac{h^2}{2}\right)^4}.$$

For proving $-\Delta_h u_h \geq 0$, one has to show that

$$\left((x+h)^2+y^2+\frac{h^2}{2}\right)\left((x-h)^2+y^2+\frac{h^2}{2}\right)\left(x^2+(y+h)^2+\frac{h^2}{2}\right)$$
$$\cdot\left(x^2+(y-h)^2+\frac{h^2}{2}\right)\geq\left(x^2+y^2+\frac{h^2}{2}\right)^4.$$

This inequality follows from the identity

$$\begin{array}{l}\left((x+h)^2+y^2+\frac{h^2}{2}\right)\left((x-h)^2+y^2+\frac{h^2}{2}\right)\left(x^2+(y+h)^2+\frac{h^2}{2}\right)\left(x^2+(y-h)^2+\frac{h^2}{2}\right)\\ -\left(x^2+y^2+\frac{h^2}{2}\right)^4 \;=\; 5h^8+4h^6x^2+4h^6y^2+16h^4x^2y^2 \;\geq\; 0.\end{array}$$

(d) Since $s_h(\mathbf{x},\boldsymbol{\xi}) = u_h(\mathbf{x}-\boldsymbol{\xi})$, the parts (b) and (c) show that

$$-\left(\Delta_h s_h(\cdot,\boldsymbol{\xi})\right)(\mathbf{x})\geq 0 \quad\text{and}\quad -\left(\Delta_h s_h(\cdot,\mathbf{x})\right)(\mathbf{x}) = h^{-2}.$$

The grid function $\phi_{\boldsymbol{\xi}} := s_h(\cdot,\boldsymbol{\xi}) - g_h(\cdot,\boldsymbol{\xi})$ satisfies

$$-\Delta_h\phi_{\boldsymbol{\xi}}(\mathbf{x}) = -\Delta_h\left[s_h(\cdot,\boldsymbol{\xi})-g_h(\cdot,\boldsymbol{\xi})\right](\mathbf{x}) \quad \left\{\begin{array}{ll}= h^{-2}-h^{-2}=0 & \text{for } \mathbf{x}=\boldsymbol{\xi}\in\Omega_h\,,\\ \geq 0-0=0 & \text{for } \mathbf{x}\neq\boldsymbol{\xi}\in\Omega_h\,.\end{array}\right.$$

According to part (a), the values at boundary points $\mathbf{x}\in\Gamma_h$ are

$$\phi_{\boldsymbol{\xi}}(\mathbf{x}) = s_h(\mathbf{x},\boldsymbol{\xi}) - g_h(\mathbf{x},\boldsymbol{\xi}) = s_h(\mathbf{x},\boldsymbol{\xi})\geq 0.$$

Theorem 4.38b implies $\phi_{\boldsymbol{\xi}}\geq 0$ in Ω_h, hence $g_h(\mathbf{x},\boldsymbol{\xi})\leq s_h(\mathbf{x},\boldsymbol{\xi})$, proving the inequality (4.42). ∎

Let φ_h be defined as in (4.13b). The solution of the discrete potential equation

$$-\Delta_h u_h = 0 \quad\text{in } \Omega_h, \qquad u_h = \varphi \quad\text{on } \Gamma_h$$

is given by $u_h := L_h^{-1}\varphi_h$ (if one continues the grid function, at first only defined on Ω_h, through φ on Γ_h; cf. Remark 4.7). The representation with the aid of g_h reads

$$u_h(\mathbf{x}) = h^2\sum_{\boldsymbol{\xi}\in\Omega_h} g_h(\boldsymbol{\xi},\mathbf{x})\varphi_h(\boldsymbol{\xi}) \qquad (\mathbf{x}\in\Omega_h).$$

Since $\varphi_h(\mathbf{x})$ vanishes at all interior points $\mathbf{x}$, it suffices to extend the sum over the near-boundary points. Summing over the boundary points neighboured to

$$\Gamma_h' := \{\boldsymbol{\xi} \in \Gamma_h : \boldsymbol{\xi} \text{ is not a corner point } (0,0)\,,\ (1,0)\,,\ (0,1)\,,\ (1,1)\}$$

instead of over the boundary points, then the definition of φ_h results

$$u_h(\mathbf{x}) = -h \sum_{\boldsymbol{\xi}\in\Gamma_h'} \partial_n^- \bar{g}_h(\boldsymbol{\xi}, \mathbf{x})\varphi(\boldsymbol{\xi}) \qquad (\mathbf{x} \in \Omega_h), \tag{4.44}$$

where ∂_n^- is the backward difference in the direction of the normal $\mathbf{n}$:

$$\partial_n^- \bar{g}_h(\boldsymbol{\xi}, \mathbf{x}) = h^{-1}\left[\bar{g}_h(\boldsymbol{\xi}, \mathbf{x}) - \bar{g}_h(\boldsymbol{\xi} - h\mathbf{n}, \mathbf{x})\right]$$

(note that $\bar{g}_h(\boldsymbol{\xi}, \mathbf{x}) = 0$ for $\boldsymbol{\xi} \in \Gamma_h' \subset \Gamma_h$). The variable $\boldsymbol{\xi} - h\mathbf{n}$ ranges over all near-boundary points.

Remark 4.43. Equation (4.44) corresponds to the representation in Theorem 3.16. The summation $h\sum_{\boldsymbol{\xi}\in\Gamma_h'}$ approximates the integral $\int_\Gamma$.

Finally, we want to take a closer look at the estimate for the solution $u_h = L_h^{-1} q_h$ through

$$\|u_h\|_\infty \le \|L_h^{-1}\|_\infty \|q_h\|_\infty \le \|q_h\|_\infty / 8$$

(cf. (4.36c)). According to (4.13b), $q_h = f_h + \varphi_h$ contains the right-hand side $f_h(\mathbf{x}) = f(\mathbf{x})$ of the discrete Poisson equation and the boundary values $\varphi(\mathbf{x}), \mathbf{x} \in \Gamma_h$, hidden in φ_h. The following theorem gives a bound in which these components are separated.

Theorem 4.44. *According to (4.13b), let $q_h = f_h + \varphi_h$ be constructed from f and φ. The discrete solution $u_h = L_h^{-1} q_h$ of the Poisson boundary-value problem can be bounded by*

$$\|u_h\|_\infty \le \|L_h^{-1}\|_\infty \max_{\mathbf{x}\in\Omega_h} |f(\mathbf{x})| + \max_{\boldsymbol{\xi}\in\Gamma_h'} |\varphi(\boldsymbol{\xi})| \le \frac{1}{8} \max_{\mathbf{x}\in\Omega_h} |f(\mathbf{x})| + \max_{\boldsymbol{\xi}\in\Gamma_h'} |\varphi(\boldsymbol{\xi})|\,. \tag{4.45}$$

Proof. Set $u_h' := L_h^{-1} f_h$ and $u_h'' := L_h^{-1}\varphi_h$. The estimate for the first term in $u_h = u_h' + u_h''$ results in the first term in (4.45). To bound u_h'' use the inequality (4.39a) with $u_h^1 = u_h''$, $\varphi^1 = \varphi$ and $u_h^2 = 0$, $\varphi^2 = 0$. ∎

The corresponding inequality

$$\|u_h\|_\infty \le C\,\|f\|_\infty + \|\varphi\|_\infty$$

for the solution of the boundary-value problem (4.9a,b) has not been mentioned until now, but will be proved in a more general context in §5.1.3.

The maximum norm $\|\cdot\|_\infty$ in (4.45) can be replaced by the Euclidean norms

$$\|u_h\|_{2,\Omega_h} := \sqrt{h^2 \sum_{\mathbf{x}\in\Omega_h} |u_h(\mathbf{x})|^2}, \qquad \|\varphi\|_{2,\Gamma_h'} := \sqrt{h^2 \sum_{\boldsymbol{\xi}\in\Gamma_h'} |\varphi(\boldsymbol{\xi})|^2}.$$

Here, $h^2\sum_{\Omega_h}$ and $\int_\Omega$, $h\sum_{\Gamma_h'}$ and $\int_\Gamma$ correspond to each other.

Theorem 4.45. *Under the assumptions of Theorem 4.44 there holds*

$$\|u_h\|_{2,\Omega_h} \le \|L_h^{-1}\|_2 \, \|f_h\|_{2,\Omega_h} + \frac{1}{\sqrt{2}} \, \|\varphi\|_{2,\Gamma_h'} \le \frac{1}{16} \, \|f_h\|_{2,\Omega_h} + \frac{1}{\sqrt{2}} \, \|\varphi\|_{2,\Gamma_h'} \, .$$

Proof. (a) It suffices to consider the case of the potential equation (i.e., $f = 0$). Let the restriction of φ on Γ_h' result in the grid function $\phi_h : \ \phi_h(\mathbf{x}) = \varphi(\mathbf{x})$ for $\mathbf{x} \in \Gamma_h'$. Let the mapping $\phi_h \to u_h = L_h^{-1}\varphi_h$ be given by the rectangular matrix A: $u_h = A\phi_h$. According to equation (4.44) the entries of A read

$$a_{\mathbf{x}\boldsymbol{\xi}} = -h\partial_n^- \bar{g}_h(\boldsymbol{\xi}, \mathbf{x}) = g_h(\boldsymbol{\xi} - h\mathbf{n}, \mathbf{x}) = g_h(\mathbf{x}, \boldsymbol{\xi} - h\mathbf{n}) \quad \text{for } \mathbf{x} \in \Omega_h, \ \boldsymbol{\xi} \in \Gamma_h'.$$

Since $A > O$ as stated in Remark 4.41, one obtains the row-sum norm $\|A\|_\infty := \max_{\mathbf{x}} \sum_{\boldsymbol{\xi}} a_{\mathbf{x}\boldsymbol{\xi}}$ as $\|A\phi_h\|_\infty$ for the choice $\phi_h(\mathbf{x}) = 1$ in all $\mathbf{x} \in \Gamma_h'$. The solution $v_h = A\phi_h$ then reads $v_h = \mathbf{1}$, so it follows that

$$\|A\|_\infty = \|A\phi_h\|_\infty = \|\mathbf{1}\|_\infty = 1.$$

(b) The column sums of A are $s(\boldsymbol{\xi}) := \sum_{\mathbf{x}\in\Omega_h} a_{\mathbf{x}\boldsymbol{\xi}} = \sum_{\mathbf{x}\in\Omega_h} g_h(\mathbf{x}, \boldsymbol{\xi} - h\mathbf{n})$ for $\boldsymbol{\xi} \in \Gamma_h'$. The grid function $v_h := h^{-2}L_h^{-1}\mathbf{1}$ at the near-boundary points $\boldsymbol{\xi} - h\mathbf{n}$ has the values

$$s(\boldsymbol{\xi}) = v_h(\boldsymbol{\xi} - h\mathbf{n}) \qquad (\boldsymbol{\xi} \in \Gamma_h', \ \mathbf{n} \text{ normal direction}),$$

as is implied by Remark 4.40. Let $\boldsymbol{\xi} = (\xi_1, \xi_2) \in \Gamma_h'$ be a point of the left or right boundary (i.e., $\xi_1 = 0$ or 1). As mentioned in the proof of (4.36c), $L_h w_h \ge \mathbf{1}$ holds for $w_h(x, y) := x(1-x)/2$. Since $L_h^{-1} \ge O$, then so is $w_h \ge L_h^{-1}\mathbf{1}$ and hence $v_h \le h^{-2}w_h$. In particular we have the following estimate

$$s(\boldsymbol{\xi}) = v_h(\boldsymbol{\xi} - h\mathbf{n}) \le h^{-2}w_h(\boldsymbol{\xi} - h\mathbf{n}) = h^{-2}h\,(1-h)\,/2 \le h^{-1}/2$$

at the near-boundary point $\boldsymbol{\xi} - hn = (h, \xi_2)$ or $(1-h, \xi_2)$. For a point $\boldsymbol{\xi}$ from the upper or from the lower boundary one obtains the same estimate if one uses $w_h(x, y) := y(1-y)/2$. Since the column sums $s(\boldsymbol{\xi})$ are the row sums of A^T, we have proved

$$\|A^\mathsf{T}\|_\infty = \max\{s(\boldsymbol{\xi}) : \boldsymbol{\xi} \in \Gamma_h'\} \le h^{-1}/2.$$

(c) We have $\|A^\mathsf{T}A\|_2 = \rho(A^\mathsf{T}A) \le \|A^\mathsf{T}A\|_\infty \le \|A^\mathsf{T}\|_\infty \|A\|_\infty \le \frac{1}{2h}$ (cf. Exercise 4.28a, (4.31b), (4.31a)) so that the solution $u_h = L_h^{-1}\varphi_h = A\phi_h$ satisfies the following estimate

$$\begin{aligned}
\|u_h\|_{2,\Omega_h}^2 &= h^2 \sum\nolimits_{\mathbf{x}\in\Omega_h} |u_h(\mathbf{x})|^2 = h^2 \sum\nolimits_{\mathbf{x}\in\Omega_h} |(A\phi_h)(\mathbf{x})|^2 \\
&= h^2 \sum\nolimits_{\mathbf{x}\in\Gamma_h'} \phi_h(\mathbf{x})(A^\mathsf{T}A\phi_h)(\mathbf{x}) \le h^2 \left\|A^\mathsf{T}A\right\|_2 \sum\nolimits_{\mathbf{x}\in\Gamma_h'} \phi_h^2(\mathbf{x}) \\
&\le \frac{h}{2} \sum\nolimits_{\mathbf{x}\in\Gamma_h'} \phi_h^2(\mathbf{x}) = \frac{1}{2}\|\phi_h\|_{2,\Gamma_h'}^2 .
\end{aligned}$$

Since $\|\varphi\|_{2,\Gamma_h'} = \|\phi_h\|_{2,\Gamma_h'}$ the assertion follows. ∎

4.5 Convergence

Let U_h be the vector space of the grid functions on $\overline{\Omega}_h$. The discrete solution $u_h \in U_h$ and the continuous[1] solution $u \in C^0(\overline{\Omega})$ cannot be compared directly because of the different domains of definition. In difference methods it is customary to compare both functions on the grid $\overline{\Omega}_h$. To this end one must map the solution u by means of a *restriction*

$$u \mapsto R_h u \in U_h \qquad (u\text{: continuous solution})$$

to U_h. In the following we choose R_h as restriction on $\overline{\Omega}_h$:

$$(R_h u)(\mathbf{x}) = u(\mathbf{x}) \qquad \text{for all } \mathbf{x} \in \overline{\Omega}_h. \tag{4.46}$$

The limit $h \to 0$ is made precise as follows. Let $H \subset \mathbb{R}_+$ be a subset with accumulation point zero: $0 \in \overline{H}$. For example, the step sizes considered so far form the set $H = \{1/n : n \in \mathbb{N}\}$. For each $h \in H$ let U_h be equipped with the norm $\|\cdot\|_h$.

Definition 4.46 (convergence). The discrete solutions $u_h \in U_h$ converge (with respect to the family of norms $\|\cdot\|_h$, $h \in H$) to u if

$$\|u_h - R_h u\|_h \to 0.$$

We have *convergence of order* k if

$$\|u_h - R_h u\|_h = \mathcal{O}(h^k).$$

The proof of convergence is usually carried out with the aid of the concepts of 'stability' and 'consistency' (more details in Hackbusch [139]). Below we consider inequalities as $\|u_h - R_h u\|_h \leq \ldots$, which are called *error estimates* since $u_h - R_h u$ is the *discretisation error*.

The discretisation $\{L_h : h \in H\}$ is said to be *stable* with respect to $\|\cdot\|_\infty$ if

$$\sup_{h \in H} \|L_h^{-1}\|_\infty < \infty.$$

For the discretisation defined in §4.2, the stability has been proved in (4.36c) with respect to the row-sum norm, and in (4.36d) with respect to the spectral norm.

The grid function f_h in $-\Delta_h u_h = f_h$ is the restriction

$$f_h = \tilde{R}_h f$$

of f, where in (4.13b) $\tilde{R}_h$ was chosen as the restriction to Ω_h:

$$(\tilde{R}_h f)(\mathbf{x}) = f(\mathbf{x}) \qquad \text{for all } \mathbf{x} \in \Omega_h. \tag{4.47}$$

The notation $\tilde{R}_h$ indicates that a choice $\tilde{R}_h \neq R_h$ is possible.

[1] Here, 'continuous' is used as counterpart of 'discrete'.

We recall Remark 4.7: $D_h u_h = f_h$ describes the discretisation of the differential equation $Lu = f$. Together with the boundary condition $u = \varphi$ on Γ, $D_h u_h = f_h$ and the system $L_h u_h = q_h$ are equivalent; more precisely, the definition of q_h (cf. (4.13b)) gives $D_h u_h - f_h = L_h u_h - q_h$. In the following it is easier to estimate $D_h u_h - f_h$ since we can make use of the equivalents $u_h \leftrightarrow u$ and $f_h \leftrightarrow f$.

The discretisation described by the triple $(D_h, R_h, \tilde{R}_h)$ is said to be *consistent of order k* with respect to $\|\cdot\|_\infty$ (consistent with L) if

$$\|D_h R_h u - \tilde{R}_h L u\|_\infty \leq K h^k \|u\|_{C^{k+2}(\overline{\Omega})} \tag{4.48}$$

for all $u \in C^{k+2}(\overline{\Omega})$. Here K is independent of h and u.

Remark 4.47. Let R_h and $\tilde{R}_h$ be given by (4.46) and (4.47). The five-point formula Δ_h is consistent of order 2: estimate (4.48) holds with $k = 2$ and $K = 1/6$.

Proof. The expansion (4.3) can be applied in the x and y directions and yields

$$\Delta_h R_h u(x,y) = \Delta u(x,y) + h^2(R_{4,x} + R_{4,y}) \text{ with } |R_{4,x}|, |R_{4,y}| \leq \frac{\|u\|_{C^4(\overline{\Omega})}}{12}. \blacksquare \tag{4.49}$$

The general concept for proving convergence is based on

$$L_h R_h u - q_h = D_h R_h u - f_h = D_h R_h u - \tilde{R}_h f = \Big(D_h R_h - \tilde{R}_h L\Big) u =: \varepsilon_{\text{consistent}}(u)$$

and

$$L_h (u_h - R_h u) = L_h u_h - L_h R_h u = q_h - L_h R_h u = -\varepsilon_{\text{consistent}}(u).$$

Application of L_h^{-1} yields

$$\|u_h - R_h u\| = \|L_h^{-1} \varepsilon_{\text{consistent}}(u)\| \leq \|L_h^{-1}\| \|\varepsilon_{\text{consistent}}(u)\|$$

with respect to some norm $\|\cdot\|$ and the associated matrix norm. Here we used the maximum and row-sum norm $\|\cdot\|_\infty$. An immediate consequence of this consideration is the next theorem.

Theorem 4.48. *Let the discretisation $(D_h, R_h, \tilde{R}_h)$ be consistent of order k. Let the matrix L_h associated to the difference operator D_h be stable. Then the method is convergent of order k if $u \in C^{k+2}(\overline{\Omega})$.*

In the case of $L = \Delta$, $D_h = \Delta_h$ and the norm $\|\cdot\|_\infty$, the values $\|L_h^{-1}\|_\infty \leq \frac{1}{8}$ (cf. (4.36c)) and $K = \frac{1}{6}$ (cf. Remark 4.47) prove the following corollary.

Corollary 4.49. Let the continuous solution of the boundary-value problem (4.9a,b) belong to $C^4(\overline{\Omega})$. Let u_h be the discrete solution defined in (4.11a,b). Then the convergence of u_h to u is of second order:

$$\|u_h - R_h u\|_\infty \leq \frac{h^2}{48} \|u\|_{C^4(\overline{\Omega})} . \tag{4.50}$$

The assumption $u \in C^4(\overline{\Omega})$ can be weakened.

Corollary 4.50. Under the condition $u \in C^{3,1}(\overline{\Omega})$ we also have

$$\|u_h - R_h u\|_\infty \leq \frac{h^2}{48} \|u\|_{C^{3,1}(\overline{\Omega})} .$$

Proof. The remainder term R_4 in (4.4) may also be written as

$$R_4 = h^{-4} \int_x^{x \pm h} [u'''(\xi) - u'''(x)] \frac{(x \pm h - \xi)^2}{2!} \mathrm{d}\xi .$$

The Lipschitz estimate $|u'''(\xi) - u'''(x)| \leq |\xi - x| \, \|u\|_{C^{3,1}(\overline{\Omega})}$ implies the estimate $R_4 \leq \|u\|_{C^{3,1}(\overline{\Omega})} / 4!$ so that in (4.49), (4.48), and (4.50) the norm $\|u\|_{C^4(\overline{\Omega})}$ can be replaced by $\|u\|_{C^{3,1}(\overline{\Omega})}$. ∎

If, however, one further weakens $u \in C^{3,1}(\overline{\Omega})$ to $u \in C^s(\overline{\Omega})$, $2 < s < 4$, one obtains a weaker order of convergence.

Corollary 4.51. Under the condition $u \in C^s(\overline{\Omega})$, $2 < s < 4$, u_h converges of order $s - 2$:

$$\|u_h - R_h u\|_\infty \leq K_s h^{s-2} \|u\|_{C^s(\overline{\Omega})} ,$$

where $K_s = \frac{1}{2s(s-1)}$ for $2 \leq s < 3$, $K_3 = \frac{1}{24}$ and $K_s = \frac{2^{s-5}}{s(s-1)(s-2)}$ for $3 < s < 4$.

The proof results from $\|L_h^{-1}\|_\infty \leq \frac{1}{8}$ and the following consistency estimate.

Exercise 4.52. Show

$$\|\Delta_h R_h u - \tilde{R}_h \Delta u\|_\infty \leq 8 K_s h^{s-2} \|u\|_{C^s(\overline{\Omega})} , \quad 2 < s < 4,$$

with the above constant K_s.

Even though the proofs of convergence are simple, the results remain unsatisfactory. As can be seen from Example 2.3, the continuous solution of the boundary-value problem (4.9a,b) generally does not even satisfy $u \in C^2(\overline{\Omega})$, although one needs at least $u \in C^s(\overline{\Omega})$ with $s > 2$ in Corollary 4.51 for convergence. Stronger results can be obtained by an analysis which will be discussed in §9.3. That errors of the order of magnitude of §9.3 occur even under weaker conditions, is shown next.

Example 4.53. If one solves the difference equation (4.11a,b) for

$$-\Delta u = 1 \text{ in } \Omega = (0,1)\times(0,1) \qquad \text{and } u = 0 \text{ on } \Gamma,$$

one obtains at the centre $x = y = \frac{1}{2}$ the values $u_h(\frac{1}{2},\frac{1}{2})$, which are shown in the first column of Table 4.1. The exact solution $u(\frac{1}{2},\frac{1}{2}) = 0.0736713\ldots$ results from a representation that one can find in Example 8.18. The quotients $\varepsilon_h/\varepsilon_{2h}$ of the errors $\varepsilon_h = u(\frac{1}{2},\frac{1}{2}) - u_h(\frac{1}{2},\frac{1}{2})$ approximate $1/4$. This proves $u_h(\frac{1}{2},\frac{1}{2}) = u(\frac{1}{2},\frac{1}{2}) + \mathcal{O}(h^2)$, although $u \notin C^2(\overline{\Omega})$. At $(\frac{1}{2},\frac{1}{2})$ u_h has furthermore the asymptotic expansion

$$u_h\big(\tfrac{1}{2},\tfrac{1}{2}\big) = u\big(\tfrac{1}{2},\tfrac{1}{2}\big) + h^2 e\big(\tfrac{1}{2},\tfrac{1}{2}\big) + \mathcal{O}(h^4).$$

The error term $e(\frac{1}{2},\frac{1}{2})$ independent of h, is eliminated by using the *Richardson extrapolation*

$$u_{h,2h}\big(\tfrac{1}{2},\tfrac{1}{2}\big) := \frac{1}{3}\left[4u_h\big(\tfrac{1}{2},\tfrac{1}{2}\big) - u_{2h}\big(\tfrac{1}{2},\tfrac{1}{2}\big)\right]$$

(cf. Richardson–Gaunt [236]). The extrapolated values are already very accurate for $h = \frac{1}{16}$ (see the last column of Table 4.1).

The reason for this favourable behaviour is the inner regularity of the discrete solution (cf. §9.3.6).

h	$u_h(\frac{1}{2},\frac{1}{2})$	$u_{h,2h}(\frac{1}{2},\frac{1}{2})$	ε_h	Quotient	$\varepsilon_{h,2h}$	Quotient
1/8	0.0727826		8.89_{10}-4			
1/16	0.07344576	0.0736668	2.26_{10}-4	0.250	4.5_{10}-6	
1/32	0.0736147373	0.07367106	5.66_{10}-5	0.251	2.7_{10}-7	0.06
1/64	0.0736571855	0.07367133	1.41_{10}-5	0.249		

Table 4.1 Solution of Example 4.53; $\varepsilon_{h[,2h]} := \big|u(\frac{1}{2},\frac{1}{2}) - u_{h[,2h]}(\frac{1}{2},\frac{1}{2})\big|$.

4.6 Discretisations of Higher Order

The five-point formula (4.19) is of second order. Even if the solution u belongs to $C^s(\overline{\Omega})$ with $s > 4$, no better bound for $\Delta_h R_h u - R_h \Delta u$ than $\mathcal{O}(h^2)$ would result. An obvious method for constructing difference methods of higher order is the following. As an ansatz for the discretisation of the second derivative u'' choose

$$(D_h u_h)(x) = h^{-2} \sum_{\nu=-k}^{k} c_\nu u_h(x+\nu h).$$

The Taylor expansion provides

$$(D_h R_h u)(x) = \sum_{\mu=0}^{2k} a_\mu h^{\mu-2} u^{(\mu)}(x) + \mathcal{O}(h^{2k-1}), \qquad a_\mu = \frac{1}{\mu!}\sum_{\nu=-k}^{k} c_\nu \nu^k.$$

The $2k+1$ equations $a_0 = a_1 = a_3 = a_4 = \ldots = a_{2k} = 0$ and $a_2 = 1$ form a linear system for the $2k+1$ unknown coefficients c_ν. For $k=1$ one obtains the usual difference formula (4.3); for $k=2$ a difference of fourth order results:

$$h^2 (D_h u_h)(x) = -\frac{1}{12}[u_h(x-2h) + u_h(x+2h)] + \frac{4}{3}[u_h(x-h) + u_h(x+h)] - \frac{5}{2}u_h(x).$$

If one applies this approximation for u'' to the x and y coordinates, one obtains for $-\Delta$ the difference star

$$\frac{h^{-2}}{12}\begin{bmatrix} & & 1 & & \\ & & -16 & & \\ 1 & -16 & 60 & -16 & 1 \\ & & -16 & & \\ & & 1 & & \end{bmatrix} \tag{4.51}$$

(cf. (4.20)). The difference scheme (4.51) is of fourth order, but presents difficulties at points near the boundary. To set up the difference formula at $(h,h) \in \Omega_h$, for example, one needs the values $u_h(-h,h)$ and $u_h(h,-h)$ outside $\overline{\Omega}_h$ (cf. Figure 4.4). One possibility would be to use scheme (4.51) only at points far from the boundary and to use the five-point formula (4.19) at points near the boundary. Another is the extrapolation from far-boundary points of $\overline{\Omega}_h$.

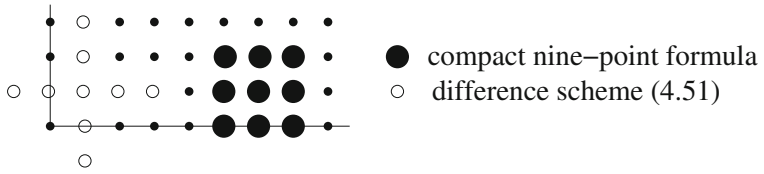


Fig. 4.4 Difference scheme (4.51) and compact nine-point formula (4.52).

The above complications do not occur if one limits oneself to *compact nine-point formulae*; by this one means difference methods (4.20) which are characterised by

$$c_{\alpha\beta} \neq 0 \quad \text{only for } -1 \leq \alpha, \beta \leq 1$$

(cf. Figure 4.4). An ansatz with the nine free parameters $c_{\alpha,\beta}$ $(-1 \leq \alpha, \beta \leq 1)$, leads, however, to a negative result: there is no compact nine-point formula with $D_h u = -\Delta u + \mathcal{O}(h^3)$. In this sense the five-point formula is already optimal. Nevertheless, nine-point schemes of fourth order can be obtained if one also selects the right-hand side f_h, of the system of equations (4.13a,b) in a suitable manner.

If one applies the compact nine-point scheme

$$D_h := \frac{h^{-2}}{6}\begin{bmatrix} -1 & -4 & -1 \\ -4 & 20 & -4 \\ -1 & -4 & -1 \end{bmatrix} \tag{4.52}$$

to a sufficiently smooth u, the Taylor expansion results in

$$D_h u = -\Delta u - \frac{h^2}{12}\Delta^2 u - \frac{h^4}{360}\left[\frac{\partial^4}{\partial x^4} + 4\frac{\partial^4}{\partial x^2 \partial y^2} + \frac{\partial^4}{\partial y^4}\right]\Delta u + \mathcal{O}(h^6). \quad (4.53)$$

Here it is crucial that the error term can be expressed by $-\Delta u$ and hence by f. For the special choice of the restriction $\tilde{R}_h$ via

$$f_h = \tilde{R}_h f := \frac{1}{12}\begin{bmatrix} & 1 & \\ 1 & 8 & 1 \\ & 1 & \end{bmatrix} f,$$

i.e.,

$$\begin{aligned} f_h(x,y) &= (\tilde{R}_h f)(x,y) \qquad (4.54)\\ &:= \frac{1}{12}\left[f(x+h,y) + f(x-h,y) + f(x,y+h) + f(x,y-h) + 8f(x,y)\right] \end{aligned}$$

one obtains the expansion

$$f_h(x,y) = f(x,y) + \frac{h^2}{12}\Delta f(x,y) + \frac{h^4}{144}\left(\frac{\partial^4 f(x,y)}{\partial x^4} + \frac{\partial^4 f(x,y)}{\partial y^4}\right) + \mathcal{O}(h^4), \quad (4.55)$$

which, because $f = -\Delta u$, agrees with (4.53) up to $\mathcal{O}(h^4)$.

The matrix L_h of the system of equations which results after the elimination of the boundary values $u_h(\mathbf{x}) = \varphi(\mathbf{x})$ in $\mathbf{x} \in \Gamma_h$, has the entries

$$L_{\mathbf{x}\boldsymbol{\xi}} = h^{-2}\begin{cases} 20/6 & \text{if } \mathbf{x} = \boldsymbol{\xi}, \\ -1/6 & \text{if } \mathbf{x} - \boldsymbol{\xi} = (\pm h, \pm h) \text{ or } \mathbf{x} - \boldsymbol{\xi} = (\pm h, \mp h), \\ -4/6 & \text{if } \mathbf{x} - \boldsymbol{\xi} = (\pm h, 0) \quad \text{or } \mathbf{x} - \boldsymbol{\xi} = (0, \pm h), \\ 0 & \text{otherwise.} \end{cases} \quad (4.56)$$

The right-hand side of the system of equations $L_h u_h = q_h$ is

$$q_h := f_h + \varphi_h, \quad f_h = \tilde{R}_h f \text{ according to (4.54)}, \quad \varphi_h := \sum_{\boldsymbol{\xi} \in \Gamma_h} L_{\mathbf{x}\boldsymbol{\xi}}\varphi(\boldsymbol{\xi}). \quad (4.57)$$

The discretisation $(D_h, R_h, \tilde{R}_h)$ with D_h from (4.52), R_h from (4.46), $\tilde{R}_h$ from (4.54) is called the *mehrstellen method* (cf. Collatz [74, §V.2.5]).

Exercise 4.54. Let D_h and L_h be defined respectively by (4.52) and (4.56). Prove:
(a) L_h is an M-matrix;
(b) $\|L_h^{-1}\|_\infty \le 1/8$ (stability), $\|L_h\|_\infty \le 20h^{-2}/3$.

Theorem 4.55 (convergence of the mehrstellen method). *Let Ω_h be defined as in §4.2. Let u_h be the solution provided by the mehrstellen method $L_h u_h = q_h$ with L_h from (4.56), and q_h from (4.57). Let the solution u of the boundary-value problem*

(4.9a,b) belong to $C^6(\overline{\Omega})$. Then the following error estimates hold in the respective cases:

$$\|u_h - R_h u\|_\infty \le \frac{7}{2880} h^4 \|f\|_{C^4(\overline{\Omega})} + o(h^4) \tag{4.58}$$

as well as

$$\|u_h - R_h u\|_\infty \le \frac{1}{360} h^4 \|u\|_{C^6(\overline{\Omega})} + o(h^4).$$

In the case of the potential equation, i.e., $f = 0$, we even have

$$\|u_h - R_h u\|_\infty \le K h^6 \|u\|_{C^8(\overline{\Omega})}, \qquad \|u_h - R_h u\|_\infty \le K h^6 \|u\|_{C^{7,1}(\overline{\Omega})},$$

of $u \in C^8(\overline{\Omega})$, resp. $u \in C^{7,1}(\overline{\Omega})$.

Proof. Since $f = -\Delta u$, the combination of the h^4 remainder terms in (4.53) and (4.55) yield

$$\begin{aligned} & h^4 \left[\frac{1}{360} \left(\frac{\partial^4}{\partial x^4} + 4 \frac{\partial^4}{\partial x^2 \partial y^2} + \frac{\partial^4}{\partial y^4} \right) - \frac{1}{144} \left(\frac{\partial^4}{\partial x^4} + \frac{\partial^4}{\partial y^4} \right) \right] f \\ &= \frac{h^4}{720} \left(-3 \frac{\partial^4}{\partial x^4} + 8 \frac{\partial^4}{\partial x^2 \partial y^2} - 3 \frac{\partial^4}{\partial y^4} \right) f \\ &= \frac{h^4}{720} \left(3 \frac{\partial^6}{\partial x^6} - 5 \frac{\partial^6}{\partial x^4 \partial y^2} - 5 \frac{\partial^6}{\partial x^2 \partial y^4} + 3 \frac{\partial^6}{\partial y^6} \right) u. \end{aligned}$$

The above error of consistency is to be multiplied by the stability constant $\|L_h^{-1}\|_\infty \le 1/8$ in Exercise 4.54b (cf. Theorem 4.48). Inequality (4.58) follows from $\frac{3+8+3}{720}\frac{1}{8} = \frac{7}{2880}$, resp. $\frac{3+5+5+3}{720}\frac{1}{8} = \frac{1}{360}$. If $f = 0$ also the $\mathcal{O}(h^4)$ term in (4.53) vanishes. ∎

4.7 The Discretisation of the Neumann Boundary-Value Problem

The Dirichlet boundary values $u(\mathbf{x}) = \varphi(\mathbf{x})$ were used directly in the difference method; a discretisation was not necessary. A different situation arises for the Neumann boundary-value problem

$$-\Delta u = f \quad \text{in } \Omega = (0,1) \times (0,1), \qquad \frac{\partial u}{\partial n} = \varphi \quad \text{on } \Gamma. \tag{4.59}$$

The normal derivative, which reads explicitly

$$\begin{aligned} \frac{\partial u}{\partial n} &= -u_y \text{ for } \mathbf{x} = (x_1, 0) \in \Gamma, & \frac{\partial u}{\partial n} &= u_y \text{ for } \mathbf{x} = (x_1, 1) \in \Gamma, \\ \frac{\partial u}{\partial n} &= -u_x \text{ for } \mathbf{x} = (0, x_2) \in \Gamma, & \frac{\partial u}{\partial n} &= u_x \text{ for } \mathbf{x} = (1, x_2) \in \Gamma, \end{aligned}$$

like the Laplace operator, must be replaced by a difference. We will investigate three different discretisations.

4.7.1 One-Sided Difference for $\partial u/\partial n$

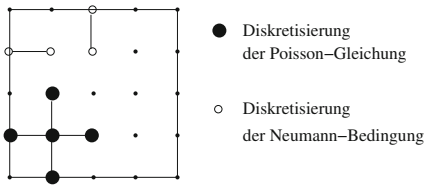


Fig. 4.5 Discretisation of the differential equation and the Neumann boundary condition.

The Poisson equation leads to the $(n-1)^2 = (1/h-1)^2$ equations

$$\left(-\Delta_h u_h\right)(\mathbf{x}) = f(\mathbf{x}) \qquad (4.60a)$$

for all $\mathbf{x} \in \Omega_h$, which require the values of $u_h(\mathbf{x})$ for all $\mathbf{x} \in \overline{\Omega}_h'$, where

$$\overline{\Omega}_h' := \Omega_h \cup \Gamma_h', \qquad (4.60b)$$
$$\Gamma_h' := \Gamma_h \backslash \{(0,0),(0,1),(1,0),(1,1)\}.$$

To obtain a further $4(n-1)$ equations for $\{u_h(\mathbf{x}) : \mathbf{x} \in \Gamma_h'\}$, we replace, at all $\mathbf{x} \in \Gamma_h'$, the normal derivative $\partial u/\partial n = \varphi$ by the *backward difference*

$$\left(\partial_n^- u_h\right)(\mathbf{x}) := \frac{1}{h}\left[u_h(\mathbf{x}) - u_h(\mathbf{x} - h\mathbf{n})\right] = \varphi(\mathbf{x}) \qquad \text{for } \mathbf{x} \in \Gamma_h'. \qquad (4.61)$$

If one inserts the corresponding normal directions $\mathbf{n}$ for the four sides of the square one obtains

$$\left.\begin{array}{l} \frac{1}{h}\left[u_h(x,0) - u_h(x,h)\right] = \varphi(x,0) \\ \frac{1}{h}\left[u_h(x,1) - u_h(x,1-h)\right] = \varphi(x,1) \end{array}\right\} \text{for } x = h, 2h, \ldots, 1-h,$$

$$\left.\begin{array}{l} \frac{1}{h}\left[u_h(0,y) - u_h(h,y)\right] = \varphi(0,y) \\ \frac{1}{h}\left[u_h(1,y) - u_h(1-h,y)\right] = \varphi(1,y) \end{array}\right\} \text{for } y = h, 2h, \ldots, 1-h.$$

Equations (4.60a) and (4.61) yield $(n+1)^2 - 4$ equations for as many unknowns.

Exercise 4.56. After a rescaling of equations (4.61) to $h^{-1}\partial_n^- u_h(\mathbf{x}) = h^{-1}\varphi(\mathbf{x})$, $\mathbf{x} \in \Gamma_h'$, these equations, together with (4.60a), form a system $L_h u_h = q_h$. Show that L_h is symmetric and satisfies (4.21a).

As in the Dirichlet problem the variables $u_h(\mathbf{x})$, $\mathbf{x} \in \Gamma_h'$, can be eliminated with the aid of (4.61) in (4.60a). At the near-boundary point (h, y), for example, equation (4.60a) becomes

$$\frac{1}{h^2}\left[3u_h(h,y) - u_h(h,y-h) - u_h(h,y+h) - u_h(2h,y)\right] = f(h,y) + \frac{1}{h}\varphi(0,y).$$

The star $h^{-2}\begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix}$ thus becomes $h^{-2}\begin{bmatrix} & 0 & \\ -1 & 3 & -1 \\ & -1 & \end{bmatrix}$, $h^{-2}\begin{bmatrix} & -1 & \\ -1 & 3 & 0 \\ & -1 & \end{bmatrix}$, $h^{-2}\begin{bmatrix} & 0 & \\ -1 & 3 & -1 \\ & -1 & \end{bmatrix}$, $h^{-2}\begin{bmatrix} & -1 & \\ -1 & 3 & -1 \\ & 0 & \end{bmatrix}$ near, respectively, the left, right, upper, and lower boundaries. At the corner points one even has to replace two boundary values, so that, for example, at $(h,h) \in \Omega_h$ the star reads $h^{-2}\begin{bmatrix} & -1 & \\ 0 & 2 & -1 \\ & 0 & \end{bmatrix}$. Except for the

special case $h = \frac{1}{2}$, one obtains for the $(n-1)^2$ values $u_h(\mathbf{x})$, $\mathbf{x} \in \Omega_h$, the system of equations

$$L_h u_h = q_h \qquad \text{with} \tag{4.62a}$$

$$L_{\mathbf{xx}} = \begin{cases} 4/h^2 & \text{if } \mathbf{x} \in \Omega_h \text{ is a far-boundary point,} \\ 2/h^2 & \text{if } \mathbf{x} \in \{(h,h), (h,1-h), (1-h,h), (1-h,1-h)\}, \\ 3/h^2 & \text{otherwise,} \end{cases} \tag{4.62b}$$

$$L_{\mathbf{x}\boldsymbol{\xi}} = \begin{cases} -1/h^2 & \text{if } \mathbf{x}, \boldsymbol{\xi} \in \overline{\Omega}_h \text{ are neighbours,} \\ 0 & \text{otherwise for } \mathbf{x} \neq \boldsymbol{\xi}, \end{cases}$$

$$q_h = f_h + \varphi_h \quad \text{with} \begin{cases} f_h(\mathbf{x}) = (\tilde{R}_h f)(\mathbf{x}) := f(\mathbf{x}), \\ \varphi_h(\mathbf{x}) = -h \sum_{\boldsymbol{\xi} \in \Gamma_h} L_{\mathbf{x}\boldsymbol{\xi}} \varphi(\boldsymbol{\xi}). \end{cases} \tag{4.62c}$$

Remark 4.57. (a) L_h is symmetric and satisfies the sign condition (4.21a). (b) With lexicographical arrangement of the grid points of Ω_h, L_h has the form

$$L_h = h^{-2} \begin{bmatrix} T-I & -I & & & \\ -I & T & -I & & \\ & \ddots & \ddots & \ddots & \\ & & -I & T & -I \\ & & & -I & T-I \end{bmatrix}, \qquad T = \begin{bmatrix} 3 & -1 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -1 & 3 \end{bmatrix}.$$

The matrix L_h is singular because the system $L_h u_h = q_h$, like the continuous boundary-value problem (4.59), is, in general, not solvable. The analogue of Theorem 3.28 reads as follows.

Theorem 4.58. *The system of equations (4.62a) is solvable if and only if*

$$-h^2 \sum_{\mathbf{x} \in \Omega_h} f(\mathbf{x}) = h \sum_{\mathbf{x} \in \Gamma_h'} \varphi(\mathbf{x}). \tag{4.63}$$

Any two solutions u_h^1, u_h^2 of (4.62a) can only differ by a constant:

$$u_h^1 - u_h^2 = c\,\mathbf{1}, \qquad c \in \mathbb{R}.$$

Proof. Evidently, $L_h \mathbf{1} = 0$ holds, i.e., $\mathbf{1} \in \ker(L_h)$. Furthermore, Theorem 4.59 will then imply $\dim(\ker(L_h)) = 1$. This proves

$$\ker(L_h) = \{c\mathbf{1} : c \in \mathbb{R}\} \tag{4.64}$$

and thus the second part of the assertion. (4.62a) is solvable if and only if the scalar product $\langle v, q_h \rangle$ vanishes for all $v \in \ker(L_h^{\mathsf{T}}) = \mathrm{range}(L_h)^{\perp}$. Because of $L_h^{\mathsf{T}} = L_h$ and (4.64)

$$\langle \mathbf{1}, q_h \rangle = 0, \qquad \text{i.e., } \sum\nolimits_{\mathbf{x}\in\Omega_h} q_h(\mathbf{x}) = 0 \tag{4.65}$$

is sufficient and necessary. According to Definition (4.62c), equations (4.63) and (4.65) agree. ∎

Let condition (4.63) be satisfied. System (4.62a) can be solved as follows. Select an arbitrary $\mathbf{x}_0 \in \Omega_h$ and normalise the solution u_h (determined except for one constant) by

$$u_h(\mathbf{x}_0) = 0. \tag{4.66}$$

Let $\hat{u}_h$ be the vector u_h without the component $u_h(\mathbf{x}_0)$. Let $\hat{L}_h$ be the principal submatrix of L_h in which the row and column with index $\mathbf{x}_0$ have been left out. Let $\hat{q}_h$ be constructed likewise. Then

$$\hat{L}_h \hat{u}_h = \hat{q}_h \tag{4.67}$$

is a system with $(n-1)^2 - 1$ equations and unknowns.

Theorem 4.59. *The system of equations (4.67) is solvable; in particular, $\hat{L}_h$ is a symmetric M-matrix. Under condition (4.63), $\hat{u}_h = \hat{L}_h^{-1}\hat{q}_h$, supplemented by (4.66), yields the solution u_h of system (4.62a).*

Proof. (a) As a principal submatrix of L_h, $\hat{L}_h$ is symmetric. In $\Omega_h \backslash \{\mathbf{x}_0\}$ any two grid points can be connected by a chain of neighbouring points so that $\hat{L}_h$ is irreducible. For all $\mathbf{x} \in \Omega_h \backslash \{\mathbf{x}_0\}$ there holds (4.26b); at neighbouring points of $\mathbf{x}_0$ we even have (4.26a), so that $\hat{L}_h$ is irreducibly diagonally dominant. According to Criterion 4.18, $\hat{L}_h$ is an M-matrix, thus nonsingular.

(b) If u_h is the solution of (4.62a), one can assume (4.66) without loss of generality, so that u_h restricted to $\Omega_h \backslash \{\mathbf{x}_0\}$ also solves equation (4.67) and has to agree with the unique solution $\hat{u}_h$. ∎

As a corollary of Theorem 4.59 one obtains that $\operatorname{rank}(L_h) \geq \operatorname{rank}(\hat{L}_h) = (n-1)^2 - 1$, i.e., $\dim(\ker(L_h)) = 1$ and hence (4.64) holds.

Another possibility for the solution of equation (4.62a) is to pass to an *extended system of equations*

$$\bar{L}_h \bar{u}_h = \bar{q}_h \quad \text{with} \tag{4.68a}$$

$$\bar{L}_h = \begin{bmatrix} L_h & \mathbf{1} \\ \mathbf{1}^{\mathsf{T}} & 0 \end{bmatrix}, \qquad \bar{u}_h = \begin{bmatrix} u_h \\ \lambda \end{bmatrix}, \qquad \bar{q}_h = \begin{bmatrix} q_h \\ \sigma \end{bmatrix}, \tag{4.68b}$$

where σ can be prescribed arbitrarily.

Theorem 4.60. *Equation (4.68a) is always solvable. If for the last component of the solution $\bar{u}_h$ we have $\lambda = 0$, then condition (4.63) is satisfied and u_h represents the solution of system (4.62a) which is normalised by $\mathbf{1}^{\mathsf{T}} u_h = \sum_{\mathbf{x}\in\Omega_h} u_h(\mathbf{x}) = \sigma$. However, if $\lambda \neq 0$ holds one can interpret u_h as solution of $L_h u_h = \tilde{q}_h$, where $\tilde{q}_h = q_h - \lambda \mathbf{1}$ belongs to $\tilde{f}(\mathbf{x}) := f(\mathbf{x}) - \lambda$ and $\tilde{f}$ and φ satisfy condition (4.63).*

Proof. The vector $\mathbf{1}$ is linearly independent of the columns of L_h, so that $\operatorname{rank}[L_h, \mathbf{1}] = \operatorname{rank}(L_h) + 1 = (n-1)^2$. Likewise, $(\mathbf{1}^\mathsf{T}, 0)$ is linearly independent of the rows of $[L_h, \mathbf{1}]$ so that $\operatorname{rank}(\bar{L}_h) = (n-1)^2 + 1$, i.e., $\bar{L}_h$ is nonsingular. The other statements can be read from (4.68b). ∎

Recommendation. One should either use equation (4.68a,b) or equation (4.67), after first replacing q_h by $\tilde{q}_h := q_h - (\mathbf{1}^\mathsf{T} q_h \,/\, \mathbf{1}^\mathsf{T}\mathbf{1})\,\mathbf{1}$.

As a justification of this recommendation note that the condition for solvability of the continuous problem is $\int_\Omega f \mathrm{d}\mathbf{x} + \int_\Gamma \varphi \mathrm{d}\Gamma = 0$ and that this does not at all imply the discrete solvability condition (4.63). For smooth functions f and φ equation (4.63) can be shown to hold up to a remainder of order $\mathcal{O}(h)$. Thus it is generally unavoidable to replace f_h and q_h by $f_h - \lambda\mathbf{1}$ and $q_h - \lambda\mathbf{1}$ ($\lambda = \mathbf{1}^\mathsf{T} q_h / \mathbf{1}^\mathsf{T}\mathbf{1}$). In the case of equation (4.68a,b) this correction is carried out implicitly. If, however, (4.67) is used without any correction, the resulting solution can be interpreted as a solution of $L_h u_h = \tilde{q}_h$ with $\tilde{q}_h(\mathbf{x}) = q_h(\mathbf{x})$ for $\mathbf{x} \neq \mathbf{x}_0$ and $\tilde{q}_h(\mathbf{x}_0) := -\sum_{\mathbf{x}\neq\mathbf{x}_0} q_h(\mathbf{x})$. Here too, an implicit correction of q_h is carried out, with the difference that the correction is not distributed over all components as before, but is concentrated on $q_h(\mathbf{x}_0)$. If equation (4.63) is satisfied up to order $\mathcal{O}(h)$, then $q_h(\mathbf{x}_0)$ and $\tilde{q}_h(\mathbf{x}_0)$ differ by $\mathcal{O}(h^{-1})$. Therefore the solution $\hat{u}_h$ of equation (4.67) contains a singularity at the point $\mathbf{x}_0$.

Theorem 4.61 (convergence). *Let $u \in C^{3,1}(\overline{\Omega})$ be the solution of the Neumann problem (4.59). Let $\bar{u}_h = \begin{bmatrix} u_h \\ \lambda \end{bmatrix}$ be the solution of equation (4.68a). Then there exists a $c \in \mathbb{R}$ and constants C, C' independent of u and h such that*

$$\begin{aligned} |\lambda| &\le C' h \,\|u\|_{C^{0,1}(\overline{\Omega})}\,, \\ \|u_h - R_h u - c\mathbf{1}\|_\infty &\le C\left[h\,\|u\|_{C^{1,1}(\overline{\Omega})} + h^2\,\|u\|_{C^{3,1}(\overline{\Omega})}\right]. \end{aligned} \tag{4.69}$$

Proof. (a) We have $\lambda = \mathbf{1}^\mathsf{T} q_h / \mathbf{1}^\mathsf{T}\mathbf{1} = [h^2 \sum_{\Omega_h} f(\mathbf{x}) + h \sum_{\Gamma'_h} \varphi(\mathbf{x})] h^{-2}(n-1)^2$, where the bracket $[\ldots] = \mathcal{O}(h\,\|u\|_{C^{0,1}(\overline{\Omega})})$ is the quadrature error $\int_\Omega f \mathrm{d}\mathbf{x} + \int_\Gamma \varphi \mathrm{d}\Gamma = 0$.

(b) According to Theorem 4.60, u_h is the solution of $L_h u_h = \tilde{q}_h := q_h - \lambda\mathbf{1}$. This corresponds to the difference equations

$$-\Delta_h u_h = \tilde{f}_h := f_h - \lambda\mathbf{1} \quad \text{in } \Omega_h, \qquad \partial_n^- u_h = \varphi \quad \text{on } \Gamma'_h.$$

The difference $w_h := u_h - R_h u$ satisfies

$$\begin{aligned} -\Delta_h w_h &= -\Delta_h u_h + \Delta_h R_h u = \Delta_h R_h u + f_h - \lambda\mathbf{1} && \text{(4.70a)} \\ &= \Delta_h R_h u - \tilde{R}_h \Delta u - \lambda\mathbf{1} =: c_h \qquad \text{in } \Omega_h, \\ \partial_n^- w_h &= \partial_n^- u_h - \partial_n^- R_h u = \varphi - \partial_n^- R_h u = \frac{\partial u}{\partial n} - \partial_n^- R_h u =: \psi \quad \text{on } \Gamma'_h. && \text{(4.70b)} \end{aligned}$$

The errors of consistency are

$$\|c_h\|_\infty \le \frac{1}{6}h^2 \|u\|_{C^{3,1}(\overline{\Omega})} \qquad \text{(cf. Remark 4.47)},$$
$$|\psi(\mathbf{x})| \le \frac{1}{2}h \|u\|_{C^{1,1}(\overline{\Omega})} + |\lambda| \qquad \text{(cf. Lemma 4.1)}.$$

Since the solution w_h exists, condition (4.63) is satisfied with c_h and ψ:

$$h^2 \sum_{\Omega_h} c_h(\mathbf{x}) + h \sum_{\Gamma_h'} \psi(\mathbf{x}) = 0.$$

Then $\tilde{w}_h := w_h - c\mathbf{1}$ with $c = \mathbf{1}^\mathsf{T} w_h / \mathbf{1}^\mathsf{T}\mathbf{1}$ is the solution of (4.70a,b) normalised by $\mathbf{1}^\mathsf{T} w_h = 0$. The application of the following Theorem 4.62 to equation (4.70a,b) provides the inequality (4.69) (cf. Remark 4.47). ∎

Theorem 4.62 (stability). *Let condition (4.63) be satisfied. Let the solution u_h of (4.60a), (4.61) be normalised by $\mathbf{1}^\mathsf{T} u_h = 0$. Then there exist constants C_1, C_2 independent of u and h such that*

$$\|u_h\|_\infty \le C_1 \max_{\mathbf{x}\in\Omega_h} |f(\mathbf{x})| + C_2 \max_{\mathbf{x}\in\Gamma_h'} |\varphi(\mathbf{x})| \,. \tag{4.71}$$

The proof of this theorem, which corresponds to Theorem 4.44, will be supplied in §4.7.4.

4.7.2 Symmetric Difference for $\partial u/\partial n$

As can be seen from Theorem 4.61, the one-sided difference ∂_n^- causes the error term $\mathcal{O}(h)$. It seems obvious to replace ∂_n^- by a symmetric difference. To this end the five-point discretisation is set up at all points $\mathbf{x} \in \overline{\Omega}_h = \Omega_h \cup \Gamma_h$ (cf. (4.8c)):

$$-\Delta_h u_h = f_h := \tilde{R}_h f \qquad \text{in } \overline{\Omega}_h, \tag{4.72a}$$

where $\tilde{R}_h$ is the restriction to $\overline{\Omega}_h$. The symmetric difference ∂_n^0 is defined by

$$\left(\partial_n^0 u_h\right)(\mathbf{x}) := \frac{1}{2h}\left[u_h(\mathbf{x}+h\mathbf{n}) - u_h(\mathbf{x}-h\mathbf{n})\right] = \varphi(\mathbf{x}) \quad \text{for } \mathbf{x} \in \Gamma_h. \tag{4.72b}$$

Here, we assign two normal directions to the corner points, so that for each of the corner points two equations of the form (4.72b) can be set up. In the corner $\mathbf{x} = (0,0)$ one has, for example, the normals $\mathbf{n} = \left[\begin{smallmatrix}0\\-1\end{smallmatrix}\right]$, $\mathbf{n} = \left[\begin{smallmatrix}-1\\0\end{smallmatrix}\right]$. The corresponding equations (4.72b) also contain different (!) values

$$\varphi(0+,0) = \lim_{x\searrow 0} \varphi(x,0) \qquad \text{and} \quad \varphi(0,0+) = \lim_{y\searrow 0} \varphi(0,y).$$

For $\mathbf{x} \in \Gamma_h$, the difference formula (4.72a) needs the values in points $\mathbf{x} + h\mathbf{n}$ outside of $\overline{\Omega}_h$. These can be eliminated with the aid of (4.72b) so that a system of equations $L_h u_h = q_h$ remains for the $(n+1)^2$ components $u_h(\mathbf{x})$, $\mathbf{x} \in \overline{\Omega}_h$.

Exercise 4.63. (a) If the grid points of $\overline{\Omega}_h$ are arranged in lexicographical order, L_h has the form

$$L_h = h^{-2} \begin{bmatrix} T & -2I & & & \\ -I & T & -I & & \\ & \ddots & \ddots & \ddots & \\ & & -I & T & -I \\ & & & -2I & T \end{bmatrix}, \qquad T = \begin{bmatrix} 4 & -2 & & & \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ & & & -2 & 4 \end{bmatrix}.$$

(b) L_h is not symmetric, but $D_h L_h$ with $D_h = \operatorname{diag}\{d(x)d(y)\}$, $d(0) = d(1) = \frac{1}{2}$ and $d(\cdot) \equiv 1$ otherwise, is symmetric.

The analogue of Theorem 4.58 reads as follows.

Theorem 4.64. *Equation (4.72a,b) is solvable if and only if $\mathbf{1}^{\mathsf{T}} D_h q_h = 0$ for D_h in Exercise 4.63b. Any two solutions may differ by only a constant. The formulation of $\mathbf{1}^{\mathsf{T}} D_h q_h = 0$ with the aid of f and φ reads:*

$$-h^2 \sum_{(x,y)\in\overline{\Omega}_h} d(x)d(y)f(x,y) = 2h \sum_{(x,y)\in\Gamma_h} d(x)d(y)\varphi(x,y), \tag{4.73}$$

where the term for the corner points occurs twice in the second sum, and both the different limits for φ are taken into account.

Remark 4.65. The sums in (4.73) represent summed trapezoidal formulae[2]. Thus, from $\int_\Omega f \mathrm{d}\mathbf{x} + \int_\Gamma \varphi \, \mathrm{d}\Gamma = 0$ follows equation (4.73), except for a remainder $\mathcal{O}(h^2 \|u\|_{C^{1,1}(\overline{\Omega})})$.

Theorems 4.59 and 4.60 can be transferred without difficulty. Theorem 4.61 becomes the following convergence theorem.

Theorem 4.66. *Let $u \in C^{3,1}(\overline{\Omega})$ be a solution of (4.59). Let $\bar{u}_h = \begin{bmatrix} u_h \\ \lambda \end{bmatrix}$ be the solution of $L_h \bar{u}_h = \begin{bmatrix} D_h q_h \\ 0 \end{bmatrix}$ with $\bar{L}_h = \begin{bmatrix} D_h L_h & \mathbf{1} \\ \mathbf{1}^{\mathsf{T}} & 0 \end{bmatrix}$. We have convergence of second order:*

$$|\lambda| \le C' h^2 \|u\|_{C^{1,1}(\overline{\Omega})}, \qquad \|u_h - R_h u - c\mathbf{1}\|_\infty \le C h^2 \|u\|_{C^{3,1}(\overline{\Omega})}. \tag{4.74}$$

Proof. The proof is essentially the same as for Theorem 4.61. An additional technical difficulty is the fact that the consistency error $\Delta_h R_h u - R_h \Delta u$ must also be determined in $\mathbf{x} \in \Gamma_h$ although u is only defined in $\overline{\Omega}$. Instead of treating the difference equation and the boundary discretisation separately as in (4.70a,b) one should directly analyse the equations $L_h u_h = q_h$ from which the values $u_h(\mathbf{x}+h\mathbf{n})$ outside of $\overline{\Omega}$ have already been eliminated. ∎

[2] See Stoer [274, §3.1].

4.7.3 Symmetric Difference for $\partial u/\partial n$ on an Offset Grid

If we offset the above grid by $h/2$ in the x and y directions, we obtain the grid

$$\Omega_h := \left\{(x,y) \in \Omega : \frac{x}{h} - \frac{1}{2} \in \mathbb{Z} \text{ and } \frac{y}{h} - \frac{1}{2} \in \mathbb{Z}\right\}$$

in Figure 4.6. The near-boundary points of Ω_h are at a distance $h/2$ from Γ. We set

$$\Gamma_h := \left\{(x,y) \in \Gamma : \frac{x}{h} - \frac{1}{2} \in \mathbb{Z} \text{ or } \frac{y}{h} - \frac{1}{2} \in \mathbb{Z}\right\}$$

(cf. Figure 4.6). To each near-boundary point $\mathbf{x} - h\mathbf{n}/2$ ($\mathbf{x} \in \Gamma_h$) corresponds an outlying neighbour $\mathbf{x} + h\mathbf{n}/2$. The discretisation of the Neumann problem (4.59) is

$$\begin{aligned} -\Delta_h u_h(\mathbf{x}) &= f(\mathbf{x}) && \text{in } \mathbf{x} \in \Omega_h, \\ h^{-1}\left[u_h(\mathbf{x} + h\mathbf{n}/2) - u_h(\mathbf{x} - h\mathbf{n}/2)\right] &= \varphi(\mathbf{x}) && \text{for } \mathbf{x} \in \Gamma_h\,. \end{aligned} \tag{4.75}$$

The difference (4.75) is symmetric with respect to the boundary point $\mathbf{x}$ and nevertheless agrees with the backward difference ∂_n^- at the grid point $\mathbf{x} + \frac{h}{2}\mathbf{n}$.

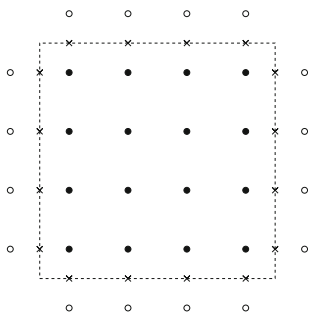


Fig. 4.6 Offset grid, •: grid point of Ω_h, o: outlying grid point, ×: boundary point in Γ_h.

Remark 4.67. After eliminating the values

$$u_h(\mathbf{x} + h\mathbf{n}/2) \quad \text{for } \mathbf{x} \in \Gamma_h$$

we obtain a system of equations $L_h u_h = q_h$, where Remark 4.57 is also valid for this matrix L_h. In contrast to §4.7.1, L_h is of size $n^2 \times n^2$.

The Theorems 4.58, 4.59, 4.60, and 4.62 hold analogously. Theorem 4.61 holds with the inequality (4.74) instead of (4.69).

4.7.4 Proof of the Stability Theorem 4.62

In the case of Dirichlet boundary values the stability statement of Theorem 4.44 follows immediately from the maximum principle and the bound for L_h^{-1}. The corresponding statement of Theorem 4.62 for Neumann boundary values, however, cannot be proved that easily. In the literature one can only find weaker estimates which on the right-hand side of equation (4.71) contain an additional factor $|\log h|$. However, this factor is not to be avoided if one uses equation (4.67), $\hat{L}_h \hat{u}_h = \hat{q}_h$, without condition (4.63) being satisfied.

Let the *discrete Green function (of the second kind)* $g_h(\mathbf{x},\boldsymbol{\xi})$, $\mathbf{x} \in \overline{\Omega}'_h$, $\boldsymbol{\xi} \in \Omega_h$, be defined by

$$-\Delta_h g_h(\mathbf{x},\boldsymbol{\xi}) = \delta(\mathbf{x},\boldsymbol{\xi}) := \left\{ \begin{matrix} h^{-2} & \text{if } \mathbf{x} = \boldsymbol{\xi} \\ 0 & \text{if } \mathbf{x} \neq \boldsymbol{\xi} \end{matrix} \right\} - \left\{ \begin{matrix} \frac{1}{h(4-4h)} & \text{if } \mathbf{x} \in \Gamma'_h \\ 0 & \text{if } \mathbf{x} \in \Omega_h \end{matrix} \right\}, \quad (4.76a)$$

$$\left(\partial_n^- g_h\right)(\mathbf{x},\boldsymbol{\xi}) = 0 \qquad \text{for } \mathbf{x} \in \Gamma'_h \quad (4.76b)$$

where Δ_h and ∂_n^- act on $\mathbf{x}$. g_h exists since $\sum_{\mathbf{x}\in\Omega_h} \delta(\mathbf{x},\boldsymbol{\xi}) = 0$ proves the solvability condition (4.65). The sets Ω_h, $\overline{\Omega}'_h$, and Γ'_h are defined in (4.60b).

Lemma 4.68. *For arbitrary q_h with $\mathbf{1}^\mathsf{T} q_h = 0$ (cf. (4.65)),*

$$u_h(\mathbf{x}) := h^2 \sum_{\boldsymbol{\xi}\in\Omega_h} q_h(\boldsymbol{\xi}) g_h(\mathbf{x},\boldsymbol{\xi}) \quad (4.77)$$

represents a solution of $L_h u_h = q_h$.

Proof. At far-boundary points $\mathbf{x} \in \Omega_h$, $(L_h u_h)(\mathbf{x}) = -\Delta_h u_h(\mathbf{x}) = q_h(\mathbf{x})$. In near-boundary points $\mathbf{x} \in \Omega_h$, from $\partial_n^- u_h = 0$ and (4.65), one has the identity

$$(L_h u_h)(\mathbf{x}) = (-\Delta_h u_h)(\mathbf{x}) = q_h(\mathbf{x}) - \frac{1/h}{4-4h} \sum_{\boldsymbol{\xi}\in\Omega_h} q_h(\boldsymbol{\xi}) = q_h(\mathbf{x}). \qquad \blacksquare$$

Theorem 4.69. *The equations (4.76a,b) determine g_h up to a constant. The Green function $g_h(\mathbf{x},\boldsymbol{\xi})$ can be so selected that*

$$|g_h(\mathbf{x},\boldsymbol{\xi})| \le C\,[\,1 + |\log(|\mathbf{x}-\boldsymbol{\xi}| + h)|\,] \qquad \textit{for } \mathbf{x},\boldsymbol{\xi} \in \Omega_h. \quad (4.78)$$

This inequality corresponds to the bound (4.42) in the Dirichlet case. Before Theorem 4.69 is proved by Lemma 4.70, we want to show that Theorem 4.62 follows from it.

Proof of Theorem 4.62. Analogously to $\int_\Omega [\,1 + |\log|\mathbf{x}-\boldsymbol{\xi}||\,]\,\mathrm{d}\mathbf{x} \le K_1$ and $\int_\Gamma [\,1 + |\log|\mathbf{x}-\boldsymbol{\xi}||\,]\,\mathrm{d}\Gamma_x \le K_2$ we obtain

$$h^2 \sum_{\boldsymbol{\xi}\in\Omega_h} [1 + |\log(|\mathbf{x}-\boldsymbol{\xi}| + h)|] \le K'_1\,, \quad h \sum_{\mathbf{x}\in\Gamma'_h} [1 + |\log(|\mathbf{x}-\boldsymbol{\xi}| + h)|] \le K'_2\,.$$

From (4.77), (4.78), and $q_h = f_h + \varphi_h$ (cf. (4.62c)) thus follows the estimate

$$|u_h(\mathbf{x})| \le K_1 \|f_h\|_\infty + hK_2 \|\varphi_h\|_\infty \le K_1 \max_{\mathbf{x}\in\Omega_h} |f(\mathbf{x})| + 2K_2 \max_{\mathbf{x}\in\Gamma'_h} |\varphi(\mathbf{x})|.$$

If $\tilde{u}_h := u_h - \mathbf{1}^\mathsf{T} u_h / \mathbf{1}^\mathsf{T}\mathbf{1}$ is the solution of $L_h u_h = q_h$ normalised by $\mathbf{1}^\mathsf{T}\tilde{u}_h = 0$, one can see that

$$\|\tilde{u}_h\|_\infty \le 2 \inf\{\|u_h - c\mathbf{1}\|_\infty : c \in \mathbb{R}\} \le 2\,\|\tilde{u}_h\|_\infty\,,$$

so that the inequality (4.69) and Theorem 4.62 are proved. $\blacksquare$

It remains to prove Theorem 4.69. The construction of the function G_h postulated in the next lemma will follow after its proof.

Lemma 4.70. *Let* $\mathbf{e}$ *be one of the unit vectors* $\left[\begin{smallmatrix}1\\0\end{smallmatrix}\right]$ *or* $\left[\begin{smallmatrix}0\\1\end{smallmatrix}\right]$. *Let* $\boldsymbol{\xi} \in \Omega_h$ *be such that* $\boldsymbol{\xi} + h\mathbf{e} \in \Omega_h$ *also. For all these* $\mathbf{e}$ *and* $\boldsymbol{\xi}$ *let there exist a function* $G_h(\mathbf{x}) = G_h(\mathbf{x}; \boldsymbol{\xi}, \mathbf{e})$ *with the properties*

$$-\Delta_h G_h(\mathbf{x}) = h^{-2} \left\{ \begin{array}{c} 1 \ \textit{if } \mathbf{x} = \boldsymbol{\xi} \\ -1 \ \textit{if } \mathbf{x} = \boldsymbol{\xi} + h\mathbf{e} \\ 0 \ \textit{otherwise} \end{array} \right\} \qquad \textit{in } \Omega_h, \tag{4.79a}$$

$$\partial_n^- G_h = 0 \qquad \textit{on } \Gamma_h', \tag{4.79b}$$

$$|G_h(\mathbf{x})| \le h\, C' / \left(|\mathbf{x} - \boldsymbol{\xi}| + h \right), \tag{4.79c}$$

where C' *does not depend on* $\mathbf{e}$ *and* $\boldsymbol{\xi}$. *Then Theorem 4.69 holds, and thus also Theorem 4.62.*

Proof. For $\boldsymbol{\xi}, \boldsymbol{\xi}' \in \Omega_h$ let $g_h(\mathbf{x}, \boldsymbol{\xi}, \boldsymbol{\xi}')$ be defined as a solution of

$$\partial_n^- g_h(\cdot, \boldsymbol{\xi}, \boldsymbol{\xi}') = 0 \quad \text{on } \Gamma_h', \quad -\Delta_h g_h(\cdot, \boldsymbol{\xi}, \boldsymbol{\xi}') = h^{-2} \left\{ \begin{array}{c} 1 \ \text{if } \mathbf{x} = \boldsymbol{\xi} \\ -1 \ \text{if } \mathbf{x} = \boldsymbol{\xi}' \\ 0 \ \text{otherwise} \end{array} \right\} \quad \text{in } \Omega_h.$$

For $\boldsymbol{\xi}' = \boldsymbol{\xi} + h\mathbf{e}$, $g_h(\cdot, \boldsymbol{\xi}, \boldsymbol{\xi}')$ agrees with G_h in Lemma 4.70. For arbitrary $\boldsymbol{\xi}, \boldsymbol{\xi}'$, one finds a connection $\boldsymbol{\xi} = \boldsymbol{\xi}^0, \boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^\ell = \boldsymbol{\xi}'$ with $\boldsymbol{\xi}^{k+1} - \boldsymbol{\xi}^k = \pm h\mathbf{e}^k$, $\mathbf{e}^k = (1,0)$ or $(0,1)$. Since $g_h(\cdot, \boldsymbol{\xi}, \boldsymbol{\xi}') = \sum_{k=1}^{\ell} g_h(\cdot, \boldsymbol{\xi}^{k-1}, \boldsymbol{\xi}^k)$, (4.79c) implies the estimate

$$\left| g_h(\mathbf{x}, \boldsymbol{\xi}, \boldsymbol{\xi}') \right| \le h \sum_{k=1}^{\ell} C' / \left(|\mathbf{x} - \boldsymbol{\xi}^{k-1}| + h \right).$$

Considering first the case $x_1 = \xi_1 = \xi_1'$, $x_2 \le \xi_2 < \xi_2'$ and applying

$$\sum_{k=k_1}^{k_2} \frac{1}{k} \le \text{const} \cdot \big(1 + \log(k_2 h) - \log(k_1 h) \big),$$

one obtains

$$\left| g_h(\mathbf{x}, \boldsymbol{\xi}, \boldsymbol{\xi}') \right| \le C'' \left[1 + \left| \log(|\mathbf{x} - \boldsymbol{\xi}| + h) \right| + \left| \log\left(\left| \mathbf{x} - \boldsymbol{\xi}' \right| + h \right) \right| \right].$$

In the general case, this bound also comes out, but one must choose the connection $\boldsymbol{\xi}^k$ such that $|\mathbf{x} - \boldsymbol{\xi}^k| \ge \min\{ |\mathbf{x} - \boldsymbol{\xi}|, \left| \mathbf{x} - \boldsymbol{\xi}' \right| \}$. Since we know that

$$h \sum_{\boldsymbol{\xi}' \in \Gamma_h'} \left| \log\left(\left| \mathbf{x} - \boldsymbol{\xi}' \right| + h \right) \right| \le \text{const}$$

for all $\mathbf{x} \in \Omega_h$, the new function

$$g_h(\mathbf{x}, \boldsymbol{\xi}) := \frac{h}{4 - 4h} \sum_{\boldsymbol{\xi}' \in \Gamma_h'} g_h(\mathbf{x}, \boldsymbol{\xi}, \boldsymbol{\xi}')$$

satisfies the inequality (4.78). Because $\sum_{\Gamma_h'} 1 = (4-4h)/h$ the equations (4.76a,b) are also satisfied, i.e., $g_h(\mathbf{x},\boldsymbol{\xi})$ is the Green function sought in Theorem 4.69. ∎

We shall construct the function G_h, needed in Lemma 4.70 explicitly. The building block is the *discrete singularity function* $s_h(\mathbf{x},\boldsymbol{\xi})$ on the infinite grid Q_h in $\mathbb{R}^2$:

$$Q_h := \left\{(x,y) \in \mathbb{R}^2 : \; x/h,\, y/h \in \mathbb{Z}\right\}.$$

Lemma 4.71. *The singularity function defined by* $s_h(\mathbf{x},\boldsymbol{\xi}) := \sigma_h(\mathbf{x}-\boldsymbol{\xi})$ *and*

$$\begin{aligned}\sigma_h(\mathbf{x}) &:= \frac{1}{16\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \frac{\mathrm{e}^{\mathrm{i}(x_1\eta_1+x_2\eta_2)/h}-1}{\sin^2(\eta_1/2)+\sin^2(\eta_2/2)}\mathrm{d}\eta_1\mathrm{d}\eta_2 \\ &= \frac{1}{8\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} \frac{\sin^2\left((x_1\eta_1+x_2\eta_2)\,/\,(2h)\right)}{\sin^2(\eta_1/2)+\sin^2(\eta_2/2)}\mathrm{d}\eta_1\mathrm{d}\eta_2\end{aligned}$$

for all grid points $\mathbf{x},\boldsymbol{\xi} \in Q_h$ *has the property* $-\Delta_h s_h(\mathbf{x},\boldsymbol{\xi}) = h^{-2}$ *for* $\mathbf{x}=\boldsymbol{\xi}$ *and* $-\Delta_h s_h = 0$ *otherwise.*

Proof. Let $e(\mathbf{x},\boldsymbol{\eta}) := \exp(\mathrm{i}(x_1\eta_1+x_2\eta_2)/h)$. Note that

$$-\Delta_h\left[\,e(\mathbf{x},\boldsymbol{\eta})-1\,\right] = 4h^{-2}e(\mathbf{x},\boldsymbol{\eta})\left[\sin^2(\eta_1/2)+\sin^2(\eta_2/2)\right]$$

and $\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} e(\mathbf{x},\boldsymbol{\eta})\mathrm{d}\boldsymbol{\eta} = \left\{\begin{array}{c} 4\pi^2 \text{ for } \mathbf{x}=0 \\ 0 \quad \text{for } \mathbf{x}\neq 0\end{array}\right\}$. ∎

For multi-indices $\alpha = (\alpha_1,\alpha_2) \in \mathbb{N}^2$ with $\alpha_1,\alpha_2 \geq 0$ one defines the partial difference operators of order $|\alpha| = \alpha_1+\alpha_2$ by

$$\partial^\alpha = \left(\partial_x^+\right)^{\alpha_1}\left(\partial_y^+\right)^{\alpha_2}.$$

Starting with the representation in Lemma 4.71, Thomée [285, Theorem 3.1] proves the next inequality.

Lemma 4.72. $|(\partial^\alpha\sigma_h)(\mathbf{x})| \leq C\left(|\mathbf{x}|+h\right)^{-|\alpha|}$ *for all* $\mathbf{x} \in Q_h$, $|\alpha| \geq 1$.

For the construction of the function G_h in Lemma 4.70 we select, without loss of generality, $\mathbf{e} = (1,0)$ and keep $\boldsymbol{\xi} \in \Omega_h$ with $\boldsymbol{\xi}+h\mathbf{e} \in \Omega_h$. The function

$$G_h'(\mathbf{x}) := h\left(\partial_x^-\sigma_h\right)(\mathbf{x}-\boldsymbol{\xi}) = \sigma_h(\mathbf{x}-\boldsymbol{\xi}) - \sigma_h(\mathbf{x}-\boldsymbol{\xi}-h\mathbf{e})$$

satisfies the difference equation (4.79a) but not the boundary condition (4.79b). Let the symmetrisation operators S_x and S_y be defined by

$$S_x u_h(x,y) := \frac{u_h(x,y)+u_h(h-x,y)}{2}, \;\; S_y u_h(x,y) := \frac{u_h(x,y)+u_h(x,h-y)}{2}$$

for $(x,y) \in Q_h$. The function

$$G_h'' := 4S_xS_yG_h'$$

is symmetric with respect to the axes $y = \frac{h}{2}$ and $x = \frac{h}{2}$. Thus we have (4.79b) on the left and lower boundary:

$$\partial_n^- G_h''(0, y) = \partial_n^- G_h''(x, 0) = 0. \tag{4.80}$$

Furthermore, G_h'' satisfies condition (4.79a) just as G_h' does. For each $\boldsymbol{\beta} \in \mathbb{N}^2$ we define the operator P_β by

$$(P_\beta u_h)(x, y) = \frac{1}{4} \begin{bmatrix} u_h(x + \beta_1 L, y + \beta_2 L) + u_h(x - \beta_1 L, y + \beta_2 L) \\ + u_h(x + \beta_1 L, y - \beta_2 L) + u_h(x - \beta_1 L, y - \beta_2 L) \end{bmatrix},$$

where $L = 2 - 2h$, and set

$$G_h^\beta(\mathbf{x}) := (P_\beta G_h'')(\mathbf{x}) - (P_\beta G_h'')(\mathbf{0}).$$

Lemma 4.73. *For $\boldsymbol{\beta} = (\beta_1, \beta_2)$ with $\|\boldsymbol{\beta}\|_\infty \geq 2$ there holds*

$$\left|\partial^\alpha G_h^\beta(\mathbf{x})\right| \leq hK / |\boldsymbol{\beta}|^3 \qquad \textit{for all } |\alpha| \leq 2,\ \mathbf{x} \in \Omega_h. \tag{4.81}$$

Here, K is independent of the choice of the points $\boldsymbol{\xi}, \boldsymbol{\xi} + h\mathbf{e} \in \Omega_h$.

Proof. According to the definition we have

$$G_h^\beta(0, 0) = 0. \tag{4.82a}$$

The operator P_β preserves the symmetry, i.e., $u_h = S_x u_h$ implies $P_\beta u_h = S_x P_\beta u_h$. Thus we have that $G_h^\beta = S_x G_h^\beta = S_y G_h^\beta$ and consequently

$$\partial_x^+ G_h^\beta(0, 0) = \partial_y^+ G_h^\beta(0, 0) = 0. \tag{4.82b}$$

$G_h^\beta(\mathbf{x})$ is the linear combination of $h\partial_x^- \sigma_h(\tilde{\mathbf{x}})$ for different grid points $\tilde{\mathbf{x}}$ with $|\tilde{\mathbf{x}}| + h \geq K' |\beta|$ if $\mathbf{x} \in \Omega_h$. According to Lemma 4.72,

$$\left|\partial^\alpha G_h^\beta(\mathbf{x})\right| \leq hCK'' / |\boldsymbol{\beta}|^3 \qquad \text{for all } |\alpha| = 2,\ \mathbf{x} \in \Omega_h, \tag{4.82c}$$

so that (4.81) follows for $|\alpha| = 2$. Let $|\alpha| = 1$. $\partial^\alpha G_h^\beta(\mathbf{x})$ can be written as

$$\partial^\alpha G_h^\beta(\mathbf{0}) + \sum_{k=1}^{\ell} \left[\partial^\alpha G_h^\beta(\mathbf{x}^k) - \partial^\alpha G_h^\beta(\mathbf{x}^{k-1})\right] \qquad \text{with } \begin{cases} \mathbf{x}^0 = \mathbf{0},\ \mathbf{x}^\ell = \mathbf{x}, \\ \|\mathbf{x}^k - \mathbf{x}^{k-1}\|_\infty = h. \end{cases}$$

Each term has the form $\pm h\partial^\gamma G_h^\beta$ with $|\gamma| = 2$ so that (4.82b,c) lead to the estimate

$$\left|\partial^\alpha G_h^\beta(\mathbf{x})\right| \leq 2hCK'' / |\boldsymbol{\beta}|^3 \qquad \text{for } |\alpha| = 1,\ \mathbf{x} \in \Omega_h. \tag{4.82d}$$

Likewise one infers from (4.82a) and (4.82d) the inequality (4.81) for $|\alpha| = 0$, which proves Lemma 4.73. ∎

Since $\sum_{\beta\in\mathbb{Z}^2\setminus(0,0)} |\boldsymbol{\beta}|^{-3} < \infty$, the infinite sum

$$G_h(\mathbf{x}) := \sum\nolimits_{\beta\in\mathbb{Z}^2} G_h^\beta(\mathbf{x}) \tag{4.83}$$

exists. Since $-\Delta_h G_h^\beta(\mathbf{x}) = 0$ in Ω_h for all $\boldsymbol{\beta} \neq (0,0)$, G_h also satisfies equation (4.79a). As already mentioned in the proof of (4.82b), $G_h = S_x G_h = S_y G_h$ holds so that G_h also satisfies equation (4.80): $\partial_n^- G_h = 0$ at the left and lower boundary. The proof of $\partial_n^- G_h = 0$ at the other boundaries requires the following statement.

Lemma 4.74. *G_h is L-periodic: $G_h(x,y) = G_h(x+L,y) = G_h(x,y+L)$.*

Proof. Let $G_h''(x+\beta_1 L, y+\beta_2 L) - G_h''(\beta_1 L, \beta_2 L)$ be abbreviated to $\gamma_\beta(x,y)$. Definition (4.83) says that[3]

$$G_h = \lim_{k\to\infty} \sum_{|\beta_1|\le k} \sum_{|\beta_2|\le k} \gamma_\beta. \tag{4.84a}$$

Now γ_β can be written as a sum over the differences

$$\begin{aligned} & h\partial_h^+ G_h''(x+\beta_1 L, \beta_2 L + \nu h) && \text{from } \nu h = 0 \text{ to } \nu h = y \\ \text{and} \quad & h\partial_h^+ G_h''(\beta_1 L + \nu h, \beta_2 L) && \text{from } \nu h = 0 \text{ to } \nu h = x. \end{aligned}$$

For $|\beta_1| \ge 2$ the distance between the arguments to $\boldsymbol{\xi}$ and $\boldsymbol{\xi} + h\mathbf{e}$ is always $\ge (|\beta_1|-1)L$. Each term is thus, by Lemma 4.72, bounded by $\mathcal{O}(\frac{h^2}{(|\beta_1|-1)^2}) = \mathcal{O}(\frac{h^2}{\beta_1^2})$. As a sum of such terms then γ_β can be estimated by

$$|\gamma_\beta| \le \mathcal{O}(h/\beta_1^2) \qquad \text{for } |\beta_1| \ge 2. \tag{4.84b}$$

We want now to show that the following equation (4.84c) also holds:

$$G_h = \lim_{k\to\infty} \sum_{|\beta_1-1|\le k} \sum_{|\beta_2-1|\le k} \gamma_\beta = \lim_{k\to\infty} \sum_{1-k\le\beta_1\le 1+k} \sum_{|\beta_2-1|\le k} \gamma_\beta. \tag{4.84c}$$

The sums (4.84a) and (4.84c) differ by

$$D_h := \sum\nolimits_{|\beta_2|\le k} \big[\, \gamma_{(1+k,\beta_2)} - \gamma_{(-k,\beta_2)} \,\big].$$

The values $\beta_1 = 1+k$ and $-k$ appearing in the bracket satisfy $|\beta_1| \ge 2$ and $|\beta_1| \le k+1$. The absolute value of the sum D_h is then bounded, from (4.84b), by $\sum_{|\beta_2|\le k} \mathcal{O}(h/k^2) = (2k+1)\mathcal{O}(h/k^2) = \mathcal{O}(h/k)$, so that the limits of the double sums in (4.84a) and (4.84c) are the same. Changing the variable β_1 to $\beta_1 - 1$ then transforms (4.84c) into

$$G_h(x,y) = \lim_{k\to\infty} \sum_{-k\le\beta_1,\beta_2\le k} \big[\, G_h''(x+\beta_1 L+L, y+\beta_2 L) - G_h''(\beta_1 L+L, \beta_2 L) \,\big]. \tag{4.84d}$$

[3] The partial sums $\sum_{|\beta_1|\le k} \sum_{|\beta_2|\le k} G_h^\beta(\mathbf{x})$ and $\sum_{|\beta_1|\le k} \sum_{|\beta_2|\le k} \gamma_\beta(\mathbf{x})$ are identical. Note that the sum (4.83) is an absolutely convergent. In contrast $\sum \gamma_\beta$ only converges conditionally. (4.84a) describes the convergence of $\sum \gamma_\beta$ for a special ordering of the terms. Because of that the identity in (4.84c) is nontrivial.

The first part of the identity

$$\begin{aligned}
&\sum_{\beta_1,\beta_2=-k}^{k} [G_h''(\beta_1 L, \beta_2 L) - G_h''(\beta_1 L + L, \beta_2 L)] \\
&= \sum_{\beta_2=-k}^{k} [G_h''(-kL, \beta_2 L) - G_h''(kL + L, \beta_2 L)] \\
&= \sum_{\beta_2=-k}^{k} [G_h''(h + kL, \beta_2 L) - G_h''(kL + L, \beta_2 L)] \qquad (4.84e)
\end{aligned}$$

is elementary; the second results from the symmetry $G_h'' = S_x G_h''$. As above, the terms of the last sum can be bounded by $\mathcal{O}(h/k^2)$ so that (4.84e) vanishes for $k \to \infty$. Together with (4.84d) one obtains

$$\begin{aligned}
G_h(x, y) &= \lim_{k\to\infty} \sum_{-k\le\beta_1,\beta_2\le k} [G_h''(x + \beta_1 L + L, y + \beta_2 L) - G_h''(\beta_1 L, \beta_2 L)] \\
&= \lim_{k\to\infty} \sum_{-k\le\beta_1,\beta_2\le k} \gamma_\beta(x + L, y) \;=\; G_h(x + L, y).
\end{aligned}$$

The proof of $G_h(x, y) = G_h(x, y + L)$ is analogous. ∎

Proof of Theorem 4.69. Since $L = 2 - 2h$, the symmetry $G_h = S_x G_h$ and the periodicity yield

$$G_h(1, y) = G_h(h - 1, y) = G_h(h - 1 + L, y) = G_h(1 - h, y),$$

i.e., $\partial_n^- G_h(1, y) = 0$ on the upper boundary $(1, y) \in \Gamma_h'$. Likewise we show that $\partial_n^- G_h(x, 1) = 0$ on the right boundary. Thus (4.79b) is also proved. It remains to show (4.79c):

$$|G_h(\mathbf{x})| \le hC/(|\mathbf{x} - \boldsymbol{\xi}| + h) \qquad \text{in } \Omega_h.$$

The G_h^β for $\|\beta\|_\infty \le 1$ are linear combinations of $h\partial_x^- \sigma_h(\mathbf{x} - \tilde{\boldsymbol{\xi}})$ for $\tilde{\boldsymbol{\xi}} = \boldsymbol{\xi}$ and other $\tilde{\boldsymbol{\xi}} \notin \Omega_h$, which are generated by S_x, S_y, and P_β. For all $\tilde{\boldsymbol{\xi}}$ there holds $|\mathbf{x} - \tilde{\boldsymbol{\xi}}| \ge |\mathbf{x} - \boldsymbol{\xi}|$, so that from Lemma 4.72 follows

$$\left|G_h^\beta(\mathbf{x})\right| \le hC'/(|\mathbf{x} - \boldsymbol{\xi}| + h) \qquad \text{for } \|\beta\|_\infty \le 1,\ \mathbf{x} \in \Omega_h.$$

On the other hand, Lemma 4.73 shows that

$$\left|\sum\nolimits_{\|\beta\|_\infty\ge 2} G_h^\beta(\mathbf{x})\right| \le h \sum\nolimits_{\|\beta\|_\infty\ge 2} \frac{K}{|\beta|^3} = hK' \le hK''/(|\mathbf{x} - \boldsymbol{\xi}| + h).$$

The assumptions of Lemma 4.70 are hence satisfied: G_h with (21a–c) exists. Thus Theorems 4.69 and 4.62 are proved. ∎

4.8 Discretisation in an Arbitrary Domain

4.8.1 Shortley–Weller Approximation

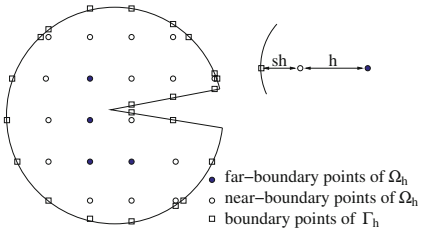


Fig. 4.7 Ω_h and Γ_h.

Let the boundary-value problem

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ u &= \varphi \quad \text{on } \Gamma \end{aligned} \tag{4.85}$$

be given for an arbitrary domain Ω. If one places a square grid of step size h over Ω one obtains

$$\Omega_h := \left\{(x,y) \in \Omega : \tfrac{x}{h} \in \mathbb{Z},\ \tfrac{y}{h} \in \mathbb{Z}\right\}$$

as the set of grid points. By contrast, the set Γ_h of boundary points must be defined differently from the case of the square. The left neighbour point of $(x,y) \in \Omega_h$ reads $(x-h,y)$. If the connecting segment $\{(x-\vartheta h, y) : \vartheta \in (0,1]\}$ does not lie completely in Ω, there exists a boundary point

$$(x-sh,y) \in \Gamma \quad \text{with} \quad \begin{cases} (x,y) \in \Omega_h, \quad s \in (0,1], \\ (x-\vartheta h, y) \in \Omega \text{ for all } \vartheta \in [0,s), \end{cases} \tag{4.86a}$$

which now, instead of $(x-h,y)$ is called the left neighbour point of (x,y) (cf. Figure 4.7). Likewise the right, lower, and upper neighbour points can be boundary points of the following form:

$$(x+sh,y) \in \Gamma \quad \text{with} \quad \begin{cases} (x,y) \in \Omega_h, \quad s \in (0,1], \\ (x+\vartheta h, y) \in \Omega \text{ for all } \vartheta \in [0,s), \end{cases} \tag{4.86b}$$

$$(x,y-sh) \in \Gamma \quad \text{with} \quad \begin{cases} (x,y) \in \Omega_h, \quad s \in (0,1], \\ (x, y-\vartheta h) \in \Omega \text{ for all } \vartheta \in [0,s), \end{cases} \tag{4.86c}$$

$$(x,y+sh) \in \Gamma \quad \text{with} \quad \begin{cases} (x,y) \in \Omega_h, \quad s \in (0,1], \\ (x, y+\vartheta h) \in \Omega \text{ for all } \vartheta \in [0,s). \end{cases} \tag{4.86d}$$

We set

$$\Gamma_h := \{\text{boundary points which satisfy (4.86a–d)}\}$$

A grid point $(x,y) \in \Omega_h$, possessing a neighbour from Γ_h, is said to be a *near-boundary point*. All other points of Ω_h are said to be far-boundary points. As can be seen in Figure 4.7, (x,y) may be a near-boundary point although $(x \pm h, y)$ and $(x, y \pm h)$ belong to Ω_h, (namely if Ω is not convex and not all connecting segments lie completely in Ω).

If one wishes to approximate the second derivative $u''(x)$ with the aid of the values of u at $x' < x < x''$, one can use Newton's divided differences:

$$u''(x) = 2\left[\frac{u(x'') - u(x)}{x'' - x} - \frac{u(x) - u(x')}{x - x'}\right] / (x'' - x') + \text{Rem}\,. \tag{4.87}$$

Exercise 4.75. Show that (a) the Taylor expansion implies

$$|\text{Rem}| \le \frac{(x''-x)^2+(x-x')^2}{3(x''-x')}\,\|u\|_{C^3([x',x''])} \le \frac{\max\{x''-x,x-x'\}}{3}\,\|u\|_{C^3([x',x''])} \tag{4.88}$$

for the remainder in equation (4.87) if $u \in C^3([x', x''])$.

(b) In (4.88) one can replace the norm of $C^3([x', x''])$ by that of $C^{2,1}([x', x''])$.

(c) If $x'' = x + h$ and $x' = x - h$ the difference in (4.87) agrees with the usual second difference $\partial^-\partial^+ u(x)$.

To set up the difference equation for $-\Delta u = f$ at $(x, y) \in \Omega_h$ we use the four neighbouring points

$$(x - s_\ell h, y)\,, \quad (x + s_r h, y)\,, \quad (x, y - s_u h)\,, \quad (x, y + s_o h) \in \Omega_h \cup \Gamma_h,$$

as defined above with factors $s_* \in (0, 1]$ ($\star \in \{\ell, r, u, o\}$ for left, right, under, and over). For far-boundary points $s_\star = 1$ holds; for near-boundary points (x, y) at least one neighbour lies on Γ_h and the corresponding distance $s_\star h$ may be smaller than h. Equation (4.87) with $x' = x - s_\ell h$ and $x'' = x + s_r h$ provides an approximation for u_{xx}. Analogously one can replace u_{yy} by a divided difference. One obtains the *difference scheme of Shortley and Weller* [264]:

$$\begin{aligned} -D_h u(x,y) := \frac{1}{h^2}\Bigg[&\left(\frac{2}{s_\ell s_r} + \frac{2}{s_u s_o}\right) u(x,y) \\ &- \frac{2}{s_o\,(s_o + s_u)}\,u(x, y + s_o h) - \frac{2}{s_\ell\,(s_\ell + s_r)}\,u(x - s_\ell h, y) \\ &- \frac{2}{s_r\,(s_\ell + s_r)}\,u(x + s_r h, y) - \frac{2}{s_u\,(s_o + s_u)}\,u(x, y - s_u h)\Bigg]\,. \end{aligned} \tag{4.89}$$

Remark 4.76. If $s_\ell = s_r = s_u = s_o = 1$, D_h agrees with the standard five-point formula Δ_h (cf. Exercise 4.75c).

The discrete boundary-value problem assumes the form

$$\begin{aligned} -D_h u_h = f_h := \tilde{R}_h f \quad &\text{on } \Omega_h \qquad \text{with } (\tilde{R}_h f)(\mathbf{x}) := f(\mathbf{x}), \\ u_h = \varphi \quad &\text{on } \Gamma_h\,. \end{aligned} \tag{4.90}$$

The five coefficients on the right-hand side of equation (4.89) define the matrix elements $L_{\mathbf{x}\boldsymbol{\xi}}$ for $\boldsymbol{\xi} = \mathbf{x}$ and for the four neighbours $\boldsymbol{\xi}$ of $\mathbf{x}$. Otherwise we set

$L_{\mathbf{x}\boldsymbol{\xi}} = 0$. The right-hand side of the system of equations

$$L_h u_h = q_h,$$

in which u_h is interpreted as a grid function on Ω_h (not $\Omega_h \cup \Gamma_h$) is given by

$$q_h = f_h + \varphi_h, \qquad \varphi_h(\mathbf{x}) := -\sum_{\boldsymbol{\xi}\in\Gamma_h} L_{\mathbf{x}\boldsymbol{\xi}}\, \varphi(\boldsymbol{\xi}).$$

Again, $\varphi_h(\mathbf{x}) = 0$ holds for far-boundary points $\mathbf{x} \in \Omega_h$.

Theorem 4.77. *Let Ω be bounded and contained in a strip $(x_0, x_0 + d) \times \mathbb{R}$ or $\mathbb{R} \times (y_0, y_0 + d)$ of width d. For the matrix L_h belonging to the Shortley–Weller discretisation the following holds:*

(a) L_h is generally not symmetric;

(b) L_h is M-matrix with

$$\|L_h^{-1}\|_\infty \le d^2/8. \tag{4.91}$$

Proof. (a) Let $\mathbf{x} = (x, y) \in \Omega_h$ be a near-boundary point with $(x - s_\ell h, y) \in \Gamma_h$, but its neighbour $\mathbf{x}' = (x + h, y) \in \Omega_h$ be a far-boundary point. Then it holds that $L_{\mathbf{x}\mathbf{x}'} = \frac{-2h^{-2}}{1+s_\ell} \neq -h^{-2} = L_{\mathbf{x}'\mathbf{x}}$ if $s_\ell < 1$. Other than in Exercise 4.56, in general no scaling can be found so that $D_h L_h$ (D_h diagonal) becomes symmetric.

(b) L_h need not necessarily be irreducible and hence irreducibly diagonally dominant. But the weaker condition in Exercise 4.16 is satisfied and proves the M-matrix property.

(c) For the proof of (4.91) we use Theorem 4.24. If the domain Ω lies in the strip $(x_0, x_0 + d) \times \mathbb{R}$ we select $w_h(x, y) := R_h w$, $w := (x - x_0)(x_0 + d - x)/2$. The remainder in equation (4.87) contains only third derivatives that vanish for w. Thus $D_h w_h$ agrees with $\Delta w = -1$:

$$-D_h u_h = \mathbf{1} \quad \text{in } \Omega_h, \qquad w_h \ge 0 \quad \text{on } \Gamma_h.$$

The corresponding system of equations reads $L_h w_h = q_h := f_h + \varphi_h$ with $f_h = \mathbf{1}$ and $q_h \ge 0$. Thus we have $L_h w_h \ge \mathbf{1}$ and Theorem 4.24 proves $\|L_h^{-1}\|_\infty \le \|w_h\|_\infty \le (\frac{d}{2})^2/2 = d^2/8$. ∎

Exercise 4.78. Prove the analogue of the inequality (4.45):

$$\|u_h\|_\infty \le \left\|L_h^{-1}\right\|_\infty \|f_h\|_\infty + \max_{\boldsymbol{\xi}\in\Gamma_h} |\varphi(\boldsymbol{\xi})| \le \frac{d^2}{8} \max_{\mathbf{x}\in\Omega_h} |f(\mathbf{x})| + \max_{\boldsymbol{\xi}\in\Gamma_h} |\varphi(\boldsymbol{\xi})|\,.$$

With (4.91) stability has been proved. The order of consistency however is only 1, for at near-boundary points

$$c_h := D_h R_h u - \tilde{R}_h \Delta u$$

is of the order of magnitude $\mathcal{O}(h^1)$. Here, R_h is the restriction to $\Omega_h \cup \Gamma_h$. $\tilde{R}_h$ has been defined in (4.90). We want to show that nevertheless there is convergence of

order 2. The difference $w_h := u_h - R_h u$ between the discrete solution $u_h = L_h^{-1} q_h$ and the solution $u \in C^{3,1}(\overline{\Omega})$ of (4.85) satisfies

$$-D_h w_h = -D_h u_h + D_h R_h u = \tilde{R}_h f + D_h R_h u = D_h R_h u - \tilde{R}_h \Delta u = c_h \quad \text{in } \Omega_h,$$
$$w_h = 0 \qquad \text{on } \Gamma_h,$$

so that $w_h = L_h^{-1} c_h$ follows. c_h can be written as $c_h^x + c_h^y$, where c_h^x (resp. c_h^y) is the error of discretisation of the x difference (resp. y difference). In turn, c_h^x is split into $c_h^x = c_h^{x,1} + c_h^{x,2}$:

$$c_h^{x,2}(x,y) := \left\{ \begin{array}{ll} c_h^x(x,y), & \text{if } s_\ell = s_r = 1 \\ 0 & \text{otherwise} \end{array} \right\}, \qquad c_h^{x,1} := c_h^x - c_h^{x,2}.$$

Analogously one defines $c_h^{y,1}$ and $c_h^{y,2}$ and sets

$$c_h^1 = c_h^{x,1} + c_h^{y,1}, \qquad c_h^2 = c_h^{x,2} + c_h^{y,2}, \qquad w_h^i := L_h^{-1} c_h^i \quad (i = 1, 2)\,.$$

The errors c_h^2 are described by (4.49):

$$\|w_h^2\|_\infty \le \left\|L_h^{-1}\right\|_\infty \|c_h^2\|_\infty, \qquad \|c_h^2\|_\infty \le \frac{1}{6} h^2 \|u\|_{C^{3,1}(\overline{\Omega})}\,. \tag{4.93a}$$

With $K := \frac{1}{3} h^3 \|u\|_{C^{2,1}(\overline{\Omega})}$ define

$$v_h = K\mathbf{1} \quad \text{in } \Omega_h, \qquad v_h = 0 \quad \text{on } \Gamma_h, \qquad \tilde{c}_h := L_h v_h.$$

$\tilde{c}_h(\mathbf{x}) = 0$ holds for far-boundary points $\mathbf{x} \in \Omega_h$; for near-boundary points $\mathbf{x} \in \Omega_h$, however, we have

$$\tilde{c}_h(\mathbf{x}) = K \sum_{\boldsymbol{\xi} \in \Omega_h} L_{\mathbf{x}\boldsymbol{\xi}} = -K \sum_{\xi \in \Gamma_h} L_{\mathbf{x}\xi} \qquad (\mathbf{x} \in \Omega_h \text{ near the boundary})\,.$$

Consider, for example, the near-boundary case $\mathbf{x} = (x, y) \in \Omega_h$ with $\boldsymbol{\xi} = (x - s_\ell h, y) \in \Gamma_h$. According to Exercise (4.75a,b) the x difference has the error

$$|c_h^{x,1}(\mathbf{x})| \le \frac{h}{3} \frac{s_r^2 + s_\ell^2}{s_r + s_\ell} \|u\|_{C^{2,1}(\overline{\Omega})} = h^{-2} \frac{s_r^2 + s_\ell^2}{s_r + s_\ell} K \le K \frac{2h^{-2}}{s_\ell\,(s_r + s_\ell)} = -K L_{\mathbf{x}\boldsymbol{\xi}}.$$

The analogous estimate for $c_h^{y,1}(\mathbf{x})$ gives $\left|c_h^1(\mathbf{x})\right| \le \tilde{c}_h(\mathbf{x})$. Because $c_h^1(\mathbf{x}) = \tilde{c}_h(\mathbf{x}) = 0$ at far-boundary points $\mathbf{x} \in \Omega_h$ one has $-\tilde{c}_h \le c_h h^1 \le \tilde{c}_h$. Since $L_h^{-1} \ge 0$ it follows that $-v_h \le w_h^1 \le v_h$, i.e.,

$$\|w_h^1\|_\infty \le K = \frac{1}{3} h^3 \|u\|_{C^{2,1}(\overline{\Omega})}\,. \tag{4.93b}$$

Equations (4.93a,b) together with $w_h^1 + w_h^2 = w_h = u_h - R_h u$ prove the following theorem.

Theorem 4.79 (convergence of the Shortley–Weller method). *Let Ω satisfy the assumption of Theorem 4.77. The convergence for the Shortley–Weller method is of second order if $u \in C^{3,1}(\overline{\Omega})$:*

$$\begin{aligned}\|u_h - R_h u_h\|_\infty &\le \frac{1}{3}h^3 \|u\|_{C^{2,1}(\overline{\Omega})} + \|L_h^{-1}\|_\infty \frac{1}{6}h^2 \|u\|_{C^{3,1}(\overline{\Omega})} \\ &\le \left(\frac{1}{3}h^3 + \frac{d^2}{48}h^2\right) \|u\|_{C^{3,1}(\overline{\Omega})} . \end{aligned} \tag{4.94}$$

Exercise 4.80. Show that if one uses the Shortley–Weller discretisation for all near-boundary points, but the mehrstellen method from Section 4.6 for all far-boundary points, the convergence is of third order: $\|u_h - R_h u\|_\infty = \mathcal{O}(h^3)$.

Approximations of fourth order are described by van Linde [294].

4.8.2 Interpolation in Near-Boundary Points

Instead of discretising the Poisson equation at near-boundary points $\mathbf{x} \in \Omega_h$ by a difference scheme, one could also try to determine $u_h(\mathbf{x})$ by interpolation from the neighbouring points. If, for example, $\mathbf{x} = (x, y) \in \Omega_h$ is near the boundary, $(x - s_\ell h, y) \in \Gamma_h$ and $(x + s_r h, y) \in \Omega_h \cup \Gamma_h$, then linear interpolation yields

$$u_h(x, y) = \left[s_\ell u_h(x + s_r h, y) + s_r u_h(x - s_\ell h, y)\right] / (s_r + s_\ell) .$$

Thus, at the point $\mathbf{x}$ we set up the equation

$$(s_r + s_\ell)\, u_h(x, y) - s_\ell u_h(x + s_r h, y) - s_r u_h(x - s_\ell h, y) = 0. \tag{4.95a}$$

Since $\boldsymbol{\xi} = (x - s_\ell h, y)$ should be a boundary point, one can replace $u_h(\boldsymbol{\xi})$ by $\varphi(\boldsymbol{\xi})$. If, however, $(x, y + s_o h)$ or $(x, y - s_u h)$ is a boundary point, we choose interpolation in the y direction:

$$(s_u + s_o)\, u_h(x, y) - s_u u_h(x + s_o h, y) - s_o u_h(x - s_u h, y) = 0. \tag{4.95b}$$

At any far-boundary point $\mathbf{x} \in \Omega_h$ the five-point formula (4.10) is used:

$$-\left(\Delta_h u_h\right)(\mathbf{x}) = f_h(\mathbf{x}) = \tilde{R}_h f(\mathbf{x}) = f(\mathbf{x}) \qquad (\mathbf{x} \text{ far-boundary point}) . \tag{4.95c}$$

Theorem 4.81. *Let Ω satisfy the assumptions of Theorem 4.77. Let the discretisation be given by (4.95a–c) with the choice between (4.95a) and (4.95b) being made in such a way that always (at least) one boundary point is used for the interpolation. Let the system of equations resulting after the elimination of the boundary values $u_h(\boldsymbol{\xi}) = \varphi(\boldsymbol{\xi})$ ($\boldsymbol{\xi} \in \Gamma_h$) be $L_h u_h = q_h$. L_h is an (in general unsymmetric) M-matrix which satisfies the estimate (4.91). The discrete solutions u_h converge with the order 2 if $u \in C^{3,1}(\overline{\Omega})$:*

$$\|u_h - R_h u_h\|_\infty \le h^2 \|u\|_{C^{1,1}(\overline{\Omega})} + \frac{1}{6}h^2 \|L_h^{-1}\|_\infty \|u\|_{C^{3,1}(\overline{\Omega})} .$$

Proof. (i) The M-matrix property and (4.91) are proved as in Theorem 4.77.

(ii) Let $\mathbf{x} \in \Omega_h$ be a near-boundary point at which (4.95a) is used. The error of interpolation is

$$c_h^1(x,y) := (s_r + s_\ell)\, R_h u(x,y) - s_\ell R_h u(x + s_r h, y) - s_r R_h u(x - s_\ell h, y),$$
$$\left|c_h^1(x,y)\right| \le \frac{1}{2} s_r s_\ell\, (s_r + s_\ell)\, h^2 \left\|u\right\|_{C^{1,1}(\overline{\Omega})}.$$

With this c_h^1 and $K := h^2 \left\|u\right\|_{C^{1,1}(\overline{\Omega})}$ one can essentially just repeat the proof of Theorem 4.79. ∎

By rescaling and adding the equations (4.95a,b), one obtains

$$\frac{1}{h^2}\left\{\left(\frac{s_r + s_\ell}{s_r s_\ell} + \frac{s_o + s_u}{s_o s_u}\right) u_h(x,y) - \frac{1}{s_\ell} u_h(x - s_\ell h, y) \right. \tag{4.96}$$
$$\left. - \frac{1}{s_r} u_h(x + s_r h, y) - \frac{1}{s_o} u_h(x, y + s_o h) - \frac{1}{s_u} u_h(x, y - s_u h)\right\} = 0.$$

Using this device one can obtain a symmetric matrix L_h, even at arbitrary Ω.

Exercise 4.82. Show that the discretisation (4.95c), (4.96) leads to a symmetric M-matrix. The estimate of convergence reads

$$\left\|u_h - R_h u_h\right\|_\infty \le 2h^2 \left\|u\right\|_{C^{1,1}(\overline{\Omega})} + \frac{1}{6} d^2 h^2 \left\|u\right\|_{C^{3,1}(\overline{\Omega})}.$$

In (4.95a,b) linear interpolation was chosen because the values at the neighbouring points were sufficient for it. Constant interpolation by

$$u(x,y) = u_h(x - s_\ell h, y) = \varphi(x - s_\ell h, y) \quad \text{if } (x,y) \in \Omega_h,\ (x - s_\ell h, y) \in \Gamma_h,$$

is less desirable since it only provides first-order convergence : $\left\|u_h - R_h u\right\|_\infty = \mathcal{O}(h)$. By contrast, interpolation of higher order is very applicable indeed (cf. Pereyra—Proskurowski—Widlund [219]). However, it is described by an equation which also contains points at an distance of $\ge 2h$. Higher boundary approximations are in particular then necessary if one wants to apply extrapolation methods (cf. Marchuk–Shaidurov [199, pages 162ff]).

Chapter 5
General Boundary-Value Problems

Abstract Section 5.1 introduces the general elliptic linear differential equation of second order together with the Dirichlet boundary values. An important statement is the maximum-minimum principle in §5.1.2. In §5.1.3 sufficient conditions for the uniqueness of the solution and the continuous dependence on the data are proved. The discretisation of the general differential equation in a square is described in §5.1.4. **Section 5.2** treats alternative boundary conditions replacing the Dirichlet data. Examples are the Neumann condition, the conormal derivative and the Robin boundary condition. Their discretisation (cf. §5.2.2) for general domains is rather laborious. **Section 5.3** discusses differential equations of higher order. In particular, the biharmonic equation of fourth order is described in §5.3.1 followed by equations of order $2m$ in §5.3.2. The discretisation of the biharmonic equation is in §5.3.3.

5.1 Dirichlet Boundary-Value Problems for Linear Differential Equations

5.1.1 Posing the Problem

In Section 1.2 we have already formulated the general linear differential equation of second order:

$$Lu = f \quad \text{in } \Omega \quad \text{with } L = \sum_{i,j=1}^{n} a_{ij}(\mathbf{x}) \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^{n} a_i(\mathbf{x}) \frac{\partial}{\partial x_i} + a(\mathbf{x}). \quad (5.1)$$

We mentioned that, without loss of generality, one can assume

$$a_{ij}(\mathbf{x}) = a_{ji}(\mathbf{x}) \qquad (1 \le i, j \le n,\ \mathbf{x} \in \Omega)$$

so that the matrix

$$A(\mathbf{x}) = (a_{ij}(\mathbf{x}))_{i,j=1,\ldots,n} \quad (5.2)$$

W. Hackbusch, *Elliptic Differential Equations*, Springer Series
in Computational Mathematics 18, DOI 10.1007/978-3-662-54961-2_5

is symmetric. A differential operator apparently more general than (5.1) is

$$L = \sum_{i,j=1}^{n} \left[a_{ij}^{I} \frac{\partial^2}{\partial x_i \partial x_j} + \frac{\partial}{\partial x_j} a_{ij}^{II} \frac{\partial}{\partial x_i} + \frac{\partial^2}{\partial x_i \partial x_j} a_{ij}^{III} \right] + \sum_{i=1}^{n} \left[a_i^{I} \frac{\partial}{\partial x_i} + \frac{\partial}{\partial x_i} a_i^{II} \right] + a(\mathbf{x}). \tag{5.3}$$

But since, for example,

$$\frac{\partial}{\partial x_j} \left(a_{ij}^{II} \frac{\partial}{\partial x_i} u \right) = a_{ij}^{II} u_{x_i x_j} + \left(\partial a_{ij}^{II} / \partial x_j \right) u_{x_i},$$

the operator (5.3) can be described in the form (5.1), provided that the coefficients are sufficiently often differentiable. According to Definition 1.14, equation (5.1) is elliptic in Ω if all eigenvalues of $A(\mathbf{x})$ have the same sign. One can assume without loss of generality that all eigenvalues are negative so that $A(\mathbf{x})$ is negative definite (cf. Exercise 4.30a). Thus, L is elliptic in Ω if

$$-\sum_{i,j=1}^{n} a_{ij}(\mathbf{x}) \xi_i \xi_j > 0 \qquad \text{for all } \mathbf{x} \in \Omega,\ 0 \neq \boldsymbol{\xi} \in \mathbb{R}^n. \tag{5.4a}$$

The choice of the negative sign corresponds to Footnote 1 on page 13. For any $\mathbf{x} \in \Omega$ there exists

$$c(\mathbf{x}) := \min \left\{ -\sum_{i,j=1}^{n} a_{ij}(\mathbf{x}) \xi_i \xi_j : |\boldsymbol{\xi}| = 1 \right\}$$

and it must be positive ($c(\mathbf{x})$ is the smallest eigenvalue of $A(\mathbf{x})$). Hence one can also write (5.4a) in the form (5.4b):

$$-\sum_{i,j=1}^{n} a_{ij}(\mathbf{x}) \xi_i \xi_j \geq c(\mathbf{x}) \, |\boldsymbol{\xi}|^2, \quad c(\mathbf{x}) > 0 \text{ for all } \mathbf{x} \in \Omega,\ 0 \neq \boldsymbol{\xi} \in \mathbb{R}^n. \tag{5.4b}$$

Definition 5.1. The equation (5.1), or the operator L, is defined to be *uniformly elliptic* in Ω if

$$\inf\{c(\mathbf{x}) : \mathbf{x} \in \Omega\} > 0 \qquad (\, c(\mathbf{x}) \text{ from (5.4b)} \,). \tag{5.4c}$$

Exercise 5.2. Let L have continuous coefficients in the domain Ω and let it be elliptic. Then in each compact set $K \subset \Omega$, L is uniformly elliptic.

On $\Gamma = \partial\Omega$ we impose the following Dirichlet boundary-value condition:

$$u = \varphi \qquad \text{on } \Gamma. \tag{5.5}$$

Exercise 5.3. Let $\mathbf{x} \mapsto \Phi(\mathbf{x}) := \hat{\mathbf{x}} := \mathbf{x}_0 + \mathbf{S}\mathbf{x}$ be an affine map of Ω onto $\hat{\Omega}$, where $\mathbf{S}$ is a regular $n \times n$ matrix. Show that the operator $\hat{L}$ of the transformed differential equation has the same ellipticity properties and that the last coefficient in

$$\hat{L} = \sum_{i,j=1}^{n} \hat{a}_{ij}(\hat{\mathbf{x}}) \frac{\partial^2}{\partial \hat{x}_i \partial \hat{x}_j} + \sum_{i=1}^{n} \hat{a}_i(\hat{\mathbf{x}}) \frac{\partial}{\partial \hat{x}_i} + \hat{a}(\hat{\mathbf{x}})$$

satisfies $\hat{a}(\Phi(\mathbf{x})) = a(\mathbf{x})$. *Hint:* Exercise 1.16.

5.1.2 Maximum Principle

In general, the maximum principle does not hold for the equation $Lu = f$, nor is the solution of the boundary-value problem (5.1), (5.5) uniquely determined.

Example 5.4. Let $\Omega = (0,\pi) \times (0,\pi)$, $\varphi = 0$, $f = 0$, $L = -\Delta - 2$. Then both $u = 0$ and $u(x,y) = \sin(x)\sin(y)$ are solutions of the boundary-value problem. The second solution assumes its maximum at the interior point $(\pi/2, \pi/2) \in \Omega$.

In the above example the coefficient $a(\mathbf{x}) = -2$ (cf. (5.1)) has the wrong sign. As soon as $a \geq 0$, we have the following statement due to Hopf [155].

Theorem 5.5 (maximum-minimum principle). *Assume the coefficients of the elliptic operator (5.1) are continuous in Ω. Let $u \in C^2(\Omega)$ satisfy $Lu = f$ and be nonconstant. For any compact set $K \subset \Omega$ with $a(\mathbf{x}) \geq 0$ we have:*

(a) if $f \geq 0$ and $u \leq 0$ in K, the minimum of u is taken on ∂K;

(b) if $f \leq 0$ and $u \geq 0$ in K, the maximum of u is taken on ∂K.

Proof. (i) It is sufficient to prove part (a), since (b) coincides with the statement (a) for $-u$ and $-f$.

(ii) Since K is compact, u takes a minimum in K. For an indirect proof assume that the minimum of u is attained at an *interior* point $\mathbf{x}^\star \in K \backslash \partial K$ and that

$$u(\mathbf{x}^\star) < \min_{\mathbf{x} \in \partial K} u(\mathbf{x}) \leq 0. \tag{5.6a}$$

Consequently we have $u_{x_i}(\mathbf{x}^\star) = 0$, and the *Hessian matrix* $B := (u_{x_i x_j}(\mathbf{x}^\star))_{i,j=1}^n$ is positive semidefinite (i.e., $B = B^\mathsf{T}$ and $\langle \boldsymbol{\xi}, B\boldsymbol{\xi} \rangle \geq 0$ for all $\boldsymbol{\xi} \in \mathbb{R}^n$). Evaluation of the differential equation at $\mathbf{x}^\star$ yields

$$0 \leq f(\mathbf{x}^\star) = (Lu)(\mathbf{x}^\star) = \underbrace{\sum_{i,j=1}^{n} a_{ij}(\mathbf{x}^\star)\, u_{x_i x_j}(\mathbf{x}^\star)}_{\leq 0} + \underbrace{a(\mathbf{x}^\star)}_{\geq 0} \underbrace{u(\mathbf{x}^\star)}_{\leq 0} \leq 0. \tag{5.6b}$$

The inequality $\operatorname{trace}(A(\mathbf{x}^\star)B) = \sum_{i,j} a_{ij}(\mathbf{x}^\star) u_{x_i x_j}(\mathbf{x}^\star) \leq 0$ (first sum on the right-hand side) is the subject of Exercise 5.6c.

First we replace the inequality $f \geq 0$ by the stronger $f > 0$. Then inequality (5.6b) yields the contradiction $0 < f(\mathbf{x}^\star) = \ldots \leq 0$.

(iii) It remains to study the case $f \geq 0$. Without loss of generality, we may assume that K is contained in $\{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 \geq 1\}$ (cf. Exercise 5.3). The function $q(\mathbf{x}) := \exp(-\alpha\|\mathbf{x}\|_2^2)$ $(\alpha \in \mathbb{R})$ has the properties $q(\mathbf{x}) > 0$ and

$$Lq(\mathbf{x}) = \left(4\alpha^2 \sum_{i,j=1}^{n} a_{ij}(\mathbf{x})x_i x_j - 2\alpha \sum_{i=1}^{n} \big[a_{ii}(\mathbf{x}) + a_i(\mathbf{x})x_i\big] + a(\mathbf{x})\right) q(\mathbf{x}). \quad (5.6c)$$

Since the region K is compact, L is uniformly elliptic in K (cf. Exercise 5.2), so that $-\sum_{i,j=1}^{n} a_{ij}(\mathbf{x})x_i x_j \geq c\|\mathbf{x}\|_2^2$ with $c > 0$. Since $\|\mathbf{x}\|_2 \geq 1$ for $\mathbf{x} \in K$, we have $-\sum_{i,j=1}^{n} a_{ij}(\mathbf{x})x_i x_j \geq c > 0$. For sufficiently large $|\alpha|$, the quadratic part $4\alpha^2 \sum_{i,j=1}^{n} a_{ij}(\mathbf{x})x_i x_j$ prevails, and the bracket in (5.6c) is negative. This choice gives $Lq < 0$ in K. We set

$$v(\mathbf{x}) := u(\mathbf{x}) - \beta\, q(\mathbf{x}) \qquad \text{with } \beta > 0.$$

v satisfies the differential equation $Lv = g := f - \beta\, Lq$ and the condition $v \leq 0$. The inequalities $Lq < 0$, $\beta > 0$, and $f \geq 0$ imply $g > 0$.

The inequalities in (5.6a) are also valid for v if $\beta > 0$ is chosen sufficiently small. Hence $v \leq 0$ and $Lv = g > 0$ hold in K, but the minimum of v is not attained on $\partial K_\rho(\mathbf{x}^\star)$. This contradicts the result in part (ii), since the function f there is now $g > 0$. ∎

Exercise 5.6. The trace of a square matrix is defined by $\operatorname{trace}(A) := \sum_{i=1}^{n} a_{ii}$. Prove that:

(a) $\operatorname{trace}(AB) = \operatorname{trace}(BA) = \sum_{i,j=1}^{n} a_{ij}b_{ji}$;

(b) $a_{ii} \geq 0$ and $\operatorname{trace}(A) \geq 0$, if A is positive semidefinite;

(c) $\operatorname{trace}(AB) \geq 0$ if A and B are positive semidefinite. *Hint:* $B^{1/2}AB^{1/2}$ is positive semidefinite; cf. Exercise 4.30e.

The maximum-minimum principle in Theorem 2.13 (for the Laplace case) is stronger, since an interior extremum implies that the solution u must be constant. This is called the *strong maximum-minimum principle*.

Corollary 5.7. Under the conditions of Theorem 5.5 the strong maximum-minimum principle holds: if the extremum is taken in the interior of K, u is constant in K.

Proof. Assume $Lu = f \geq 0$ and $u \leq 0$ in K. Let the minimum be assumed at the interior point $\mathbf{x}^\star \in \mathring{K}$:

$$u(\mathbf{x}^\star) = m := \min\{u(\mathbf{x}) : \mathbf{x} \in K\}.$$

If the set

$$\omega := \{\mathbf{x} \in K : u(\mathbf{x}) > m\}$$

is empty, $u = m$ in K follows. For an indirect proof assume $\omega \neq \emptyset$.

Since ω is open (relative to K), it cannot be a subset of ∂K. Also $\partial\omega \subset \partial K$ is excluded since otherwise $\overline{\omega} = K$ in contradiction to $\mathbf{x}^\star \in \mathring{K}$ and $\mathbf{x}^\star \notin \omega$. Hence there exists $\mathbf{x}' \in \partial\omega \backslash \partial K$ with $d := \operatorname{dist}(\mathbf{x}', \partial K) > 0$. Let $\mathbf{y} \in \omega$ be a point in the $d/3$ neighbourhood of $\mathbf{x}'$. Obviously,

$$\delta := \operatorname{dist}(\mathbf{y}, \partial\omega) \leq d/3 < 2d/3 \leq \operatorname{dist}(\mathbf{y}, \partial K).$$

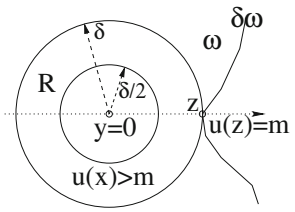


Fig. 5.1 Annulus $K_\delta \backslash \overline{K_{\delta/2}}$.

By construction of δ, the open ball $K_\delta := K_\delta(\mathbf{y})$ satisfies

$$K_\delta \subset \omega, \qquad \overline{K_\delta} \subset \mathring{K} \quad (\text{i.e., } \overline{K_\delta} \cap \partial K = \emptyset), \qquad \overline{K_\delta} \cap (\partial\omega \backslash \partial K) \neq \emptyset,$$

i.e., there exists $\mathbf{z} \in \partial K_\delta \cap \partial\omega \backslash \partial K \subset \mathring{K} \backslash \omega$. The definition of the sets implies

$$u(\mathbf{y}) > m, \qquad u(\mathbf{z}) = m.$$

Without loss of generality (after a suitable affine transformation, cf. Exercise 5.3) the situation of Figure 5.1 can be assumed:

$$\mathbf{y} = \mathbf{0}, \qquad \mathbf{z} = \delta\,\mathbf{e} \qquad (\mathbf{e}\text{: unit vector in } x_1 \text{ direction}).$$

Figure 5.1 shows the ball $K_{\delta/2} := K_{\delta/2}(\mathbf{y}) \subset K_\delta$ and the (open) annulus $R := K_\delta \backslash \overline{K_{\delta/2}}$. The boundary ∂R is the union of $\partial K_{\delta/2}$ and ∂K_δ. Because of $\partial K_{\delta/2} \subset \omega$ we have

$$\min\{u(\mathbf{x}) : \mathbf{x} \in \partial K_{\delta/2}\} > m,$$

while $u(\mathbf{x}) \geq m$ for all $\mathbf{x} \in \partial K_\delta$ with the minimum $u(\mathbf{z}) = m$ at $\mathbf{z} \in \partial K_\delta$.

The function q used in the proof of Theorem 5.5 is marginally modified:

$$p(\mathbf{x}) := \exp(-\alpha\|\mathbf{x}\|_2^2) - \exp(-\alpha\delta^2) \quad \text{with } \alpha > 0 \text{ for } \frac{\delta}{2} \leq \|\mathbf{x}\|_2 \leq \delta \text{ (i.e., } \mathbf{x} \in \overline{R})$$

Still $Lp \geq 0$ and $p \geq 0$ hold in $\overline{R}$ for sufficiently large $\alpha > 0$. Hence $w := u - cp$ with $c \geq 0$ satisfies the sign conditions $Lw =: g \leq 0$ and $w \leq 0$. According to Theorem 5.5, w as a function on $\overline{R}$ takes its minimum on $\partial R = \partial K_{\delta/2} \cup \partial K_\delta$.

For sufficiently small $c > 0$ we have

$$\min\{w(\mathbf{x}) : \mathbf{x} \in \partial K_{\delta/2}\} > m,$$

while $p(\mathbf{x}) = 0$ on ∂K_δ leads to $w|_{\partial K_\delta} = u|_{\partial K_\delta}$ and hence to

$$\min\{w(\mathbf{x}) : \mathbf{x} \in \partial K_\delta\} = \min\{u(\mathbf{x}) : \mathbf{x} \in \partial K_\delta\} = u(\mathbf{z}) = w(\mathbf{z}) = m.$$

In particular, the maximum-minimum principle $w(\mathbf{x}) \geq m$ holds in $\overline{R}$.

Since w restricted to $\overline{R}$ has a minimum at a boundary point $\mathbf{z}$, $\frac{\partial}{\partial x_1} w(\mathbf{z}) \leq 0$ must hold in the coordinate system of Figure 5.1. Inserting the definition $w = u - cp$ yields

$$\frac{\partial u(\mathbf{z})}{\partial x_1} \leq c \frac{\partial p(\mathbf{z})}{\partial x_1} = -2\, c\, \delta\, \alpha\, \mathrm{e}^{-\alpha \delta^2} < 0\,.$$

On the other hand, $\mathbf{z} \in \partial K_\delta \subset \mathring{K}$ is a local minimum of u, so that $\frac{\partial u(\mathbf{z})}{\partial x_1} = 0$ in contradiction to the previous inequality. ■

Corollary 5.8. If the coefficient $a(\mathbf{x})$ of L (cf. (5.1)) vanishes in K, the requirements $u \leq 0$ (for a minimum) or $u \geq 0$ (for a maximum) are not needed in Theorem 5.5 and Corollary 5.7.

Proof. The sign condition of u is solely needed to ensure the correct sign of the product au. Obviously, this is also guaranteed by $a = 0$. ■

Remark 5.9. The continuity of the coefficients a_{ij}, a_i, and a of L in Theorem 5.5 and Corollary 5.7 can be replaced by the assumption $a > 0$ in Ω or by the stipulation: In every compact set $K \subset \Omega$ let a_{ij}, a_i, and a be bounded and let L be uniformly elliptic.

5.1.3 Uniqueness of the Solution and Continuous Dependence

Lemma 5.10. *Let Ω be bounded, let the coefficients of L be continuous and let (5.4a) be valid and $a \geq 0$ in Ω. Let $u_1, u_2 \in C^2(\Omega) \cap C^0(\overline{\Omega})$ be solutions of the boundary-value problems*

$$Lu_i = f_i \quad \text{in } \Omega, \qquad u_i = \varphi_i \quad \text{on } \Gamma \qquad (i = 1, 2). \tag{5.7}$$

If $f_1 \leq f_2$ in Ω and $\varphi_1 \leq \varphi_2$ on Γ, then also $u_1 \leq u_2$ in Ω.

Proof. The difference $v = u_2 - u_1$ satisfies the equations $Lv = f_2 - f_1 \geq 0$ in Ω and $v = \varphi_2 - \varphi_1 \geq 0$ on Γ. Let $\omega := \{\mathbf{x} \in \Omega : v(\mathbf{x}) < 0\}$. Obviously, $v(\mathbf{x}) = 0$ holds on $\partial\omega$. Then Theorem 5.5 with $K := \overline{\omega}$ leads to a contradiction. Hence $\omega = \emptyset$ follows, i.e., $v \geq 0$ and $u_2 \geq u_1$. ■

Theorem 5.11 (uniqueness). *Under the conditions of Lemma 5.10 the solution of the boundary-value problem $Lu = f$ in Ω and $u = \varphi$ on Γ is uniquely determined.*

Proof. Let u_1, u_2 be two solutions. Lemma 5.10 with $f_1 = f_2 = f$, $\varphi_1 = \varphi_2 = \varphi$ shows that $u_1 \leq u_2$ as well as $u_2 \leq u_1$. Thus, $u_1 = u_2$. ■

The next theorem states that the solution depends Lipschitz-continuously on f and φ.

Theorem 5.12. *Let L be uniformly elliptic in Ω. Under the conditions of Lemma 5.10 the following holds:*

$$\|u_1 - u_2\|_\infty \le \|\varphi_1 - \varphi_2\|_\infty + M \|f_1 - f_2\|_\infty \tag{5.8}$$

for solutions u_1, u_2 of (5.7). In this inequality the number M depends only on

$$K := \sup\{|a_{ij}(\mathbf{x})|, |a_i(\mathbf{x})|, |a(\mathbf{x})| : \mathbf{x} \in \Omega\},$$

on the ellipticity constant defined by $m := \inf\{c(\mathbf{x}) : \mathbf{x} \in \Omega\} > 0$ (cf. (5.4c)) and on the diameter of the domain Ω.

Proof. Let $\Omega \subset K_R(\mathbf{z})$. For all $\mathbf{x} \in \Omega$ we have $z_1 - R \le x_1 \le z_1 + R$. We select $\alpha \ge 0$ so that $m\alpha^2 - K(\alpha + 1) \ge 1$ and we define

$$w(\mathbf{x}) := \|\varphi_1 - \varphi_2\|_\infty + \left(\mathrm{e}^{2R\alpha} - \mathrm{e}^{\alpha(x_1 - z_1 + R)}\right) \|f_1 - f_2\|_\infty .$$

We compare w with the solution $v := u_1 - u_2$ of $Lv = f_1 - f_2$ in Ω, $v = \varphi_1 - \varphi_2$ on Γ. From the choice of $K_R(\mathbf{z})$ we have

$$w(\mathbf{x}) \ge v(\mathbf{x}) \qquad \text{for } \mathbf{x} \in \Gamma.$$

Furthermore, the selection made of α ensures:

$$\begin{aligned}
(Lw)(\mathbf{x}) &= \underbrace{a(\mathbf{x}) \|\varphi_1 - \varphi_2\|_\infty}_{\ge 0} \\
&\quad + \Big\{ \underbrace{a\, \mathrm{e}^{2R\alpha}}_{\ge 0} - \underbrace{\mathrm{e}^{\alpha(x_1 - z_1 + R)}}_{\ge 1} [\underbrace{a_{11}(\mathbf{x})}_{\le -m} \alpha^2 + \underbrace{a_1(\mathbf{x})}_{\le K} \alpha + \underbrace{a(\mathbf{x})}_{\le K}] \Big\} \|f_1 - f_2\|_\infty \\
&\ge \left\{ m\alpha^2 - K(\alpha + 1) \right\} \|f_1 - f_2\|_\infty \ge \|f_1 - f_2\|_\infty \\
&\ge f_1(\mathbf{x}) - f_2(\mathbf{x}) = (Lv)(\mathbf{x}).
\end{aligned}$$

Using Lemma 5.10 with $u_1 = v$, $u_2 = w$, the result is $v \le w$ in Ω, i.e.,

$$u_1(\mathbf{x}) - u_2(\mathbf{x}) = v(\mathbf{x}) \le w(\mathbf{x}) \le \|w\|_\infty \le \|\varphi_1 - \varphi_2\|_\infty + M \|f_1 - f_2\|_\infty ,$$

where $M := \mathrm{e}^{2R\alpha}$. Analogously one proves $-w \le v$ so that (5.8) follows. ∎

Exercise 5.13. (a) Let Ω be bounded; let the coefficients of L be continuous in Ω. Now show that L is uniformly elliptic in Ω if and only if L is elliptic in $\overline{\Omega}$.

(b) Theorem 5.12 holds for the special case $f_1 = f_2$ without the assumption of uniform ellipticity.

(c) Let a strip in $\mathbb{R}^n$ be described by $\Sigma = \{\mathbf{x} \in \mathbb{R}^n : 0 \le \langle \eta, \mathbf{x} - \mathbf{x}^* \rangle \le \delta\}$, where $\mathbf{x}^\star \in \mathbb{R}^n$ is a boundary point of Σ, $\eta \in \mathbb{R}^n$ with $|\eta| = 1$ is a unit vector, and δ is the strip width. Show that inequality (5.8) holds with $M := \mathrm{e}^{\delta\alpha}$, if $\Omega \subset \Sigma$ and α are as in the proof of Theorem 5.12.

(d) Under the conditions of Theorem 5.12 show that

$$\|u\|_\infty \le \|\varphi\|_\infty + M\,\|f\|_\infty \ .$$

Exercise 5.14. Let the coordinate transformation $\Phi : \mathbf{x} \in \overline{\Omega} \mapsto \boldsymbol{\xi} \in \overline{\Omega'}$ and its inverse $\Phi^{-1} : \overline{\Omega'} \to \overline{\Omega}$ be continuously differentiable. To the operator L (in the x coordinates) let L' correspond in the ξ coordinates. Show that if L satisfies the conditions of Lemma 5.10, or of Theorem 5.12, then so does L'.

The right-hand side f and the boundary values φ are not the only parameters on which the solution u depends. We next investigate how the solution depends on the coefficients a_{ij}, a_i, and a of the differential operator L.

Theorem 5.15. *Let the coefficients of L^I and L^{II} be a^I_{ij}, a^I_j, a^I resp. $a^{II}_{ij}, a^{II}_j, a^{II}$. Let u^I and u^{II} be solutions of*

$$L^I u^I = L^{II} u^{II} = f \quad \text{in } \Omega, \qquad u^I = u^{II} = \varphi \quad \text{on } \Gamma.$$

Let L^I satisfy the conditions of Theorem 5.12 and let u^{II} belong to $C^2(\overline{\Omega})$. With M from (5.8) we then have

$$\|u^I - u^{II}\|_\infty \le M \Bigg\{ \sum_{i,j=1}^{n} \|a^I_{ij} - a^{II}_{ij}\|_\infty \|u^{II}\|_{C^2(\overline{\Omega})} + \sum_{i=1}^{n} \|a^I_i - a^{II}_i\|_\infty \|u^{II}\|_{C^1(\overline{\Omega})} + \|a^I_i - a^{II}_i\|_\infty \|u^{II}\|_\infty \Bigg\}. \tag{5.9}$$

If $a^I_{ij} = a^{II}_{ij}$, the condition $u^{II} \in C^1(\overline{\Omega}) \cap C^2(\Omega)$ is sufficient; if also $a^I_i = a^{II}_i$ then just $u^{II} \in C^0(\overline{\Omega}) \cap C^2(\Omega)$ will do.

Proof. Set $f' := L^I u^{II}$. Then $\|f' - f\|_\infty = \|(L^I - L^{II})u^{II}\|_\infty$ can be bounded by the bracket on the right-hand side of (5.9). Theorem 5.12 applied to $L^I u^{II} = f'$ and $L^I u^{II} = f$ implies $\|u^I - u^{II}\|_\infty \le M\|f' - f\|_\infty$. ∎

5.1.4 Difference Methods for the General Differential Equation of Second Order

For notational reasons we limit ourselves to the two-dimensional case $n = 2$.

General domains $\Omega \in \mathbb{R}^2$ require special discretisations at the boundary as explained in Section 4.8. Here we only want to discuss the difference formulae at interior points. Therefore it will suffice to base our comments on the unit square $\Omega = (0,1) \times (0,1)$.

Let L be given by (5.1). If one wishes to obtain a difference method of consistency order 2, the following choice suggests itself, which will later be modified for reasons of stability:

$$
\begin{aligned}
a_{11}(x,y)\partial_x^+\partial_x^- + 2a_{12}\partial_x^0\partial_y^0 + a_{22}(x,y)\partial_y^+\partial_y^- + a_1(x,y)\partial_x^0 + a_2(x,y)\partial_y^0 + a(x,y) \\
= h^{-2}\begin{bmatrix} -a_{12}(x,y)/2 & a_{22}(x,y) & a_{12}(x,y)/2 \\ a_{11}(x,y) & -2\,[a_{11}(x,y)+a_{22}(x,y)] & a_{11}(x,y) \\ a_{12}(x,y)/2 & a_{22}(x,y) & -a_{12}(x,y)/2 \end{bmatrix} \\
+\frac{1}{2}h^{-1}\begin{bmatrix} 0 & a_2(x,y) & 0 \\ -a_1(x,y) & 0 & a_1(x,y) \\ 0 & -a_2(x,y) & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & a(x,y) & 0 \\ 0 & 0 & 0 \end{bmatrix}.
\end{aligned} \tag{5.10}
$$

Remark 5.16. The difference method (5.10) is a nine-point scheme of consistency order 2.

Let L_h be the matrix of the system of difference equations Ω_h associated to (5.10) (cf. Remark 4.7):

$$L_h u_h = q_h\,. \tag{5.11}$$

The solvability of equation (5.11), for arbitrary q_h, is equivalent to uniqueness. In the continuous case, uniqueness was essentially given by the condition $a \geq 0$ and the ellipticity condition $a_{11}a_{22} > a_{12}^2$ (cf. Theorem 5.11). These conditions are in general not sufficient to guarantee the solvability of equation (5.11). We thus replace (5.10) with another discretisation.

Theorem 5.17. *For the coefficients of L let (5.4a,b) hold, and assume that*

$$|a_{12}(x,y)| \leq \min\{-a_{11}(x,y), -a_{22}(x,y)\} \tag{5.12}$$

and $a(x,y) \geq 0$ in Ω. Then there exists for L a seven-point difference method of consistency order 1 such that the associated matrix L_h is an M-matrix. In particular, the resulting equation (5.11) is solvable.

Note that Condition (5.12) follows from the ellipticity inequality (5.4a) only if $a_{11} = a_{22} < 0$.

Proof. At the grid point $(x,y) \in \Omega_h$ (cf. (4.8a)) we abbreviate the coefficients $a_{ij}(x,y)$, $a_i(x,y)$, and $a(x,y)$ to a_{ij}, a_i, and a. The principal part $\sum a_{ij}\partial^2/\partial x_i \partial x_j$ $(x_1 = x,\ x_2 = y)$ is discretised by the following difference stars:

$$
\begin{aligned}
\frac{\partial^2}{\partial x_1^2}: \quad & h^{-2}\begin{bmatrix} & 0 & \\ 1 & -2 & 1 \\ & 0 & \end{bmatrix}, \qquad \frac{\partial^2}{\partial x_2^2}: \quad h^{-2}\begin{bmatrix} & 1 & \\ 0 & -2 & 0 \\ & 1 & \end{bmatrix}, \\
\frac{\partial^2}{\partial x_1 \partial x_2}: \quad & \begin{cases} \frac{1}{2}h^{-2}\begin{bmatrix} 0 & -1 & 1 \\ -1 & 2 & -1 \\ 1 & -1 & 0 \end{bmatrix}, & \text{if } a_{12} \leq 0, \\ \frac{1}{2}h^{-2}\begin{bmatrix} -1 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -1 \end{bmatrix}, & \text{if } a_{12} \geq 0. \end{cases}
\end{aligned} \tag{5.13}
$$

For $a_{12} \leq 0$, we thus obtain

$$\sum_{i,j=1}^{2} a_{ij}\frac{\partial^2}{\partial x_1 \partial x_2}: \quad h^{-2}\begin{bmatrix} 0 & a_{22}-a_{12} & a_{12} \\ a_{11}-a_{12} & 2\,(a_{12}-a_{11}-a_{22}) & a_{11}-a_{12} \\ a_{12} & a_{22}-a_{12} & 0 \end{bmatrix}.$$

Introducing $a_{12}^{+} := \max\{a_{12}, 0\}$ and $a_{12}^{-} := \min\{a_{12}, 0\}$, we then get the seven-point star for any sign of a_{12}:

$$h^{-2}\begin{bmatrix} -a_{12}^{+} & a_{22}+|a_{12}| & a_{12}^{-} \\ a_{11}+|a_{12}| & 2\,(-|a_{12}|-a_{11}-a_{22}) & a_{11}+|a_{12}| \\ a_{12}^{-} & a_{22}+|a_{12}| & -a_{12}^{+} \end{bmatrix}. \tag{5.14}$$

Exercise 5.18. Let L_h be the matrix belonging to the difference formula (5.14) (note that the coefficients in (5.14) depend on the spatial coordinates). Assume (5.4a) and (5.12) hold. Prove that L_h satisfies the conditions of Exercise 4.16a.

The first derivative terms $a_i \partial/\partial x_i$ are replaced by the forward and backward differences, respectively, $a_i \partial_{x_i}^{\pm}$ if $a_i > 0$ or $a_i \le 0$:

$$\sum_{i=1}^{2} a_i \frac{\partial}{\partial x_i} + a: \quad h^{-1}\begin{bmatrix} & a_2^{-} & \\ -a_1^{+} & |a_1|+|a_2| & a_1^{-} \\ & -a_2^{+} & \end{bmatrix} + \begin{bmatrix} & a & \end{bmatrix}, \tag{5.15}$$

where $a_i^{\pm}$ is defined in the same way as $a_{12}^{\pm}$. Because $a \ge 0$, the diagonal element is positive, the others nonpositive. If one adds these terms to (5.14), the resulting matrix L_h satisfies the conditions of Exercise 4.16. Therefore L_h is an M-matrix. ■

It is easy to see that the difference formula (5.14) is of second order of consistency; (5.15), however, contains the one-sided differences so that the total discretisation is only of first order.

The condition (5.12) can be avoided if one allows larger difference stars, which also contain the values $u(x_1 \pm \nu h, x_2 \pm \mu h)$ for fixed $\nu, \mu \in \mathbb{N}$ (but in general $|\nu|, |\mu| > 1$) (cf. Bramble–Hubbard [49]). Layton–Morley [185] point out that with weaker conditions than (5.12) one may still obtain a matrix L_h, which, though not an M-matrix, does have a positive inverse.

To obtain a method of consistency order 2, one must discretise $\sum a_i \partial/\partial x_i$ as in (5.10). The following corollary shows that L_h is also an M-matrix when ha_i is sufficiently small.

Corollary 5.19. In addition to the assumptions of Theorem 5.17 let the following hold:

$$-a_{ii} > |a_{12}| + \frac{h}{2}|a_i| \qquad (i = 1, 2). \tag{5.16}$$

Then the discretisation of $a_i \frac{\partial}{\partial x_i}$ from (5.10) together with (5.14) leads to a seven-point difference method of second order of consistency such that L_h is an M-matrix.

Proof. L_h satisfies (4.21a) and is irreducibly diagonally dominant. ■

Exercise 5.20. The condition $-a_{ii} \geq |a_{12}| + h\,|a_i|\,/2$ instead of (5.16) is not sufficient. Construct a counterexample with $a_{11} = a_{22} = 1$, $a_{12} = 0$, $h = 1/3$, $a_i(x, y)$ variable, and $|a_i| = 6$ so that L_h is singular.

Considerably weaker conditions for the regularity of L_h than in Theorem 5.17 and Corollary 5.19 will be discussed in Section 9.3 (cf. Exercise 9.40, Corollary 11.38).

In general, L_h is not a symmetric matrix. Symmetry of L_h is to be expected only if L is also symmetric: $L = L'$. Here the formally *adjoint differential operator* L' which is associated to L in (5.3), is defined by

$$L' = \sum_{i,j=1}^{n} \left[a_{ij}^{III} \frac{\partial^2}{\partial x_i \partial x_j} + \frac{\partial}{\partial x_i} a_{ij}^{II} \frac{\partial}{\partial x_j} + \frac{\partial^2}{\partial x_i \partial x_j} a_{ij}^{I} \right] - \sum_{i=1}^{n} \left[a_i^{II} \frac{\partial}{\partial x_i} + \frac{\partial}{\partial x_i} a_i^{I} \right] + a. \tag{5.17}$$

It is easy to see that a *symmetric operator* can always be written in the form

$$L = \sum_{i,j=1}^{n} \frac{\partial}{\partial x_i} a_{ij}(\mathbf{x}) \frac{\partial}{\partial x_j} + a(\mathbf{x}), \qquad a_{ij}(\mathbf{x}) = a_{ji}(\mathbf{x}). \tag{5.18}$$

A difference method for this is given for the case $n = 2$ and $a_{12} = 0$ by the five-point star

$$h^{-2} \begin{bmatrix} 0 & a_{22}(x, y + \frac{h}{2}) & 0 \\ a_{11}(x - \frac{h}{2}, y) & \left\{ \begin{matrix} -a_{11}(x - \frac{h}{2}, y) - a_{11}(x + \frac{h}{2}, y) \\ -a_{22}(x, y + \frac{h}{2}) - a_{22}(x, y - \frac{h}{2}) \end{matrix} \right\} & a_{11}(x + \frac{h}{2}, y) \\ 0 & a_{22}(x, y - \frac{h}{2}) & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{5.19}$$

Theorem 5.21. *Let $a_{11}, a_{22} \in C^{2,1}(\overline{\Omega})$. The difference method (5.19) is consistent of order 2. The associated matrix L_h is symmetric. If $a_{ii} < 0$ (ellipticity) and $a \geq 0$, then L_h is a positive-definite M-matrix.*

Proof. (a) For the proof of consistency, expand

$$v(x + \frac{h}{2}) := a_{11}(x + \frac{h}{2}, y)\,[u(x + h, y) + u(x, y)] \qquad (y \text{ fixed})$$

around $x + h/2$, and then expand $v(x + h/2) - v(x - h/2)$ around x.

(b) The symmetry results from $L_{\mathbf{x}\boldsymbol{\xi}} = L_{\boldsymbol{\xi}\mathbf{x}}$ for $\mathbf{x}, \boldsymbol{\xi} \in \Omega_h$.

(c) L_h is irreducibly diagonally dominant so that Criteria 4.18 and 4.32 are applicable. ∎

The mixed term $\frac{\partial}{\partial x} a_{12}(x, y) \frac{\partial}{\partial y} + \frac{\partial}{\partial y} a_{12}(x, y) \frac{\partial}{\partial x}$ can be discretised by

$$\frac{h^{-2}}{2} \begin{bmatrix} -a_{12}^{A} - a_{12}^{B} & a_{12}^{A} + a_{12}^{C} & 0 \\ a_{12}^{B} + a_{12}^{D} & -a_{12}^{C} - a_{12}^{D} - a_{12}^{E} - a_{12}^{F} & a_{12}^{H} + a_{12}^{F} \\ 0 & a_{12}^{E} + a_{12}^{G} & -a_{12}^{H} - a_{12}^{G} \end{bmatrix}, \tag{5.20}$$

where the upper index indicates the evaluation at $\mathbf{x} + \delta h$ with δ specified below:

	A	B	C	D	E	F	G	H
δ	$(-\frac{1}{2}, 1)$	$(-1, \frac{1}{2})$	$(0, \frac{1}{2})$	$(-\frac{1}{2}, 0)$	$(0, -\frac{1}{2})$	$(\frac{1}{2}, 0)$	$(\frac{1}{2}, -1)$	$(1, -\frac{1}{2})$

Another way of writing (5.20) is:

$$\frac{1}{2}\left[\partial_y^+ b(\cdot - \frac{h}{2}, \cdot)\partial_x^- + \partial_x^+ b(\cdot, \cdot - \frac{h}{2})\partial_y^- + \partial_y^- b(\cdot + \frac{h}{2}, \cdot)\partial_x^+ + \partial_x^- b(\cdot, \cdot + \frac{h}{2})\partial_y^+\right],$$

where $b := a_{12}$.

Exercise 5.22. (a) For the coefficients of L in (5.18) let the following hold: $a_{ij} \in C^1(\Omega)$, $a_{11} < 0$, $a_{22} < 0$, $a_{12} = a_{21} \geq 0$, $a_{11} + a_{12} < 0$, $a_{22} + a_{12} < 0$, $a \geq 0$. Show that the difference scheme which is described by (5.19) and (5.20) has consistency order 2 and that for sufficiently small h the associated matrix L_h is a symmetric, irreducibly diagonally dominant and positive-definite M-matrix.
(b) What is the suitable discretisation for the case $a_{12} = a_{21} < 0$?

Exercise 5.23. The difference formula from Theorem 5.17 for the operator L reads

$$a_{11}\partial_x^+\partial_x^- + a_{12}(\partial_x^+\partial_y^+ + \partial_x^-\partial_y^-) + a_{22}\partial_y^+\partial_y^- + a_1\partial_x^+ + a_2\partial_y^+ + a,$$

when $a_{12} \leq 0$, $a_1 \leq 0$, $a_2 \leq 0$. Let the associated matrix be L_h. Then prove that the transposed matrix L_h^T describes a difference method for the adjoint operator L' and also possesses consistency order 1.

In general it is possible to show for regular difference methods that L_h^T is a discretisation of L'. The role of regularity is demonstrated in the next example.

Example 5.24. Let $Lu := -u'' + au'$ in $\Omega = (-1, 1)$ with $a(x) \leq 0$ for $x \leq 0$ and $a(x) \geq 0$ for $x > 0$. According to (5.15) au' is discretised for $x \leq 0$ by $a(x)\partial^+ u(x)$ and for $x > 0$ by $a(x)\partial^- u(x)$. Let the associated matrix be $L_h = L_{h,2} + L_{h,1}$, where $L_{h,2}$ and $L_{h,1}$ correspond to the terms $-u''$ and au' respectively. According to the above, $L_{h,1}^\mathsf{T} v_h$ should be a discretisation of $-(av)'$. But the differences $L_{h,1}^\mathsf{T} v_h$ at $x = 0$ and $x = h$ are

$$\begin{aligned}
\tfrac{1}{h}\left[a(-h)v_h(-h) - a(0)v_h(0) - a(h)v_h(h)\right] &= -(av)' - \tfrac{1}{h}a(0)v_h(0) + \mathcal{O}(h),\\
\tfrac{1}{h}\left[a(0)v_h(0) - a(h)v_h(h) - a(2h)v_h(2h)\right] &= -(av)' + \tfrac{1}{h}a(0)v_h(0) + \mathcal{O}(h),
\end{aligned}$$

thus they are not consistent. Nevertheless, L_h^T is a possible discretisation of $L'v = -v'' - (av)'$, for it can be shown that the error $v_h - R_h v$ has order of magnitude $\mathcal{O}(h)$.

To prove stability one has to show $\|L_h^{-1}\|_\infty \leq \text{const}$. Obviously it is sufficient to prove this inequality for sufficiently small h. In the proof of Theorem 5.12 we used the fact that

$$Lw \geq 1 \quad \text{in } \Omega, \qquad w \geq 0 \quad \text{on } \Gamma$$

for

$$w(\mathbf{x}) := \exp(2R\alpha) - \exp(\alpha(x_1 - z_1 + R)).$$

Let $D_h u_h(\mathbf{x})$ be the difference equations from which $L_h u_h$ results after elimination of the boundary values. We set $w_h := 2R_h w$; i.e., $w_h(\mathbf{x}) = 2w(\mathbf{x})$ for $\mathbf{x} \in \overline{\Omega}_h$. The following holds:

$$D_h w_h = 2\left(D_h R_h - \tilde{R}_h L\right) w + 2\tilde{R}_h L w \geq 2 + 2\left(D_h R_h - \tilde{R}_h L\right) w.$$

Each consistent difference method satisfies $\|(D_h R_h - \tilde{R}_h L)w\|_\infty \to 0$. For sufficiently small h_0 we thus have

$$D_h w_h(\mathbf{x}) \geq 1 \qquad (\mathbf{x} \in \Omega_h,\ h \leq h_0)\,.$$

This inequality agrees for far-boundary points $\mathbf{x} \in \Omega_h$ with $L_h w_h(\mathbf{x}) \geq 1$. For near-boundary points $\mathbf{x} \in \Omega_h$, $D_h w_h(\mathbf{x})$ also contains the sum $\sum_{\boldsymbol{\xi} \in \Gamma_h} L_{\mathbf{x}\boldsymbol{\xi}} w_h(\boldsymbol{\xi})$, which is not contained in $L_h w_h(\mathbf{x})$. For the discretisations from Theorem 5.17, Corollary 5.19, Theorem 5.21, and Exercise 5.22, however, $L_{\mathbf{x}\boldsymbol{\xi}} \leq 0$ ($\mathbf{x} \in \Omega_h,\ \boldsymbol{\xi} \in \Gamma_h$) holds, so that because $w_h \geq 0$ on Γ_h,

$$L_h w_h \geq D_h w_h \geq \mathbf{1} \qquad (h \leq h_0)$$

holds for all grid points. Theorem 4.24 yields the next theorem.

Theorem 5.25. *The discretisations described in Theorem 5.17, Corollary 5.19, Theorem 5.21, and Exercise 5.22 are stable under the conditions posed there, i.e.,* $\|L_h^{-1}\|_\infty \leq \text{const}$ *for all* $h \in H = \{1/n : n \in \mathbb{N}\}$. *According to Theorem 4.48 the methods converge. The order of convergence agrees with the corresponding order of consistency.*

5.1.5 Green's Function

The idea of representing the solution by the Green function can be repeated for the general differential equation (5.1). The Green function (of the first kind) $g(\boldsymbol{\xi}, \mathbf{x})$ is singular at $\boldsymbol{\xi} = \mathbf{x}$ and satisfies

$$\begin{aligned} L_{\mathbf{x}} g(\boldsymbol{\xi}, \mathbf{x}) = 0, \qquad & L'_{\boldsymbol{\xi}} g(\boldsymbol{\xi}, \mathbf{x}) = 0 \qquad \text{for } \mathbf{x}, \boldsymbol{\xi} \in \Omega,\ \mathbf{x} \neq \boldsymbol{\xi}, \\ g(\boldsymbol{\xi}, \mathbf{x}) = 0 \qquad & \text{for } \mathbf{x} \in \Gamma \text{ or } \boldsymbol{\xi} \in \Gamma. \end{aligned}$$

Here, L' is the adjoint differential operator (5.17). If $L \neq L'$, then g is no longer symmetric: $g(\mathbf{x}, \boldsymbol{\xi}) \neq g(\boldsymbol{\xi}, \mathbf{x})$. The singularity of $g(\boldsymbol{\xi}, \mathbf{x})$ at $\boldsymbol{\xi} = \mathbf{x}$ is such that under suitable conditions the solution of (5.1), (5.5) can be represented as

$$u(\mathbf{x}) = \int_\Omega g(\boldsymbol{\xi}, \mathbf{x})\, f(\boldsymbol{\xi})\, \mathrm{d}\boldsymbol{\xi} + \int_\Gamma \varphi(\boldsymbol{\xi})\, B_{\boldsymbol{\xi}} g(\boldsymbol{\xi}, \mathbf{x})\, \mathrm{d}\Gamma_\xi$$

where

$$B = B_\xi = \sum_{i,j=1}^{n} n_j \frac{\partial}{\partial \xi_i} a_{ij}$$

is a so-called *boundary differential operator* (n_j are the components of the normal vector $\mathbf{n} = \mathbf{n}(\boldsymbol{\xi}),\ \boldsymbol{\xi} \in \Gamma$) describing the *conormal derivative*. Only when the principal part of L agrees with $-\Delta$ is B the normal derivative.

In the discrete case the inverse L_h^{-1} again corresponds to the Green function $g_h(\cdot,\cdot)$.

5.2 General Boundary Conditions

5.2.1 Formulating the Boundary-Value Problem

Let the differential equation be given by (5.1). The Dirichlet boundary condition (5.5) can be written in the form

$$Bu = \varphi \qquad \text{on } \Gamma \tag{5.21a}$$

where B is the identity (to be precise: the trace on Γ). In more general settings B can be an operator — a so-called *boundary differential operator* — of order 1:

$$B = \sum_{i=1}^{n} b_i(\mathbf{x}) \frac{\partial}{\partial x_i} + b_0(\mathbf{x}) \qquad (\mathbf{x} \in \Gamma)\,. \tag{5.21b}$$

Introducing the vector $\mathbf{b}(\mathbf{x}) = (b_1(\mathbf{x}), \ldots, b_n(\mathbf{x}))^\mathsf{T}$, we can write Bu in the form

$$Bu = \langle \mathbf{b}(\mathbf{x}), \nabla u(\mathbf{x}) \rangle + b_0(\mathbf{x}) u(\mathbf{x}) \qquad \text{or} \quad B = \mathbf{b}^\mathsf{T} \nabla + b_0.$$

Example 5.26. (a) From $\mathbf{b} = 0$, $b_0(\mathbf{x}) \neq 0$ there results what is known as the *Dirichlet condition* $u = \varphi/b_0$ on Γ, also known as the boundary condition of the *first kind*.

(b) The choice $\mathbf{b} = \mathbf{n}$, $b_0 = 0$ characterises the *Neumann condition*, also called the boundary condition of the *second kind*.

(c) Equation (5.21a) with $\langle \mathbf{b}, \mathbf{n} \rangle \neq 0$, $b_0 \neq 0$ is known as the boundary condition of the *third kind* or the *Robin boundary condition.*[1]

Remark 5.27. The case $\langle \mathbf{b}, \mathbf{n} \rangle = 0$ is excluded in general. For $\langle \mathbf{b}, \mathbf{n} \rangle = 0$, $\mathbf{b}^\mathsf{T} \nabla$ is a tangential derivative. The boundary condition $Bu = \varphi$ is then very similar to a Dirichlet condition. The condition "$u = \varphi$ on Γ" implies "$Bu = \tilde{\varphi} := B\varphi$ on Γ".

[1] Details about the eponym Victor Gustave Robin (1855–1897) can be found in Gustafson–Abe [126, 127].

The normal derivative $B = \partial/\partial n$ (i.e., $\mathbf{b} = \mathbf{n}$) is important in connection with $L = -\Delta$. For the general operator L in (5.1) the so-called *conormal derivative* B with

$$\mathbf{b} = A\mathbf{n}$$

(A as in (5.2)) is of greater importance, as we will see in Section 7.4.

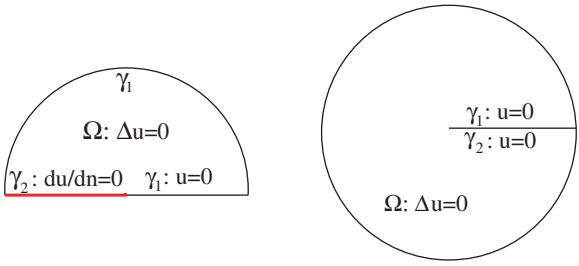


Fig. 5.2 (a) Boundary-value problem with changing boundary condition type, (b) Dirichlet problem in a disk with a cut.

Statements about existence and uniqueness of the solution always depend on L *and* B. We have seen already that for $L = -\Delta$, $B = I$ (Dirichlet condition) uniqueness is guaranteed (cf. Theorem 3.2), while the problem associated to $L = -\Delta$, $B = \mathbf{n}^{\mathsf{T}}\nabla = \partial/\partial n$ is, in general, not solvable (cf. Theorem 3.28).

The coefficients of B depend on position. Of course, $B(\mathbf{x}) = 0$, i.e., $b(\mathbf{x}) = 0$ and $b_0(\mathbf{x}) = 0$, must not occur for any $\mathbf{x} \in \Gamma$. But it is possible that $\mathbf{b}(\mathbf{x}) = 0$ (and $b_0 \neq 0$) in $\gamma \subset \Gamma$ and $b(\mathbf{x}) \neq 0$ in $\Gamma\backslash\gamma$. Then there is a Dirichlet condition $u = \varphi/b_0$ on the piece γ and a boundary condition of first order on the remaining boundary piece $\Gamma\backslash\gamma$. At the points of contact between γ and $\Gamma\backslash\gamma$ the solution generally is not smooth (it has singularities in the derivatives).

Example 5.28. Let Ω be the upper semicircle around $x = y = 0$ with radius 1. Let the differential equation and boundary conditions be given as in Figure 5.2a. The boundary condition changes its order at $x = y = 0$. The solution in polar coordinates reads: $u = \sqrt{r}\sin(\varphi/2)$ (cf. (2.3a)). Check that $u_x = \mathcal{O}(1/\sqrt{r})$ and $u_y = \mathcal{O}(1/\sqrt{r})$.

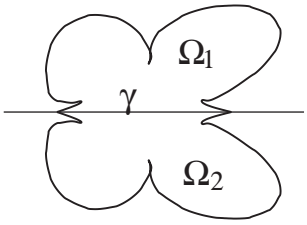


Fig. 5.3 A domain symmetric with respect to γ.

The same singularity as in Example 5.28 occurs in the problem described in Figure 5.2b; the solution is also $r^{1/2}\sin(\varphi/2)$. This Dirichlet problem and Example 5.28 are closely connected with each other.

Example 5.29. Let $\Omega = \Omega_1 \cup \Omega_2 \cup \gamma$ with $\Omega_1 \cap \Omega_2 = \emptyset$ be as in Figure 5.3: let the reflection of Ω_1 in γ result in Ω_2. If one seeks a solution of

$$Lu = f \text{ in } \Omega, \qquad Bu = \varphi \text{ on } \partial\Omega,$$

and if together with u the function $\bar{u}$ reflected in γ is also a solution, one expects $u = \bar{u}$. This solution then satisfies

$$Lu = f \text{ in } \Omega_1, \quad Bu = \varphi \text{ on } \partial\Omega_1\backslash\gamma, \quad \partial u/\partial n = 0 \text{ on } \gamma.$$

Exercise 5.30. (a) Let Ω be symmetric with respect to the x_1-axis as in Figure 5.3. Which conditions must the coefficients of L in (5.1) satisfy so that $Lu = 0$ implies that also $u^-(x_1, x_2) := u(x_1, -x_2)$ is a solution of $Lu = 0$?

(b) Let the conditions of part (a) hold and assume that u is a solution of $Lu = 0$ with Dirichlet values $u = \varphi$ on Γ which satisfy the symmetry condition $\varphi(x_1, x_2) = \varphi(x_1, -x_2)$. Which conditions guarantee that u is symmetric?

While the boundary condition $Bu = \varphi$ on $\partial\Omega$ may be of physical origin, Example 5.29 shows that a Neumann condition may also have a geometric basis. Another geometrically justified boundary condition is the following. Let Ω be given as in Figure 5.4: γ_1 and γ_2 are parts of $\Gamma = \partial\Omega$ with

$$\gamma_i = \{(x_i, y) : y_1 \le y \le y_2\} \qquad (i = 1, 2).$$

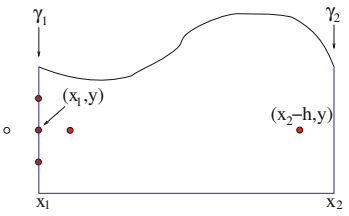


Fig. 5.4 Discretisation for periodic boundary conditions.

Then, in addition to $Bu = \varphi$ on $\Gamma\backslash(\gamma_1 \cup \gamma_2)$, we can require the *periodic boundary condition*

$$u(x_1, y) = u(x_2, y), \qquad u_x(x_1, y) = u_x(x_2, y) \qquad \text{for } y_1 \le y \le y_2 \quad (5.22)$$

on γ_1 and γ_2. The solution is periodically continuable in the x direction (with period $x_2 - x_1$). The origin of periodic boundary conditions is discussed in Example 5.31.

Example 5.31. (a) Let Ω' be an annulus which is described by the polar coordinates

$$r \in (r_1, r_2), \qquad \varphi \in [0, 2\pi).$$

Transformation of the differential equation to polar coordinates gives as the image domain the rectangle $\Omega = (r_1, r_2) \times (0, 2\pi)$. The original boundary conditions on Ω' become boundary conditions at the upper and lower boundaries while equation (5.22) describes the periodicity of the angular variable $x \in (0, 2\pi)$.

(b) Instead of on $\Omega' \subset \mathbb{R}^2$, one can also define a boundary-value problem on a part of the 2-dimensional surface of a 3-dimensional body. If Ω' lies, for example, on the surface of the cylinder

$$\{\boldsymbol{\xi} \in \mathbb{R}^3 : \xi_1^2 + \xi_2^2 \le r^2,\ \xi_3 \in \mathbb{R}\},$$

the unfolding of Ω' results in a domain Ω as in Figure 5.4. Here too, equation (5.22) is justified by the fact that x plays the role of an angle variable.

(c) If one seeks solutions in the unbounded strip $\Omega' = \mathbb{R} \times (y_1, y_2)$ one can instead look for periodic solutions in $\Omega = (0, 2\pi) \times (y_1, y_2)$ with the boundary condition (5.22), since those (after periodic continuation) are also solutions of the original problem. Solutions with periodic boundary conditions in the x *and* y directions can even be continued onto $\Omega' = \mathbb{R}^2$.

5.2.2 *Difference Methods for General Boundary Conditions*

The boundary conditions posing the least difficulties are the periodic boundary conditions. Define Ω_h as the set of grid points in Ω and on γ_1 (but not on γ_2; cf. Figure 5.4). The difference equation at a grid point $(x_1, y) \in \gamma_1$ has a left neighbour $(x_1 - h, y)$ outside $\overline{\Omega}$. Replace $u_h(x_1 - h, y)$ in the difference equation by $u_h(x_2 - h, y)$. By doing so we have transferred the periodicity onto the difference solution without explicitly discretising equation (5.22). Of course, the step size must be chosen so that $x_2 - x_1$ is a multiple of h.

Exercise 5.32. For periodic boundary conditions the structure of the matrix L_h changes. Let $L = -\Delta$ in the square $\Omega = (0,1) \times (0,1)$. Assume Dirichlet conditions for the upper and lower boundaries, and for the lateral boundaries the periodicity condition (5.22). In analogy to (4.16) exhibit the form of the matrix L_h for a lexicographical arrangement of the grid points.

In Section 4.7 we described the discretisation of the boundary condition $B = \mathbf{n}^\mathsf{T}\nabla = \partial/\partial n$; for the case that Ω is a rectangle and Γ coincides with the grid. In the same situation one can discretise the general boundary condition (5.21a,b) as follows. Let

$$(\bar{x}, \bar{y}) \in \Gamma_h = \overline{\Omega}_h \cap \Gamma$$

be a grid point on Γ. The point of intersection $(x + h, \tilde{y})$ drawn in Figure 5.5 is given by

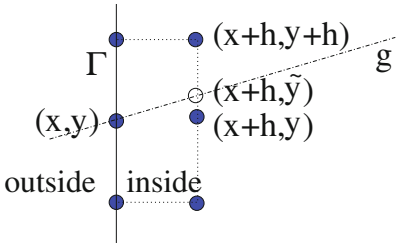


Fig. 5.5 Discretisation of $Bu = \varphi$; g is the line $b_1(x - \xi) = b_2(y - \eta)$.

$$\tilde{y} = y + h\, b_2(x,y)/b_1(x,y).$$

Note that $b_1 \neq 0$, since $\mathbf{b}^\mathsf{T}\nabla$ is not to be a tangential derivative. For $u \in C^2(\overline{\Omega})$ the directional derivative $\mathbf{b}^\mathsf{T}\nabla u = \langle \mathbf{b}, \nabla u \rangle$ in (x, y) is approximated by

$$\frac{u(x,y) - u(x+h,\tilde{y})}{\sqrt{h^2 + \delta^2}} = \big(\mathbf{b}^\mathsf{T}\nabla u\big)(x,y) + \mathcal{O}(h) = \varphi(x,y) - b_0(x,y)u(x,y) + \mathcal{O}(h), \tag{5.23a}$$

where $\delta = \tilde{y} - y = \dfrac{h b_2(x,y)}{b_1(x,y)}$. On the other hand, one obtains $u(x + h, \tilde{y})$ under the same conditions by interpolation up to order $\mathcal{O}(h^2)$:

$$u(x+h,\tilde{y}) = \frac{\delta}{h}u(x+h, y+h) + \frac{h-\delta}{h}u(x+h,y) + \mathcal{O}(h^2). \tag{5.23b}$$

The combination of (5.23a) and (5.23b) leads to the difference formula

$$\left(B_h u_h\right)(x,y) := c_0 u_h(x,y) - c_1 u_h(x+h,y) - c_2 u_h(x+h,y+h) = \varphi(x,y) \tag{5.23c}$$

with

$$c_0 = \frac{1}{\rho} + b_0(x,y), \quad c_1 = \frac{\delta}{h\rho}, \quad c_2 = \frac{h-\delta}{h\rho}, \quad \rho := \sqrt{h^2+\delta^2}.$$

Lemma 5.33. *Let R_h be the restriction to the grid points $\overline{\Omega}_h$. For a function $u \in C^2(\overline{\Omega})$ one has*

$$B_h(u_h - R_h u)(x,y) = \varphi(x,y) - B_h R_h u(x,y) = \mathcal{O}(h\,\|u\|_{C^2(\overline{\Omega})}),$$

i.e., the discretisation of the boundary condition has consistency order 1.

We assume that the difference equations

$$\sum_{\boldsymbol{\xi}\in\overline{\Omega}_h} L_{\mathbf{x}\boldsymbol{\xi}} u_h(\boldsymbol{\xi}) = f_h(\mathbf{x})$$

(before the elimination of the values $u_h(\boldsymbol{\xi})$, $\boldsymbol{\xi} \in \Gamma_h$) satisfy the inequalities

$$L_{\mathbf{x}\boldsymbol{\xi}} \le 0 \quad (\mathbf{x} \ne \boldsymbol{\xi}), \qquad L_{\mathbf{xx}} \ge -\sum_{\boldsymbol{\xi}\ne\mathbf{x}} L_{\mathbf{x}\boldsymbol{\xi}} \qquad \left(\mathbf{x}\in\Omega_h,\ \boldsymbol{\xi}\in\overline{\Omega}_h\right). \tag{5.24a}$$

Here $L_{\mathbf{xx}} > 0$ and $L_{\mathbf{x}\boldsymbol{\xi}} \le 0$ correspond to the sign condition (4.21a). The second inequality in (5.24a) agrees with (4.26b). The corresponding inequalities for (5.23c) read:

$$c_1 \ge 0, \qquad c_2 \ge 0, \qquad c_0 \ge c_1 + c_2. \tag{5.24b}$$

Let the difference equations in Ω_h, and the boundary equations (5.23c) given for $(x,y) \in \Gamma_h$ be combined into $A_h u_h = g_h$. (5.24a,b) implies

$$a_{\mathbf{x}\boldsymbol{\xi}} \le 0 \quad (\mathbf{x},\boldsymbol{\xi}\in\overline{\Omega}_h), \qquad a_{\mathbf{xx}} \ge -\sum_{\boldsymbol{\xi}\ne\mathbf{x}} a_{\mathbf{x}\boldsymbol{\xi}} \quad (\mathbf{x}\in\overline{\Omega}_h).$$

Therefore, A_h is an M-matrix if A_h is irreducible and (4.26a) holds for at least one $\mathbf{x} \in \overline{\Omega}_h$.

The above considerations explain why the boundary discretisation should satisfy the conditions (5.24b). (5.24b) holds for the case $b_0 = 0$ if and only if $b_2/b_1 \in [0,1]$. If $b_2/b_1 \in [-1,0]$, (5.24b) can be satisfied if one interpolates between y and $y-h$ (instead of y and $y+h$). If, however, $|b_2| > |b_1|$, i.e., if the tangential component is greater, one could interpolate between $(x+h,y)$ and $(x+h, y \pm kh)$, where $|b_2/b_1| \le k$. For the case $b_2/b_1 \ge 1$, however, it is more practical to interpolate at the point $(\tilde{x}, y+h)$ between $(x,y+h)$ and $(x+h,y+h)$. Generally one should choose as interpolation point the point of intersection of the line with the dotted straight line in Figure 5.5.

In the preceding discussion we started with the case $\Omega = (0,1)\times(0,1)$. If Ω is a general region, two discretisation techniques offer themselves (cf. Sections 4.8.1 and 4.8.2).

First option for discretisation. Let Ω_h and Γ_h be chosen as in §4.8.1. At near-boundary points the differential equation is approximated by a difference scheme which in the case of $L = -\Delta$ corresponds to the Shortley–Weller method. For this one needs the values of u_h at the boundary points $\boldsymbol{\xi} \in \Gamma_h$. As in Figure 5.6 let

$$(x, y) = (\hat{x} - s_\ell h, \hat{y}) \in \Gamma_h$$

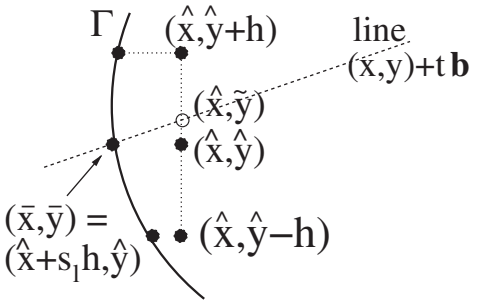


Fig. 5.6 First boundary discretisation.

be a boundary point. Again, we can set up an equation analogous to (5.23a–c). In Figure 5.6 $(\hat{x}, \tilde{y})$ has the same position as $(x + h, \tilde{y})$ in Figure 5.5. In general one has to use the point of intersection of the line which is also presented by $(x, y) + t\mathbf{b}$ $(t \in \mathbb{R})$, and the dotted straight line in Figure 5.6.

Second option for discretisation. Let $\overline{\Omega}_h$ be the grid Ω_h, used above. Now let Ω_h consist of all interior points of $\overline{\Omega}_h$. For all $(x, y) \in \Omega_h$, difference equations (with equidistant step size) are declared which involve $\{u_h(\boldsymbol{\xi}) : \boldsymbol{\xi} \in \overline{\Omega}_h\}$. For each point

$$(\hat{x}, \hat{y}) \in \Gamma_h := \overline{\Omega}_h \backslash \Omega_h$$

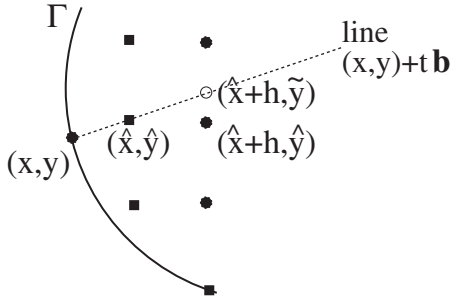


Fig. 5.7 Second boundary discretisation.

a boundary discretisation must be found (cf. Figure 5.7). Equation (5.23a) can be set up with $(\hat{x}, \hat{y})$ and $(\hat{x} + h, \tilde{y})$ instead of (x, y) and $(x + h, \tilde{y})$. The arguments of the coefficients b_1 and b_2 are (x, y). This point results implicitly from

$$b_2(x, y)\ (x - \hat{x}) = b_1(x, y)\ (y - \hat{y})\,, \qquad (x, y) \in \Gamma. \tag{5.25}$$

Remark 5.34. (a) At near-boundary points, the first discretisation requires a difference scheme for $Lu = f$ with nonequidistant step sizes. The second discretisation requires an approximation of the nonlinear problem (5.25). From a programming viewpoint both procedures are undesirable because of the case distinctions.
(b) If the vector $\mathbf{b}$ from B approaches the tangential direction, both methods fail since the straight line no longer intersects the dotted straight line from Figures 5.6 and 5.7. Here, the second discretisation fails earlier than the first one.

Another option for avoiding the difficulties described above consists in using variational difference equations, at least near the boundary (cf. Remark 8.73).

Occasionally it is also possible to simplify the boundary conditions by using coordinate transformations. Let $Bu = \varphi$ be prescribed on $\gamma \subset \Gamma$. If one finds a transformation $(x, y) \leftrightarrow (\xi, \eta)$ such that the equations $\xi = 0$ and $b_2 x_\xi = b_1 y_\xi$ are satisfied on γ, one obtains for the transformed problem

$$Bu \equiv \sigma \frac{\partial u}{\partial n} + b_0 u = \varphi$$

on a vertical boundary piece ($\partial/\partial n = -\partial/\partial\xi$). The discretisation may be carried out as described in Section 4.7.

The same reasoning as in the preceding sections proves stability and convergence.

Remark 5.35. Let the difference method $D_h u_h = f_h$ satisfy the conditions of Theorem 5.25 so that

$$D_h \tilde{w}_h \geq \mathbf{1} \qquad \text{for } \tilde{w}_h := K\mathbf{1} + w_h$$

($K > 0$ is constant; w_h is as in the proof of Theorem 5.25). Let the boundary discretisation (5.23c) satisfy

$$c_0 - c_1 - c_2 = b_0(\mathbf{x}) \geq \varepsilon > 0 \qquad \text{for all } \mathbf{x} \in \Gamma_h\,.$$

For sufficiently large K, one then obtains $(B_h \tilde{w}_h)(\mathbf{x}) \geq 1$ for $\mathbf{x} \in \Gamma_h$. Hence it follows that $A_h \tilde{w}_h \geq \mathbf{1}$, where A_h is the M-matrix defined immediately after equation (5.24b). Since

$$\|A_h^{-1}\|_\infty \leq \text{const}$$

(cf. Theorem 4.3.16), stability has been proved. The order of convergence is the minimum of 1 (order of consistency of B_h; cf. Lemma 5.33) and the order of consistency of $D_h u_h = f_h$.

In general it is not possible to approximate $\mathbf{b}^\mathsf{T}\nabla$ by symmetric differences. To construct a discretisation of $Bu = \varphi$ with consistency order 2 despite this fact, set:

$$B_h u_h(x, y) := c_0 u_h(x, y) - \sum_{i=1}^{3} c_i u_h(x_i, y_i) = g(x, y), \qquad (x, y) \in \Gamma_h,$$

where (x_i, y_i) are three points in Ω_h. If, in order to remove the first two terms in the Taylor expansion, one uses the differential equation $Lu = f$ in (x, y) and the tangential derivative of $Bu = \varphi$, one obtains a discretisation of order 2. For the special case $L = \Delta$, $B = \partial/\partial n + b_0$, Bramble–Hubbard [50] proves that the three points (x_i, y_i) may be chosen such that the inequalities $c_i \geq 0$, $c_0 \geq c_1 + c_2 + c_3$ hold and guarantee stability. However, in general, the (x_i, y_i) do not lie in the direct neighbourhood of $(x, y) \in \Gamma_h$. Nonetheless, the construction of the discretisation seems to be too complicated to be recommended for practical purposes.

5.3 Boundary Problems of Higher Order

5.3.1 The Biharmonic Differential Equation

In elastomechanics the free vibration of rods leads to (ordinary) differential equations of second order if longitudinal vibrations (= compression waves) or torsional vibrations are involved. By contrast, transversal vibrations (= bending waves) result in an equation of fourth order. Correspondingly, the bending vibration of a plate leads to a partial differential equation of fourth order. This is the *biharmonic equation* (or *plate equation*)

$$\Delta^2 u = f \qquad \text{in } \Omega \tag{5.26}$$

($\Delta^2 = \partial^4/\partial x^4 + 2\partial^4/\partial x^2\partial y^2 + \partial^4/\partial y^4$). Here u describes the deflection of the plate perpendicular to the surface. If the plate is firmly clamped at the edge, one obtains the boundary conditions

$$u = \varphi_1 \quad \text{and} \quad \frac{\partial u}{\partial n} = \varphi_2 \qquad \text{on } \Gamma \tag{5.27}$$

with $\varphi_1 = \varphi_2 = 0$. A biharmonic problem (5.26), (5.27) also results from a transformation of the Stokes equations in $\Omega \subset \mathbb{R}^2$ (cf. Remark 12.5).

The differential equation (5.26) may be combined with other boundary values than (5.27). An example is:

$$u = \varphi_1 \quad \text{and} \quad \Delta u = \varphi_2 \qquad \text{on } \Gamma \tag{5.28}$$

(simply supported plate). Other examples can be found in (7.30b,e).

Exercise 5.36. Show that if one solves the Poisson equations $-\Delta v = f$ in Ω, $v = -\varphi_2$ on Γ and, subsequently, $-\Delta u = v$ in Ω and $u = \varphi_1$ on Γ, then u is the solution of the boundary-value problem (5.26), (5.28). Why can Problem (5.26), (5.27) not be handled likewise?

Remark 5.37. The solutions of $\Delta^2 u = 0$ do not satisfy a maximum-minimum principle (counterexample: $u = x^2 + y^2$ in $\Omega = K_R(0)$).

5.3.2 General Linear Differential Equations of Order $2m$

The partial derivative D^α ($\alpha \in \mathbb{N}_0^n$: multi-index) of order $|\alpha| = \alpha_1 + \ldots + \alpha_n$ is defined in (3.11b). A differential operator of order $2m$ has the form[2]

$$L = \sum_{|\alpha| \le 2m} a_\alpha(\mathbf{x}) D^\alpha \qquad (\mathbf{x} \in \Omega) \tag{5.29}$$

[2] The notation $\sum_{|\alpha|\le 2m}$ (or $\sum_{|\alpha|=2m}$ etc.) means the summation over all multiindices $\alpha \in \mathbb{N}_0^d$ with the indicated side condition.

and defines the differential equation of order $2m$:

$$Lu = f \qquad \text{in } \Omega.$$

Ellipticity has been explained thus far only for equations of second order (cf. Definition 1.14).

Definition 5.38. The differential operator L (with real-valued coefficients a_α) is said to be *elliptic* (of order $2m$) at $\mathbf{x} \in \Omega$ if

$$\sum_{|\alpha|=2m} a_\alpha(\mathbf{x})\boldsymbol{\xi}^\alpha \neq 0 \qquad \text{for all } 0 \neq \boldsymbol{\xi} \in \mathbb{R}^n. \tag{5.30a}$$

Here, $\boldsymbol{\xi}^\alpha$ is an abbreviation for the polynomial $\xi_1^{\alpha_1}\xi_2^{\alpha_2}\dots\xi_n^{\alpha_n}$ of degree $|\alpha|$. We set $P(\mathbf{x},\boldsymbol{\xi}) := \sum_{|\alpha|=2m} a_\alpha(\mathbf{x})\boldsymbol{\xi}^\alpha$. Evidently (5.30a) is equivalent to $P(\mathbf{x},\boldsymbol{\xi}) \neq 0$ for all $\boldsymbol{\xi} \in \mathbb{R}^n$ with $|\boldsymbol{\xi}| = 1$. For reasons of continuity either $P(\mathbf{x},\boldsymbol{\xi}) > 0$ or $P(\mathbf{x},\boldsymbol{\xi}) < 0$ must hold. Without loss of generality, we may assume that $P(\mathbf{x},\boldsymbol{\xi}) > 0$; otherwise we scale with the factor -1 (changing from $Lu(\mathbf{x}) = f(\mathbf{x})$ to $-Lu(\mathbf{x}) = -f(\mathbf{x})$). Since the set $\{\boldsymbol{\xi} \in \mathbb{R}^n : |\boldsymbol{\xi}| = 1\}$ is compact it follows that $c(\mathbf{x}) := \min\{P(\mathbf{x},\boldsymbol{\xi}) : |\boldsymbol{\xi}| = 1\} > 0$ and this justifies the formulation of (5.30a) as

$$\sum_{|\alpha|=2m} a_\alpha(\mathbf{x})\boldsymbol{\xi}^\alpha \geq c(\mathbf{x})\,|\boldsymbol{\xi}|^{2m} \qquad \text{for all } \boldsymbol{\xi} \in \mathbb{R}^n \text{ with } c(\mathbf{x}) > 0. \tag{5.30b}$$

Definition 5.39. The differential operator L is said to be *uniformly elliptic* in Ω if

$$\inf\{c(\mathbf{x}) : \mathbf{x} \in \Omega\} > 0 \quad \text{for } c(\mathbf{x}) \text{ from (5.30b).}$$

Exercise 5.40. (a) Translate L from (5.1) into the notation of (5.29). What are the coefficients a_α for $L = \Delta^2$?

(b) Prove that the biharmonic operator Δ^2 is uniformly elliptic.

(c) Let a_α be real-valued. Why are there no elliptic operators L of odd order?

(d) If the coefficients a_α are sufficiently smooth one can write L from (5.29) in the form $L = \sum_{|\alpha|\leq m}\sum_{|\beta|\leq m}(-1)^{|\beta|}D^\beta a_{\alpha\beta}(\mathbf{x})D^\alpha$ (cf. (5.1) and (5.3)).

For $m = 1$ (equation of order 2) we have used one boundary condition; for the biharmonic equation ($m = 2$) two boundary conditions occur. In general one needs m boundary conditions

$$(B_j u)(\mathbf{x}) := \sum_{|\alpha|\leq m_j} b_{j\alpha}(\mathbf{x})D^\alpha u(\mathbf{x}) = \varphi_j(\mathbf{x}) \qquad \text{for } \mathbf{x} \in \Gamma,\ 1 \leq j \leq m,$$

with boundary differential operators B_j of order $0 \leq m_j < 2m$.

Remark 5.41. The boundary operators B_j cannot be chosen arbitrarily, but must be independent of each other (they must form a so-called *normal system*; cf. Lions–Magenes [194, page 113] and Wloka [308, Definition 14.1]). In particular the orders m_j must differ pairwise.

We have a *Dirichlet boundary condition* if

$$B_j = \partial^{j-1}/\partial n^{j-1} = (\partial/\partial n)^{j-1} \quad \text{for } j = 1, \ldots, m \quad \text{(cf. (5.27))}.$$

The representation of the solution using the Green function will not be discussed at this point. However, it is remarkable that the Green function (and in particular the singularity function) for L is continuous whenever $2m > n$. For $L = \Delta^2$ in $\mathbb{R}^2$, for example, the singularity function is $\frac{1}{8\pi} |\mathbf{x} - \mathbf{y}|^2 \log |\mathbf{x} - \mathbf{y}|$ (cf. Wloka [308, Exercise 21.9]).

Theorem 2.29 raises the question regarding the continuous dependence of the solution on variations of the domain Ω. The given result applies to differential equations of second order. Boundary-value problems of fourth order require a smoother variation: a suitable metric must take into account the curvature of the boundary. In the case of the (weaker) Hausdorff norm, the so-called *Babuška paradox* is a prominent counterexample. Consider the biharmonic equation (5.26) with the boundary condition $u = \Delta u = 0$ on the regular n-gon Ω_n with corners on the unit circle Ω. Obviously Ω_n converges to Ω with respect to the Hausdorff norm. However, the solutions u_n in Ω_n do not converge to the solution u in Ω (cf. Babuška [14] and Rannacher [231]).

5.3.3 Discretisation of the Biharmonic Differential Equation

The simplest difference formula for $L = \Delta^2$ is the 13-point difference method

$$D_h = \Delta_h^2 = h^{-4} \begin{bmatrix} & & 1 & & \\ & 2 & -8 & 2 & \\ 1 & -8 & 20 & -8 & 1 \\ & 2 & -8 & 2 & \\ & & 1 & & \end{bmatrix}, \tag{5.31}$$

which can be represented as the square of the five-point star Δ_h from (4.10). Let the grid $\Omega_h \subset \Omega$ and the boundaries $\Gamma_h \subset \Gamma$, $\gamma_h \subset \mathbb{R}^2 \backslash \overline{\Omega}$ be defined as in Figure 5.8 ($\bullet$ for inner grid points, o for outer points, $\times$ for boundary points). The difference equation

$$D_h u_h = f_h \qquad \text{in } \Omega_h \tag{5.32a}$$

requires the values of u_h in $\overline{\Omega}_h := \Omega_h \cup \Gamma_h \cup \gamma_h$. These are determined by the boundary conditions

$$u_h(\mathbf{x}) = \varphi_1(\mathbf{x}) \qquad \text{for } \mathbf{x} \in \Gamma_h, \tag{5.32b}$$

$$\partial_n^0 u_h(\mathbf{x}) := \frac{u_h(\mathbf{x} + h\mathbf{n}) - u_h(\mathbf{x} - h\mathbf{n})}{2h} = \varphi_2(\mathbf{x}) \quad \text{for } \mathbf{x} \in \Gamma_h,\ \mathbf{x} + h\mathbf{n} \in \gamma_h \tag{5.32c}$$

(cf. (5.27)). Note that two ($m = 2$) boundary layers occur, Γ_h and γ_h, and that two boundary conditions are given.[3]

Remark 5.42. If with the aid of (5.32b,c) one eliminates in equation (5.32a) the unknowns $\{u_h(\boldsymbol{\xi}) : \boldsymbol{\xi} \in \Gamma_h \cup \gamma_h\}$, one obtains a system of equations $L_h u_h = q_h$ for $\{u_h(\boldsymbol{\xi}) : \boldsymbol{\xi} \in \Omega_h\}$. L_h is not an M-matrix since the sign condition (4.21a) is violated. But L_h^{-1} also is in general not $\geq O$ (cf. (4.21b)).

Thus the methods of the proof in Chapter 4 cannot be used here. Instead we will prove the analogue of Theorem 4.45.

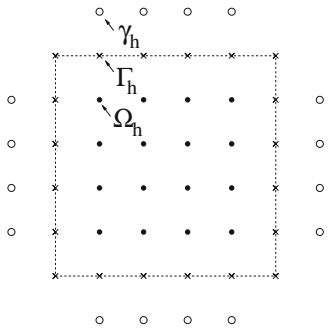


Fig. 5.8 Grid for the biharmonic equation in $\Omega = (0,1)^2$.

Theorem 5.43. *Let u_h be the discrete solution of the biharmonic equation in the square $\Omega = (0,1)\times(0,1)$ defined by (5.32a–c). The matrix L_h described in Remark 5.42 is symmetric and positive definite. It satisfies*

$$\|L_h\|_2 \leq 64h^{-4}, \quad \|L_h^{-1}\|_2 \leq \tfrac{1}{256}. \quad (5.33a)$$

If $\varphi_1 = 0$ on Γ_h (cf. (5.32b)), then u_h can be estimated by

$$\|u_h\|_{\Omega_h} \leq \frac{1}{256}\,\|f_h\|_{\Omega_h} + \frac{h^{-1/2}}{16\sqrt{2}}\,|\varphi_2|_{\Gamma_h} \quad (5.33b)$$

where the norms are defined as follows:

$$\|u_h\|_{\Omega_h} := \sqrt{h^2 \sum_{\mathbf{x}\in\Omega_h} |u_h(\mathbf{x})|^2}, \qquad |\varphi|_{\Gamma_h} := \sqrt{h \sum_{\mathbf{x}\in\Gamma_h} |\varphi(\mathbf{x})|^2}. \quad (5.33c)$$

Let the sum $\sum_{\mathbf{x}\in\Gamma_h}$ in $|\cdot|_{\Gamma_h}$ contain the points of Γ_h without the corner points (cf. Footnote 3).

Proof. We limit ourselves to the most important part, the inequality (5.33b). In the infinite grid $Q_h := \{(x,y) \in \mathbb{R}^2 : x/h, y/h \in \mathbb{N}\}$ we set

$$\bar{u}_h(\mathbf{x}) := u_h(\mathbf{x}) \quad \text{for } \mathbf{x} \in \Omega_h, \qquad \bar{u}_h(\mathbf{x}) := 0 \quad \text{for } \mathbf{x} \in Q_h \backslash \Omega_h.$$

After partial summation one obtains

$$\sum_{\mathbf{x}\in\Omega_h} u_h(\mathbf{x}) f_h(\mathbf{x}) = \sum_{\mathbf{x}\in Q_h} \bar{u}_h(\mathbf{x}) \Delta_h^2 u_h(\mathbf{x}) = \sum_{\mathbf{x}\in Q_h} \Delta_h \bar{u}_h(\mathbf{x}) \Delta_h u_h(\mathbf{x})$$

(First consider the identities

[3] The definition of γ_h does not involve the outer neighbours of the four corner points. Alternatively, one can add the corner points and require Equation (5.32c). As in Equation (4.72b) we assign to the corner points two normal directions each. Correspondingly, to each corner point belong two equations (5.32c).

$$\sum v(x)\partial^+\partial^- u(x) = -\sum(\partial^- v)(x)(\partial^- u)(x) = \sum(\partial^+\partial^- v)(x)u(x)$$

in the one-dimensional case.) From the definition of $\bar{u}_h$ it follows that $\Delta\bar{u}_h = 0$ on $Q_h\backslash(\Omega_h \cup \Gamma_h)$. On Γ_h one has $u_h = \bar{u}_h = 0$ and

$$\bar{u}_h(\mathbf{x}-h\mathbf{n}) = u_h(\mathbf{x}-h\mathbf{n}), \quad \bar{u}_h(\mathbf{x}+h\mathbf{n}) = 0, \quad u_h(\mathbf{x}+h\mathbf{n}) = u_h(\mathbf{x}-h\mathbf{n}) + 2h\varphi_2(\mathbf{x})$$

for $\mathbf{x} \in \Gamma_h$. If one sets $\partial_n^- u_h(\mathbf{x}) := [u_h(\mathbf{x}) - u_h(\mathbf{x}-h\mathbf{n})]/h = -u_h(\mathbf{x}-h\mathbf{n})/h$ at $\mathbf{x} \in \Gamma_h$, then

$$\Delta_h\bar{u}_h(\mathbf{x}) = -\frac{1}{h}\partial_n^- u_h(\mathbf{x}), \quad \Delta_h u_h(\mathbf{x}) = -\frac{2}{h}\partial_n^- u_h(\mathbf{x}) + \frac{2}{h}\varphi_2(\mathbf{x}) \quad \text{in } \mathbf{x} \in \Gamma_h$$

implies that

$$(u_h, L_h u_h)_{\Omega_h} = (u_h, f_h)_{\Omega_h} = \|\Delta_h u_h\|^2_{\Omega_h} + \frac{2}{h}\left|\partial_n^- u_h\right|^2_{\Gamma_h} - \frac{2}{h}(\partial_n^- u_h, \varphi_2)_{\Gamma_h},$$

where $(\cdot,\cdot)_{\Gamma_h}$ and $(\cdot,\cdot)_{\Omega_h}$ are the scalar products belonging to the norms (5.33c). According to

$$ab \le a^2 + b^2/4 \qquad (a, b \in \mathbb{R}) \tag{5.34}$$

one estimates

$$(\partial_n^- u_h, \varphi_2)_{\Gamma_h} \le \left|\partial_n^- u_h\right|_{\Gamma_h} |\varphi_2|_{\Gamma_h} \le \left|\partial_n^- u_h\right|^2_{\Gamma_h} + |\varphi_2|^2_{\Gamma_h}/4.$$

Thus one obtains

$$\begin{aligned}\|\Delta_h u_h\|^2_{\Omega_h} + 2h^{-1}\left|\partial_n^- u_h\right|^2_{\Gamma_h} &\le (u_h, f_h)_{\Omega_h} + 2h^{-1}(\partial_n^- u_h, \varphi_2)_{\Gamma_h} \\ &\le \|u_h\|_{\Omega_h}\|f_h\|_{\Omega_h} + \frac{2}{h}\left|\partial_n^- u_h\right|^2_{\Gamma_h} + \frac{1}{2h}|\varphi_2|^2_{\Gamma_h}.\end{aligned} \tag{5.35}$$

Theorem 4.45 shows that $\|u_h\|_{\Omega_h} \le \|\Delta_h u_h\|_{\Omega_h}/16$. Thus for $\varphi_2 = 0$, (5.35) implies $\|u_h\|^2_{\Omega_h} \le 16^{-2}\|\Delta_h u_h\|^2_{\Omega_h} \le 16^{-2}\|u_h\|_{\Omega_h}\|f_h\|_{\Omega_h}$, so that

$$\|u_h\|_{\Omega_h} \le 16^{-2}\|f_h\|_{\Omega_h}. \tag{5.36a}$$

For $f_h = 0$, (5.35) implies the inequality $\|\Delta_h u_h\|^2_{\Omega_h} \le (2h)^{-1}|\varphi_2|^2_{\Gamma_h}$, therefore

$$\|u_h\|_{\Omega_h} \le \frac{1}{16}\|\Delta_h u_h\|_{\Omega_h} \le \frac{1}{16}(2h)^{-1/2}|\varphi_2|_{\Gamma_h}. \tag{5.36b}$$

In the general case we write u_h as $u_h^I + u_h^{II}$, where $\varphi_2^I = 0$, $\varphi_2^{II} = \varphi_2$, $f_h^I = f_h$, and $f_h^{II} = 0$. $\|u_h\|_{\Omega_h} \le \|u_h^I\|_{\Omega_h} + \|u_h^{II}\|_{\Omega_h}$ together with (5.36a,b) implies (5.33b). ■

Theorem 5.44 (convergence). *Let the solution of the biharmonic boundary-value problem (5.26), (5.27) in $\Omega = (0,1)\times(0,1)$ satisfy $u \in C^{5,1}(\overline{\Omega})$. Then the discrete solution of equations (5.32a–c) is convergent (at least) of order $\frac{3}{2}$:*

$$\|u_h - R_h u\|_{\Omega_h} \le C_1 h^2 \|u\|_{C^{5,1}(\overline{\Omega})} + C_2 h^{3/2} \|u\|_{C^{2,1}(\overline{\Omega})}, \tag{5.37}$$

where R_h is the restriction to $\Omega_h \cup \Gamma_h$.

Proof. The Taylor expansion shows that

$$\Delta_h R_h u = \Delta u + \frac{h^2}{4!}(u_{xxxx} + u_{yyyy}) + h^4 R, \quad |R| \le \frac{1}{360}\|u\|_{C^{5,1}(\overline{\Omega})} \text{ in } \Omega_h. \tag{5.38}$$

Outside Ω_h one can specify $R_h u$ arbitrarily since these values do not appear in (5.37). In γ_h we set

$$R_h u(\mathbf{x}+h\mathbf{n}) := u(\mathbf{x}-h\mathbf{n}) + 2h\varphi_2(\mathbf{x}) + \frac{2h^3}{3!}\frac{\partial^3 u(\mathbf{x})}{\partial n^3} - \frac{2h^4}{4!}\frac{\partial^4 u(\mathbf{x})}{\partial n^4} + \frac{2h^5}{5!}\frac{\partial^5 u(\mathbf{x})}{\partial n^5}$$

($\mathbf{x} \in \Gamma_h$). The choice is made so that (5.38) also holds for $\mathbf{x} \in \Gamma_h$. Applying Δ_h to $\Delta_h R_h u$ yields

$$\Delta_h^2 R_h u = D_h R_h u = \tilde{R}_h \Delta^2 u + \mathcal{O}(h^2 \|u\|_{C^{5,1}(\overline{\Omega})}) \qquad \text{in } \Omega_h$$

($\tilde{R}_h$ is the restriction to Ω_h). The difference $w_h := u_h - R_h u$ (defined in Ω_h) satisfies

$$\begin{aligned}
D_h w_h &= D_h u_h - D_h R_h u = \tilde{R}_h\left(f - \Delta^2 u\right) + \mathcal{O}(h^2\|u\|_{C^{5,1}(\overline{\Omega})}) =: g_h,\\
w_h(\mathbf{x}) &= u_h(\mathbf{x}) - (R_h u)(\mathbf{x}) = \varphi_1(\mathbf{x}) - u(\mathbf{x}) = 0 \qquad \text{for } \mathbf{x} \in \Gamma_h,\\
\partial_n^0 w_h(\mathbf{x}) &= \partial_n^0 u_h(\mathbf{x}) - \partial_n^0 R_h u(\mathbf{x})\\
&= \varphi_2(\mathbf{x}) - \left[\varphi_2(\mathbf{x}) + \mathcal{O}(h^2\|u\|_{C^{2,1}(\overline{\Omega})})\right] =: \psi_h \quad \text{for } \mathbf{x} \in \Gamma_h.
\end{aligned}$$

If one sets $g_h = \mathcal{O}(h^2\|u\|_{C^{5,1}(\overline{\Omega})})$ and $\psi_h = \mathcal{O}(h^2\|u\|_{C^{2,1}(\overline{\Omega})})$ instead of f_h and φ_2 in Theorem 5.43 (note that $\varphi_1 = 0$!) then (5.33b) implies assertion (5.37). ∎

By using more complicated methods one can replace the factor $h^{-1/2}$ in (5.37) by $C(\varepsilon)h^{-\varepsilon}$ ($\varepsilon > 0$ arbitrary) so that the order of convergence is $\mathcal{O}(h^{2-\varepsilon})$, i.e., almost 2.

Remark 5.45. Inequality (5.33a) shows $\operatorname{cond}_2(L_h) := \|L_h\|_2 \|L_h^{-1}\|_2 = \mathcal{O}(h^{-4})$. In general a difference method for a differential equation of order $2m$ leads to the condition

$$\operatorname{cond}_2(L_h) = \mathcal{O}(h^{-2m}).$$

This indicates greater sensitivity to round-off errors at higher orders $2m$ (cf. Quarteroni–Sacco–Saleri [230, §3.1]).

Difference methods for boundary-value problems of fourth order with variable coefficients and general domains Ω are discussed in a paper by Zlámal [321]. There too convergence of order $\mathcal{O}(h^{3/2})$ is shown. By contrast, $\mathcal{O}(h^2)$-convergence is proved by Bramble [48] for the 13-point star (5.31) in a general domain Ω if the boundary conditions are suitably discretised.

Chapter 6
Tools from Functional Analysis

Abstract Here we collect those definitions and statements which are needed in the next chapters. **Section 6.1** introduces the normed spaces, Banach and Hilbert spaces as well as the operators as linear and bounded mappings between these spaces. In most of the later applications these spaces will be function spaces, containing for instance the solutions of the differential equations. It will turn out that the Sobolev spaces from **Section 6.2** are well suited for the solutions of boundary value problems. The Sobolev space $H^k(\Omega)$ and $H_0^k(\Omega)$ for nonnegative integers k as well as $H^s(\Omega)$ for real $s \geq 0$ are introduced. The definition of the trace (restriction to the boundary Γ) will be essential in §6.2.5 for the interpretation of boundary values. To this end the Sobolev spaces $H^s(\Gamma)$ of functions on the boundary Γ must be defined (cf. Theorem 6.57). Sobolev's Embedding Theorem 6.48 connects Sobolev spaces and classical spaces. **Section 6.3** introduces dual spaces and dual mappings. Compactness properties are important for statements about the unique solvability. Compact operators and the Riesz–Schauder theory are presented in **Section 6.4**. The weak formulation $a(u, v) = f(v)$ of the boundary-value problem is based on bilinear forms described in **Section 6.5**. The inf-sup condition in Lemma 6.94 is a necessary and sufficient criterion for the solvability of the weak formulation.

6.1 Banach Spaces and Hilbert Spaces

6.1.1 Normed Spaces

Let X be a linear space (alternative term: vector space) over $\mathbb{K}$, where $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$. In the following the normal case $\mathbb{K} = \mathbb{R}$ is always intended. $\mathbb{K} = \mathbb{C}$ occurs for instance in connection with Fourier transforms.

The notion of a norm $\|\cdot\| : X \to [0, \infty)$ is explained in Definition 4.20. The linear space X equipped with a norm $\|\cdot\|$ is called a *normed space* and is denoted

W. Hackbusch, *Elliptic Differential Equations*, Springer Series
in Computational Mathematics 18, DOI 10.1007/978-3-662-54961-2_6

by the pair $(X, \|\cdot\|)$. Whenever it is clear which norm belongs to X, this norm is called $\|\cdot\|_X$ and one writes X instead of $(X, \|\cdot\|_X)$.

Example 6.1. (a) The Euclidean norm $|\cdot|$ from (4.34) and the maximum norm (4.23) are norms on $\mathbb{R}^n$.

(b) The continuous functions on $\overline{\Omega}$ form the (infinite-dimensional) space $C^0(\overline{\Omega})$. If Ω is bounded, all $u \in C^0(\overline{\Omega})$ are bounded so that the supremum norm (2.21) is defined and satisfies the norm axioms. If Ω is unbounded, the bounded, continuous functions form a proper subset $(C^0(\overline{\Omega}), \|\cdot\|_{C^0(\overline{\Omega})})$ of $C^0(\overline{\Omega})$. Instead of $\|\cdot\|_{C^0(\overline{\Omega})}$, we use the traditional notation $\|\cdot\|_\infty$.

(c) The Hölder-continuous functions introduced in Definition 3.14 form the normed space $(C^s(\overline{\Omega}), \|\cdot\|_{C^s(\overline{\Omega})})$.

The norm defines a topology on X: $A \subset X$ is open if for all $x \in A$ there exists an $\varepsilon > 0$ so that the ball $K_\varepsilon(x) = \{y \in X : \|x - y\| < \varepsilon\}$ is contained in A. We write

$$x_n \to x \quad \text{or} \quad x = \lim_{n\to\infty} x_n, \qquad \text{if } \|x_n - x\| \to 0.$$

Example 6.2. Let $f_n, f \in C^0(\overline{\Omega})$. The limit process $f_n \to f$ (with respect to $\|\cdot\|_\infty$) denotes the *uniform convergence* known from analysis.

Exercise 6.3. The norm $\|\cdot\| : X \to [0, \infty)$ is continuous; in particular the *reversed triangle inequality* holds:

$$\Big|\, \|x\| - \|y\| \,\Big| \le \|x - y\| \qquad \text{for all } x, y \in X. \tag{6.1}$$

As Example 6.1a shows, several norms can be defined on X. Two norms $\|\cdot\|$ and $|||\cdot|||$ on X are said to be *equivalent* if there exists a number $0 < C < \infty$ such that

$$\frac{1}{C}\,\|x\| \le |||x||| \le C\,\|x\| \qquad \text{for all } x \in X. \tag{6.2}$$

Exercise 6.4. Equivalent norms lead to the same topology on X.

6.1.2 Operators

Let X and Y be normed spaces with the norms $\|\cdot\|_X$ resp. $\|\cdot\|_Y$. A linear mapping $T : X \to Y$ is called an *operator*. If the *operator norm*

$$\|T\|_{Y \leftarrow X} := \sup \{\|Tx\|_Y \,/\, \|x\|_X : 0 \ne x \in X\} \tag{6.3}$$

is finite, T is said to be bounded (cf. (4.30)).

Exercise 6.5. (a) T is bounded if and only if it is continuous. (b) $\|\cdot\|_{Y\leftarrow X}$ is a norm.

With the addition $(T_1 + T_2)\,x = T_1\,x + T_2\,x$ and the scalar multiplication $\lambda\,T$, the bounded operators form a linear space which is denoted by $L(X,Y)$. $(L(X,Y), \|\cdot\|_{Y\leftarrow X})$ is a normed space. With the additional multiplication $(T_1T_2)\,x = T_1(T_2\,x)$, $L(X,X)$ even forms an algebra with unit I, since it is bounded by $\|I\|_{X\leftarrow X} = 1$. I always denotes the identity: $Ix = x$.

Exercise 6.6. Prove that

$$\|Tx\|_Y \le \|T\|_{Y\leftarrow X}\,\|x\|_X \qquad \text{for all } x \in X,\ T \in L(X,Y).$$

If $T_1 \in L(Y,Z)$ and $T_2 \in L(X,Y)$, then $T_1T_2 \in L(X,Z)$ and

$$\|T_1T_2\|_{Z\leftarrow X} \le \|T_1\|_{Z\leftarrow Y}\,\|T_2\|_{Y\leftarrow X}\,.$$

We write $T_n \to T$ for $T, T_n \in L(X,Y)$ if $\|T - T_n\|_{Y\leftarrow X} \to 0$ (convergence in the operator norm).

6.1.3 Banach Spaces

A sequence $\{x_n \in X : n \ge 1\}$ is said to be *Cauchy convergent*, or is called a *Cauchy sequence*, if

$$\sup\{\|x_n - x_m\|_X : n, m \ge k\} \to 0 \text{ for } k \to \infty.$$

A space X is said to be *complete* if any Cauchy-convergent sequence converges to an $x \in X$. A *Banach space* is a complete normed space.

Example 6.7. (a) $\mathbb{R}^n$ is complete, but $\mathbb{Q}^n$ is not ($\mathbb{Q}$: rational numbers).
(b) Let $D \subset \mathbb{R}^n$. $(C^0(D), \|\cdot\|_\infty)$ is a Banach space..
(c) $(C^k(D), \|\cdot\|_{C^k(D)})$ for $k \in \mathbb{N}$ and the Hölder spaces $(C^s(D), \|\cdot\|_{C^s(D)})$ for $s > 0$ as well as $(C^{k,1}(D), \|\cdot\|_{C^{k,1}(D)})$ are complete, thus Banach spaces.

Proof of (b). If $\{u_n\}$ is Cauchy convergent then there exists a limit $u^\star(x) := \lim u_n(x)$ for all $x \in D$. Since u_n converges uniformly to $u^\star$, $u^\star$ must be continuous, i.e., $u^\star \in C^0(D)$. ∎

Denote by $L^\infty(D)$ the set of functions that are bounded and locally integrable on D. Here we do not distinguish between functions which agree almost everywhere. In this case the supremum norm is defined by

$$\|u\|_{L^\infty(D)} := \inf\big\{\sup\{|u(x)| : x \in D\backslash A\} : A \text{ set of measure zero}\big\}.$$

$(L^\infty(D), \|\cdot\|_{L^\infty(D)})$ is a Banach space.

Exercise 6.8. Let X be a normed space and Y a Banach space. Show that $L(X,Y)$ is a Banach space.

A set A is said to be *dense* in $(X, \|\cdot\|_X)$ if $A \subset X$ and $\bar{A} = X$, i.e., every $x \in X$ is the limit of a sequence $a_n \in A$. If $(X, \|\cdot\|_X)$ is a normed, but not complete, space, one calls $(\tilde{X}, \|\cdot\|_{\tilde{X}})$ the *completion* of X, if X is dense in $\tilde{X}$ and $\|x\|_{\tilde{X}} = \|x\|_X$ for all $x \in X$. The completion is uniquely determined up to isomorphism, and can be constructed via equivalence classes of Cauchy sequences.

Lemma 6.9. *Let $(X, \|\cdot\|_X)$ be a normed linear space and subspace of a Banach space $(Y, \|\cdot\|_Y)$ with $\|x\|_Y \le C\,\|x\|_X$ for all $x \in X \subset Y$. Then there exists a completion $\tilde{X}$ of X in Y: $X \subset \tilde{X} \subset Y$.*

Proof. A Cauchy sequence in $(X, \|\cdot\|_X)$ is also a Cauchy sequence in $(Y, \|\cdot\|_Y)$. ■

Frequently it is not easy to describe the image Tx of an operator $T \in L(X,Y)$ for all elements x of the Banach space X. The following theorem permits a considerable simplification: It suffices to investigate T on a dense set $X_0 \subset X$.

Theorem 6.10. *Let X_0 be a dense subspace (or just a dense subset) of the normed space X. Let Y be a Banach space.*

(a) An operator $T_0 \in L(X_0, Y)$ defined on X_0 with $\|T_0\|_{Y \leftarrow X_0} = \sup\limits_{0 \neq x \in X_0} \frac{\|T_0 x\|_Y}{\|x\|_X}$ has a unique continuation $T \in L(X,Y)$, i.e., $Tx = T_0 x$ for all $x \in X_0$.

(b) For $x_n \to x$ $(x_n \in X_0,\ x \in X)$ holds $Tx = \lim_{n\to\infty} T_0 x_n$.

(c) $\|T\|_{Y \leftarrow X} = \|T_0\|_{Y \leftarrow X_0}$.

Proof. For $x \in X_0$, T is defined by $Tx = T_0 x$, while for $x \in X \backslash X_0$ there exists a sequence

$$x_n \to x, \quad x_n \in X_0 \text{ and } Tx = \lim_{n\to\infty} Tx_n = \lim_{n\to\infty} T_0 x_n.$$

It remains to show that $\lim T_0 x_n$ exists and is independent of the choice of the sequence $x_n \in X_0$. Since $\|T_0 x_n - T_0 x_m\|_Y \le \|T_0\|_{Y \leftarrow X} \|x_n - x_m\|_X$, the sequence $T_0 x_n$ is Cauchy convergent. Because of the completeness of Y there exists $y \in Y$ with $T_0 x_n \to y$. Similarly, for a second sequence $x'_n \in X_0$ with $x'_n \to x$ there exists for the same reason a $y' \in Y$ such that $T_0 x'_n \to y'$. Since $y' - y = \lim(T_0 x'_n - T_0 x_n)$ and

$$\|T_0 x'_n - T_0 x_n\|_Y \le \|T_0(x'_n - x_n)\|_Y \le \|T_0\|_{Y \leftarrow X_0} \|x'_n - x_n\|_X \to 0,$$

Tx is well defined by (b). $\|Tx\|_Y \,/\, \|x\|_X = \lim \|T_0 x_n\|_Y \,/\, \|x_n\|_X$ for $x_n \to x$, $x_n \in X_0$ proves part (c). ■

Exercise 6.11. Let X_0 be dense in $(X, \|\cdot\|)$. Let $|||\cdot|||$ be a second norm on X_0, that is equivalent with $\|\cdot\|$ on X_0. Show that the completion of X_0 with respect to $|||\cdot|||$ results in $(X, |||\cdot|||)$ and $\|\cdot\|$ and $|||\cdot|||$ are also equivalent on X.

A corollary of the *open mapping theorem* (also called the theorem of Banach–Schauder; cf. Yosida [312, §II.5]) is the following, important result.

Theorem 6.12. *Let X, Y be Banach spaces. Let $T \in L(X,Y)$ be bijective.*[1] *Then $T^{-1} \in L(Y,X)$ also holds.*

Let X and Y be Banach spaces with $X \subset Y$. Obviously the inclusion denoted by $I : x \in X \mapsto x \in Y$ is a linear mapping. If it is bounded, i.e.,

$$I \in L(X,Y), \qquad \text{i.e., } \|x\|_Y \le C\,\|x\|_X \quad \text{for all } x \in X, \tag{6.4}$$

then X is said to be *continuously embedded* in Y. If furthermore X is dense in $(Y, \|\cdot\|_Y)$, then X is said to be *densely and continuously embedded* in Y.

Exercise 6.13. Let $X \subset Y \subset Z$ be Banach spaces. Show that if X is [densely and] continuously embedded in Y and Y is [densely and] continuously embedded in Z, then X is [densely and] continuously embedded in Z.

6.1.4 Hilbert Spaces

A mapping $(\cdot,\cdot) : X \times X \to \mathbb{K}$ ($\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$) is called a *scalar product* on X if

$$\begin{array}{ll} (x,x) > 0 & \text{for all } 0 \neq x \in X, \\ (\lambda x + y, z) = \lambda\,(x,z) + (y,z) & \text{for all } \lambda \in \mathbb{K},\ x,y,z \in X, \\ (x,y) = \overline{(y,x)} & \text{for all } x,y \in X, \end{array} \tag{6.5}$$

where $\bar{\lambda}$ denotes the complex conjugate of $\lambda \in \mathbb{C}$.

Exercise 6.14. Show that (a) $\|x\| := \sqrt{(x,x)}$ is a norm on X. For the proof use (b) the *Schwarz inequality*:

$$|(x,y)| \le \|x\|\,\|y\| \qquad \text{for all } x,y \in X. \tag{6.6}$$

Prove (6.6) without using $(x,x) > 0$ for $x \neq 0$. *Hint:* consider $\|x - \lambda y\|^2 \ge 0$ for $\|x\| = \|y\| = 1$ and $\lambda = \frac{1}{(y,x)}$, if $(x,y) \neq 0$.

(c) $(\cdot,\cdot) : X \times X \to \mathbb{K}$ is continuous.

A Banach space $(X, \|\cdot\|_X)$ is called a *Hilbert space* if there exists a scalar product $(\cdot,\cdot)_X$ on X such that $\|x\|_X = \sqrt{(x,x)_X}$ for all $x \in X$.

$x \in X$ and $y \in X$ are said to be *orthogonal* ($x \perp y$) if $(x,y)_X = 0$. If $A \subset X$ is a subset of the Hilbert space X then the orthogonal space

$$A^\perp := \{x \in X :\ (x,a)_X = 0 \text{ for all } a \in A\}$$

[1] $T \in L(X,Y)$ is *injective* if $T(x) = T(x')$ implies $x = x'$. It is *surjective* if $\mathrm{range}(T) = Y$. T is *bijective* if it is injective and surjective.

defines a closed subspace of X (the closedness follows from Exercise 6.14c).

Lemma 6.15. *Let U be a closed subspace of a Hilbert space X. Then X can be decomposed into the* direct sum $X = U \oplus U^\perp$*, i.e., any $x \in X$ has a unique decomposition $x = u + v$ with $u \in U$, $v \in U^\perp$. Furthermore, $\|x\|_X^2 = \|u\|_X^2 + \|v\|_X^2$.*

Proof. (i) $U \cap U^\perp = \{0\}$ holds, since any $u \in U \cap U^\perp$ satisfies $\|u\|^2 = (u,u) = 0$ because $u \in U^\perp$. We assume $U \neq X$ since otherwise the statement is trivial.

(ii) It remains to show $U + U^\perp = X$. We select an arbitrary $0 \neq x \in X \backslash U$ and want to prove that $x = u + u^\perp$ with $u \in U$ and $u^\perp \in U^\perp$. The distance $\delta := \inf_{u \in U} \|x - u\|$ must be positive, since otherwise there exists a sequence $u_\nu \in U$ with $\lim u_\nu = x$, and $x \in U$ follows from the closedness of U in contradiction to the choice of x. Now we choose a sequence $u_\nu \in U$ with $\lim \|x - u_\nu\| = \delta > 0$. From $\|u_\mu - u_\nu\| = \|(x - u_\nu) - (x - u_\mu)\|$ and the *parallelogram identity*

$$\|a+b\|^2 + \|a-b\|^2 = 2\left(\|a\|^2 + \|b\|^2\right)$$

we conclude that

$$\begin{aligned}\|u_\mu - u_\nu\|^2 &= 2\left(\|x-u_\nu\|^2 + \|x-u_\mu\|^2\right) - 4\|x - \tfrac{1}{2}(u_\nu + u_\mu)\|^2 \\ &\leq 2\left(\|x-u_\nu\|^2 + \|x-u_\mu\|^2\right) - 4\delta^2,\end{aligned}$$

since $\frac{1}{2}(u_\nu + u_\mu) \in U$. Then $\|u_\mu - u_\nu\|^2 \to 2(\delta^2 + \delta^2) - 4\delta^2 = 0$ for $\nu, \mu \to \infty$, i.e., $\{u_\nu\}$ is a Cauchy sequence. Since the Hilbert space is complete and U is closed, $\delta = \|x - u\|$ follows for $u := \lim u_\nu \in U$.

For an indirect proof assume that there is a $v \in U$ with $\lambda := (x - u, v) \neq 0$. Then $x - \hat{u}$ with $\hat{u} := u + \frac{\lambda}{\|v\|^2} v$ has the squared norm $\|x - \hat{u}\|^2 = \delta^2 - \lambda^2 / \|v\|^2$ in contradiction to the minimality of δ. Hence $(x - u, v) = 0$ holds for all $v \in U$ and therefore $x - u \in U^\perp$. Setting $u^\perp := x - u$, we have proved the unique decomposition $x = u + u^\perp$. ∎

Corollary 6.16. Let X be a Hilbert space.

(a) For any subspace $U \subset X$, $(U^\perp)^\perp = \overline{U}$ is the closure of U.

(b) Let U be a closed subspace. The distance of $x \in X$ to U is a minimum taken for some $u^* \in U$: $\inf_{u \in U} \|x - u\| = \|x - u^*\|$.

Proof. (a) $U^\perp$ is closed. Hence $X = U^\perp \oplus (U^\perp)^\perp$ holds as well as $X = \overline{U} \oplus U^\perp$.
(b) The decomposition $x = u + u^\perp$ yields $u^* := u$. ∎

Exercise 6.17. Let A be a subset of the Hilbert space X. Prove the equivalence of the following two statements:

(a) A is dense in X;

(b) for any $0 \neq x \in X$ there exists some $a \in A$ with $(a, x)_X \neq 0$.

6.2 Sobolev Spaces

In the following Ω is always an open subset of $\mathbb{R}^n$.

6.2.1 $L^2(\Omega)$

$L^2(\Omega)$ consists of all Lebesgue-measurable functions whose squares on Ω are Lebesgue-integrable. Two functions $u, v \in L^2(\Omega)$ are considered to be equal ($u = v$) if $u(\mathbf{x}) = v(\mathbf{x})$ for almost all $\mathbf{x} \in \Omega$, i.e., for all $\mathbf{x} \in \Omega \backslash A$, where the exceptional set A has Lebesgue measure $\mu(A) = 0$.

Theorem 6.18. *$L^2(\Omega)$ forms a Hilbert space with the scalar product*

$$(u, v)_0 := \int_\Omega u(\mathbf{x})\, \overline{v(\mathbf{x})}\, \mathrm{d}\mathbf{x} \qquad \textit{for all } u, v \in L^2(\Omega) \tag{6.7}$$

and the norm

$$|u|_0 := \|u\|_{L^2(\Omega)} := \sqrt{\int_\Omega |u(x)|^2 \, \mathrm{d}\mathbf{x}}. \tag{6.8}$$

Lemma 6.19. *The spaces $C^\infty(\Omega) \cap L^2(\Omega)$ and $C_0^\infty(\Omega)$ are dense in $L^2(\Omega)$. Here*

$$C_0^\infty(\Omega) := \{u \in C^\infty(\Omega) : \operatorname{supp}(u) \subset\subset \Omega\}. \tag{6.9}$$

$\operatorname{supp}(u) := \overline{\{\mathbf{x} \in \Omega : u(\mathbf{x}) \neq 0\}}$ *denotes the* support *of u. The double inclusion*

$$\omega \subset\subset \Omega$$

indicates that $\overline{\omega}$ is compact and lies in the interior of Ω (i.e., $\overline{\omega} \subset \Omega$ is bounded and $\operatorname{dist}(\omega, \partial\Omega) > 0$).

Let D^α be the partial derivative operator (3.11b). In the following we need so-called *weak derivatives* $D^\alpha u$ which are defined in a nonclassical way.

Definition 6.20. $u \in L^2(\Omega)$ has a (weak) derivative[2] $v := D^\alpha u \in L^2(\Omega)$ if for the latter $v \in L^2(\Omega)$ holds:

$$(w, v)_0 = (-1)^{|\alpha|} (D^\alpha w, u)_0 \qquad \text{for all } w \in C_0^\infty(\Omega). \tag{6.10}$$

Exercise 6.21. Show the following: (a) Let u have a weak derivative $D^\alpha u \in L^2(\Omega)$. If the classical derivative $D^\alpha u$ exists in $\Omega' \subset \Omega$, it coincides there (almost everywhere) with the weak one.

[2] More generally, one can replace the basic set $L^2(\Omega)$ by distributions.

(b) If u has the weak derivative $v_\alpha = D^\alpha u \in L^2(\Omega)$, and if v_α has the weak derivative $v_{\alpha+\beta} = D^\beta v_\alpha \in L^2(\Omega)$ then $v_{\alpha+\beta}$ is also the weak $D^{\alpha+\beta}$-derivative of u; i.e., $D^{\alpha+\beta}u = D^\beta(D^\alpha u)$.

(c) Let $\Omega \subset \mathbb{R}^n$ be bounded and $0 \in \Omega$. $u(\mathbf{x}) := |\mathbf{x}|^\sigma$ has weak first derivatives in $L^2(\Omega)$ if $\sigma = 0$ or $2\sigma + n > 2$.

(d) For $u_\nu \in C^\infty(\Omega)$ let $u_\nu \to u \in L^2(\Omega)$ and $D^\alpha u_\nu \to v \in L^2(\Omega)$ in the $L^2(\Omega)$ norm. Then $v = D^\alpha u$.

For later applications to finite elements we consider the next example.

Example 6.22. Let $\overline{\Omega} = \bigcup_{i=1}^{N} \overline{\Omega}_i$, where the bounded subdomains Ω_i are disjoint and have piecewise smooth boundaries. Let $k \in \mathbb{N}$. A function $u \in C^{k-1}(\overline{\Omega})$ with restrictions $u_{\Omega_i} \in C^k(\overline{\Omega}_i)$ $(1 \le i \le N)$ has a (weak) k-th derivative $v_\alpha = D^\alpha u \in L^2(\Omega)$, $|\alpha| \le k$, which coincides with the classical one in $\bigcup_{i=1}^N \Omega_i$.

Proof. (i) The $(k-1)$-th derivatives $v_\beta := D^\beta u \in C(\overline{\Omega})$ exist as classical derivatives. Hence the assertion need only be shown for $k = 1$ and applied to v_β.

(ii) Let $D^\alpha = \frac{\partial}{\partial x_j}$, $w \in C_0^\infty(\Omega)$ and integrate by parts:

$$
\begin{aligned}
-(D^\alpha w, u)_0 &= -\left(w_{x_j}, u\right)_0 = -\int_\Omega w_{x_j} u \mathrm{d}\mathbf{x} = -\sum_{i=1}^N \int_{\Omega_i} w_{x_j} u \mathrm{d}\mathbf{x} \\
&= \sum_{i=1}^N \left[\int_{\Omega_i} w\, u_{x_j} \mathrm{d}\mathbf{x} - \int_{\partial\Omega_i} w\, u\, n_j^{(i)} \mathrm{d}\Gamma \right],
\end{aligned}
$$

where $n_j^{(i)}$ is the j-th components of the normal vector $\mathbf{n}^{(i)}$ of Ω_i. Note that we have opposite normals $\mathbf{n}^{(i)} = -\mathbf{n}^{(k)}$ at $\mathbf{x} \in \partial\Omega_i \cap \partial\Omega_k$ $(i \ne k)$. Therefore all contributions in

$$
\sum_{i=1}^N \int_{\partial\Omega_i} w\, u\, n_j^{(i)} \,\mathrm{d}\Gamma
$$

belonging to inner edges cancel. Moreover, boundary terms on $\partial\Omega$ vanish since $w = 0$ on $\partial\Omega$. Define $v_\alpha \in L^2(\Omega)$ by $v_\alpha = u_{x_j}$ on all subdomains Ω_i and arbitrarily on the remaining set $\overline{\Omega} \setminus \bigcup \Omega_i$ of measure zero. Then

$$
-(D^\alpha w, u)_0 = \sum_{i=1}^N \int_{\Omega_i} w u_{x_j} \mathrm{d}\mathbf{x} = \sum_{i=1}^N \int_{\Omega_i} w v_\alpha \mathrm{d}\mathbf{x} = \int_\Omega w v_\alpha \mathrm{d}\mathbf{x} = (w, v_\alpha)_0 ,
$$

proves that v_α is the weak derivative $D^\alpha u$. ∎

The Schwarz inequality $|(u,v)_0| \le |u|_0 \, |v|_0$ (cf. (6.6)) reads explicitly

$$
\left| \int_\Omega u(\mathbf{x}) v(\mathbf{x}) \mathrm{d}\mathbf{x} \right| \le \sqrt{\int_\Omega |u(\mathbf{x})|^2 \,\mathrm{d}\mathbf{x} \int_\Omega |v(\mathbf{x})|^2 \,\mathrm{d}\mathbf{x}} . \tag{6.11a}
$$

For $v = 1$ the result is

$$\left| \int_\Omega u(\mathbf{x}) \mathrm{d}\mathbf{x} \right| \leq \sqrt{\mu(\Omega)} \sqrt{\int_\Omega |u(\mathbf{x})|^2 \, \mathrm{d}\mathbf{x}} \qquad (\mu(\Omega) : \text{ measure of } \Omega) . \quad (6.11b)$$

For $a \in L^\infty(\Omega)$ and $u, v \in L^2(\Omega)$ one now has

$$\left| \int_\Omega a(\mathbf{x}) u(\mathbf{x}) v(\mathbf{x}) \mathrm{d}\mathbf{x} \right| \leq \|a\|_{L^\infty(\Omega)} \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} = \|a\|_\infty |u|_0 |v|_0 . \quad (6.11c)$$

6.2.2 $H^k(\Omega)$ and $H_0^k(\Omega)$

Let $k \in \mathbb{N}_0$. Let $H^k(\Omega) \subset L^2(\Omega)$ be the set of all functions having weak derivatives $D^\alpha u \in L^2(\Omega)$ for all $|\alpha| \leq k$:

$$H^k(\Omega) = \left\{ u \in L^2(\Omega) : D^\alpha u \in L^2(\Omega) \text{ for } |\alpha| \leq k \right\} .$$

The Sobolev space denoted here by $H^k(\Omega)$ is denoted by $W_2^k(\Omega)$ or $W^{k,2}(\Omega)$ in some other places.

Theorem 6.23. *$H^k(\Omega)$ forms a Hilbert space with the scalar product*

$$(u, v)_k := (u, v)_{H^k(\Omega)} := \sum_{|\alpha| \leq k} (D^\alpha u, D^\alpha v)_{L^2(\Omega)} \quad (6.12)$$

and the (Sobolev) norm

$$|u|_k := \|u\|_{H^k(\Omega)} := \sqrt{\sum_{|\alpha| \leq k} \|D^\alpha u\|^2_{L^2(\Omega)}} . \quad (6.13)$$

Exercise 6.24. Let $\Omega = \left\{ (x, y) \in \mathbb{R}^2 : x^2 + y^2 < 1 \right\}$ be the unit circle. Prove that the function $u(x, y) = \log(\log \frac{2}{\sqrt{x^2+y^2}})$ belongs to $H^1(\Omega)$, although it is discontinuous.

Lemma 6.25. *$C^\infty(\Omega) \cap H^k(\Omega)$ lies densely in $H^k(\Omega)$.*

Lemma 6.25, whose proof can be found, for example, in Wloka [308, Theorem 3.5] permits a second definition of the Sobolev space $H^k(\Omega)$. The precise meaning of the 'completion in $L^2(\Omega)$' is explained in Lemma 6.9.

Remark 6.26. Let $X_0 := \{u \in C^\infty(\Omega) : |u|_k < \infty\}$. The completion of X_0 in $L^2(\Omega)$ with respect to norm (6.13) results in $H^k(\Omega)$.

Definition 6.27. The completion of $C_0^\infty(\Omega)$ in $L^2(\Omega)$ with respect to the norm (6.13) is denoted by $H_0^k(\Omega)$.

Theorem 6.28. *The Hilbert space $H_0^k(\Omega)$ is a subspace of $H^k(\Omega)$ with the same scalar product (6.12) and same norm (6.13). $C_0^\infty(\Omega)$ is dense in $H_0^k(\Omega)$. For $k = 0$ there holds*

$$H_0^0(\Omega) = H^0(\Omega) = L^2(\Omega). \tag{6.14}$$

Proof. (i) $C_0^\infty(\Omega) \subset X_0$ (cf. Remark 6.26) implies $H_0^k(\Omega) \subset H^k(\Omega) \subset L^2(\Omega)$.

(ii) According to the definition $C_0^\infty(\Omega)$ is dense in $H_0^k(\Omega)$.

(iii) For $k = 0$ the norms (6.8) and (6.13) coincide. Lemma 6.19 proves $H_0^0(\Omega) = L^2(\Omega)$. Because of $H_0^k(\Omega) \subset H^k(\Omega) \subset L^2(\Omega)$ (k arbitrary) (6.14) follows. ∎

Lemma 6.29. *For Ω bounded, $\|\cdot\|_{H^k(\Omega)}$ and*

$$|u|_{k,0} := \sqrt{\sum_{|\alpha|=k} \|D^\alpha u\|^2_{L^2(\Omega)}} \tag{6.15}$$

are equivalent norms in $H_0^k(\Omega)$.

Proof. (i) Evidently, $|\cdot|_{k,0} \le \|\cdot\|_{H^k(\Omega)}$.

(ii) Let $u \in C_0^\infty(\Omega)$. For an α with $|\alpha| = k-1$ set $v := D^\alpha u \in C_0^\infty(\Omega)$. Set $u = v = 0$ on $\mathbb{R}^n \backslash \overline{\Omega}$. There exists an R with $\Omega \subset K_R(0)$. For each $\mathbf{x} \in \Omega$ with $x_1 \in [-R, R]$; thus

$$\begin{aligned} |v(\mathbf{x})|^2 &= \left| \int_{-R}^{x_1} v_{x_1}(\xi, x_2, \ldots, x_n) \mathrm{d}\xi \right|^2 \underset{(6.11\mathrm{b})}{\le} (x_1 + R) \int_{-R}^{x_1} |v_{x_1}(\xi, x_2, \ldots, x_n)|^2 \mathrm{d}\xi \\ &\le 2R \int_{-R}^{R} |v_{x_1}(\xi, x_2, \ldots, x_n)|^2 \mathrm{d}\xi. \end{aligned}$$

Integration over $\mathbf{x} \in \Omega$ yields $|v|_0^2 \le 4R^2 |v_{x_1}|_0^2 \le 4R^2 |u|_{k,0}^2$, since v_{x_1} is k-fold derivative. Summation over all multi-indices α with $|\alpha| = k-1$ yields $|u|_{k-1,0}^2 \le C_{k-1} |u|_{k,0}^2$. Now $|u|_{j-1,0}^2 \le C_{j-1} |u|_{j,0}^2$ follows likewise for all $1 \le j \le k$, and thus $|u|_{j,0}^2 \le C_j C_{j+1} \cdots C_{k-1} |u|_{k,0}^2$. Since $|u|_k^2 = \sum_{j=0}^k |u|_{j,0}^2 \le C |u|_{k,0}^2$, for all $u \in C_0^\infty(\Omega)$, the statement follows from Exercise 6.11. ∎

$c_\Omega = c_{\Omega,k}$ involved in the Poincaré–Friedrichs inequality $\|u\|_{L^2(\Omega)} \le c_\Omega |u|_{k,0}$ is called the Poincaré–Friedrichs constant (cf. Rektorys [235, §18]).

Exercise 6.30. Show that (a) for bounded Ω and $k \ge 1$, $H^k(\Omega)$ and $H_0^k(\Omega)$ are different. *Hint*: Consider the constant function $u(\mathbf{x}) = 1$ and use Lemma 6.29.

(b) Lemma 6.29 holds even if Ω is bounded in one direction, i.e., if Ω lies on a strip $\{\mathbf{x} \in \mathbb{R}^n : |x_\nu| < R\}$ ($\nu \in \{1, \ldots, n\}$) or more generally $\{\mathbf{x} \in \mathbb{R}^n : |(\mathbf{x}, \mathbf{t})| < R\}$ for some $\mathbf{t} \in \mathbb{R}^n$ with $|\mathbf{t}| = 1$.

Theorem 6.31. *Let $m \geq 1$. There exist constants $C = C(m)$ and $\eta(\varepsilon) = \eta(\varepsilon, m)$ such that*

$$|u|_k \leq C\,|u|_m^{k/m}\,|u|_0^{(m-k)/m} \qquad \text{for all } 0 \leq k \leq m,\ u \in H_0^m(\Omega), \tag{6.16a}$$

$$|u|_k \leq \varepsilon\,|u|_m + \eta(\varepsilon)\,|u|_0 \qquad \text{for all } \varepsilon > 0,\ 0 \leq k < m,\ u \in H_0^m(\Omega). \tag{6.16b}$$

Proof. (i) Let α be a multi-index with $|\alpha| = 1$. Partial integration yields

$$|D^\alpha u|_0^2 = (D^\alpha u, D^\alpha u)_0 = -(D^{2\alpha} u, u)_0 \leq |D^{2\alpha} u|_0\,|u|_0 \leq |u|_2\,|u|_0\,.$$

Since also $|u|_0^2 \leq |u|_2\,|u|_0$, we have $|u|_1^2 \leq (n+1)\,|u|_2\,|u|_0$. Replacing u by $D^\beta u$ with $|\beta| = 1$, it follows that

$$|u|_\ell^2 \leq \tilde{C}\,|u|_{\ell+1}\,|u|_{\ell-1} \qquad \text{for all } 0 \leq \ell < m,\ u \in H_0^m(\Omega). \tag{6.16c}$$

(ii) Let $u \neq 0$ be fixed. Set $\eta_\ell := \log |u|_\ell$, $\xi_\ell := \frac{\ell\eta_m + (m-\ell)\eta_0}{m}$, and $\zeta_\ell := \eta_\ell - \xi_\ell$. (6.16c) leads to $2\zeta_\ell - \zeta_{\ell-1} - \zeta_{\ell+1} \leq \tilde{c} := \log \tilde{C}$, where $\zeta_0 = \zeta_m = 0$. For $z = (\zeta_1, \ldots, \zeta_{m-1})^{\mathsf{T}}$ one obtains $Az \leq \tilde{c}\,\mathbf{1}$. Here A is the M-matrix (4.7b) for $h = 1$. $A^{-1} \geq 0$ proves that $z \leq c := \tilde{c}\,A^{-1}\mathbf{1}$. $\eta_\ell = \zeta_\ell + \xi_\ell \leq c_\ell + \xi_\ell$ $(1 \leq \ell < m)$ implies $|u|_\ell = \exp(\eta_\ell) \leq \exp(c_\ell + \zeta_\ell)$, i.e., (6.16a) with $C := \exp(\tilde{c}\,\|A^{-1}\mathbf{1}\|_\infty)$.

(iii) Elementary calculation shows that for each $\varepsilon > 0$, $0 \leq \Theta \leq \Theta_0 < 1$ there exists an $\eta(\varepsilon) = \eta(\varepsilon, \Theta_0)$ such that

$$a^\Theta b^{1-\Theta} \leq \varepsilon a + \eta(\varepsilon) b \qquad \text{for all } a, b \geq 0. \tag{6.16d}$$

Formula (6.16b) is a corollary of (6.16a,d). ∎

By similar means, together with (5.34), one proves

$$\left(D^\alpha u, D^\beta u\right)_0 \leq \varepsilon\,|u|_m^2 + \frac{1}{4\varepsilon}\,|u|_k^2 \qquad \text{for } \begin{cases} \varepsilon > 0, & u \in H^m(\Omega), \\ |\alpha| \leq m, & |\beta| \leq k \leq m, \end{cases} \tag{6.16e}$$

$$\left(D^\alpha u, D^\beta u\right)_0 \leq \varepsilon\,|u|_m^2 + \eta(\varepsilon)\,|u|_0^2 \qquad \text{for } \begin{cases} \varepsilon > 0, & u \in H_0^m(\Omega), \\ |\alpha| \leq m, & |\beta| < m. \end{cases} \tag{6.16f}$$

The following statement is of interest for unbounded domains Ω.

Remark 6.32. The set $\{u \in C^\infty(\Omega) : \operatorname{supp}(u) \text{ compact}, |u|_k < \infty\}$ is dense in $H^k(\Omega)$.

Proof. Let $u \in H^k(\Omega)$, $\varepsilon > 0$. According to Lemma 6.25 there exists a function $u_\varepsilon \in C^\infty(\Omega)$ with $\|u - u_\varepsilon\|_{H^k(\Omega)} \leq \varepsilon/3$. For sufficiently large R, one has $\|u_\varepsilon\|_{H^k(\Omega \setminus K_R(\mathbf{0}))} \leq \varepsilon/3$ and there exists $a \in C^\infty(\mathbb{R}^n)$ $a(\mathbf{x}) = 1$ for $|\mathbf{x}| \leq R$ and $a(\mathbf{x}) = 0$ for $|x| \geq 2R$ and $\|u_\varepsilon - a u_\varepsilon\|_{H^k(\Omega) \cap K_{2R}(0)} \leq \frac{\varepsilon}{3}$. Thus there exists $v = a u_\varepsilon \in C_0^\infty(\Omega)$ with $|u - v|_k \leq \varepsilon$. ∎

Since "$\operatorname{supp}(u)$ compact" already implies "$\operatorname{supp}(u) \subset\subset \mathbb{R}^n$" we obtain the following corollary.

Corollary 6.33. $H_0^k(\mathbb{R}^n) = H^k(\mathbb{R}^n)$ for all $k \geq 0$.

The Leibniz rule for derivatives of products proves the following.

Theorem 6.34. $\|au\|_{H^k(\Omega)} \leq C_k \|a\|_{C^k(\overline{\Omega})} \|u\|_{H^k(\Omega)}$ *for all* $a \in C^k(\overline{\Omega})$ *and* $u \in H^k(\Omega)$.

Theorem 6.34 together with the substitution rule for volume integrals shows the next theorem.

Theorem 6.35 (transformation theorem). *Let* $T \in C^{\max(k,1)}(\overline{\Omega}) : \Omega \to \Omega'$ *be a bijective transformation with* $\left|\det \frac{dT}{d\mathbf{x}}\right| \geq \delta > 0$ *in* Ω. *We write* $v = u \circ T$ *for* $v(\mathbf{x}) = u(T(\mathbf{x}))$. *Then* $u \in H^k(\Omega')$ $[u \in H_0^k(\Omega')]$ *also implies* $u \circ T \in H^k(\Omega)$ $[u \circ T \in H_0^k(\Omega)]$ *and*

$$\|u \circ T\|_{H^k(\Omega)} \leq C_k \|T\|_{C^k(\overline{\Omega})} \|u\|_{H^k(\Omega)} / \sqrt{\delta}\,. \tag{6.17}$$

6.2.3 Fourier Transformation and $H^k(\mathbb{R}^n)$

For $u \in C_0^\infty(\mathbb{R}^n)$ one defines the Fourier-transformed function $\hat{u}$ by

$$\hat{u}(\boldsymbol{\xi}) := (\mathcal{F}u)(\boldsymbol{\xi}) := (2\pi)^{-n/2} \int_{\mathbb{R}^n} \mathrm{e}^{-\mathrm{i}\langle \boldsymbol{\xi}, \mathbf{x}\rangle} u(\mathbf{x}) \mathrm{d}\mathbf{x} \qquad (\boldsymbol{\xi} \in \mathbb{R}^n)\,. \tag{6.18}$$

Note that $\hat{u}$ is described by a proper integral since the support of u is bounded.

Lemma 6.36. *Let* $u \in C_0^\infty(\mathbb{R}^n)$. *For* $R \to \infty$

$$I_R(u; \mathbf{y}) := (2\pi)^{-n} \int_{|\boldsymbol{\xi}|_\infty \leq R} \left[\int_{\mathbb{R}^n} \mathrm{e}^{-\mathrm{i}\langle \boldsymbol{\xi}, \mathbf{x}-\mathbf{y}\rangle} u(\mathbf{x}) \mathrm{d}\mathbf{x} \right] \mathrm{d}\boldsymbol{\xi}$$

converges uniformly to $u(\mathbf{y})$ *on* $\operatorname{supp}(u)$. *More precisely,* $I_R(u; \mathbf{y}) = u(\mathbf{y}) + \mathcal{O}(\frac{1}{R})$.

Proof. It suffices to discuss the case $n = 1$ (by Fubini's theorem). Integration with respect to $\boldsymbol{\xi}$ results in

$$I_R(u; y) = \frac{1}{\pi} \int_{\mathbb{R}} (x - y)^{-1} \sin\big(R(x - y)\big)\, u(x)\, \mathrm{d}x.$$

Then $I_R(1; y) = \frac{1}{\pi} \int_{\mathbb{R}} t^{-1} \sin(t) \mathrm{d}t = 1$ for all $R > 0$. Since $u \in C^\infty(\mathbb{R})$, then also $w(x, y) := \frac{u(x) - u(y)}{x - y} \in C^\infty(\mathbb{R}^{2n})$. The estimates

$$w(x, y) = \mathcal{O}\Big(\frac{1}{1 + |x - y|}\Big) \quad \text{and} \quad w_x(x, y) = \mathcal{O}(1/\,[\,1 + |x - y|\,]^2)$$

hold uniformly for $y \in \operatorname{supp}(u)$. Partial integration yields

$$I_R(u(\cdot)-u(y);y) = I_R((\cdot-y)w(\cdot,y);y) = \frac{1}{\pi}\int_{\mathbb{R}} \sin(R(x-y))w(x,y)\mathrm{d}x$$
$$= -\frac{1}{\pi R}\int_{\mathbb{R}} \cos(R(x-y))\, w_x(x,y)\,\mathrm{d}x = \mathcal{O}(1/R).$$

The statement follows from $I_R(u;y) = u(y)\, I_R(1;y) + I_R(u(\cdot)-u(y);y) = u(y)+\mathcal{O}(\frac{1}{R})$. ∎

Lemma 6.37. *$\hat{u} \in L^2(\mathbb{R}^n)$ and $|\hat{u}|_0 = |u|_0$ for all $u \in C_0^\infty(\mathbb{R}^n)$.*

Proof. Lemma 6.36 shows that

$$\int_{|\boldsymbol{\xi}|_\infty \le R} |\hat{u}(\boldsymbol{\xi})|^2\,\mathrm{d}\boldsymbol{\xi} = (2\pi)^{-n} \int_{|\boldsymbol{\xi}|_\infty \le R} \left[\int_{\mathbb{R}^n} \mathrm{e}^{-\mathrm{i}\langle\boldsymbol{\xi},\mathbf{x}\rangle} u(\mathbf{x})\mathrm{d}\mathbf{x}\right]\left[\int_{\mathbb{R}^n} \mathrm{e}^{\mathrm{i}\langle\boldsymbol{\xi},\mathbf{y}\rangle}\overline{u(\mathbf{y})}\mathrm{d}\mathbf{y}\right]\mathrm{d}\boldsymbol{\xi}$$
$$= \int_{\mathbb{R}^n} I_R(u;\mathbf{y})\overline{u(\mathbf{y})}\mathrm{d}\mathbf{y}$$

converges to $\int_{\mathbb{R}^n} |u(\mathbf{y})|^2\,\mathrm{d}\mathbf{y}$. ∎

Lemma 6.38. *The inverse Fourier transformation $\mathcal{F}^{-1}\hat{u} = u$ is defined for $u \in C_0^\infty(\mathbb{R}^n)$ by*

$$(\mathcal{F}^{-1}\hat{u})(\mathbf{x}) := \frac{1}{(2\pi)^{\frac{n}{2}}}\int_{\mathbb{R}^n} \mathrm{e}^{\mathrm{i}\langle\boldsymbol{\xi},\mathbf{x}\rangle}\hat{u}(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} := \frac{1}{(2\pi)^{\frac{n}{2}}}\lim_{R\to\infty}\int_{|\boldsymbol{\xi}|_\infty \le R} \mathrm{e}^{\mathrm{i}\langle\boldsymbol{\xi},\mathbf{x}\rangle}\hat{u}(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi}\,. \quad (6.19)$$

Proof. $\frac{1}{(2\pi)^{\frac{n}{2}}}\int_{|\boldsymbol{\xi}|_\infty \le R} \mathrm{e}^{\mathrm{i}\langle\boldsymbol{\xi},\mathbf{x}\rangle}\hat{u}(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} = \frac{1}{(2\pi)^n}\int_{|\boldsymbol{\xi}|_\infty \le R}\int_{\mathbb{R}^n} \mathrm{e}^{\mathrm{i}\langle\boldsymbol{\xi},\mathbf{x}-\mathbf{y}\rangle}u(\mathbf{y})\mathrm{d}\mathbf{y}\mathrm{d}\boldsymbol{\xi} = u(\mathbf{x})+\mathcal{O}(\frac{1}{R})$ follows from Lemma 6.36. ∎

Theorem 6.39. *(a) The mappings $\mathcal{F}, \mathcal{F}^{-1} \in L(L^2(\mathbb{R}^n), L^2(\mathbb{R}^n))$ let the L^2 norm invariant: $\|\mathcal{F}\|_{L^2(\mathbb{R}^2)\leftarrow L^2(\mathbb{R}^n)} = \|\mathcal{F}^{-1}\|_{L^2(\mathbb{R}^n)\leftarrow L^2(\mathbb{R}^n)} = 1$.*
(b) The scalar product satisfies $(u,v)_0 = (\hat{u},\hat{v})_0$ for all $u, v \in L^2(\mathbb{R}^n)$.

Proof. (a) Since $C_0^\infty(\mathbb{R}^n)$ is dense in $L^2(\mathbb{R}^n)$ (cf. Lemma 6.19), $\mathcal{F}$ can be continued to $\mathcal{F} : L^2(\mathbb{R}^n) \to L^2(\mathbb{R}^n)$ (cf. Theorem 6.10). The norm estimate follows from Lemma 6.37. The roles of $\mathcal{F}$ and $\mathcal{F}^{-1}$ are interchangeable (cf. (6.18) and (6.19)); thus $\mathcal{F}^{-1} \in L(L^2(\mathbb{R}^n), L^2(\mathbb{R}^n))$ also holds.

(b) $(u,v)_0 = \frac{1}{2}\big(|u+v|_0^2 - |u|_0^2 - |v|_0^2\big) = \frac{1}{2}\big(|\hat{u}+\hat{v}|_0^2 - |\hat{u}|_0^2 - |\hat{v}|_0^2\big) = (\hat{u},\hat{v})_0$ follows from Lemma 6.37. ∎

Exercise 6.40. Prove that: (a) With $\boldsymbol{\xi}^\alpha = \xi_1^{\alpha_1}\ldots\xi_n^{\alpha_n}$, there holds

$$\mathcal{F}(D^\alpha u)(\boldsymbol{\xi}) = \mathrm{i}^{|\alpha|}\boldsymbol{\xi}^\alpha\hat{u}(\boldsymbol{\xi}) \qquad \text{for } u \in C_0^\infty(\mathbb{R}^n).$$

(b) There exists $C = C(k)$ such that $\frac{1}{C}(1+|\boldsymbol{\xi}|^2)^k \le \sum_{|\alpha|\le k} |\boldsymbol{\xi}^\alpha|^2 \le C(1+|\boldsymbol{\xi}|^2)^k$ for all $\boldsymbol{\xi} \in \mathbb{R}^n$.

Lemma 6.41. *(a)* $|u|_k = \left|\sqrt{\sum_{|\alpha|\le k} |\boldsymbol{\xi}^\alpha|^2}\,\hat{u}(\boldsymbol{\xi})\right|_0$ *holds for all* $u \in H^k(\mathbb{R}^n)$.
(b) A norm on $H^k(\mathbb{R}^n)$ *equivalent to* $|\cdot|_k$ *is*

$$|u|_k^\wedge := \left|(1+|\boldsymbol{\xi}|^2)^{k/2}\hat{u}(\boldsymbol{\xi})\right|_0 . \tag{6.20}$$

Proof. (a) It suffices to show the statement for $u \in C_0^\infty(\mathbb{R}^n)$ (cf. Corollary 6.33, Theorem 6.10):

$$|u|_k^2 = \sum_{|\alpha|\le k} |D^\alpha u|_0^2 = \sum_{|\alpha|\le k} |\mathcal{F}D^\alpha u|_0^2 = \sum_{|\alpha|\le k} |\boldsymbol{\xi}^\alpha \hat{u}(\boldsymbol{\xi})|_0^2 = \left|\sqrt{\sum_{|\alpha|\le k} |\boldsymbol{\xi}^\alpha|^2}\,\hat{u}(\boldsymbol{\xi})\right|_0 .$$

(b) The statement follows from Exercise 6.40b. ∎

Lemma 6.42. *Let* $\partial_{h,j}$ *be the difference operator*

$$\partial_{h,j}u(x) := \frac{1}{h}\left[u\Big(x+\frac{h}{2}\mathbf{e}_j\Big) - u\Big(x-\frac{h}{2}\mathbf{e}_j\Big)\right], \quad h>0, \quad \mathbf{e}_j : j\text{-th unit vector}.$$

If $u \in H^k(\mathbb{R}^n)$ *and* $|\partial_{h,j}u|_k \le C$ *for all* $h>0$, $1\le j\le n$, *then* $u \in H^{k+1}(\mathbb{R}^n)$ *holds. Conversely,*

$$|\partial_{h,j}u|_k \le |u|_{k+1} \qquad \text{for all } u \in H^{k+1}(\mathbb{R}^n).$$

Proof. (a) $\widehat{u(\cdot+\delta\mathbf{e}_j)} = \exp(-\mathrm{i}\delta\xi_j)\hat{u}(\boldsymbol{\xi})$ shows $\widehat{\partial_{h,j}u}(\boldsymbol{\xi}) = \frac{2\mathrm{i}}{h}\sin(\xi_j\frac{h}{2})\hat{u}(\boldsymbol{\xi})$. Hence, since

$$4h^{-2}\sin^2\big(\xi_j\tfrac{h}{2}\big) \ge \xi_j^2 \qquad \text{for } h \le 1/|\boldsymbol{\xi}|,$$

the inequality $|\widehat{\partial_{h,j}u}|^2 \ge \xi_j^2\,|\hat{u}|^2$ follows for $|\boldsymbol{\xi}| \le 1/h$. Using $(1+|\boldsymbol{\xi}|^2)^{k+1} = (1+|\boldsymbol{\xi}|^2)^k + (1+|\boldsymbol{\xi}|^2)^k\,|\boldsymbol{\xi}|^2$, summation over j and integration over $\boldsymbol{\xi}$ then gives

$$\begin{aligned}
\big(|u|_{k+1}^\wedge\big)^2 &= \int_{\mathbb{R}^n} (1+|\boldsymbol{\xi}|^2)^{k+1}\,|\hat{u}(\boldsymbol{\xi})|^2\,\mathrm{d}\boldsymbol{\xi} \\
&= \int_{\mathbb{R}^n} (1+|\boldsymbol{\xi}|^2)^k\,|\hat{u}(\boldsymbol{\xi})|^2\,\mathrm{d}\boldsymbol{\xi} + \int_{\mathbb{R}^n} (1+|\boldsymbol{\xi}|^2)^k\,|\boldsymbol{\xi}|^2\,|\hat{u}(\boldsymbol{\xi})|^2\,\mathrm{d}\boldsymbol{\xi} \\
&\le \big[\,|u|_k^\wedge\big]^2 + \int_{|\boldsymbol{\xi}|\ge 1/h} (1+|\boldsymbol{\xi}|^2)^k\,|\boldsymbol{\xi}|^2\,|\hat{u}(\boldsymbol{\xi})|^2\,\mathrm{d}\boldsymbol{\xi} + \int_{|\boldsymbol{\xi}|\le 1/h} (1+|\boldsymbol{\xi}|^2)^k\,\Big|\widehat{\partial_{h,j}u}(\boldsymbol{\xi})\Big|^2\,\mathrm{d}\boldsymbol{\xi} \\
&\le \big[\,|u|_k^\wedge\big]^2 + \big[\,|\partial_{h,j}u|_k^\wedge\big]^2 + \int_{|\boldsymbol{\xi}|\ge 1/h} \dots .
\end{aligned}$$

The integral over $|\boldsymbol{\xi}| \ge 1/h$ vanishes for $h \to 0$ so that the statement follows.

(b) For the converse use $(\partial_{hj}u)(\mathbf{x}) = \int_{-1/2}^{1/2} u_{x_j}(\mathbf{x}+th\mathbf{e}_j)\mathrm{d}t$. ∎

6.2.4 $H^s(\Omega)$ for Real $s \geq 0$

Let $s \geq 0$. For $\Omega = \mathbb{R}^n$ one can define the following scalar product (6.21a) and the Sobolev norm (6.21b) for all $u \in C_0^\infty(\mathbb{R}^n)$:

$$(u,v)_s^\wedge := \int_{\mathbb{R}^n} (1+|\boldsymbol{\xi}|^2)^s \, \hat{u}(\boldsymbol{\xi}) \, \overline{\hat{v}(\boldsymbol{\xi})} \, \mathrm{d}\boldsymbol{\xi}, \tag{6.21a}$$

$$|u|_s^\wedge := \|(1+|\boldsymbol{\xi}|^2)^{s/2} \, \hat{u}(\boldsymbol{\xi})\|_{L^2(\mathbb{R}^n)}. \tag{6.21b}$$

The completion in $L^2(\mathbb{R}^n)$ also defines the Sobolev space $H^s(\mathbb{R}^n)$ for noninteger order s. On the basis of Lemma 6.41b and Exercise 6.1.6.11 the newly defined space $H^s(\mathbb{R}^n)$ for $s \in \mathbb{N}_0$ agrees with the Sobolev spaces used until this point.

Let $\Omega \subset \mathbb{R}^n$ (it may also be a surface; cf. §7.5). The noninteger number $s > 0$ can be decomposed as $s = k + \lambda$ with $k \in \mathbb{N}_0$ and $0 < \lambda < 1$. We define

$$(u,v)_s := \sum_{|\alpha| \leq k} \left[\int_\Omega D^\alpha u(\mathbf{x}) D^\alpha v(\mathbf{x}) \mathrm{d}\mathbf{x} \right. \tag{6.22a}$$

$$\left. + \iint_{\Omega\times\Omega} \frac{[D^\alpha u(\mathbf{x}) - D^\alpha u(\mathbf{y})]\,[D^\alpha v(\mathbf{x}) - D^\alpha v(\mathbf{y})]}{|\mathbf{x}-\mathbf{y}|^{n+2\lambda}} \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y} \right],$$

$$|u|_s := \|u\|_{H^s(\Omega)} := \sqrt{(u,u)_s} \qquad (s = k+\lambda,\ 0<\delta<1)\,. \tag{6.22b}$$

The norm $|\cdot|_s$ is called the *Sobolev–Slobodeckii norm*. One can define the Hilbert spaces $H^s(\Omega)$ and $H_0^s(\Omega)$ in the same way as in the case $s = k \in \mathbb{N}$. The properties of these spaces are summarised in the following theorem (cf. Adams [1], Wloka [308, §3]).

Theorem 6.43. *Let $s \geq 0$. (a) For $\Omega = \mathbb{R}^n$ the norms (6.21b) and (6.22b) are equivalent, i.e., both norms define the same space $H^s(\mathbb{R}^n) = H_0^s(\mathbb{R}^n)$.*

(b) $\{u \in C^\infty(\Omega) : \mathrm{supp}(u)$ compact, $|u|_s < \infty\}$ is dense in $H^s(\Omega)$.

(c) $C_0^\infty(\Omega)$ is dense in $H_0^s(\Omega)$.

(d) $aD^\alpha(bu) \in H^{s-|\alpha|}(\Omega)$, if $|\alpha| \leq s$, $u \in H^s(\Omega)$, $a \in C^{t-|\alpha|}(\overline{\Omega})$, $b \in C^t(\overline{\Omega})$, where $t = s \in \mathbb{N}_0$ or $t > s$.

(e) $H^s(\Omega) \subset H^t(\Omega)$, $H_0^s(\Omega) \subset H_0^t(\Omega)$ for $s \geq t$.

(f) In (6.16a,b) k and m can be noninteger real numbers.

(g) (6.17) holds with $\|T\|_{C^k(\overline{\Omega})}$ replaced by $\|T\|_{C^t(\overline{\Omega})}$ with $t > k$ if $k \notin \mathbb{N}$.

Exercise 6.44. Check, using the norm (6.21b), as well as the Sobolev–Slobodeckii norm (6.22b), that the characteristic function $u(x) = 1$ in $[-1,1]$, $u(x) = 0$ for $|x| > 1$ belongs to $H^s(\mathbb{R})$ if and only if $0 \leq s < 1/2$.

6.2.5 Trace and Extension Theorems

The restriction of a function to a lower-dimensional manifold is called its *trace* on this set.

The nature of boundary-value problems requires that one can form boundary values $u|_\Gamma$ (i.e., the trace of u on $\Gamma = \partial\Omega$) in a meaningful way. As can be seen easily, a Hölder-continuous function $u \in C^s(\overline{\Omega})$ has a restriction $u|_\Gamma \in C^s(\Gamma)$ if only Γ is sufficiently smooth. But, from $u \in H^s(\Gamma)$ does not necessarily follow $u|_\Gamma \in H^s(\Gamma)$. Since the equality $u = v$ on $H^s(\Omega)$ only means that $u(x) = v(x)$ almost everywhere in Ω, and Γ is a set of measure zero, $u(x) \neq v(x)$ may hold everywhere on Γ. Also, the boundary value $u(x)$ $(x \in \Gamma)$ cannot be defined by a continuous extension either since, for example, $u \in H^1(\Omega)$ need not be continuous (cf. Exercise 6.21c).

The inverse problem for the definition of the trace $u|_\Gamma$ is extension: does there exist, for a given boundary value φ on Γ, a function $u \in H^s(\Omega)$ such that φ and u coincide on Γ? If the answer is negative there exists no solution $u \in H^s(\Omega)$ to the Dirichlet boundary-value problem.

First we study these problems on the half-space $\mathbb{R}^n_+$:

$$\Omega = \mathbb{R}^n_+ := \{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_n > 0\} \quad \text{with } \Gamma = \partial\Omega = \mathbb{R}^{n-1} \times \{0\}. \tag{6.23}$$

Functions $u \in H^s(\mathbb{R}^n_+)$ will first be continued to $\bar{u} \in H^s(\mathbb{R}^n)$, and then $\bar{u}$ restricted to Γ.

Theorem 6.45 (extension operator). *Let $s \geq 0$. There exists an extension operator $\phi_s \in L(H^s(\mathbb{R}^n_+), H^s(\mathbb{R}^n))$ such that for all $u \in H^s(\mathbb{R}^n_+)$ the extension $\bar{u} = \phi_s u$ coincides with u on $\mathbb{R}^n_+$, i.e., $\bar{u}|_{\mathbb{R}^n_+} = u$.*

Proof. For $s = 0$, set $\bar{u} = \left\{ \begin{matrix} u & \text{on } \mathbb{R}^n_+ \\ \bar{u} = 0 & \text{otherwise} \end{matrix} \right\}$. Since $\|\bar{u}\|_{L^2(\mathbb{R}^n)} = \|u\|_{L^2(\mathbb{R}^n_+)}$, the mapping defined by $\phi_0 u := \bar{u}$ is bounded: $\|\phi_0\|_{L^2(\mathbb{R}^n) \leftarrow L^2(\mathbb{R}^n_+)} = 1$.

For $s \leq 1$ define $\phi_s u = \bar{u}$ by reflection on Γ:

$$\bar{u} = u \text{ on } \mathbb{R}^n_+, \quad \bar{u}(x_1, \ldots, x_{n-1}, -x_n) = \bar{u}(x_1, \ldots, x_{n-1}, x_n) \text{ for } x_n > 0\,.$$

For $s = 1$ one obtains, e.g., $|\bar{u}|_1 = \sqrt{2}\,|u|_1$, i.e., $\phi_s \in L(H^s(\mathbb{R}^n_+), H^s(\mathbb{R}^n))$.

For larger s one uses higher interpolation formulae for $\bar{u}(\ldots, -x_n)$ (cf. Exercise 9.16 and Wloka [308, page 101]). ∎

In the following the restriction $u|_\Gamma$ is written in the form γu. At first, the operator γ is defined only on $C_0^\infty(\mathbb{R}^n)$:

$$\gamma : C_0^\infty(\mathbb{R}^n) \to C_0^\infty(\Gamma) \subset L^2(\mathbb{R}^{n-1}), \quad (\gamma u)(\mathbf{x}) := u(\mathbf{x}) \text{ for all } \mathbf{x} \in \Gamma\,. \tag{6.24}$$

We write $\mathbf{x} = (\mathbf{x}', x_n)$ with $\mathbf{x}' = (x_1, \ldots, x_{n-1}) \in \mathbb{R}^{n-1}$. The boundary $\Gamma = \{(\mathbf{x}', 0) : \mathbf{x}' \in \mathbb{R}^{n-1}\}$ is identified with $\mathbb{R}^{n-1}$: $H^s(\Gamma) = H^s(\mathbb{R}^{n-1})$.

Theorem 6.46. *Let $s > \frac{1}{2}$. Then the trace operator γ from (6.24) can be continued to $\gamma \in L(H^s(\mathbb{R}^n), H^{s-\frac{1}{2}}(\mathbb{R}^{n-1}))$. Thus we have in particular*

$$|\gamma u|_{s-\frac{1}{2}} \le C_s \, |u|_s \qquad \text{for } u \in H^s(\mathbb{R}^n).$$

In the case $n = 1$, i.e., $\gamma u = u(0)$, we have $|\gamma u| \le C_s \, |u|_s$.

Proof. It suffices to show $|\gamma u|_{s-1/2} \le C'_s \, |u|_s$ for $u \in C_0^\infty(\mathbb{R}^n)$ (cf. Theorems 6.10 and 6.43a). Let the Fourier transforms of $u \in C_0^\infty(\mathbb{R}^n)$ and $w := \gamma u \in C_0^\infty(\mathbb{R}^{n-1})$ be $\hat{u} = \mathcal{F}_n u$ and $\hat{w} = \mathcal{F}_{n-1} w$ ($\mathcal{F}_k$: k-dimensional Fourier transform). $\mathcal{F}_n$ can be written as product $\mathcal{F}_1 \circ \mathcal{F}_{n-1}$, where $\mathcal{F}_{n-1}$ acts on $\mathbf{x}' \in \mathbb{R}^{n-1}$ and $\mathcal{F}_1$ on x_n. Therefore $\hat{W}(\cdot, x_n) := \mathcal{F}_{n-1} u(\cdot, x_n)$ has the properties $\hat{u}(\boldsymbol{\xi}', \cdot) = \mathcal{F}_1 \hat{W}(\boldsymbol{\xi}', \cdot)$ and $\hat{w} = \hat{W}(\cdot, 0)$. According to Lemma 6.38,

$$\hat{w}(\boldsymbol{\xi}') = \hat{W}(\boldsymbol{\xi}', 0) = [\mathcal{F}_1^{-1} \hat{u}(\boldsymbol{\xi}', \cdot)]|_{x_n = 0}$$

has the representation

$$\hat{w}(\boldsymbol{\xi}') = (2\pi)^{-1/2} \int_{\mathbb{R}^n} \hat{u}(\boldsymbol{\xi}', \xi_n) \, \mathrm{d}\xi_n \qquad \text{for all } \boldsymbol{\xi}' \in \mathbb{R}^{n-1}. \tag{6.25}$$

Since $u \in H^s(\mathbb{R}^n)$, $\hat{U}(\boldsymbol{\xi}', \xi_n) := (1 + \left|\boldsymbol{\xi}'\right|^2 + \xi_n^2)^{\frac{s}{2}} \hat{u}(\boldsymbol{\xi}', \xi_n)$ lies in $L^2(\mathbb{R}^n)$ (cf. Lemma 6.41b). Inequality (6.11a) yields

$$2\pi \left|\hat{w}(\boldsymbol{\xi}')\right|^2 = \left| \int_{\mathbb{R}^n} \hat{u}(\boldsymbol{\xi}', \xi_n) \mathrm{d}\xi_n \right|^2 \le \int_{\mathbb{R}^n} (1 + \left|\boldsymbol{\xi}'\right|^2 + \xi_n^2)^{-s} \mathrm{d}\xi_n \int_{\mathbb{R}^n} |\hat{U}(\boldsymbol{\xi}', \xi_n)|^2 \mathrm{d}\xi_n.$$

The first integral has the value $K_s (1 + |\boldsymbol{\xi}'|^2)^{\frac{1}{2} - s}$ with $K_s = \int_{\mathbb{R}} \frac{\mathrm{d}x}{(1+x^2)^s} < \infty$, since $s > 1/2$. The second is the square of $V(\boldsymbol{\xi}') := \|\hat{U}(\boldsymbol{\xi}', \cdot)\|_{L^2(\mathbb{R})} \in L^2(\mathbb{R}^{n-1})$ because $\|V\|_{L^2(\mathbb{R}^{n-1})} = \|\hat{U}\|_{L^2(\mathbb{R}^n)}$ (Fubini's theorem). Together we have:

$$\left(1 + \left|\boldsymbol{\xi}'\right|^2\right)^{s-1/2} \left|\hat{w}(\boldsymbol{\xi}')\right|^2 \le \frac{K_s}{2\pi} V(\boldsymbol{\xi}')^2 \qquad \text{for } \boldsymbol{\xi}' \in \mathbb{R}^{n-1}.$$

Integration over $\boldsymbol{\xi}' \in \mathbb{R}^{n-1}$ results in

$$\begin{aligned} \int_{\mathbb{R}^{n-1}}^{s-\frac{1}{2}} \left(1 + \left|\boldsymbol{\xi}'\right|^2\right) \left|\hat{w}(\boldsymbol{\xi}')\right|^2 \mathrm{d}\boldsymbol{\xi}' &\le \frac{K_s}{2\pi} |V|_0^2 \;=\; \frac{K_s}{2\pi} |\hat{U}|_0^2 \\ &= \frac{K_s}{2\pi} \int_{\mathbb{R}^n} \left(1 + |\boldsymbol{\xi}|^2\right)^s |\hat{u}(\boldsymbol{\xi})|^2 \, \mathrm{d}\boldsymbol{\xi}. \end{aligned}$$

Thus $|w|_{s-1/2}^\wedge \le C'_s \, |u|_s^\wedge$ is proved with $C'_s = \sqrt{K_s/(2\pi)}$ (cf. (6.21b)). If $n = 1$, $\hat{w}(\boldsymbol{\xi}')$ already represents $\gamma u = u(0)$, and the integration over $\boldsymbol{\xi}'$ is not required. ∎

Theorem 6.46 describes the restriction $u(\cdot, 0) = \gamma u$ to $x_n = 0$. Evidently, similarly we have $|u(\cdot, x_n)|_{s-1/2} \le C_s \, |u|_s$ for any other $x_n \in \mathbb{R}$ with the same

constant C_s. The mapping $x_n \mapsto u(\cdot, x_n)$ is continuous [resp. Hölder- or Lipschitz-continuous] in the following sense.

Theorem 6.47. *For $s > 1/2$ the following statements hold:*

$$\lim_{y_n \to x_n} \|u(\cdot, x_n) - u(\cdot, y_n)\|_{H^{s-\frac{1}{2}}(\mathbb{R}^{n-1})} = 0 \quad \textit{for all } x_n \in \mathbb{R},\ u \in H^s(\mathbb{R}^n), \tag{6.26}$$

$$\|u(\cdot, x_n) - u(\cdot, y_n)\|_{H^{s-1/2}(\mathbb{R}^{n-1})} \le K_{s,\lambda} |x_n - y_n|^\lambda \|u\|_{H^s(\mathbb{R}^n)}$$
$$\textit{for } u \in H^s(\mathbb{R}^n),\ 0 \le \lambda < 1 \ (\textit{and for } \lambda = 1 \textit{ if } s > 3/2).$$

Proof. (i) Let $u_\nu \in C_0^\infty(\mathbb{R}^n)$ be a sequence with $u_\nu \to u \in H^s(\mathbb{R}^n)$ and set $\varphi_\nu(x) := \|u_\nu(\cdot, x)\|_{H^{s-1/2}(\mathbb{R}^{n-1})}$. The function φ_ν is continuous in $\mathbb{R}$ and converges uniformly to $\|u(\cdot, x)\|_{H^{s-1/2}(\mathbb{R}^{n-1})}$ since $|u_\nu(\cdot, x) - u(\cdot, x)|_{s-1/2} \le C_s |u_\nu - u|$ for all $x \in \mathbb{R}$. Thus (6.26) follows.

(ii) $u_\varepsilon(\cdot, x_n) := u(\cdot, x_n + \varepsilon) - u(\cdot, x_n)$ has the Fourier transform

$$\hat{u}_\varepsilon(\boldsymbol{\xi}) = \hat{u}_\varepsilon(\boldsymbol{\xi}', \xi_n) = [\exp(\mathrm{i}\xi_n \varepsilon) - 1]\, \hat{u}(\boldsymbol{\xi}),$$

so that $|\hat{u}_\varepsilon(\boldsymbol{\xi})|^2 = 4 \sin^2(\xi_n \varepsilon/2)\ |\hat{u}(\boldsymbol{\xi})|^2$. As in the proof for Theorem 6.46 set $W(\cdot, x_n) := \mathcal{F}_{n-1} u_\varepsilon(\cdot, x_n)$, $\hat{w} = \hat{W}(\cdot, 0)$. The first integral in the estimate of $2\pi |\hat{w}(\boldsymbol{\xi}')|^2$ now reads

$$\int_{\mathbb{R}^n} (1 + |\boldsymbol{\xi}'|^2 + \xi_n^2)^{-s} \sin^2(\xi_n \tfrac{\varepsilon}{2})\, \mathrm{d}\xi_n = (1 + |\boldsymbol{\xi}'|^2)^{\frac{1}{2} - s} \int_{\mathbb{R}^n} \left(1 + t^2\right)^{-s} \sin^2(\eta t) \mathrm{d}t$$

with $\eta = \frac{\varepsilon}{2}(1 + |\boldsymbol{\xi}'|^2)^{1/2}$. Decomposing the last integral into subintegrals over $|t| \le 1/\eta$ and $|t| \ge 1/\eta$ shows

$$\int_{\mathbb{R}} (1 + t^2)^{-s} \sin^2(\eta t) \mathrm{d}t \le C_{s,\lambda} \eta^{2\lambda}.$$

The remainder of the argument follows the same lines as in the proof of Theorem 6.46. ∎

Up to this point we have obtained $H^s(\mathbb{R}^n)$ by completion of $C_0^\infty(\mathbb{R}^n)$ in $L^2(\mathbb{R}^n)$. The next theorem shows that for sufficiently large s one can also complete in $C^0(\mathbb{R}^n) \cap L^2(\mathbb{R}^n)$ so that $H^s(\mathbb{R}^n)$ contains only classical functions (i.e., continuous, Hölder-continuous, [Hölder] continuously differentiable functions) (cf. Sobolev [266, §I.8]).

Theorem 6.48 (Sobolev's embedding). *$H^s(\mathbb{R}^n) \subset C^k(\mathbb{R}^n)$ holds for $k \in \mathbb{N}_0$, $s > k + \frac{n}{2}$ and $H^s(\mathbb{R}^n) \subset C^t(\mathbb{R}^n)$ for $0 < t \notin \mathbb{N}$, $s \ge t + \frac{n}{2}$.*

Proof. (i) Let $s \ge t + n/2$, $0 < t < 1$, $u \in H^s(\mathbb{R}^n)$. For given $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ we want to show $|u(\mathbf{x}) - u(\mathbf{y})| \le C |\mathbf{x} - \mathbf{y}|^t$ with C independent of $\mathbf{x}, \mathbf{y}$. The coordinates of the $\mathbb{R}^n$ can be rotated so that $\mathbf{x} = (x_1, 0, \ldots, 0)$, $\mathbf{y} = (y_1, 0, \ldots, 0)$. An $(n-1)$-fold application of Theorem 6.46 to $u(\cdot)$, $u(\cdot, 0)$, $u(\cdot, 0, 0)$, etc., results

in $u(\cdot,0,0,\ldots,0) \in H^{s-(n-1)/2}(\mathbb{R})$. Theorem 6.47 provides the desired estimate with $C = K_{s-(n-1)/2,t}\,|u(\cdot,0,\ldots,0)|_{s-(n-1)/2} \le C'\,|u|_s$; thus $u \in C^t(\mathbb{R}^n)$ and $\|\cdot\|_{C^t(\mathbb{R}^n)} \le C\,\|\cdot\|_{H^s(\mathbb{R}^n)}$. Then $u \in C^0(\mathbb{R}^n)$ follows from $C^t(\mathbb{R}^n) \subset C^0(\mathbb{R}^n)$.

(ii) Let $s \ge t + n/2$, $1 < t < 2$, $u \in H^s(\mathbb{R}^n)$. Part (a) is applicable to $D^\alpha u \in H^{s-1}(\mathbb{R}^n)$ (cf. Theorem 6.43d,e) with $|\alpha| \le 1 : D^\alpha u \in C^{t-1}(\mathbb{R}^n)$ for $|\alpha| \le 1$. Thus $u \in C^t(\mathbb{R}^n)$, etc. ∎

We return to the statements of Theorems 6.45 and 6.46. For all $u \in C_0^\infty(\mathbb{R}^n_+)$ the restriction $\gamma u = u(\cdot,0)$ agrees with $\gamma\phi_s u$. Completion in $H^s(\mathbb{R}^n_+)$ yields

$$\|\gamma\|_{H^{s-1/2}(\mathbb{R}^{n-1})\leftarrow H^s(\mathbb{R}^n_+)} \le \|\gamma\|_{H^{s-1/2}(\mathbb{R}^{n-1})\leftarrow H^s(\mathbb{R}^n)}\,\|\phi_s\|_{H^s(\mathbb{R}^n)\leftarrow H^s(\mathbb{R}^n_+)}\,.$$

This proves the following corollary.

Corollary 6.49. Let $s > 1/2$. $\gamma \in L(H^s(\mathbb{R}^n_+), H^{s-1/2}(\mathbb{R}^{n-1}))$ holds for the restriction $\gamma u := u(\cdot,0)$.

With the restriction to $x_n = 0$ one evidently loses half an order of differentiability. Conversely one gains half an order if one continues $w \in H^{s-1/2}(\mathbb{R}^{n-1})$ suitably in $\mathbb{R}^n$.

Theorem 6.50. *Let $s > 1/2$ and $w \in H^{s-1/2}(\mathbb{R}^{n-1})$. There exists a function $u \in H^s(\mathbb{R}^n)$ [or $u \in H^s(\mathbb{R}^n_+)$] such that $|u|_s \le C_s\,|w|_{s-1/2}$, and $\gamma u = w$, i.e., $w = u(\cdot,0)$.*

Proof. Let $\hat{u} = \mathcal{F}_n u$ and $\hat{w} = \mathcal{F}_{n-1} w$ be the Fourier transforms. $\gamma u = w$ is equivalent to (6.25). For

$$\hat{u}(\boldsymbol{\xi}) = \hat{u}(\boldsymbol{\xi}', \xi_n) := \hat{w}(\boldsymbol{\xi}')\,\frac{(1+|\boldsymbol{\xi}'|^2)^{s-1/2}}{K_s(1+|\boldsymbol{\xi}'|^2+\xi_n^2)^s}\,, \quad K_s = \int_{\mathbb{R}^n} \left(1+t^2\right)^{-s} \mathrm{d}t\,,$$

one checks that (6.25) and $|u|_s^\wedge = K_s^{-1/2}\,|w|_{s-1/2}^\wedge$ hold. Restriction of $u \in H^s(\mathbb{R}^n)$ to $\mathbb{R}^n_+$ proves the parenthetical addition. ∎

If one replaces $\mathbb{R}^n_+$ with a general domain $\Omega \subset \mathbb{R}^n$, then $\mathbb{R}^{n-1} \cong \mathbb{R}^{n-1} \times \{0\} = \partial\mathbb{R}^n_+$ becomes $\Gamma = \partial\Omega$, and the necessity arises of defining the Sobolev space $H^s(\Gamma)$. We begin with the following definition.

Definition 6.51. Let $0 < t \in \mathbb{R} \cup \{\infty\}$ [resp. $k \in N_0$]. We write $\Omega \in C^t$ [resp. $\Omega \in C^{k,1}$], if for every $x \in \Gamma := \partial\Omega$ there exists a neighbourhood $U \subset \mathbb{R}^n$ such that there exists a bijective mapping $\phi : U \to K_1(0) = \{\boldsymbol{\xi} \in \mathbb{R}^n : |\boldsymbol{\xi}| < 1\}$ with

$$\phi \in C^t(\bar{U}),\ \phi^{-1} \in C^t(\overline{K_1(0)}) \quad [\,\phi \in C^{k,1}(\bar{U}),\ \phi^{-1} \in C^{k,1}(\overline{K_1(0)})\,], \tag{6.27a}$$

$$\phi(U \cap \Gamma) = \{\boldsymbol{\xi} \in K_1(0) : \xi_n = 0\}, \tag{6.27b}$$

$$\phi(U \cap \Omega) = \{\boldsymbol{\xi} \in K_1(0) : \xi_n > 0\}, \tag{6.27c}$$

$$\phi(U \cap (\mathbb{R}^n \backslash \Omega)) = \{\boldsymbol{\xi} \in K_1(0) : \xi_n < 0\}. \tag{6.27d}$$

Here $K_1(0)$ is a ball if $|\cdot|$ is the Euclidean norm. For the maximum norm $|\cdot|_\infty$, $K_1(0)$ is a cube. Likewise, $K_1(0)$ can be replaced by another ball $K_R(\mathbf{z})$ or any cuboid $(x_1', x_2'') \times \ldots \times (x_n', x_n'')$.

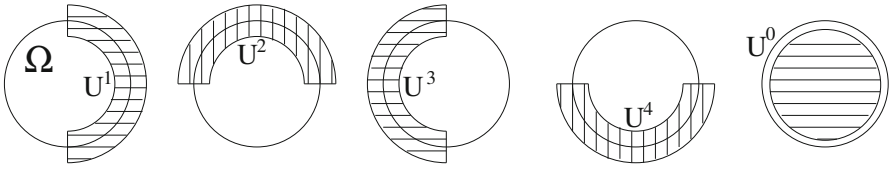


Fig. 6.1 Covering neighbourhoods of Γ and Ω.

Example 6.52. Let Ω be the circle $K_1(0) \subset \mathbb{R}^2$. A neighbourhood of $\mathbf{x}^\star = (1,0)$ is U^1 from Figure 6.1. The mapping $\mathbf{x} \in U^1 \mapsto \phi(\mathbf{x}) := \boldsymbol{\xi} \in (-1,1) \times (-1,1)$ with

$$x_1 = (1 - \xi_2/2)\cos(\pi\xi_1/2), \quad x_2 = (1 - \xi_2/2)\sin(\pi\xi_1/2)$$

is bijective and satisfies (6.27a–d) with $t = \infty$. The same holds for any $\mathbf{x} \in \Gamma$. Thus $\Omega \in C^\infty$.

Exercise 6.53. (a) The rectangle $\Omega = (x_1', x_1'') \times (x_2', x_2'')$ and the L-shaped domain from Example 2.4 are domains in $C^{0,1}$, also called *Lipschitz domains*.

(b) The cut circle in Figure 5.2b does not belong to $C^{0,1}$.

Lemma 6.54. *Let $\Omega \in C^t$ $[\Omega \in C^{k,1}]$ be a bounded domain. Then there exists $N \in \mathbb{N}$, U^i $(0 \le i \le N)$, U_i, α_i $(1 \le i \le N)$ with*

$$U^i \text{ open, bounded } (0 \le i \le N), \quad \bigcup_{i=0}^{N} U^i \supset \overline{\Omega}, \quad U^0 \subset\subset \Omega, \tag{6.28a}$$

$$U_i := U^i \cap \Gamma \ (1 \le i \le N), \qquad \bigcup_{i=0}^{N} U_i = \Gamma, \tag{6.28b}$$

$$\alpha_i : U_i \to \alpha_i(U_i) \subset \mathbb{R}^{n-1} \text{ bijective for all } i = 1, \ldots, N, \tag{6.28c}$$

$$\alpha_i \circ \alpha_j^{-1} \in C^t(\overline{\alpha_j(U_i \cap U_j)}) \quad [\text{resp. } \alpha_i \circ \alpha_j^{-1} \in C^{k,1}(\overline{\alpha_j(U_i \cap U_j)})]. \tag{6.28d}$$

On U^i $(1 \le i \le N)$ are defined mappings ϕ_i with the properties (6.27a–d).

Proof. For every $\mathbf{x} \in \Gamma$ there exist $U = U(\mathbf{x})$ and $\phi = \phi_x$ according to Definition 6.51. Let α_x be the restriction of ϕ_x to $U(\mathbf{x}) \cap \Gamma$. Let V^i $(i \in \mathbb{N})$ be the open sets $\{\mathbf{x} \in \Omega : \operatorname{dist}(\mathbf{x}, \Gamma) > 1/i\} \subset\subset \Omega$. Let $\bigcup_{\mathbf{x}\in\Gamma} U(\mathbf{x}) \cup \bigcup_i V^i$ be an open covering of the compact set $\overline{\Omega}$. Therefore there exists a finite covering through $U^i := U(\mathbf{x}_i)$ $(1 \le i \le N)$ and at most one V^j which is denoted by U^0. If one sets $U_i := U^i \cap \Gamma$, $\alpha_i = \alpha_{x_i}$, and $\phi_i = \phi_{x_i}$, the statements follow from (6.28a–d). ∎

A set of pairs $\{(U_i, \alpha_i) : 1 \le i \le N\}$ fulfilling the conditions (6.28b–d) is called a C^t- respectively $C^{k,1}$-coordinate system for Γ.

In Example 6.52 one has $N = 4$. The maps inverse to α_1 and α_4 are given by $\alpha_1^{-1}(\xi_1) = (\cos\frac{\pi\xi_1}{2}, \sin\frac{\pi\xi_1}{2}) \in U_1$, $\alpha_4^{-1}(\xi_1) = (\cos\frac{(3+\xi_1)\pi}{2}, \sin\frac{(3+\xi_1)\pi}{2}) \in U_4$, where in each case $-1 < \xi_1 < 1$. On $\alpha_1(U_1 \cap U_4)$ one obtains $\alpha_4(\alpha_1^{-1}(\xi_1)) = \xi_1 + 1$.

Lemma 6.55 (partition of unity). *Let $\{U^i : 0 \le i \le N\}$ satisfy (6.28a). There exist functions $\sigma_i \in C_0^\infty(\mathbb{R}^n)$, $0 \le i \le N$, with*

$$\operatorname{supp}(\sigma_i) \subset U^i, \qquad \sum_{i=0}^{N} \sigma_i^2(\mathbf{x}) = 1 \quad \textit{for all } \mathbf{x} \in \overline{\Omega}\,. \tag{6.29}$$

The general construction of the σ_i can be found, for example, in Wloka [308, §1.2]. In the special case of Figure 6.1 one may proceed as follows. Let $\sigma(t) := 0$ for $|t| \ge 1$ and $\sigma(t) := \exp(1/(t^2-1))$ for $t \in (-1,1)$. Then $\sigma \in C_0^\infty(\mathbb{R})$ and $\operatorname{supp}(\sigma) = [-1,1]$. In U^i from Figure 6.1 one defines, for example,

$$\psi_0(\mathbf{x}) := \sigma\left(\tfrac{9}{4}\,|\mathbf{x}|^2\right), \quad \psi_1(\mathbf{x}) := \sigma(2r-2)\sigma(2\varphi/\pi) \text{ for } \mathbf{x} = r\begin{pmatrix}\cos\varphi\\ \sin\varphi\end{pmatrix}, \quad \text{etc.}$$

Then the functions $\sigma_i(\mathbf{x}) := \psi_i(\mathbf{x})/\sqrt{\sum \psi_i^2(\mathbf{x})}$ satisfy (6.29).

A function u on Γ can be written in the form $\sum \sigma_i^2 u$. Each term $\sigma_i^2 u$ is parametrisable over $\alpha_i(U_i) \subset \mathbb{R}^{n-1}$: $(\sigma_i^2 u) \circ \alpha_i^{-1} : \alpha_i(U_i) \subset\subset \mathbb{R}^{n-1} \to \mathbb{R}$. This enables the following definition.

Definition 6.56. Let $\Omega \in C^t$ $[\in C^{k,1}]$. Assume (U_i, α_i) and σ_i satisfy (6.28b–d) and (6.29). Let $s \le t \in \mathbb{N}$ $[s \le k+1]$ or $s < t \notin \mathbb{N}$, $t > 1$. The Sobolev space $H^s(\Gamma)$ is defined as the set of all functions $u : \Gamma \to \mathbb{R}$ such that $(\sigma_i u) \circ \alpha_i^{-1} \in H_0^s(\mathbb{R}^{n-1})$ $(1 \le i \le N)$.

Theorem 6.57. *(a) $H^s(\Gamma)$ is a Hilbert space with the scalar product*

$$(u,v)_s := (u,v)_{H^s(\Gamma)} := \sum_{i=1}^{N} \left((\sigma_i u) \circ \alpha_i^{-1}, (\sigma_i v) \circ \alpha_i^{-1}\right)_{H^s(\mathbb{R}^{n-1})}.$$

(b) If $\{(\tilde{U}_i, \tilde{\alpha}_i) : 1 \le i \le N\}$ is another C^t- $[C^{k,1}$-$]$ coordinate system of Γ and $\{\tilde{\sigma}_i\}$ another partition of unity, then the space $\tilde{H}^s(\Gamma)$ defined by this is equal to $H^s(\Gamma)$ as a set. The norms of $H^s(\Gamma)$ and $\tilde{H}^s(\Gamma)$ are equivalent.

Proof of (b). Use the transformation Theorem 6.35 [resp. 6.43g]. For $\Omega \in C^{0,1}$ we refer to Wloka [308, Lemma 4.5]. ∎

The trace and extension theorems (Corollary 6.49 and Theorem 6.50) can be extended to any domain with a sufficiently smooth boundary. γ now denotes the restriction to $\Gamma = \partial\Omega$: $\gamma u = u|_\Gamma$.

Theorem 6.58. *Let $\Omega \in C^t$ with $1/2 < s < t \in \mathbb{N}$ or $1/2 < s < t$ [resp. $\Omega \in C^{k,1}$, $1/2 < s = k+1 \in \mathbb{N}$].*

(a) The trace of $u \in H^s(\Omega)$ belongs to $H^{s-1/2}(\Gamma)$: $\gamma \in L(H^s(\Omega), H^{s-1/2}(\Gamma))$.

(b) For each $w \in H^{s-\frac{1}{2}}(\Gamma)$ there exists an extension $u \in H^s(\Omega)$ with $w = \gamma u$, $|u|_s \le C_s |w|_{s-\frac{1}{2}}$.

(c) For each $w \in H^s(\Omega)$ there exists a continuation $Ew \in H^s(\mathbb{R}^n)$ with $w = \gamma Ew$: $E \in L(H^s(\Omega), H^s(\mathbb{R}^n))$.

Proof. The proofs follow the same pattern. Let U^i, U_i and ϕ_i, α_i be as in Lemma 6.54. The term $u_i = \sigma_i^2 u$ from $u = \sum \sigma_i^2 u$ [resp. $w_i = \sigma_i^2 w$] has its support in U^i [resp. U_i] and can be mapped via ϕ_i (resp. α_i) onto $\mathbb{R}^n_+$ [resp. $\mathbb{R}^{n-1} \cong \mathbb{R}^{n-1} \times \{0\} = \partial\mathbb{R}^n$]. There Corollary 6.49 and Theorem 6.50 hold. The restriction to $\mathbb{R}^{n-1}$ [resp. continuation to $\mathbb{R}^n_+$ or $\mathbb{R}^n$] can be mapped back again. The first statement of the theorem is proved in detail as follows.

Let $u_i := \sigma_i^2 u$ and $\tilde{u}_i := u_i \circ \phi_i^{-1}$. By Theorems 6.35, 6.43g the function $\tilde{u}_i$ belongs to $H^s(\mathbb{R}^n_+)$. Thus the restriction $\gamma_+ \tilde{u}_i := u_i(\cdot, 0)$ lies in $H^{s-1/2}(\mathbb{R}^{n-1})$ (cf. Corollary 6.49) and has $\alpha_i(U_i)$ as support. Set $w_i := (\gamma_+ \tilde{u}_i) \circ \alpha_i$ on U_i, $w_i := 0$ on $\Gamma \backslash U_i$. According to Definition 6.56, $w := \sum w_i$ belongs to $H^{s-1/2}(\Gamma)$. Since α_i represents the restriction of ϕ_i on U_i one finds for all $u \in C^t(\overline{\Omega})$:

$$w = \sum_i w_i = \sum_i \gamma_+ \tilde{u}_i(\alpha_i) = \sum_i \tilde{u}_i(\phi_i) = \sum_i \left(\sigma_i^2 u\right) = u \qquad \text{on } \Gamma.$$

Note that $C^t(\overline{\Omega})$ is dense in $H^s(\Omega)$. Since all the partial mappings are bounded, one finds that $|\gamma u|_{s-1/2} = |w|_{s-1/2} \le C_s |u|_s$. ∎

Remark 6.59. (a) Under the conditions of Theorem 6.58 and the additional condition $s > |\alpha| + 1/2$ there exists a restriction $\gamma D^\alpha u \in H^{s-|\alpha|-1/2}(\Gamma)$ of the derivative $D^\alpha u$ of $u \in H^s(\Omega)$.

(b) For each $u \in H_0^s(\Omega)$ with $s < \ell + 1/2$ one has $\gamma D^\alpha u = 0$ if $|\alpha| \le \ell$.

Theorem 6.60. *For $\Omega \in C^{0,1}$ holds $H_0^1(\Omega) = \{u \in H^1(\Omega) : u|_\Gamma = 0\}$.*

Proof. (a) $u \in H_0^1(\Omega)$ is the limit of $u_\nu \in C_0^\infty(\Omega)$, all of which satisfy $u_\nu|_\Gamma = 0$. Since $u_\nu|_\Gamma \to u|_\Gamma$ in $H^{1/2}(\Gamma)$, it follows that $u \in H^1(\Omega)$ and $u|_\Gamma = 0$.

(b) Conversely, let $u \in H^1(\Omega)$ and $u|_\Gamma = 0$. The proof that $u \in H_0^1(\Omega)$ can be divided up as follows:

(ba) By using the partition of unity over $\{U^i : 0 \le i \le N\}$ (cf. Lemma 6.55), the statement reduces to the case $\Omega = \mathbb{R}^n_+$.

(bb) Without loss of generality one can assume $n = 1$: $\Omega = \mathbb{R}_+$.

(bc) For each $\eta > 0$ there exists $\varphi_\eta \in C^\infty(\mathbb{R}_+)$ with

$$\varphi_\eta(x) = 0 \text{ for } x \le \frac{\eta}{2}, \quad \varphi_\eta(x) = 1 \text{ for } x \ge \eta, \quad \left|\varphi_\eta'(x)\right| \le \frac{3}{\eta}, \quad 0 \le \varphi_\eta(x) \le 1.$$

According to Remark 6.32, there exists a $\tilde{u}_\eta \in C^\infty(\mathbb{R}_+)$ with finite support such that $|u-\tilde{u}_\eta|_1 \le \eta$. The function $u_\eta := \tilde{u}_\eta - (1-\varphi_\eta)\,\tilde{u}_\eta(0)$ satisfies $u_\eta \in C^\infty(\mathbb{R}_+)$, $u_\eta(0)=0$, and $\operatorname{supp}(u_\eta)$ is bounded. Since

$$|\tilde{u}_\eta(0)| = |\tilde{u}_\eta(0)-u(0)| \le \mathcal{O}(|u-\tilde{u}_\eta|_1) = \mathcal{O}(\eta) \qquad \text{(cf. Corollary 6.49)}$$

and $|1-\varphi_\eta|_1 = \mathcal{O}(\eta^{-1/2})$, we have $|u_\eta-\tilde{u}_\eta|_1 = \mathcal{O}(\eta^{1/2})$ and therefore $|u-u_\eta|_1 \to 0$. Thus $X := \{v \in C^\infty(\mathbb{R}_+) : v(0)=0,\ \operatorname{supp}(v) \subset\subset \mathbb{R}_+\}$ is dense in $\{u \in H^1(\mathbb{R}_+) : u(0)=0\}$.

(bd) The statement would be proved if $C_0^\infty(\mathbb{R}_+)$ were also dense in X with respect to $|\cdot|_1$. Let $v \in X$. Evidently, $v_\eta := \varphi_\eta v \in C_0^\infty(\mathbb{R}_+)$ holds for all $\eta > 0$. Set $\Theta(\eta) := \|v'\|_{L^2(0,\eta)}$ and note that $\Theta(\eta) \to 0$ for $\eta \to 0$. Since $v(x) = v_\eta(x)$ for $x \ge \eta$, it remains to estimate $\|v-v_\eta\|_{L^2(0,\eta)}$ and $\|v'-v_\eta'\|_{L^2(0,\eta)}$. Because of $v_\eta'-v' = \varphi_\eta' v + (\varphi_\eta - 1)v'$, one obtains

$$\begin{aligned}\|v_\eta'-v'\|_{L^2(0,\eta)} &\le \|\varphi_\eta'\|_{L^\infty(0,\eta)}\|v\|_{L^2(0,\eta)} + \|\varphi_\eta-1\|_{L^\infty(0,\eta)}\|v'\|_{L^2(0,\eta)}\\ &\le \tfrac{3}{\eta}\|v\|_{L^2(0,\eta)} + \Theta(\eta).\end{aligned}$$

Since $|v(x)| \le \int_0^x v'(\zeta)\mathrm{d}\zeta$, (6.11b) implies the estimate $|v(x)| \le \sqrt{\eta}\,\Theta(\eta)$ for all $0 \le x \le \eta$ and hence $\|v\|_{L^2(0,\eta)} \le \eta\,\Theta(\eta)$. Then the statement follows from $|v-v_\eta|_1^2 \le \|v-v_\eta\|_{L^2(0,\eta)}^2 + \|v'-v_\eta'\|_{L^2(0,\eta)}^2 \le C\Theta^2(\eta) \to 0$. ∎

When $\Omega \in C^1$, the normal direction $\mathbf{n}$ exists at all boundary points. Analogously to Theorem 6.60 one proves the following corollary.

Corollary 6.61. For $\Omega \in C^1$ and $k \in \mathbb{N}$ holds

$$\begin{aligned}H_0^k(\Omega) &= \left\{u \in H^k(\Omega) : \partial^\ell u/\partial n^\ell|_\Gamma = 0 \ \text{ for all } 0 \le \ell \le k-1\right\}\\ &= \left\{u \in H^k(\Omega) : D^\alpha u|_\Gamma = 0 \ \text{ for all } \ 0 \le |\alpha| \le k-1\right\}.\end{aligned}$$

Corollary 6.62. Set $\omega_h := \{\mathbf{x} \in \Omega : \operatorname{dist}(\mathbf{x},\partial\Omega) \le h\}$ for $s > \frac{1}{2}$ with $s - \frac{1}{2} \notin \mathbb{N}$ and $h > 0$. Then

$$\|u\|_{L^2(\omega_h)} \le C_s h^s \|u\|_{H^1(\Omega)} \ \text{ for all } u \in H_0^s(\Omega).$$

Proof. Using the partition of unity, it is sufficient to show the statement for $\mathbb{R}_+^n$. This reduces to the case $\Omega = (0,\infty)$. Since $C^{s-1/2}(\Omega) \subset H_0^s(\Omega)$ (Theorem 6.48), $|u(x)| = |u(x)-u(0)| \le |x|^{s-1/2}$ holds for $1/2 < s < 3/2$. Integration of $|u(x)|^2$ over $[0,h]$ yields the desired result. For $s > 3/2$, treat first the derivatives analogously. ∎

6.3 Dual Spaces

6.3.1 Dual Space of a Normed Space

Let X be a normed, linear space over $\mathbb{R}$. As a *dual space*, X' denotes the space of all bounded, linear mappings of X onto $\mathbb{R}$:

$$X' = L(X, \mathbb{R}).$$

According to Exercise 6.8, X' is a Banach space with the norm (*dual norm*)

$$\|x'\|_{X'} := \|x'\|_{\mathbb{R}\leftarrow X} = \sup\left\{|x'(x)| \,/\, \|x\|_X : 0 \neq x \in X\right\}. \tag{6.30}$$

The elements $x' \in X'$ are called *(linear) functionals* on X. Instead of $x'(x)$ (application of x' to x) one also writes $\langle x, x'\rangle_{X\times X'}$ or $\langle x', x\rangle_{X'\times X}$, and calls $\langle\cdot,\cdot\rangle_{X\times X'}$ the *dual form* on $X \times X'$:

$$\langle x, x'\rangle_{X\times X'} = \langle x', x\rangle_{X'\times X} = x'(x).$$

X'' denotes the dual space of the dual space X' and is called the *bidual space*. X is called *reflexive* if X'' is isomorphic to X. The isomorphism $\Phi : X \to X''$ maps $x \in X$ into the functional $\xi_x \in X'$ defined by $\xi_x(\varphi) := \varphi(x)$ for all $\varphi \in X'$.

Lemma 6.63. *Let the Banach space X be embedded densely and continuously in the Banach space Y. Then Y' is continuously embedded in X'. If X is reflexive, the embedding $Y' \subset X'$ is also dense.*

Proof. (i) Let $y' \in Y'$. Because $X \subset Y$, y' is defined on X. The dense embedding ensures that the restriction $y'|_X : X \to \mathbb{R}$ and the unrestricted form $y' : Y \to \mathbb{R}$ can be identified. This identification allows to consider Y' as a subspace of X'.

(ii) Since X is a dense subspace of Y, according to Theorem 6.10 and (6.4), for each $y' \in Y'$ holds:

$$\|y'\|_{Y'} = \sup_{0\neq x\in X} |y'(x)| \,/\, \|x\|_Y \geq \frac{1}{C} \sup_{0\neq x\in X} |y'(x)| \,/\, \|x\|_X = \frac{1}{C}\,\|y'\|_{X'}\,,$$

i.e., Y' is embedded continuously in X'.

(iii) Let $Z \subset X'$ be the completion of Y' with respect to the topology of X'. For an indirect proof assume that $Z \subsetneq X'$. This implies the existence of some $0 \neq \zeta \in X'$ with $\zeta \notin Z$. The Hahn–Banach theorem (cf. Yosida [312, §IV.6]) proves the existence of a functional $\xi \in X''$ with $\xi|_Z = 0$ and $\xi(\zeta) = 1$. Since X is reflexive, there is an $x \in X$ with $\xi = \xi_x$, i.e., $\xi(x') = x'(x)$ for all $x' \in X'$. Note that $x \neq 0$ because of $1 = \xi(\zeta) = \zeta(x)$. By $X \subset Y$, also $x \in Y$ holds. There is some $y' \in Y'$ with $y'(x) \neq 0$. This is a contradiction to $\xi|_Z = 0$, since $\xi(y') = y'(x) \neq 0$. ∎

To the transposed matrix in the finite-dimensional case corresponds the *dual mapping* (or the *dual operator*).

Lemma 6.64. *Let X and Y be normed and let $T \in L(X,Y)$. For each $y' \in Y'$*

$$\langle Tx, y' \rangle_{Y \times Y'} = \langle x, x' \rangle_{X \times X'} \qquad \text{for all } x \in X \tag{6.31}$$

defines a unique $x' \in X'$. The linear mapping $y' \mapsto x'$ defines the dual operator

$$T' : Y' \to X' \text{ with } T'y' = x'.$$

$T' \in L(Y', X')$ holds since it is bounded by

$$\|T'\|_{X' \leftarrow Y'} = \|T\|_{Y \leftarrow X} . \tag{6.32}$$

Proof. (6.32) follows from the definitions of the norms:

$$\|T'\|_{X' \leftarrow Y'} = \sup_{y' \neq 0} \frac{\|T'y'\|_{X'}}{\|y'\|_{Y'}} = \sup_{x \neq 0, y' \neq 0} \frac{|\langle x, T'y' \rangle_{X \times X'}|}{\|x\|_X \, \|y'\|_{Y'}}$$
$$= \sup_{x \neq 0, y' \neq 0} \frac{|\langle Tx, y' \rangle_{Y \times Y'}|}{\|x\|_X \, \|y'\|_{Y'}} = \sup_{x' \neq 0} \frac{\|Tx\|_Y}{\|x\|_X} = \|T\|_{Y \leftarrow X}$$

and the fact that $\sup_{x \neq 0} \sup_{y' \neq 0} = \sup_{y' \neq 0} \sup_{x \neq 0}$. ∎

Example 6.65. Let $\Omega = (0,1)$, $X = (C^0(\overline{\Omega}), \|\cdot\|_\infty)$, and $\mathbf{x} \in \Omega$. The mapping $\delta_{\mathbf{x}} : u \in C^0(\overline{\Omega}) \mapsto u(\mathbf{x}) \in \mathbb{R}$ is a functional: $\delta_{\mathbf{x}} \in C^0(\overline{\Omega})'$ (the so-called delta functional or Dirac function; cf. Dirac [87]). The Laplace operator Δ belongs to $L(C^2(\overline{\Omega}), C^0(\overline{\Omega}))$. The dual mapping $\Delta' \in L(C^0(\overline{\Omega})', C^2(\overline{\Omega})')$ is applicable to $\delta_{\mathbf{x}}$: $\Delta' \delta_{\mathbf{x}}$ is characterised by $(\Delta' \delta_{\mathbf{x}})u = \Delta u(\mathbf{x})$ for all $u \in C^2(\overline{\Omega})$.

Exercise 6.66. Let $S \in L(X,Y)$ and $T \in L(Y,Z)$. Show that $(TS)' = S'T'$.

Exercise 6.67. Show that if $T \in L(X,Y)$ is surjective, then T' is injective.

6.3.2 Adjoint Operators

Let X be a Hilbert space (over $\mathbb{R}$). Every $y \in X$ defines

$$f_y(x) := (x,y)_X$$

which is a linear functional $f_y \in X'$ with $\|f_y\|_{X'} = \|y\|_X$. The converse also holds as the following theorem states (cf. Riesz–Sz.-Nagy [239, §II.30] or Yosida [312, §III.6]).

Theorem 6.68 (Riesz representation theorem). *Let X be a Hilbert space and $f \in X'$. There exists a unique $y_f \in X$ such that*

$$f(x) = (x, y_f)_X \quad \textit{for all } x \in X \qquad \textit{and} \quad \|f\|_{X'} = \|y_f\|_X \ .$$

Conclusion 6.69 (Riesz isomorphism). *Let X be a Hilbert space. (a) There exists a one-to-one correspondence (the* Riesz isomorphism*)*

$$J_X \in L(X, X') \qquad \textit{with} \quad J_X y = f_y \,,\ J_X^{-1} f = y_f \,,$$

that preserves the norm, i.e., $\|J_X\|_{X' \leftarrow X} = \|J_X^{-1}\|_{X \leftarrow X'} = 1$.

(b) X' is a Hilbert space with the scalar product $(x', y')_{X'} = (J_X^{-1} x', J_X^{-1} y')_X$. The dual norm $\|x'\|_{X'}$ from (6.30) agrees with the norm induced by $\sqrt{(x', x')_{X'}}$.

(c) One always identifies X with the bidual space X'' because $x(x') := x'(x)$. From this follows

$$J_{X'} = J_X^{-1}, \ \ J_X = J_X' \qquad \textit{and} \qquad T'' = T \qquad \textit{for } T \in L(X, Y).$$

(d) Using the Riesz isomorphism, one can identify X and X': $X = X'$, $J_X = I$.

Let X, Y be Hilbert spaces and $T \in L(X, Y)$. The mapping defined by $T^\star := J_X^{-1} T' J_Y \in L(Y, X)$ is called the operator *adjoint* to T and satisfies

$$(Tx, y)_Y = (x, T^* y)_X \ \text{ for all } x \in X,\ y \in Y, \quad \|T\|_{Y \leftarrow X} = \|T^*\|_{X \leftarrow Y} \,. \tag{6.33}$$

The adjoint and the dual operator only coincide (i.e., $T^\star = T'$) if X' is identified with X and Y' with Y. $T \in L(X, X)$ is said to be *selfadjoint* (or *symmetric*) if $T = T^\star$. $T \in L(X, X)$ is called a *projection* if $T^2 = T$. It is an *orthogonal projection* if furthermore T is selfadjoint.

Remark 6.70. Let X_0 be a closed subspace of the Hilbert space X. An orthogonal projection is given by $Tx := y \in X_0$ with y being the minimiser in

$$\|x - y\|_X := \inf \{\|x - \eta\|_X : \eta \in X_0\} \,. \tag{6.34}$$

If conversely $T \in L(X, X)$ is an orthogonal projection with the range $X_0 := \{Tx : x \in X\}$ one has (6.34) for $y = Tx$. An orthogonal projection always has the norm $\|T\|_{X \leftarrow X} \le 1$.

Proof. (i) x can be decomposed uniquely into $x = y + z$ ($y \in X_0$, $z \in X_0^\perp$) (cf. Lemma 6.15). y is the unique solution of (6.34). $x \in X_0$ implies $y = x$, thus $T^2 = T$. The analogous decomposition $x' = y' + z'$ shows

$$(x, T^\star x') = (Tx, x') = (y, x') = (y, y' + z') = (y, y') = (y + z, y') = (x, Tx'),$$

hence $T = T^\star$.

(ii) Let T be an orthogonal projection with range X_0. Let $x = y + z$ be split as above. $T^2 = T$ shows $Ty = y$. For each $y' \in X_0$ holds

$$(Tz, y') = (z, T^\star y') = (z, Ty') = (z, y') = 0,$$

thus $Tz \in X_0^\perp$. Together with $Tz \in X_0$ follows $Tz = 0$ so that $Tx = Ty + Tz = y$.

(iii) $Tx = y$ and $\|x\|_X^2 = \|y\|_X^2 + \|z\|_x^2 \geq \|y\|_X^2$ prove $\|T\|_{X \leftarrow X} \leq 1$. ∎

Lemma 6.71. *(a) Let V be a Hilbert space with closed subspace $X \subsetneqq V$. Hence V is the direct sum $X \oplus X^\perp$. Then holds $V' = X' \oplus (X^\perp)'$, where all $\varphi \in X'$ are extended on $X^\perp$ by zero. Correspondingly, $\psi \in (X^\perp)'$ is defined as the zero mapping on X.*

(b) The (differently defined) dual norms of X' and V' coincide for functionals in X'.

Proof. (i) Let $\phi \in V'$. The restriction of ϕ to X and $Y := X^\perp$ defines functionals in X' and Y', whose sum is ϕ. Hence $V' \subset X' \oplus Y'$.

(ii) Define the functionals $\varphi \in X'$ and $\psi \in Y'$ by zero on the respective complementary space. Let $v = x + x^\perp \in V$ ($x \in X$, $x^\perp \in X^\perp$). The inequalities

$$\begin{aligned} |(\varphi + \psi)(v)| &\leq \left|\varphi(x) + \psi(x^\perp)\right| \leq \|\varphi\|_{X'} \|x\|_V + \|\psi\|_{Y'} \|x^\perp\|_V \\ &\leq \sqrt{\|\varphi\|_{X'}^2 + \|\psi\|_{Y'}^2} \sqrt{\|x\|_V^2 + \|x^\perp\|_V^2} \leq \sqrt{\|\varphi\|_{X'}^2 + \|\psi\|_{Y'}^2}\, \|x\|_V , \end{aligned}$$

show that $\varphi + \psi$ is bounded, i.e., $\varphi + \psi \in V'$ and thus $X' \oplus (X^\perp)' = X' \oplus Y' \subset V'$.

(iii) The dual norm of $\varphi \in X'$ is $\|\varphi\|_{X'} = \sup_{0 \neq x \in X} \frac{|\varphi(x)|}{\|x\|_V}$, while the functional $\varphi \in V'$ has the norm $\|\varphi\|_{V'} = \sup_{0 \neq v \in V} |\varphi(v)| \,/\, \|v\|_V$. Taking the supremum over a larger set, we get $\|\varphi\|_{X'} \leq \|\varphi\|_{V'}$. Decomposing of v into $x + x^\perp$ with $x \in X$ and $x^\perp \in X^\perp$ and noting that $\varphi(v) = \varphi(x)$, we obtain

$$\|\varphi\|_{V'} = \sup_{\substack{x + x^\perp \neq 0 \\ x \in X,\, x^\perp \in X^\perp}} \frac{|\varphi(x)|}{\|x + x^\perp\|_V} \underset{\|x + x^\perp\|_V \geq \|x\|_V}{\leq} \sup_{0 \neq x \in X} \frac{|\varphi(x)|}{\|x\|_V} = \|\varphi\|_{X'} ,$$

so that also $\|\varphi\|_{X'} \geq \|\varphi\|_{V'}$ is valid. This proves part (b). ∎

6.3.3 Scales of Hilbert Spaces

We assume:

$$V \subset U \quad \text{are Hilbert spaces with a continuous and dense embedding.} \tag{6.35}$$

Lemma 6.72. *Under assumption (6.35) U' is embedded continuously and densely in V'.*

Proof. The continuity of the embedding $U' \subset V'$ is established in Lemma 6.63. Since Hilbert spaces are reflexive, U' is also dense in V'. A more elementary proof uses Exercise 6.17 with $A := U'$, $X := V'$. Let $0 \neq v' \in V'$ be arbitrary and $u := J_V^{-1} v' \in V \subset U$. According to the definition, $u' := J_U u \in U' \subset V'$ is characterised by $u'(x) = (x,u)_U$ for all $x \in U$. For $x := u = J_V^{-1} v' \in V$ follows

$$(v',u')_{V'} = \left(J_V^{-1} v', J_V^{-1} u'\right)_V = \left(u, J_V^{-1} u'\right)_V = u'(u) = (u,u)_U > 0$$

and thus $(v',u')_{V'} \neq 0$. Therefore the statement in Exercise 6.17b applies. ∎

According to Corollary 6.69d, U and U' can be identified. By this one obtains the *Gelfand triple*

$$V \subset U \subset V' \qquad (V \subset U \text{ continuously and densely embedded}). \tag{6.36}$$

Conclusion 6.73. *In a Gelfand triple (6.36) V and U are also continuously and densely embedded in V'.*

Proof. For $U \subset V'$ see Lemma 6.72, for $V \subset V'$ see Exercise 6.13. ∎

Attention. Likewise one could identify V with V' and one would obtain $U' \subset V' = V \subset U$. But it is not possible to identify U with U' and V with V' simultaneously. In the first case one interprets $x(y) = \langle y, x \rangle_{U \times U'}$ for $x, y \in U$ as $(y,x)_U$ (in particular for $x, y \in V \subset U$), in the second case as $(y,x)_V$.

Because $U = U'$ the scalar product $(x,y)_U$ can also be written in the form $y(x) = \langle x, y \rangle_{U \times U'}$. If $x \in V$, then $y(x) = \langle x, y \rangle_{V \times V'}$ also holds. That means that $(x,y)_U = \langle x, y \rangle_{V \times V'}$ for all $x \in V$, $y \in U \subset V'$. Likewise one obtains $(x,y)_U = \langle x, y \rangle_{V' \times V}$ for all $x \in U$ and $y \in V$. The dense and continuous embedding $U \subset V'$ proves the following remark.

Remark 6.74. Let $V \subset U \subset V'$ be a Gelfand triple. The continuous extension of the scalar product $(\cdot,\cdot)_U$ to $V \times V'$ $[V' \times V]$ results in the dual form $\langle \cdot, \cdot \rangle_{V \times V'}$ $[\langle \cdot, \cdot \rangle_{V' \times V}]$. Therefore the following notation is practical:

$$\langle x, y \rangle_{V \times V'} = (x,y)_U \text{ for } \left\{ \begin{array}{l} x \in V \\ y \in V' \end{array} \right\}, \quad \langle x, y \rangle_{V' \times V} = (x,y)_U \text{ for } \left\{ \begin{array}{l} x \in V' \\ y \in V \end{array} \right\}.$$

In connection with Sobolev spaces one always chooses $U := L^2(\Omega)$ so that the embeddings read as follows:

$$H_0^s(\Omega) \subset L^2(\Omega) \subset \left(H_0^s(\Omega)\right)' \qquad (s \geq 0)\,, \tag{6.37}$$

$$H^s(\Omega) \subset L^2(\Omega) \subset \left(H^s(\Omega)\right)' \qquad (s \geq 0)\,. \tag{6.38}$$

Exercise 6.75. Show that (6.37) and (6.38) are Gelfand triples.

The dual space of $H_0^s(\Omega)$ is also denoted by $H^{-s}(\Omega)$ or $H_0^{-s}(\Omega)$:

$$H_0^{-s}(\Omega) := H^{-s}(\Omega) := (H_0^s(\Omega))' \qquad (s \geq 0)\,.$$

The norm of $H^{-s}(\Omega)$ according to (6.30) reads:

$$|u|_{-s} := \sup\left\{\left|(u,v)_{L^2(\Omega)}\right| / |v|_s : 0 \neq v \in H_0^s(\Omega)\right\},$$

where $(u,v)_{L^2(\Omega)}$ is the dual form on $H_0^s(\Omega) \times H^{-s}(\Omega)$ (cf. Remark 6.74).

Remark 6.76. (a) Let $\Omega = \mathbb{R}^n$. The norm dual to $|\cdot|_s^\wedge$,

$$|u|_{-s}^\wedge := \sup_{0 \neq v \in H^s(\mathbb{R}^n)} |(u,v)_0| / |v|_s^\wedge$$

is equivalent to $|\cdot|_{-s}$ and has the representation (6.21b) with $-s$ instead of s.
(b) The Fourier transform shows

$$D^\alpha \in L(H^s(\mathbb{R}^n), H^{s-|\alpha|}(\mathbb{R}^n)) \qquad \text{for all } s \in \mathbb{R}.$$

(c) $au \in H^s(\Omega)$, if $u \in H^s(\Omega)$, $a \in C^t(\overline{\Omega})$, where $t = |s| \in \mathbb{N}_0$ or $t > |s|$.

6.4 Compact Operators

Definition 6.77. A subset K of a Banach space is said to be *compact* if each sequence $x_i \in K$ $(i \in \mathbb{N})$ contains a convergent subsequence x_{i_k} with limit in K. A subset K is *relatively compact* if the closure $\overline{K}$ is compact.

Another definition of compactness reads: Each open covering of K already contains a finite covering of K. Both definitions are equivalent in metric spaces (cf. Dieudonné [86, (3.16.1)]). There is also a definition of *precompact* sets. In our Banach space setting also the terms 'relatively compact' and 'precompact' are equivalent (cf. Dieudonné [86, (3.17.5)]).

Remark 6.78. (a) $K \subset \mathbb{R}^n$ is relatively compact [compact] if and only if K is bounded [and closed].
(b) Let X be a Banach space. The unit ball $\{x \in X : \|x\| \leq 1\}$ is compact if and only if $\dim(X) < \infty$.

Definition 6.79. Let X and Y be Banach spaces. The mapping $T \in L(X,Y)$ is said to be *compact* if $\{Tx : x \in X, \|x\|_X \leq 1\}$, the image of the unit ball in X, is relatively compact in Y.

Exercise 6.80. When is the identity $I \in L(X, X)$ compact?

The following statement is known as the Theorem of Arzelà [9] and Ascoli [10] (cf. Yosida [312, §III.3]). We recall that a set $\mathcal{F}$ of functions is *equicontinuous* if

$$\lim_{\delta \to 0} \sup_{x,x' \in X, \|x-x'\| \le \delta, f \in \mathcal{F}} |f(x) - f(x')| = 0.$$

Theorem 6.81. *Let D be compact. A family $\mathcal{F} \subset C(D)$ of uniformly bounded and equicontinuous functions is relatively compact.*

Lemma 6.82. *Let X, Y, and Z be Banach spaces.*

(a) Let one of the mappings $T_1 \in L(X, Y)$ or $T_2 \in L(Y, Z)$ be compact. Then $T_2 T_1 \in L(X, Z)$ is also compact.

(b) $T \in L(X, Y)$ is compact if and only if $T' \in L(Y', X')$ is compact.

Proof. (a) Let $K_1 := \{x \in X : \|x\|_X \le 1\}$. If T_1 is compact, i.e., $T_1(K_1)$ is relatively compact, then $T_2(T_1(K_1))$ is also relatively compact and thus $T_2 T_1$ is compact. If, however, T_2 is compact, one proves the assertion as follows. Since scaling does not change compactness, $\|T_1\|_{Y \leftarrow X} \le 1$ can be assumed without loss of generality. Hence $T_1(K_1)$ is a subset of the unit ball in Y and therefore $T_2(T_1(K_1))$ is relatively compact.

(b_1) Let $T \in L(X, Y)$ be compact and set $K_1' := \{y' \in Y' : \|y\|_{Y'} \le 1\}$. A sequence in $T'(K_1') \subset X'$ is of the form $T'(y_j')$ with $y_j' \in K_1'$. The image $Y_1 := \overline{T(K_1)} \subset Y$ is a compact set. $\{y_j'\}$ is uniformly bounded in Y_1 since

$$|y_j'(y)| \le \|y\|_Y \le \|T\|_{Y \leftarrow X},$$

and it is equicontinuous since $|y_j'(y_1) - y_j'(y_2)| \le |y_j'(y_1 - y_2)| \le \|y_1 - y_2\|_Y$. By Ascoli's Theorem 6.81 there exists a convergent subsequence $y_{j_k}' \to y' \in Y'$. From

$$\begin{aligned} \|T'(y_j' - y')\|_{X'} &= \sup_{x \in K_1} |(T'(y_j' - y'))(x)| = \sup_{x \in K_1} |(y_j' - y')(Tx)| \\ &\le \|y_j' - y'\|_{Y'} \, \|T\|_{Y \leftarrow X} \end{aligned}$$

we obtain $T'(y_{j_k}') \to T'(y')$ in X'. Hence $T' \in L(Y', X')$ is compact.

(b_2) If $T' \in L(Y', X')$ is compact, part (b_1) implies that $T'' \in L(X'', Y'')$ is compact, i.e., $T''(K_1'')$ with $K_1'' := \{x \in X'' : \|x\|_{X''} \le 1\}$ is relatively compact. Since Y is isomorphically embedded in the bidual space Y'', also $T(K_1)$ is relatively compact. ∎

Exercise 6.83. Show that $T \in L(X, Y)$ is compact if $\dim X < \infty$ or $\dim Y < \infty$.

A special type of compact mapping is a *compact embedding*.

Definition 6.84. Let $X \subset Y$ be a continuous embedding. X is said to be *compactly embedded* in Y if the *inclusion*

$$I \in L(X, Y), \qquad Ix = x$$

is compact.

Together with Definitions 6.77 and 6.79 one obtains: $X \subset Y$ is compactly embedded if every sequence $x_i \in X$ with $\|x_i\|_X \leq 1$ contains a subsequence convergent in Y.

Example 6.85. Let Ω be bounded. $C^s(\overline{\Omega}) \subset C^0(\overline{\Omega})$ is a compact embedding for $s > 0$.

Proof. Functions $u_i \in C^s(\overline{\Omega})$ with $\|u_i\|_{C^s(\overline{\Omega})} \leq 1$ are equicontinuous and uniformly bounded. The assertion follows from Theorem 6.81. ∎

Analogous results can be obtained for Sobolev spaces (cf. Adams [1, page 144], Wloka [308, Theorems 7.8–7.10]).

Theorem 6.86. *Let $\Omega \subset \mathbb{R}^n$ be open and bounded.*

(a) The embeddings $H_0^s(\Omega) \subset H_0^t(\Omega)$ $(s, t \in \mathbb{R},\ s > t)$ are compact.

(b) Further, let $\Omega \in C^{0,1}$. The embeddings $H^k(\Omega) \subset H^\ell(\Omega)$ $(k, \ell \in \mathbb{N}_0,\ k > \ell)$ are compact.

(c) Let $0 \leq t < s$ and $\Omega \in C^r$ $(r > t,\ r > 1)$ or $\Omega \in C^{k,1}$ $(k + 1 > t)$. Then the embedding $H^s(\Omega) \subset H^t(\Omega)$ is compact.

Remark 6.87. In Theorem 6.86b one can replace $\Omega \in C^{0,1}$ by the *uniform cone property* (cf. Wloka [308, Definition 2.3 and Theorem 7.2]). To ensure $\Omega \in C^{0,1}$ it is sufficient that the boundary $\partial\Omega$ is piecewise smooth and the inside angles of possible corners are smaller than 2π.

Re-entrant corners (cf. Figure 2.1) are thus permitted while a cut domain (cf. Figure 5.2b) is excluded.

In Section 6.5 the following situation will arise:

$$V \subset U \subset V' \text{ Gelfand triple}, \qquad T \in L(V', V). \tag{6.39}$$

Because of the continuous embeddings, T also belongs to $L(V', V')$, $L(U, U)$, $L(V, V)$, and $L(U, V)$.

Theorem 6.88. *Let (6.39) hold. Let $V \subset U$ be a compact embedding. Then T is a compact operator in the spaces $L(V', V')$, $L(U, U)$, $L(V, V)$, $L(V', U)$, and $L(U, V)$.*

Proof. As an example, let us do $T \in L(U, V)$. Since the inclusion $I \in L(V, U)$ is compact we see $I \in L(U, V')$ is also compact (cf. Lemma 6.82b). $T \in L(U, V)$, as the product $T \cdot I$ of the compact mapping $I \in L(U, V')$ with $T \in L(V', V)$, is compact (cf. Lemma 6.82a). ∎

The significance of compact operators $T \in L(X,X)$ lies in the fact that the equation

$$Tx - \lambda x = y \qquad (x, y \in X,\ y \text{ given, } x \text{ sought}) \tag{6.40}$$

has properties analogous to the finite-dimensional case. The following statement goes back to Riesz [238].

Theorem 6.89 (Riesz–Schauder theory). *Let $T \in L(X,X)$ be compact, where X is a Banach space.*
(a) For each $\lambda \in \mathbb{C}\backslash\{0\}$ one of the following alternatives holds:

$$\textit{(i) } (T-\lambda I)^{-1} \in L(X,X) \qquad \textit{or} \qquad \textit{(ii) } \lambda \textit{ is an eigenvalue.}$$

In case (i) the equation $Tx - \lambda x = y$ has a unique solution $x \in X$ for all $y \in X$. In case (ii) there exists a finite-dimensional eigenspace

$$E(\lambda, T) := \ker(T - \lambda I) \neq \{0\}.$$

All $x \in E(\lambda, T)$ solve the eigenvalue problem $Tx = \lambda x$.

(b) The spectrum *$\sigma(T)$ of T consists by definition of all eigenvalues and, if not $T^{-1} \in L(X,X)$, $\lambda = 0$. There exist at most countably many eigenvalues which can only accumulate at zero. Furthermore,*

$$\sigma(T) = \sigma(T') \quad \textit{and} \quad \dim(E(\lambda, T)) = \dim(E(\lambda, T')) < \infty.$$

(c) For $\lambda \in \sigma(T)\backslash\{0\}$, $Tx - Ax = y$ has at least one solution $x \in X$ if and only if $\langle y, x' \rangle_{X \times X'} = 0$ for all $x' \in E(\lambda, T')$.

In Lemma 6.110 we need the following statement (lemma of G. Ehrling).

Lemma 6.90. *Let $X \subset Y \subset Z$ be continuously embedded Banach spaces and let $X \subset Y$ be compactly embedded. Then for every $\varepsilon > 0$ there exists a C_ε such that*

$$\|x\|_Y \leq \varepsilon \|x\|_X + C_\varepsilon \|x\|_Z \qquad \textit{for all } x \in X. \tag{6.41}$$

Proof. Let $\varepsilon > 0$ be fixed. The negation of (6.41) reads: There exists $x_i \in X$ with

$$(\, \|x_i\|_Y - \varepsilon \|x_i\|_X \,) \,/\, \|x_i\|_Z \to \infty.$$

For $y_i := (\varepsilon \|x_i\|_X)^{-1} x_i \in X$ we thus have $(\|y_i\|_Y - 1) \,/\, \|y_i\|_Z \to \infty$. From this one infers $\|y_i\|_Z \to 0$ and $\|y_i\|_Y > 1$ for sufficiently large i. Since $\|y_i\|_X < 1/\varepsilon$ and $X \subset Y$ is a compact embedding, a subsequence y_{i_k} converges to $y^* \in Y$. Now, $\|y_i\|_Y > 1$ implies $\|y^*\|_Y \geq 1$, i.e., $y^* \neq 0$. On the other hand, y_{i_k} also converges in Z to y^* since $Y \subset Z$ is continuously embedded. $\|y_i\|_Z \to 0$ gives the contradiction sought: $y^* = 0$. ∎

6.5 Bilinear Forms

In the following let us assume that V is a Hilbert space. The mapping

$$a(\cdot,\cdot) : V \times V \to \mathbb{R}$$

is called a *bilinear form* if

$$a(x+\lambda y, z) = a(x,z) + \lambda a(y,z), \quad a(x, y+\lambda z) = a(x,y) + \lambda a(x,z)$$

for all $\lambda \in \mathbb{R},\ x, y, z \in V$. In the complex case, $\lambda a(x,z)$ in the second equation is to be replaced by $\bar{\lambda} a(x,z)$. Because of the antilinearity $a(x, \lambda y) = \bar{\lambda} a(x,y)$ in the second argument, $a(\cdot,\cdot) : V \times V \to \mathbb{C}$ is called a *sesquilinear form*.

$a(\cdot,\cdot)$ is said to be *continuous* (or *bounded*) if there exists a C_S such that

$$|a(x,y)| \le C_S \|x\|_V \|y\|_V \qquad \text{for all } x, y \in V. \tag{6.42}$$

Lemma 6.91. *(a) To a continuous bilinear form one can assign a unique operator $A \in L(V, V')$ such that*

$$a(x,y) = \langle Ax, y\rangle_{V' \times V} \quad \textit{for all } x, y \in V.$$

The inequality (6.42) is equivalent to

$$\|A\|_{V' \leftarrow V} \le C_S.$$

$C_S := \|A\|_{V' \leftarrow V}$ is the smallest bound in (6.42).
(b) Let V_1 and V_2 be dense in V. Let $a(\cdot,\cdot)$ be defined on $V_1 \times V_2$ and satisfy (6.42) with "$x \in V_1,\ y \in V_2$" instead of any "$x, y \in V$". Then $a(\cdot,\cdot)$ can be extended uniquely to $V \times V$ so that (6.42) holds with the same C_S for all $x, y \in V$.

Proof. (a) Let $x \in V$ be fixed. $\varphi_x(y) := a(x,y)$ defines a functional $\varphi_x \in V'$ with $\|\varphi_x\|_{V'} \le C_S \|x\|_V$. Since $x \mapsto \varphi_x$ is a linear map, one sets $Ax := \varphi_x$ for $x \in V$. $\|Ax\|_{V'} \le C_S \|x\|_V$ proves $\|A\|_{V' \leftarrow V} \le C_S$. The definitions show

$$\langle Ax, y\rangle_{V' \times V} = \langle \varphi_x, y\rangle_{V' \times V} = \varphi_x(y) = a(x,y).$$

Conversely, for each $A \in L(V, V')$, $a(x,y) := \langle Ax, y\rangle_{V' \times V}$ is also a bilinear form with

$$\langle Ax, y\rangle_{V' \times V} \le \|Ax\|_{V'} \|y\|_V \le \|A\|_{V' \leftarrow V} \|x\|_V \|y\|_V .$$

(b) According to Theorem 6.10, A is also uniquely determined if $a(\cdot,\cdot)$ is only given on $V_1 \times V_2$. Then $\langle Ax, y\rangle_{V' \times V}$ represents the continuous extension. ∎

The proof shows

$$\|A\|_{V' \leftarrow V} = \sup\{|a(x,y)| : x, y \in V,\ \|x\|_V = \|y\|_V = 1\}.$$

A is called the operator that is *associated* to $a(\cdot,\cdot)$.

The bilinear form $a^\star(\cdot,\cdot)$ *adjoint* to $a(\cdot,\cdot)$ is given by

$$a^\star(x,y) := a(y,x) \qquad (x, y \in V).$$

The bilinear form is said to be *symmetric* if $a^\star(x,y) = a(y,x)$. In the complex case, a sesquilinear form is symmetric if

$$a^\star(x,y) = \overline{a(y,x)}.$$

Exercise 6.92. Show that (a) If A is associated to $a(\cdot,\cdot)$, then the adjoint operator A' belongs to $a^\star(\cdot,\cdot)$.
(b) If $a(\cdot,\cdot)$ is symmetric, then $A = A'$.

Remark 6.93. In the sequel we have to solve the problem

$$\text{find } u \in V \text{ with } a(u,v) = f(v) \text{ for all } v \in V,$$

where $f \in V'$ is a functional. Using the associated operator $A : V \to V'$, we rewrite this problem as a linear equation in the space V':

$$Au = f.$$

If the inverse $A^{-1} \in L(V',V)$ exists, the unique solution is given by $u = A^{-1}f$.

Lemma 6.94 (inf-sup condition). *Let $A \in L(V,V')$ be the operator associated to a continuous bilinear form $a(\cdot,\cdot)$. Then the following statements (i), (ii), (iii) are equivalent:*
(i) $A^{-1} \in L(V',V)$ *exists.*
(ii) $\varepsilon, \varepsilon' > 0$ *exist such that*

$$\inf_{x\in V,\,\|x\|_V=1} \sup_{y\in V,\,\|y\|_V=1} |a(x,y)| = \varepsilon > 0, \tag{6.43a}$$

$$\inf_{y\in V,\,\|y\|_V=1} \sup_{x\in V,\,\|x\|_V=1} |a(x,y)| = \varepsilon' > 0, \tag{6.43b}$$

(iii) the inequalities (6.43a) and (6.43c) hold:

$$\sup_{x\in V,\,\|x\|_V=1} |a(x,y)| > 0 \qquad \textit{for all } 0 \neq y \in V. \tag{6.43c}$$

If one of the statements (i)–(iii) holds, then

$$\varepsilon = \varepsilon' = 1/\|A^{-1}\|_{V\leftarrow V'} \qquad (\varepsilon, \varepsilon' \textit{ from (6.43a,b)}). \tag{6.43d}$$

From (6.43a) follows

$$\inf_{x\in V,\,\|x\|_V=1} \sup_{y\in V,\,\|y\|_V=1} |a(x,y)| \geq \varepsilon > 0. \tag{6.43e}$$

Conversely, (6.43a) follows from (6.43e) with a possibly larger $\varepsilon > 0$. (6.43e) is equivalent to

$$\sup_{y\in V,\ \|y\|_V=1} |a(x,y)| \geq \varepsilon \|x\|_V \qquad \text{for all } x \in V \tag{6.43e'}$$

because (6.43e) is equal to (6.43e′) for all $x \in V$, $\|x\|_V = 1$. The scaling condition $\|x\|_V = 1$ can evidently be dropped. The left-hand side in (6.43e′) agrees with the definition of the dual norm of Ax so that (6.43e) and (6.43e′) are also equivalent to

$$\|Ax\|_{V'} \geq \varepsilon \|x\|_V \qquad \text{for all } x \in V. \tag{6.43e''}$$

(6.43a,b) are also called the *Babuška conditions*[3] or *inf-sup conditions*.

Proof of Lemma 6.94. (a) "(i) ⇒ (ii)": Let $A^{-1} \in L(V', V)$ exist. Then (6.43a) follows from

$$\begin{aligned}
\inf_{\substack{x\in V\\ \|x\|_V=1}} \sup_{\substack{y\in V\\ \|y\|_V=1}} |a(x,y)| &= \inf_{\substack{x\in V\\ x\neq 0}} \sup_{\substack{y\in V\\ y\neq 0}} \frac{|a(x,y)|}{\|x\|_V \|y\|_V} = \inf_{\substack{x\in V\\ x\neq 0}} \sup_{\substack{y\in V\\ y\neq 0}} \frac{|\langle Ax, y\rangle_{V'\times V}|}{\|x\|_V \|y\|_V} \\
&= \inf_{\substack{x'\in V'\\ x'\neq 0}} \sup_{\substack{y\in V\\ y\neq 0}} \frac{\left|\langle AA^{-1}x', y\rangle_{V'\times V}\right|}{\|A^{-1}x'\|_V \|y\|_V} \qquad \text{for } x = A^{-1}x' \\
&= \inf_{\substack{x'\in V'\\ x'\neq 0}} \left(\frac{1}{\|A^{-1}x'\|_V} \sup_{\substack{y\in V\\ y\neq 0}} \frac{|\langle x', y\rangle_{V'\times V}|}{\|y\|_V} \right) \\
&= \inf_{\substack{x'\in V'\\ x'\neq 0}} \frac{1}{\|A^{-1}x'\|_V} \|x'\|_{V'} = 1/ \sup_{\substack{x'\in V'\\ x'\neq 0}} \frac{\left\|A^{-1}x'\right\|_V}{\|x'\|_{V'}} = \frac{1}{\|A^{-1}\|_{V\leftarrow V'}} =: \varepsilon,
\end{aligned}$$

which also gives the characterisation of ε in (6.43d). In the same way one shows (6.43b) with $\varepsilon' = 1/\|A'^{-1}\|_{V\leftarrow V'}$. Because $A'^{-1} = (A^{-1})'$, (6.32), and $V'' = V$, it follows that $\varepsilon = \varepsilon'$.

(b) "(ii) ⇒ (iii)": (6.43c) is a weakening of (6.43b).

(c) "(iii) ⇒ (i)": $\varepsilon > 0$ in (6.43a) proves that A is injective. To get surjectivity, we wish to show that the image $W := \{Ax : x \in V\} \subset V'$ is closed. For a sequence $\{w_\nu\}$ with $\|w^\star - w_\nu\|_{V'} \to 0$ we must therefore show that $w^\star \in W$. According to the definition of W there exists $x_\nu \in V$ with $Ax_\nu = w_\nu$. From (6.43a) one infers via (6.43e) and (6.43e″) (with $x := x_\nu - x_\mu$) that $\|x_\nu - x_\mu\|_V \leq \|w_\nu - w_\mu\|_{V'}/\varepsilon$. Since $\{w_\nu\}$ is Cauchy convergent, this property carries over to $\{x_\nu\}$. There exists an $x^\star \in V$ with $x_\nu \to x^\star$ in V. The continuity of $A \in L(V, V')$ proves $w_\nu = Ax_\nu \to Ax^\star$ so that $w^\star = Ax^\star \in W$. According to Lemma 6.15 one can decompose V' into $W \oplus W^\perp$. If A were not surjective (thus $W \neq V'$), there would exist $w \in W^\perp$ with $w \neq 0$. Then $y := J_{V'}w = J_V^{-1}w \in V$ would satisfy $y \neq 0$ (cf. Theorem 6.68, Corollary 6.69). Since $a(x,y) = \langle Ax, y\rangle_{V'\times V} = (Ax, w)_{V'} = 0$

[3] The corresponding condition for saddle-point problems is also called *Ladyženskaja–Babuška–Brezzi condition* (abbreviated by *LBB condition*). The corresponding papers are Ladyženskaja [178] (1961), Babuška [15, Theorem 2.1] (1971), and Brezzi [53] (1974).

for all $x \in V$, a contradiction to (6.43c) would result. Therefore, A is also surjective, and Theorem 6.12 proves $A^{-1} \in L(V', V)$.

(d) Statement (6.43d) has already resulted from part (a) of the proof. ∎

It will be shown that for interesting cases conditions (6.43a) and (6.43b) are equivalent (cf. Lemma 6.109). A particularly simple case follows.

Exercise 6.95. Show that if $\dim V < \infty$, then (6.43a) implies the statement (6.43b) with $\varepsilon' = \varepsilon$ and conversely.

Definition 6.96 (V-ellipticity). A bilinear form is said to be *V-elliptic* if it is continuous on $V \times V$ and there is a constant C_E such that

$$a(x,x) \geq C_E \|x\|_V^2 \qquad \text{for all } x \in V \text{ with } C_E > 0. \tag{6.44}$$

The letter "V" in the notation "V-elliptic" is regarded as a text variable for some Hilbert space. For instance, if $V = H_0^1(\Omega)$, the bilinear form is $H_0^1(\Omega)$-elliptic.

In the complex case, (6.44) should be replaced by

$$|a(x,x)| \geq C_E \|x\|_V^2 \qquad \text{for all } x \in V \text{ with } C_E > 0. \tag{6.44$'$}$$

Lemma 6.97 (Lax–Milgram [184]). *V-ellipticity (6.44) implies (6.42) and the inf-sup conditions (6.43a,b) with $\varepsilon = \varepsilon' \geq C_E > 0$ and thus $\|A^{-1}\|_{V \leftarrow V'} \leq 1/C_E$.*

Proof. Let $x \in V$, $\|x\|_V = 1$. $\sup\{|a(x,y)| : y \in V, \|y\|_V = 1\} \geq |a(x,x)| \geq C_E$ proves (6.43a) with $\varepsilon \geq C_E$. (6.43b) follows analogously. The continuity of a is equivalent to (6.42) with a constant $C_S < \infty$. ∎

Exercise 6.98. Show that: (a) If $W \subset V$ is a Hilbert subspace with the norm equal (or equivalent) to V, then a V-elliptic bilinear form is also W-elliptic.

(b) Let $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ be continuous. If $a(x,x) \geq C_E \|x\|_V^2$ for all $x \in V_0$ where V_0 is dense in V, then (6.44) follows with the same C_E.

(c) Let $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ be continuous, symmetric, and positive (i.e., $a(x,x) > 0$ for all $0 \neq x \in V$). Then $(x,y)_a := a(x,y)$ is a scalar product and

$$\|x\|_a := \sqrt{a(x,x)}$$

a norm. *V-ellipticity* (6.44) holds if and only if the norms $\|x\|_a$ and $\|x\|_V$ are equivalent.

If $a(\cdot,\cdot)$ is symmetric, the conditions (6.43a) and (6.43b) coincide so that only one of them is to be required. If, in addition, $a(\cdot,\cdot)$ is nonnegative, the statement of Lemma 6.97 can be reversed.

Lemma 6.99. *Let the bilinear form $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ be continuous, symmetric, nonnegative (i.e., $a(x,x) \geq 0$ for all $x \in V$) and let (6.43a) be satisfied. Then $a(\cdot,\cdot)$ is V-elliptic with the constant $C_E = \varepsilon^2/C_S$ (C_S from (6.42) and ε from (6.43a)).*

Proof. According to Exercise 6.14b, also nonnegative, symmetric bilinear forms satisfy the Schwarz inequality

$$|a(x,y)| \le \sqrt{a(x,x)}\sqrt{a(y,y)}.$$

Let $\|x\|_V = \|y\|_V = 1$. From (6.42) we infer

$$|a(x,y)| \le \sqrt{a(x,x)}\sqrt{C_S}\,.$$

Applying (6.43a), we obtain

$$\varepsilon = \inf_{\|\xi\|_V=1} \sup_{\|y\|_V=1} |a(\xi,y)| \le \sup_{\|y\|_V=1} |a(x,y)| \le \sqrt{a(x,x)}\sqrt{C_S}.$$

Squaring the inequality gives $a(x,x) \ge \varepsilon^2/C_S$ for all x with $\|x\|_V = 1$. This is equivalent to (6.44) with $C_E = \varepsilon^2/C_S$. ∎

Combining Lemmata 6.91, 6.94, 6.97 together with $\|A'^{-1}\|_{V\leftarrow V'} = \|A^{-1}\|_{V\leftarrow V'}$ (cf. Lemma 6.64) proves the next theorem.

Theorem 6.100. *Let the bilinear form be V-elliptic [or satisfy (6.42), (6.43a,c)]. Then the corresponding operator A satisfies the conditions*

$$\begin{aligned} &A \in L(V,V'), && \|A'\|_{V'\leftarrow V} &&= \|A\|_{V'\leftarrow V} &&\le C_S,\\ &A^{-1} \in L(V',V), && \|A'^{-1}\|_{V\leftarrow V'} &&= \|A^{-1}\|_{V\leftarrow V'} &&\le C' \end{aligned}$$

with C_S from (6.42) and $C' = 1/C_E$ [resp. $C' = 1/\varepsilon = 1/\varepsilon'$].

With the help of the bilinear form $a(\cdot,\cdot)$ and a functional $f \in V'$ one can formulate the following *variational problem*:

$$\text{find } x \in V \text{ with } \quad a(x,y) = f(y) \quad \text{for all } y \in V. \tag{6.45}$$

According to Lemma 6.91 one can write (6.45) in the form

$$\langle Ax - f, y\rangle_{V'\times V} = 0 \text{ for all } y \in V, \quad \text{i.e., } Ax = f \text{ in } V'.$$

The equation $Ax = f$ is solvable for $f \in V'$ if and only if $A^{-1} \in L(V',V)$ (cf. Remark 6.93). Hence one obtains the next statement.

Theorem 6.101. *Let the bilinear form be continuous (cf. (6.42)) and satisfy the stability condition (6.43a,c) [it is sufficient that $a(\cdot,\cdot)$ is V-elliptic]. Then problem (6.45) has exactly one solution $x := A^{-1}f$. This satisfies*

$$\|x\|_V \le C\,\|f\|_{V'} \qquad \textit{with } C = 1/\varepsilon = 1/\varepsilon' \text{ [resp. } C = 1/C_E].$$

Corollary 6.102 (adjoint problem). Under the assumptions of Theorem 6.101 the analogous statement holds with the same estimate for the *adjoint variation problem*

$$\text{find } x^* \in V \text{ with } \quad a^*(x^*,y) = f(y) \quad \text{for all } y \in V. \tag{6.46}$$

Proof. $\|A^{-1}\|_{V\leftarrow V'} = \|A'^{-1}\|_{V\leftarrow V'}$ (cf. Exercise 6.92a, Theorem 6.100). ∎

Exercise 6.103. Let $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ be continuous. Let V_0 be dense in V. Show that the solution $x \in V$ of problem (6.45) is already uniquely determined by

$$a(x,y) = f(y) \quad \text{for all } y \in V_0.$$

The same holds for (6.46).

Problem (6.45) may be equivalent to a *minimisation problem* (often the physical background is energy minimisation).

Theorem 6.104. *Let $a(\cdot,\cdot)$ be V-elliptic and symmetric; furthermore, let $f \in V'$. Then*

$$J(x) := a(x,x) - 2f(x) \qquad (x \in V) \tag{6.47}$$

assumes its unique minimum for the solution x of equation (6.45).

Proof. Let x be the solution of (6.45). For arbitrary $z \in V$ set $y := z - x$. From

$$\begin{aligned} J(z) = J(x+y) &= a(x+y, x+y) - 2f(x+y) \\ &= a(x,x) + a(x,y) + a(y,x) + a(y,y) - 2f(x) - 2f(y) \\ &\underset{\text{symmetry}}{=} J(x) + a(y,y) + 2\,[\underbrace{a(x,y) - f(y)}_{=0,\ x \text{ is solution}}] = J(x) + a(y,y) \\ &\geq J(x) + C_E \|y\|_V^2 = J(x) + C_E \|z-x\|_V^2 \end{aligned}$$

one can read $J(z) > J(x)$ for all $z \neq x$. ∎

The term "V-elliptic" seems to indicate that to elliptic boundary-value problems correspond V-elliptic bilinear forms. In general this is not the case. Rather, V-coercive forms will be assigned to the elliptic boundary-value problems. Their definition necessitates the introduction of a Gelfand triple (cf. (6.36): $V \subset U \subset V'$, $U = U'$, $V \subset U$ continuously and densely embedded).

Definition 6.105 (V-coercivity). Let $V \subset U \subset V'$ be a Gelfand triple. A bilinear form $a(\cdot,\cdot)$ is said to be V-coercive[4] if it is continuous and if there holds the so-called Gårding inequality:

$$a(x,x) \geq C_E \|x\|_V^2 - C_K \|x\|_U^2 \quad \text{for all } x \in V \text{ with } C_K \in \mathbb{R},\ C_E > 0. \tag{6.48}$$

Obviously, V-coercivity and V-ellipticity differ by the term $C_K \|x\|_U^2$, where C_K may take any sign. It is essential that, because of $V \subset U$, the U-norm is weaker than the V-norm.

[4] Often the terms 'V-elliptic' and 'V-coercive' are used synonymously for the property (6.44). Here we follow the notation in Lions–Magenes [194, Chap. 2, §9.1] and Wloka [308, Definition 17.4].

Exercise 6.106. Set $\tilde{a}(x,y) := a(x,y) + C_K(x,y)_U$ with C_K from (6.48). Show the following:

(a) The coercivity condition (6.48) is equivalent to the V-ellipticity of $\tilde{a}$.

(b) Let $I : V \to V'$ be the inclusion. If $A \in L(V,V')$ is associated to $a(\cdot,\cdot)$, then so is $\tilde{A} := A + C_K I$ to $\tilde{a}(\cdot,\cdot)$.

The results of Riesz–Schauder theory (Theorem 6.89) transfer to A as soon as the embedding $V \subset U$ is not only continuous but also compact.

Theorem 6.107. *Let $V \subset U \subset V'$ be a Gelfand triple with compact embedding $V \subset U$. Let the bilinear form $a(\cdot,\cdot)$ be V-coercive with corresponding operator A. Let $I : V \to V'$ be the inclusion.*

(a) For each $\lambda \in \mathbb{C}$ one of the following alternatives holds:[5]

(i) $(A-\lambda I)^{-1} \in L(V',V)$ and $(A'-\lambda I)^{-1} \in L(V',V)$ or

(ii) λ is an eigenvalue.

In case (i) $Ax - \lambda x = f$ and $A'x^* - \lambda x^* = f$ are uniquely solvable for all $f \in V'$, i.e.,

$$a(x,y) - \lambda(x,y)_U = f(y) \quad \text{and} \quad a^\star(x^\star,y) - \lambda(x^\star,y)_U = f(y) \quad \text{for all } y \in V.$$

In case (ii) there exist finite-dimensional eigenspaces $E(\lambda) := \ker(A-\lambda I) \neq \{0\}$ and $E'(\lambda) := \ker(A'-\lambda I) \neq \{0\}$ such that

$$\begin{aligned} &Ax = \lambda x \quad \text{for } x \in E(\lambda), && \text{i.e., } a(x,y) = \lambda \cdot (x,y)_U \quad \text{for all } y \in V,\\ &A'x^* = \lambda x^* \quad \text{for } x^* \in E'(\lambda), && \text{i.e., } a^*(x^*,y) = \lambda \cdot (x^*,y)_U \quad \text{for all } y \in V. \end{aligned}$$

(b) The spectrum $\sigma(A)$ of A consists of at most countably many eigenvalues which cannot accumulate in $\mathbb{C}$. Furthermore $\sigma(A) = \sigma(A')$ and $\dim E(\lambda) = \dim E'(\lambda) < \infty$.

(c) For $\lambda \in \sigma(A)$, $Ax - \lambda x = f \in V'$ has at least one solution $x \in V$ if and only $f \perp E'(\lambda)$, i.e., $\langle f, x^\star\rangle_{V'\times V} = (f,x^\star)_U = 0$ for all $x^\star \in E'(\lambda)$.

Proof. With $V \subset U$, $V \subset V'$ is also a compact embedding, i.e., the inclusion $I : V \to V'$ is compact. $A + C_K I$ with C_K from (6.48) satisfies

$$A + C_K I \in L(V,V') \quad \text{and} \quad (A + C_K I)^{-1} \in L(V',V)$$

(cf. Exercise 6.106a). Lemma 6.82 shows that $K := (A + C_K I)^{-1} I : V \to V$ is compact. Hence the Riesz–Schauder theory is applicable to $K - \mu I$. Since

[5] If $a(\cdot,\cdot)$ is a sesquilinear form, the dual operator A' should be replaced by the adjoint A^*. In the finite-dimensional case the transposed matrix becomes the Hermitian transposed. The eigenvalue of the adjoint problem is $\bar{\lambda}$.

$$\begin{aligned} K - \mu I &= -\mu(I - \tfrac{1}{\mu}K) = -\mu(A + C_K I)^{-1}(A + C_K I - \tfrac{1}{\mu} I) \\ &= -\mu(A + C_K I)^{-1}(A - \lambda I) \qquad \text{with } \lambda = -C_K + \tfrac{1}{\mu}, \end{aligned}$$

the statements of Theorem 6.89 transfer via $K - \mu I$ to the shifted operator $A - \lambda I = -\mu^{-1}(A + C_K I)(K - \mu I)$. ∎

Remark 6.108. The spectrum $\sigma(A)$ has measure zero so that under the conditions of Theorem 6.107 the solvability of $Ax - \lambda x = f$ is guaranteed for almost all λ. Problem (6.45) is solvable if not "accidentally" $0 \in \sigma(A)$.

Lemma 6.109. *Under the conditions of Theorem 6.107 the inequalities (6.43a) and (6.43b) are equivalent.*

Proof. (6.43a) proves that A is injective, i.e., $0 \notin \sigma(A)$. Theorem 6.107 shows $A^{-1} \in L(V', V)$ so that (6.43b) follows from Lemma 6.94. Analogously, (6.43b) implies (6.43a). ∎

Evidently $a(\cdot,\cdot)$ remains V-coercive if one adds a multiple of $(\cdot,\cdot)_U$. A more general statement is the following.

Lemma 6.110. *Let $a(\cdot,\cdot)$ be V-coercive where $V \subset U \subset V'$. Then $a(\cdot,\cdot) + b(\cdot,\cdot)$ is also V-coercive if the bilinear form $b(\cdot,\cdot)$ satisfies one of the following three conditions:*

(a) for every $\varepsilon > 0$ exists C_ε such that

$$|b(x,x)| \le \varepsilon \|x\|_V^2 + C_\varepsilon \|x\|_U^2 \qquad \text{for all } x \in V. \tag{6.49a}$$

(b) Let the embeddings $V \subset X$ and $V \subset Y$ be continuous, with at least one of them compact. Let the following hold:

$$|b(x,x)| \le C_B \|x\|_X \|x\|_Y \qquad \text{for all } x \in V. \tag{6.49b}$$

(c) Let the embeddings $V \subset X$, $V \subset Y$ be continuous. Let (6.49b) hold. For $\|\cdot\|_X$ or $\|\cdot\|_Y$ assume that for every $\varepsilon > 0$ there exists a C'_ε such that

$$\|x\|_X \le \varepsilon \|x\|_V + C'_\varepsilon \|x\|_U \quad \text{or} \quad \|x\|_Y \le \varepsilon \|x\|_V + C'_\varepsilon \|x\|_U \quad (x \in V). \tag{6.49c}$$

Proof. (a) Select $\varepsilon = C_E/2$ in (6.49a) with C_E from (6.48). Then $a(\cdot,\cdot) + b(\cdot,\cdot)$ satisfies the V-coercivity condition with $C_E/2 > 0$ and $C_K + C_\varepsilon$ instead of C_E and C_K.

(b) Lemma 6.90 proves (6.49c).

(c) Let, e.g., the first inequality in (6.49c) hold. Since the embedding $V \subset Y$ is continuous, C_Y exists with $\|x\|_Y \le C_Y \|x\|_V$. Choose $\varepsilon' = \frac{\varepsilon}{2C_B C_Y}$ in (6.49c):

$$|b(x,x)| \underset{(6.49\text{b,c})}{\le} C_B \left(\varepsilon' \|x\|_V + C'_{\varepsilon'} \|x\|_U\right) C_Y \|x\|_V \underset{(5.34)}{\le} \frac{\varepsilon}{2} \|x\|_V^2 + \frac{1}{2\varepsilon} K^2 \|x\|_U^2$$

with $K := C_B C_Y C'_\varepsilon$. Hence (6.49a) is shown. ∎

Chapter 7
Variational Formulation

Abstract Techniques based on classical function spaces are less suited for proving the existence of a solution of a boundary-value problem. **Section 7.1** introduces another approach via a variational problem (Dirichlet's principle). Combining the variational formulation with the Sobolev spaces will be successful. In **Section 7.2** the boundary-value problem of order $2m$ with homogeneous Dirichlet conditions is transferred into the variational formulation in the space $H_0^m(\Omega)$. Existence of a solution in $H_0^m(\Omega)$ follows in Theorem 7.8 from the $H_0^m(\Omega)$-ellipticity which is discussed, e.g., in the Theorems 7.3 and 7.7. In **Section 7.3** we consider inhomogeneous Dirichlet boundary-value problems. The natural boundary condition in **Section 7.4** follows from variation in $H^m(\Omega)$ without any restrictions. In the case of the Poisson equation one obtains the Neumann condition, in the general case the conormal boundary derivative appears. We investigate how general boundary conditions can be formulated as variational problem. Complications appearing for differential equations of higher order are explained by taking the example of the biharmonic equation.

7.1 Historical Remarks About the Dirichlet Principle

In the preceding chapters it was not possible to establish even for the Dirichlet problem of the potential equation (2.1a,b) whether, or under what conditions, a classical solution $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$ exists. Green [121] took the view that his Green's function, described in 1828, always exists and that it provides the solution explicitly. This is not the case. Lebesgue [188] proved in 1913 that for certain domains the Green function does not exist.

Gauss (1840) and Kelvin (1847) offered a different line of reasoning (cf. [260, §V.2.4]). The *Dirichlet integral*

$$J(u) := \int_\Omega |\nabla u|^2 \, \mathrm{d}\mathbf{x} = \int_\Omega \sum_{i=1}^n u_{x_i}^2(\mathbf{x}) \mathrm{d}\mathbf{x} \tag{7.1}$$

W. Hackbusch, *Elliptic Differential Equations*, Springer Series
in Computational Mathematics 18, DOI 10.1007/978-3-662-54961-2_7

describes the energy in physics. With boundary values $u = \varphi$ on Γ given, one seeks to minimise $J(u)$. Let u be the minimising function and $v \neq 0$ any other function with zero boundary values: $v = 0$ on Γ. Since $u + \lambda v$ also takes the described boundary values, $j(\lambda) := J(u + \lambda v)$ as a function of $\lambda \in \mathbb{R}$ must be minimal at $\lambda = 0$. Inserting $u + \lambda v$ yields

$$j(\lambda) = J(u + \lambda v) = J(u) + 2\lambda\, I(u, v) + \lambda^2 J(v) \quad \text{with}$$
$$I(u, v) := \int_\Omega \langle \nabla u, \nabla v \rangle \, \mathrm{d}\mathbf{x} = \int_\Omega \sum_{i=1}^{n} u_{x_i}(\mathbf{x}) v_{x_i}(\mathbf{x}) \mathrm{d}\mathbf{x}.$$

Since $v \neq 0$ cannot be constant because it vanishes on the boundary, we have $j''(0) = 2J(v) > 0$, indicating a strict minimum at $\lambda = 0$. The necessary condition $j'(0) = 0$ for an extremum leads to the variational problem

$$a(u, v) \equiv \int_\Omega \sum_{i=1}^{n} u_{x_i}(\mathbf{x}) v_{x_i}(\mathbf{x}) \mathrm{d}\mathbf{x} = 0 \qquad \text{for all } v \text{ with } v = 0 \text{ on } \Gamma. \tag{7.2}$$

Green's formula (2.6a) provides[1] $a(u, v) = \int_\Omega v(-\Delta u)\mathrm{d}\mathbf{x} = 0$ for all v with $v = 0$ on Γ so that $-\Delta u = 0$. Thus, like (7.2), the variation problem $J(u) = \min$ is equivalent to the Dirichlet problem $-\Delta u = 0$ in Ω, $u = \varphi$ on Γ.

The so-called Dirichlet principle[2] states that $J(u)$, since it is bounded from below by $J(u) \geq 0$, must take a minimum for some u. Existence of a minimum was considered as evident needing no strict proof. According to the above considerations this would ensure the existence of a solution of the Dirichlet problem.

In 1870, Weierstrass argued against this line of reasoning, stating that while there may exist an infimum of $J(u)$, it need not necessarily be taken by some u. As a counterexample he showed that

$$J(u) := \int_{-1}^{1} x^2 (u'(x))^2 \mathrm{d}x \quad \text{in } \{u \in C^1([-1, 1]) : \ u(-1) = 0, \ u(1) = 1\},$$

has no minimiser taking the value $\inf J(u) = 0$.

Further, the following example due to Hadamard [148] shows that no *finite* infimum of the Dirichlet integral need exist. Let r and φ be the polar coordinates in the circle $\Omega = K_1(0)$. The function

$$u(r, \varphi) = \sum_{n=1}^{\infty} r^{n!} n^{-2} \sin(n!\varphi)$$

is harmonic in Ω and continuous in $\overline{\Omega}$ but the integral $J(u)$ does not exist.

For further remarks on the history of the calculus of variations we refer to Blanchard–Brüning [40, §0] and Stein [270].

[1] Here we assume that u allows integration by parts.

[2] The name *Dirichlet principle* is introduced by Riemann.

In the reasonings about the Dirichlet principle we have not exactly specified the set of functions in which we want to find a minimiser. The above difficulties disappear if one seeks the solutions in the more suitable Sobolev spaces instead of in $C^2(\Omega) \cap C^0(\overline{\Omega})$. The result, however, will be a solution of (7.2) called the *weak solution*. To obtain a classical solution $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$ of the Laplace problem, further conditions are to be satisfied (cf. Chapter 9).

7.2 Equations with Homogeneous Dirichlet Boundary Conditions

In the following we investigate the elliptic equation

$$Lu = g \qquad \text{in } \Omega \tag{7.3a}$$

with the differential operator

$$L = \sum_{|\alpha|\le m} \sum_{|\beta|\le m} (-1)^{|\beta|} D^\beta a_{\alpha\beta}(\mathbf{x}) D^\alpha \tag{7.3b}$$

of order $2m$ (cf. §5.3.2; Exercise 5.40d). The principal part of L is[3]

$$L_0 = (-1)^m \sum_{|\alpha|=|\beta|=m} D^\beta a_{\alpha\beta}(\mathbf{x}) D^\alpha.$$

According to Definition 5.39, L is *uniformly elliptic* in $\overline{\Omega}$ if there exists $\varepsilon > 0$ such that

$$\sum_{|\alpha|=|\beta|=m} a_{\alpha\beta}(\mathbf{x}) \boldsymbol{\xi}^{\alpha+\beta} \ge \varepsilon \, |\boldsymbol{\xi}|^{2m} \qquad \text{for all } \mathbf{x} \in \Omega \text{ and all } \boldsymbol{\xi} \in \mathbb{R}^n. \tag{7.4}$$

In the case that only $a_{\alpha\beta} \in L^\infty(\overline{\Omega})$ is assumed, one needs to replace "for all $x \in \Omega$" by "for almost all $x \in \Omega$".

7.2.1 Dirichlet Boundary Condition

We assume the *homogeneous Dirichlet boundary conditions*

$$u = 0, \quad \frac{\partial u}{\partial n} = 0, \quad \left(\frac{\partial}{\partial n}\right)^2 u = 0, \quad \ldots, \quad \left(\frac{\partial}{\partial n}\right)^{m-1} u = 0 \qquad \text{on } \Gamma, \tag{7.5}$$

[3] $D^\beta a_{\alpha\beta}(\mathbf{x}) D^\alpha$ describes the differential operator $u \mapsto D^\beta \left[a_{\alpha\beta}(\mathbf{x}) D^\alpha u\right]$. The expression $D^\beta a_{\alpha\beta}(\mathbf{x}) D^\alpha u(\mathbf{x})$ always means $D^\beta \left[a_{\alpha\beta}(\mathbf{x}) D^\alpha u(\mathbf{x})\right]$. If we want to apply D^β only to $a_{\alpha\beta}(\mathbf{x})$, we write $\left[D^\beta a_{\alpha\beta}(\mathbf{x})\right] D^\alpha u(\mathbf{x})$.

which are only meaningful if $\Gamma = \partial\Omega$ is sufficiently smooth (otherwise the normal direction $\mathbf{n}$ is not defined). Note that in the standard case $m = 1$ (an equation of second order) condition (7.5) becomes $u = 0$.

Since with $u = 0$ on Γ the tangential derivatives also vanish, not only the k-th normal derivatives ($k \le m-1$) but *all* the derivatives of order $\le m-1$ are equal to zero:

$$D^\alpha u = 0 \qquad \text{on } \Gamma \text{ for } |\alpha| \le m-1. \tag{7.5'}$$

Condition (7.5′) no longer requires the existence of a normal direction. According to Corollary 6.61, (7.5′) can also be formulated as

$$u \in H_0^m(\Omega).$$

7.2.2 Weak Formulation

Let $u \in C^{2m}(\Omega) \cap H_0^m(\Omega)$ be a classical solution[4] of (7.3a) and (7.5). To derive the variational formulation we take an arbitrary $v \in C_0^\infty(\Omega)$ and consider

$$(Lu, v)_0 = \sum_{|\alpha|\le m} \sum_{|\beta|\le m} (-1)^{|\beta|} \int_\Omega v(\mathbf{x})\, D^\beta a_{\alpha\beta}(\mathbf{x})\, D^\alpha u(\mathbf{x})\, \mathrm{d}\mathbf{x}.$$

Since $v \in C_0^\infty(\Omega)$, the integrand vanishes in the proximity of Γ so that one can integrate by parts,

$$(-1)^{|\beta|} \int_\Omega v(\mathbf{x})\, D^\beta a_{\alpha\beta}(\mathbf{x}) D^\alpha u(\mathbf{x})\, \mathrm{d}\mathbf{x} = \int_\Omega a_{\alpha\beta}(\mathbf{x})\, \left[D^\alpha u(\mathbf{x})\right]\, \left[D^\beta v(\mathbf{x})\right]\, \mathrm{d}\mathbf{x},$$

without boundary terms occurring. Thus we have found the variational formulation

$$\sum_{|\alpha|\le m} \sum_{|\beta|\le m} \int_\Omega a_{\alpha\beta}(\mathbf{x})\, \left[D^\alpha u(\mathbf{x})\right]\, \left[D^\beta v(\mathbf{x})\right]\, \mathrm{d}\mathbf{x} = \int_\Omega g(\mathbf{x})\, v(\mathbf{x})\, \mathrm{d}\mathbf{x} \qquad \text{for all } v \in C_0^\infty(\Omega) \tag{7.6}$$

since $Lu = g$.

If conversely $u \in C^{2m}(\Omega)$ with boundary conditions (7.5) satisfies condition (7.6), then the partial integration can be reversed and

$$\int_\Omega (g - Lu)\, v\, \mathrm{d}\mathbf{x} = 0 \qquad \text{for all } v \in C_0^\infty(\Omega)$$

[4] The intersection with $H_0^m(\Omega)$ is used to ensure the boundary condition (7.5) or (7.5′). If Ω is unbounded, $C^{2m}(\Omega) \cap H^m(\Omega)$ also implies that the squared derivatives of order $\le m$ have a finite integral.

proves $Lu = g$. This means that a classical solution of the variational problem (7.6) with boundary condition (7.5) is also a solution of the original boundary-value problem. Hence the differential equation (7.3a,b) and the variational formulation (7.6) are equivalent with respect to classical solutions. However, (7.6) may possess a nonclassical (weak) solution $u \in H_0^m(\Omega)$ for which the partial integration cannot be reversed.

We introduce the bilinear form

$$a(u,v) := \sum_{|\alpha|\le m} \sum_{|\beta|\le m} \int_\Omega a_{\alpha\beta}(\mathbf{x}) \; [D^\alpha u(\mathbf{x})] \; [D^\beta v(\mathbf{x})] \, \mathrm{d}\mathbf{x} \tag{7.7}$$

and the functional

$$f(v) := \int_\Omega g(\mathbf{x}) \, v(\mathbf{x}) \, \mathrm{d}\mathbf{x} \, . \tag{7.8}$$

As remarked above, the boundary condition (7.5) for classical solutions $u \in C^{2m}(\Omega) \cap H^m(\Omega)$ means that $u \in H_0^m(\Omega)$. Thus the *variational formulation* or *weak formulation* of the boundary-value problem (7.3a), (7.5) reads as follows: [5]

$$\text{find } u \in H_0^m(\Omega) \text{ with } \quad a(u,v) = f(v) \quad \text{for all } v \in C_0^\infty(\Omega). \tag{7.9}$$

A solution of problem (7.9) which, according to the definition, lies in $H_0^m(\Omega)$ but not necessarily in $C^{2m}(\Omega)$, is called a *weak solution*.

Exercise 7.1. (a) Let Ω be bounded. Show that any classical solution belonging to $u \in C^{2m}(\Omega) \cap C^m(\overline{\Omega})$ is also a weak solution.

(b) With the aid of Example 2.26 show that this statement becomes false for unbounded domains.

Theorem 7.2. *Let $a_{\alpha\beta} \in L^\infty(\Omega)$. The bilinear form defined by (7.7) is bounded on $H_0^m(\Omega) \times H_0^m(\Omega)$.*

Proof. Let $u, v \in C_0^\infty(\Omega)$. The inequality (6.11c) yields

$$|a(u,v)| \le \sum_{|\alpha|\le m} \sum_{|\beta|\le m} \|a_{\alpha\beta}\|_{L^\infty(\Omega)} \, |D^\alpha u|_0 \, \left|D^\beta v\right|_0 \le \text{const} \; |u|_m \, |v|_m \, .$$

Since $C_0^\infty(\Omega)$ is dense in $H_0^m(\Omega)$ (cf. Theorem 6.28), $a(\cdot,\cdot)$ has an continuous extension to $H_0^m(\Omega) \times H_0^m(\Omega)$ and is bounded by the same constant (cf. Lemma 6.91b). ∎

The function $f(v)$ is also defined and bounded for $v \in H_0^m(\Omega)$ if, for example, $g \in L^2(\Omega)$. According to Exercise 6.103, the variational formulation (7.9) is equivalent to the following one:

[5] In the complex case with a sesquilinear form $a(\cdot,\cdot)$, $f(\cdot)$ must be antilinear, i.e., $\overline{f(\cdot)}$ is a linear functional.

$$\text{find } u \in H_0^m(\Omega) \text{ with } \quad a(u,v) = f(v) \quad \text{for all } v \in H_0^m(\Omega). \tag{7.10}$$

One can regain the form $Lu = f$ by applying Lemma 6.91. Let $f \in H^{-m}(\Omega) = (H_0^m(\Omega))'$ and $L \in L(H_0^m(\Omega), H^{-m}(\Omega))$ be defined by

$$a(u,v) = \langle Lu, v\rangle_{H^{-m}(\Omega)\times H_0^m(\Omega)} \quad \text{and} \quad f(v) = \langle f, v\rangle_{H^{-m}(\Omega)\times H_0^m(\Omega)}$$

for all $v \in H_0^m(\Omega)$. Equation (7.10) states that

$$Lu = f. \tag{7.10$'$}$$

While (7.3a) represents an equation $Lu = g$ in $C^0(\Omega)$, (7.10′) is an equation in the dual space $H^{-m}(\Omega)$.

7.2.3 $H_0^m(\Omega)$-Ellipticity

Theorem 6.101 guarantees unique solvability of equation (7.10) if $a(\cdot,\cdot)$ is $H_0^m(\Omega)$-elliptic. We first investigate the standard case $m = 1$ (equations of order $2m = 2$).

Theorem 7.3. *Let Ω be bounded, $m = 1$, $a_{\alpha\beta} \in L^\infty(\Omega)$. Let L satisfy (7.4) (uniform ellipticity) and be equal to the principal part L_0, i.e., $a_{\alpha\beta} = 0$ for $|\alpha|+|\beta| \le 1$. Then the bilinear form $a(\cdot,\cdot)$ is $H_0^1(\Omega)$-elliptic:*

$$a(u,u) \ge \varepsilon' |u|_1^2 \qquad \textit{for all } u \in H_0^1(\Omega), \textit{ where } \varepsilon' > 0. \tag{7.11}$$

Proof. Since $|\alpha| = |\beta| = 1$ one can identify α and β according to $D^\alpha = \partial/\partial x_i$, $D^\beta = \partial/\partial x_j$ with indices $i,j \in \{1,\dots,n\}$. For fixed $\mathbf{x} \in \Omega$ use (7.4) with $\boldsymbol{\xi} = \nabla u(\mathbf{x})$:

$$\sum_{|\alpha|=|\beta|=1} a_{\alpha\beta}(\mathbf{x}) \left[D^\alpha u(\mathbf{x})\right] \left[D^\beta u(\mathbf{x})\right] = \sum_{|\alpha|=|\beta|=1} a_{\alpha\beta}(\mathbf{x}) \boldsymbol{\xi}^{\alpha+\beta}$$
$$= \sum_{i,j=1}^n a_{ij}(\mathbf{x}) \xi_i \xi_j \underset{(7.4)}{\ge} \varepsilon |\boldsymbol{\xi}|^2 = \varepsilon |\nabla u(\mathbf{x})|^2 .$$

Integration over Ω yields $a(u,u) \ge \varepsilon \int_\Omega |\nabla u|^2 \, \mathrm{d}\mathbf{x}$. Since $\int_\Omega |\nabla u|^2 \, \mathrm{d}\mathbf{x} \ge C_\Omega |u|_1^2$ (cf. Lemma 6.29), (7.11) follows with $\varepsilon' = \varepsilon C_\Omega$. ∎

Corollary 7.4. The condition "Ω bounded" may be dropped if for $\alpha = \beta = 0$ one assumes $a_{00}(\mathbf{x}) \ge \eta > 0$ (instead of $a_{00} = 0$ in Theorem 7.3).

Proof. Repeat the argument of the last proof: $a(u,u) \ge \int_\Omega \left(\varepsilon |\nabla u|^2 + \eta |u|^2\right) \mathrm{d}\mathbf{x} \ge \min\{\varepsilon, \eta\} |u|_1^2$. ∎

Example 7.5. The bilinear form $a(u,v) = \int_\Omega [\,\langle \nabla u, \nabla v\rangle + u\,v\,]\,\mathrm{d}\mathbf{x}$ of the *Helmholtz equation* $-\Delta u + u = f$ in Ω coincides with the scalar product in $H_0^1(\Omega)$ and $H^1(\Omega)$. From this we conclude the following.

(a) The $H_0^1(\Omega)$-ellipticity (Dirichlet boundary condition) as well as the $H^1(\Omega)$-ellipticity (Neumann boundary condition) holds with the constant $C_E = 1$ in (6.44).

(b) The operator $A : V \to V'$ from Lemma 6.91 in $a(u,v) = \langle Au, v\rangle_{V'\times V}$ for $V = H_0^1(\Omega)$ or $V = H^1(\Omega)$ is a symmetric, unbounded operator on $U = L^2$ with a dense domain.[6] Hence $A^{1/2}$ exists and belongs to $L(V,U)$ as well as to $L(U,V')$ (cf. Lions–Magenes [194, page 10] or Kato [168, Chap. V, §3.11]).

(c) The $H^1(\Omega)$-norm can be written as $\|u\|_{H^1(\Omega)} = \|A^{1/2}u\|_{L^2(\Omega)}$. For $|s| \le 1$, $\|A^s \cdot\|_{L^2(\Omega)}$ is equivalent to the norm $\|\cdot\|_{H^s(\Omega)}$ (if $s \ge 0$) and $\|\cdot\|_{(H^s(\Omega))'}$ (if $s \le 0$).

Exercise 7.6. Let the assumptions of Theorem 7.3 or Corollary 7.4 be satisfied, except for the fact that the coefficients $a_{\alpha 0}$ and $a_{0\beta}$ ($|\alpha| = |\beta| = 1$) of the first derivatives may be arbitrary constants. Show that inequality (7.11) holds unchanged.

Theorem 7.3 cannot easily be extended to the case $m > 1$.

Theorem 7.7. *Let the coefficients of the principal part be constants:* $a_{\alpha\beta} = \mathrm{const}$ *for* $|\alpha| = |\beta| = m$. *Furthermore, assume that* $a_{\alpha\beta} = 0$ *for* $0 < |\alpha| + |\beta| \le 2m-1$ *and* $a_{00}(\mathbf{x}) \ge 0$ *for* $\alpha = \beta = 0$. *Let* L *be uniformly elliptic (cf. (7.4)). Further let either* Ω *be bounded or* $a_{00} \ge \eta > 0$. *Then* $a(\cdot,\cdot)$ *is* $H_0^m(\Omega)$*-elliptic.*

Proof. We continue $u \in H_0^m(\Omega)$ through $u = 0$ onto $\mathbb{R}^n$. Theorem 6.39 and Exercise 6.40 show that

$$
\begin{aligned}
a(u,u) - \int_\Omega a_{00}\,u^2\,\mathrm{d}\mathbf{x} &= \sum_{|\alpha|=|\beta|=m} \int_\Omega a_{\alpha\beta}\,D^\alpha u\,D^\beta u\,\mathrm{d}\mathbf{x} \\
&= \sum_{|\alpha|=|\beta|=m} a_{\alpha\beta} \int_{\mathbb{R}^n} D^\alpha u(\mathbf{x}) D^\beta u(\mathbf{x})\,\mathrm{d}\mathbf{x} = \sum_{|\alpha|=|\beta|=m} a_{\alpha\beta}\cdot\left(D^\alpha u, D^\beta u\right)_0 \\
&= \sum_{|\alpha|=|\beta|=m} a_{\alpha\beta}\cdot\left(\widehat{D^\alpha u}, \widehat{D^\beta u}\right)_0 = \sum_{|\alpha|=|\beta|=m} a_{\alpha\beta} \int_{\mathbb{R}^n} (\mathrm{i}\xi)^\alpha\,\hat u(\boldsymbol{\xi})\,\overline{(\mathrm{i}\boldsymbol{\xi})^\beta\,\hat u(\boldsymbol{\xi})}\,\mathrm{d}\boldsymbol{\xi} \\
&= \int_{\mathbb{R}^n} \left[\sum_{|\alpha|=|\beta|=m} a_{\alpha\beta}\,\boldsymbol{\xi}^{\alpha+\beta}\right] |\hat u(\boldsymbol{\xi})|^2\,\mathrm{d}\boldsymbol{\xi} \underset{(7.4)}{\ge} \varepsilon \int_{\mathbb{R}^n} |\boldsymbol{\xi}|^{2m}\,|\hat u(\boldsymbol{\xi})|^2\,\mathrm{d}\boldsymbol{\xi}\,.
\end{aligned}
$$

Let $a_{00} \ge \eta > 0$. There exists an $\varepsilon' > 0$, so that

$$
\varepsilon\,|\boldsymbol{\xi}|^{2m} \ge \varepsilon' \sum_{|\alpha|\le m} |\boldsymbol{\xi}^\alpha|^2 - \eta \qquad \text{for all } \boldsymbol{\xi} \in \mathbb{R}^n.
$$

[6] The domain of A is $\{u \in U : Au \in U\}$.

From this follows $\varepsilon \int |\boldsymbol{\xi}|^{2m} |\hat{u}(\boldsymbol{\xi})|^2 \, \mathrm{d}\boldsymbol{\xi} \geq \varepsilon' |u|_m^2 - \eta |u|_0^2$ (cf. Lemma 6.41a) and $a(u,u) \geq \varepsilon' |u|_m^2$. If Ω is bounded, use Lemma 6.29. ∎

For real coefficients we can without loss of generality assume that the sum $\sum_{|\alpha|=|\beta|=m} a_{\alpha\beta}(\mathbf{x}) \boldsymbol{\xi}^{\alpha+\beta}$ in (7.4) is positive. This is not expected in the case of complex coefficients. The previous proof can be transferred if we assume that there is a complex number θ with $|\theta| = 1$ so that

$$\Re\mathfrak{e} \left(\theta \sum_{|\alpha|=|\beta|=m} a_{\alpha\beta}(\mathbf{x}) \, \boldsymbol{\xi}^{\alpha+\beta} \right) \geq \varepsilon \, |\boldsymbol{\xi}|^{2m} \qquad \text{for all } \mathbf{x} \in \Omega \, .$$

Applying the proof of Theorem 7.3 to $\theta \sum a_{\alpha\beta}(\mathbf{x}) \, [D^\alpha u(\mathbf{x})] \, [D^\beta u(\mathbf{x})]$ yields

$$|a(u,u)| \geq \Re\mathfrak{e} \, [\theta \, a(u,u)] \geq \varepsilon \int_\Omega |\nabla u|^2 \, \mathrm{d}\mathbf{x} \, .$$

This implies the V-ellipticity in the form (6.44′). Correspondingly, Theorem 7.7 can be transferred.

Having shown the $H_0^m(\Omega)$-ellipticity of the form $a(\cdot,\cdot)$, we are now able to apply Theorem 6.101.

Theorem 7.8 (existence and uniqueness of weak solutions). *If the bilinear form $a(\cdot,\cdot)$ is $H_0^m(\Omega)$-elliptic then there exists a solution $u \in H_0^m(\Omega)$ of problem (7.10) which satisfies*

$$|u|_m \leq \frac{1}{C_E} \, |f|_{-m} \qquad (C_E \text{ from } (6.44)) \, . \tag{7.12}$$

Since (7.12) holds for all $f \in H^{-m}(\Omega)$ and $u = L^{-1} f$ (cf. (7.10′)), inequality (7.12) is equivalent to

$$\|L^{-1}\|_{H_0^m(\Omega) \leftarrow H^{-m}(\Omega)} \leq C := 1/C_E \, . \tag{7.12′}$$

The variational problem (7.10) can be formulated as a minimisation problem.

Theorem 7.9. *Let $a(\cdot,\cdot)$ be an $H_0^m(\Omega)$-elliptic and symmetric bilinear form. Then (7.10) is equivalent to the minimisation problem*

$$\begin{aligned} &\textit{find } u \in H_0^m(\Omega) \textit{ such that} \quad J(u) \leq J(v) \quad \textit{for all } v \in H_0^m(\Omega), \\ &\textit{where} \quad J(v) := \frac{1}{2} a(v,v) - f(v). \end{aligned} \tag{7.13}$$

Attention. If $a(\cdot,\cdot)$ is either not $H_0^m(\Omega)$-elliptic or not symmetric, problem (7.10) remains meaningful although its solution does not minimise the functional $J(u)$.

Example 7.10. The *Poisson equation* $-\Delta u = f$ in Ω, $u = 0$ on Γ, leads to the bilinear form

$$a(u,v) = \int_\Omega \langle \nabla u(\mathbf{x}), \nabla v(\mathbf{x}) \rangle \mathrm{d}\mathbf{x}\,.$$

For a bounded domain Ω, $a(\cdot,\cdot)$ is $H_0^1(\Omega)$-elliptic (cf. Theorem 7.3), so that for any $f \in H^{-1}(\Omega)$ there exists exactly one (weak) solution $u \in H_0^1(\Omega)$ of the Poisson equation. This is also the solution of the minimisation problem

$$\frac{1}{2}\int_\Omega |\nabla u|^2 \,\mathrm{d}\mathbf{x} - f(u) = \min\,.$$

7.2.4 $H_0^m(\Omega)$-Coercivity

A weaker condition than $H_0^m(\Omega)$-ellipticity is the $H_0^m(\Omega)$-coercivity from (6.48): $a(u,u) \geq \varepsilon\,|u|_m^2 - C\,|u|_0^2$ with $\varepsilon > 0$.

Theorem 7.11. *Let $m = 1$, and let the coefficients $a_{\alpha\beta} \in L^\infty(\Omega)$ satisfy condition (7.4) of uniform ellipticity. Then $a(\cdot,\cdot)$ is $H_0^1(\Omega)$-coercive.*

Proof. We write L as $L = L_I + L_{II}$, with L_I satisfying the conditions of Theorem 7.3, resp. Corollary 7.4 if Ω is not bounded, and L_{II} containing only derivatives of order ≤ 1. Then we can apply the following lemma. ∎

Lemma 7.12. *Let $a(\cdot,\cdot) = a'(\cdot,\cdot) + a''(\cdot,\cdot)$ be decomposed such that $a'(\cdot,\cdot)$ is $H_0^m(\Omega)$-elliptic, or perhaps only $H_0^m(\Omega)$-coercive, while*

$$a''(u,v) = \sum_{\substack{|\alpha|\leq m,\ |\beta|\leq m\\ |\alpha|+|\beta|<2m}} \int_\Omega a_{\alpha\beta}(\mathbf{x})\,[D^\alpha u(\mathbf{x})]\,\big[D^\beta v(\mathbf{x})\big]\,\mathrm{d}\mathbf{x}$$

with $a_{\alpha\beta} \in L^\infty(\Omega)$ contains only derivatives of order $|\alpha| + |\beta| \leq 2m-1$. Then $a(\cdot,\cdot)$ is also $H_0^m(\Omega)$-coercive.

Proof. (6.49c) follows from (6.16b) with $X = H_0^{|\alpha|}(\Omega)$, $Y = H_0^{|\beta|}(\Omega)$, $V = H_0^m(\Omega)$, and $U = L^2(\Omega)$, so that Lemma 6.110c proves the assertion. ∎

The generalisation of Theorem 7.11 to arbitrary $m \geq 1$ requires stronger conditions on the coefficients of the principal part.

Theorem 7.13 (Gårding [111]). *Let L be uniformly elliptic (cf. (7.4)) and assume $a_{\alpha\beta} \in L^\infty(\Omega)$. Furthermore, let the coefficients $a_{\alpha\beta}$ with $|\alpha| = |\beta| = m$ be uniformly continuous in Ω. Then $a(\cdot,\cdot)$ is $H_0^m(\Omega)$-coercive. If conversely*

$$a_{\alpha\beta} \in C(\Omega) \text{ for } |\alpha| = |\beta| = m \quad \text{and} \quad a_{\alpha\beta} \in L^\infty(\Omega) \text{ otherwise,}$$

then from the $H_0^m(\Omega)$-coercivity follows uniform ellipticity (7.4).

Details of the proof can be found in Wloka [308, Theorem 19.2]. The proof given there also holds for unbounded Ω, since the coefficients are uniformly continuous. For the first part of the theorem one uses a partition of unity.

The significance of coercivity lies in the following statement.

Theorem 7.14. *Let Ω be bounded and $a(\cdot,\cdot)$ be $H_0^m(\Omega)$-coercive. Then one of the following alternatives holds:*

(i) Problem (7.10) has exactly one (weak) solution $u \in H_0^m(\Omega)$.

(ii) The kernels $E = \ker(L)$ and $E^\star = \ker(L')$ are k-dimensional for a $k \in \mathbb{N}$, i.e.,

$$a(e,v) = 0 \quad \textit{and} \quad a(v,e^*) = 0 \quad \textit{for all } e \in E,\ e^* \in E^*,\ v \in H_0^m(\Omega).$$

Further, the eigenvalue problem

$$a(e,v) = \lambda\,(e,v)_{L^2(\Omega)} \qquad \textit{for all } v \in H_0^m(\Omega)$$

has countably many eigenvalues which do not accumulate in $\mathbb{C}$.

Proof. Since for bounded Ω the embedding $V := H_0^m(\Omega) \subset U := L^2(\Omega)$ is compact (cf. Theorem 6.86b), Theorem 6.107 is applicable. ∎

7.3 Inhomogeneous Dirichlet Boundary Conditions

Next, we consider the boundary-value problem

$$Lu = g \quad \text{in } \Omega, \qquad u = \varphi \quad \text{on } \Gamma, \tag{7.14}$$

where L is a differential operator of second order (i.e., $m = 1$). The corresponding variational formulation of the boundary-value problem reads:

$$\begin{gathered}\text{find } u \in H^1(\Omega) \text{ with } u = \varphi \text{ on } \Gamma \text{ such that} \\ a(u,v) = f(v) \text{ for all } v \in H_0^1(\Omega).\end{gathered} \tag{7.15}$$

According to §6.2.5 the restriction $u|_\Gamma$ of $u \in H^1(\Omega)$ on Γ is well defined as a function in $H^{1/2}(\Gamma)$. Thus "$u = \varphi$ on Γ" must be understood as the equality $u|_\Gamma = \varphi$ in $H^{1/2}(\Gamma)$. In contrast to the preceding section one uses $a(\cdot,\cdot)$ in (7.15) as a bilinear form on $H^1(\Omega) \times H_0^1(\Omega)$. It is easy to see that $a(\cdot,\cdot)$ is well defined and bounded on this product.

Remark 7.15. For the solvability of Problem (7.15) it is necessary that:

$$\text{there exists a} \quad u_0 \in H^1(\Omega) \quad \text{with} \quad u_0|_\Gamma = \varphi. \tag{7.16}$$

If a function u_0 with property (7.16) is known, a second characterisation of the weak solution results:

$$\text{Let } u_0 \text{ satisfy (7.16)}, \qquad \text{find } w \in H_0^1(\Omega) \text{ such that} \tag{7.17a}$$

$$a(w,v) = f_0(v) := f(v) - a(u_0,v) \quad \text{for all } v \in H_0^1(\Omega). \tag{7.17b}$$

Remark 7.16. The variational problems (7.15) and (7.17a,b) are equivalent. If u_0 and w are the solutions of (7.17a,b), then $u = u_0 + w$ is a solution of problem (7.15). If u is a solution of (7.15), then, for example, $u_0 = u$ and $w = 0$ satisfy problem (7.17a,b).

Exercise 7.17. Show that $f_0 \in H^{-1}(\Omega)$ for f_0 from (7.17b) and

$$|f_0|_{-1} \leq |f|_{-1} + C_S \, |u_0|_1 \tag{7.18}$$

with C_S from $|a(u,v)| \leq C_S \, |u|_1 \, |v|_1$ (cf. (6.42)).

Remark 7.18. The problem (7.14) and the variational formulation (7.15) have the same classical solutions if such exist.

Proof. It suffices to assume $v \in C_0^\infty(\Omega)$ in (7.15). Integration by parts can be carried out as in Section 7.2 and proves the assertion. ∎

Theorem 7.19 (existence and uniqueness). *Let problem (7.10) (with homogeneous boundary values) be uniquely solvable for all $f \in H^{-1}(\Omega)$. Then condition (7.16) is sufficient, and necessary, for the unique solvability of problem (7.15).*

Proof. If there exists a solution $u \in H^1(\Omega)$ of (7.15) then (7.16) is satisfied. However, if (7.16) holds, one obtains via (7.17a,b) a unique solution since (7.17b) agrees with (7.10). ∎

Remark 7.20. Under the condition $\Omega \in C^{0,1}$ (Lipschitz domain), (7.16) is equivalent to $\varphi \in H^{1/2}(\Gamma)$.

Proof. If $\varphi \in H^{1/2}(\Gamma)$, then Theorem 6.58 guarantees an extension $u_0 \in H^1(\Omega)$ to Ω with $u_0|_\Gamma = \varphi$ and

$$|u_0|_1 \leq C \, |\varphi|_{1/2} \, . \tag{7.19}$$

If conversely u_0 satisfies Condition (7.16) then the same theorem shows that $\varphi \in H^{1/2}(\Gamma)$. ∎

Let inequality (7.12′) hold in the case of homogeneous boundary-values. Equation (7.18) shows

$$|u|_1 \leq |u_0|_1 + |w|_1 \leq |u_0|_1 + (\, |f|_{-1} + C' \, |u_0|_1 \,)/\varepsilon$$

for the solution of problem (7.17a,b). The estimate (7.19) proves the next theorem.

Theorem 7.21. *Let $\Omega \in C^{0,1}$. Let the bilinear form be restricted to $H^1(\Omega) \times H_0^1(\Omega)$ and let it satisfy (7.12′). Then for every $f \in H^{-1}(\Omega)$ and $\varphi \in H^{1/2}(\Gamma)$ there exists exactly one solution $u \in H^1(\Omega)$ of problem (7.15) with*

$$|u|_1 \leq C \left[|f|_{-1} + |\varphi|_{1/2} \right] .$$

Exercise 7.22. Let $a(\cdot,\cdot)$ be symmetric and $H_0^1(\Omega)$-elliptic. Show that problem (7.15) is equivalent to the minimisation problem (cf. (7.1)):

$$\text{find } u \in H^1(\Omega) \text{ with } u|_\Gamma = \varphi \text{ such that } J(u) = a(u,u) - 2f(u) \text{ is minimal.}$$

7.4 Natural Boundary Conditions

7.4.1 Variation in $H^m(\Omega)$

The bilinear form $a(\cdot,\cdot)$ defined in (7.7) is also well defined on $H^m(\Omega) \times H^m(\Omega)$. In analogy to Theorem 7.2.2 there holds the next statement.

Theorem 7.23. *Assume $a_{\alpha\beta} \in L^\infty(\Omega)$. The bilinear form defined by (7.7) is bounded on $H^m(\Omega) \times H^m(\Omega)$:*

$$|a(u,v)| \le \sum_{\alpha,\beta} \|a_{\alpha\beta}\|_{L^\infty(\Omega)} |u|_m |v|_m \quad \textit{for all} \quad u, v \in H^m(\Omega).$$

Now let f be a functional from $(H^m(\Omega))'$. Equation (7.8) with $g \in L^2(\Omega)$, for example, describes such a functional; but (7.8) is only a special case of the functional f subsequently defined in (7.20a), which we want to use as a foundation in the following.

Exercise 7.24. Let Γ be sufficiently smooth and let $g \in L^2(\Omega)$, $\varphi \in L^2(\Gamma)$ hold. Show that

$$f(v) := \int_\Omega g(\mathbf{x}) v(\mathbf{x}) \mathrm{d}\mathbf{x} + \int_\Gamma \varphi(\mathbf{x}) v(\mathbf{x}) \mathrm{d}\Gamma \qquad \big(v \in H^1(\Omega)\big) \tag{7.20a}$$

defines a functional in $(H^1(\Omega))'$ with $\|f\|_{(H^1(\Omega))'} \le C \left[\|g\|_{L^2(\Omega)} + \|\varphi\|_{L^2(\Gamma)} \right]$. This implies $f \in (H^m(\Omega))'$ for all $m \ge 1$. More precisely, the following also holds:

$$\|f\|_{(H^1(\Omega))'} \le C \left[\|g\|_{(H^1(\Omega))'} + \|\varphi\|_{H^{-1/2}(\Gamma)} \right]. \tag{7.20b}$$

Frequently, variational problems of order $2m$ have the form:

$$\text{find } u \in H^m(\Omega) \text{ such that} \quad a(u,v) = f(v) \quad \text{for all } v \in H^m(\Omega). \tag{7.21}$$

In contrast to the condition $u \in H_0^m(\Omega)$ from Section 7.2, $u \in H^m(\Omega)$ contains no boundary condition. Nevertheless, Problem (7.21) has a unique solution if $a(\cdot,\cdot)$ is $H^m(\Omega)$-elliptic. For $m = 1$ this condition is easy to satisfy.

Theorem 7.25. *Under the conditions of Theorem 7.3 or Corollary 7.4 $a(\cdot,\cdot)$ is $H^1(\Omega)$-elliptic: $a(u,u) \ge \varepsilon |u|_1^2$ for all $u \in H^1(\Omega)$. Problem (7.21) (with $m = 1$) has exactly one solution which satisfies the estimate (7.22):*

$$|u|_1 \le \tfrac{1}{\varepsilon} \|f\|_{(H^1(\Omega))'} . \tag{7.22}$$

Proof. The same as for Theorem 7.3 or Corollary 7.4, and Theorem 7.8. ∎

Corollary 7.26. (a) A unique solution which satisfies the estimate (7.22), also exists if instead of $H^1(\Omega)$-ellipticity one assumes: $a(\cdot,\cdot)$ is $H^1(\Omega)$-coercive, $\Omega \in C^{0,1}$ is bounded, $\lambda = 0$ is not an eigenvalue (i.e., $a(u,v) = 0$ for all $v \in H^1(\Omega)$ implies $u = 0$).

(b) The combination of inequalities (7.22) and (7.20b) results in

$$|u|_1 \le C \left[\|g\|_{(H^1(\Omega))'} + \|\varphi\|_{H^{-1/2}(\Gamma)} \right]$$

for the solution of (7.21), if f is defined by (7.20a) with $g \in (H^1(\Omega))'$, $\varphi \in H^{-\frac{1}{2}}(\Gamma)$.

Proof. (a) According to Theorem 6.86b, $H^1(\Omega)$ is compactly embedded in $L^2(\Omega)$ so that the statement of Theorem 7.14 can be transferred. If $\lambda = 0$ is not an eigenvalue then $L^{-1} \in L((H^1(\Omega))', H^1(\Omega))$ (cf. Theorem 6.89). ∎

7.4.2 Conormal Boundary Condition

To find out which classical boundary-value problem corresponds to the variational formulation (7.21), we assume that (7.21) has a classical solution $u \in H^m(\Omega) \cap C^{2m}(\overline{\Omega})$. Further, $v \in C^\infty(\Omega)$ can be assumed also (cf. Lemma 6.91b). For reasons of simplicity we limit ourselves to the case $m = 1$. Under the assumption $a_{\alpha\beta} \in C^1(\Omega)$ and under suitable conditions on Ω the following general Green formula is applicable:

$$\begin{aligned} a(u,v) &= \int_\Omega \left[\sum_{i,j=1}^n a_{ij}\, u_{x_i}\, v_{x_j} + \sum_{i=1}^n a_{0i}\, u\, v_{x_i} + \sum_{i=1}^n a_{i0}\, u_{x_i}\, v + a_{00} uv \right] \mathrm{d}\mathbf{x} \\ &= \int_\Omega \left[-\sum_{i,j=1}^n \left(a_{ij} u_{x_i}\right)_{x_j} - \sum_{i=1}^n \left(a_{0i} u\right)_{x_i} + \sum_{i=1}^n a_{i0}\, u_{x_i} + a_{00}\, u \right] v \,\mathrm{d}\mathbf{x} \\ &\quad + \int_\Gamma \left[\sum_{i,j=1}^n n_j\, a_{ij}\, u_{x_i} + \sum_{i=1}^n n_i\, a_{0i}\, u \right] v \,\mathrm{d}\Gamma. \end{aligned} \tag{7.23}$$

Here, the n_i are the components of the normal direction $\mathbf{n} = \mathbf{n}(\mathbf{x})$, $\mathbf{x} \in \Gamma$. We define the *boundary differential operator*

$$B := \sum_{i,j=1}^n n_j\, a_{ij}\, \frac{\partial}{\partial x_i} + \sum_{i=1}^n n_i\, a_{0i}\,, \tag{7.24}$$

when L is described by (7.3b). Equation (7.23) becomes

$$a(u,v) = \int_\Omega v\, Lu \,\mathrm{d}\mathbf{x} + \int_\Gamma v\, Bu \,\mathrm{d}\Gamma\,.$$

By the formulation of the problem, $a(u,v)$ agrees with $f(v)$ from (7.20a). If we first choose $v \in H_0^1(\Omega) \subset H^1(\Omega)$ the boundary integrals drop out and we obtain $Lu = g$ as in Section 7.2. By this the identity $a(u,v) = f(v)$ reduces to $\int_\Gamma vBu\mathrm{d}\Gamma = \int_\Gamma \varphi v \mathrm{d}\Gamma$ for all $v \in H^1(\Omega)$. According to Theorem 6.58b, $v|_\Gamma$ runs over the set $H^{1/2}(\Gamma)$ if v runs over $H^1(\Omega)$, so that we have $\int_\Gamma \psi(Bu-\varphi)\mathrm{d}\Gamma = 0$ for all $\psi = v|_\Gamma \in H^{1/2}(\Gamma)$; thus $Bu = \varphi$. This proves the next theorem.

Theorem 7.27. *Let Γ be sufficiently smooth. A classical solution of problem (7.21) with f from (7.20a) is also the classical solution of the boundary-value problem*

$$Lu = g \quad \textit{in } \Omega, \qquad Bu = \varphi \quad \textit{on } \Gamma$$

and conversely.

The condition $Bu = \varphi$ is called the *natural boundary condition*. This results from the fact that in (7.21) (as distinct from (7.10)) the function u may assume arbitrary boundary values. Note that the bilinear form determines L as well as B.

Exercise 7.28. Show that the bilinear form from Example 7.10 for $-\Delta u = g$ has as the natural boundary condition the *Neumann condition* $\partial u/\partial n = \varphi$.

Theorem 7.29. *Let $\Omega \in C^{0,1}$ be bounded and $a(\cdot,\cdot)$ be $H^m(\Omega)$-coercive. Then the statements of Theorem 7.14 hold with $H^m(\Omega)$ instead of $H_0^m(\Omega)$.*

Example 7.30. Let $\Omega \in C^{0,1}$ be a bounded domain. The bilinear form $a(u,v) = \int_\Omega [\langle \nabla u, \nabla v\rangle + cuv]\,\mathrm{d}\mathbf{x}$, associated to the *Helmholtz equation*

$$-\Delta u + cu = f \quad \text{in } \Omega \text{ with } c > 0, \qquad \partial u/\partial n = \varphi \quad \text{on } \Gamma$$

is $H^1(\Omega)$-elliptic since $a(u,u) \geq \min(1,c)\,|u|_1^2$. For $c = 0$, however, $a(\cdot,\cdot)$ is only $H^1(\Omega)$-coercive. As is known from Theorem 3.28, the Neumann boundary-value problem for the Poisson equation (i.e., for $c = 0$) is not uniquely solvable. According to alternative (ii) in Theorem 7.14, there exists a nontrivial eigenspace $E = \ker(L)$. $u \in E$ satisfies $a(u,u) = \int_\Omega \langle \nabla u, \nabla u\rangle \mathrm{d}\mathbf{x} = 0$, thus $\nabla u = 0$. Since Ω is connected, it follows that $u(\mathbf{x}) = \mathrm{const}$ and therefore $\dim E = 1$. Since $a(\cdot,\cdot)$ is symmetric, $E^\star := \ker(L')$ coincides with E. According to Theorem 7.29, the Neumann boundary-value problem $a(u,v) = f(v)$ ($v \in H^1(\Omega)$) is solvable if and only if $f \perp E$, i.e., $f(1) = 0$. Here, 1 is the function with constant value 1. If $f(v) = \int_\Omega g(x)v(x)\mathrm{d}x$, the condition $f(1) = 0$ reads as $\int_\Omega g\mathrm{d}x = 0$. If, however, f is given by (7.20a), the integrability condition reads $f(1) = \int_\Omega g\mathrm{d}x + \int_\Gamma \varphi \mathrm{d}\Gamma = 0$ (this is equation (3.17), in which f should be replaced by g).

Remark 7.31. While the classical formulation of a boundary condition such as $\frac{\partial u}{\partial n} = 0$ requires conditions on the boundary Γ, the problem (7.21) can be formulated for arbitrary measurable Ω.

7.4.3 Oblique Boundary Condition

In the following, we proceed in the opposite direction: does there exist, for a classically formulated boundary-value problem $Lu = g$ in Ω, $Bu = \varphi$ on Γ, with given L and B, a bilinear form $a(\cdot,\cdot)$ such that (7.21) is the corresponding variational formulation? This would mean that the freely prescribed boundary operator B represents the natural boundary condition.

For $m = 1$ the general form of the boundary operator reads

$$B = \sum_{i=1}^{n} b_i(\mathbf{x}) \frac{\partial}{\partial x_i} + b_0(\mathbf{x}) \qquad (\mathbf{x} \in \Gamma). \tag{7.25}$$

With $\mathbf{b}^\mathsf{T} = (b_1, \ldots, b_n)$ one can also write $B = \mathbf{b}^\mathsf{T}\nabla + b_0$ (cf. (5.21b,b′)). Here $\mathbf{b}^\mathsf{T}\nabla$ is not allowed to be a tangential derivative (cf. Remark 5.27):

$$\langle \mathbf{b}(\mathbf{x}), \mathbf{n}(\mathbf{x}) \rangle \neq 0 \qquad \text{for all } \mathbf{x} \in \Gamma. \tag{7.26}$$

Remark 7.32. Let $m = 1$. Let (7.26) hold. Let $A(\mathbf{x})$ be the matrix $A = (a_{ij})$ (cf. (5.2)). By passing from $Bu = \varphi$ to the equivalent scaled equation $\sigma B u = \sigma\varphi$ with

$$\sigma(\mathbf{x}) = \langle \mathbf{n}(\mathbf{x}), A(\mathbf{x})\mathbf{n}(\mathbf{x}) \rangle \,/\, \langle \mathbf{b}(\mathbf{x}), \mathbf{n}(\mathbf{x}) \rangle \,,$$

one can ensure that $\langle \mathbf{n}, \sigma\mathbf{b} \rangle = \langle \mathbf{n}, A\mathbf{n} \rangle$. Thus in the following it is always assumed that $\mathbf{b}$ already satisfies $\langle \mathbf{n}, \mathbf{b} \rangle = \langle \mathbf{n}, A\mathbf{n} \rangle$.

Proof. Because of (7.26) σ is well defined. For $Bu = \varphi$ and $\sigma B u = \sigma\varphi$ to be equivalent, $\sigma \neq 0$ is required. This is guaranteed by the uniform ellipticity: $\langle \mathbf{n}, A\mathbf{n} \rangle \geq \varepsilon \left|\mathbf{n}\right|^2 = \varepsilon$. ∎

Theorem 7.33 (construction of the bilinear form). *Let $m = 1$. Let L and B be given by (7.3b) and (7.25), with $\mathbf{b}$ satisfying condition (7.26). Then there exists a bilinear form $a(\cdot,\cdot)$ on $H^1(\Omega) \times H^1(\Omega)$ such that to the variational problem (7.21) corresponds the classical formulation $Lu = g$ in Ω, $Bu = \varphi$ on Γ.*

Proof. The bilinear form we seek is not uniquely determined . We shall give two possibilities for its construction. First we discuss the absolute term in (7.25). On the basis of Remark 7.32 we assume $\langle \mathbf{n}, \mathbf{b} \rangle = \langle \mathbf{n}, A\mathbf{n} \rangle$.

(i) Let the vector function $\boldsymbol{\beta}(\mathbf{x}) := (\beta_1(\mathbf{x}), \ldots, \beta_n(\mathbf{x})) \in C^1(\overline{\Omega})$ be arbitrary. The differential operator

$$L_1 := -\sum_{i=1}^{n} \frac{\partial}{\partial x_i}\beta_i + \sum_{i=1}^{n} \beta_i \frac{\partial}{\partial x_i} + \sum_{i=1}^{n} (\beta_i)_{x_i}$$

maps every $u \in C^1(\Omega)$ into zero: $L_1 u = 0$ (use the product rule of differentiation). Thus the operator L can be replaced by $L + L_1$ without changing the boundary-value problem. Let $a(\cdot,\cdot)$ be constructed according to (7.7) from the coefficients of $L + L_1$. Equation (7.24) shows that the boundary operator associated with $a(\cdot,\cdot)$ has the absolute term

$$\sum_{i=1}^{n} n_i \left(a_{0i} + \beta_i\right). \tag{7.27}$$

If $b_0 = 0$, the choice of $\beta_i = -a_{0i}$ is successful. Otherwise, two other options are available.

(ia) Select β_i such that on Γ the following holds:

$$\beta_i = b_0 n_i - a_{0i} \qquad \text{on } \Gamma\,.$$

Since $|\mathbf{n}| = 1$, the term (7.27) then agrees with $b_0(\mathbf{x})$. The practical difficulty in this method consists in the need to construct a smooth continuation on Ω of the boundary values $\beta_i(\mathbf{x})\,,\ \mathbf{x} \in \Gamma$.

(ib) Set $\beta_i = -a_{0i}$ and add a suitable boundary integral:

$$a(u,v) := \int_\Gamma b_0\, u\, v \,\mathrm{d}\Gamma \tag{7.28}$$
$$+ \int_\Omega \left\{ \sum_{i,j=1}^{n} a_{ij} u_{x_i} v_{x_j} + \sum_{i=1}^{n} \left(a_{i0} - a_{0i}\right) u_{x_i} v + \left[a_{00} + \sum_{i=1}^{n} \left(a_{0i}\right)_{x_i} \right] uv \right\} \mathrm{d}\mathbf{x}.$$

The integration by parts described above shows that the boundary operator associated to (7.28) reads

$$\tilde{B} = \sum_{i,j=1}^{n} n_j\, a_{ij} \frac{\partial}{\partial x_i} + b_0\,. \tag{7.29}$$

(ii) The operator (7.29) can be written in the form $\tilde{B} = \tilde{\mathbf{b}}^{\mathsf{T}} \nabla + b_0$ with $\tilde{\mathbf{b}} = A\mathbf{n}$. Since $\langle \mathbf{n}, \tilde{\mathbf{b}} \rangle = \langle \mathbf{n}, A\mathbf{n} \rangle = \langle \mathbf{n}, \mathbf{b} \rangle$ has been assumed already, $\mathbf{d} := \mathbf{b} - \tilde{\mathbf{b}}$ is orthogonal to $\mathbf{n}$. To change $\tilde{\mathbf{b}}$ to $\mathbf{b}$ there are again two options.

(iia) Define the $n \times n$-matrix A^s on Γ by

$$A^s = \mathbf{d}\mathbf{n}^{\mathsf{T}} - \mathbf{n}\mathbf{d}^{\mathsf{T}}, \quad \text{i.e., } a^s_{ij} = d_i n_j - n_i d_j\,.$$

This $A^s(x)$ is skew-symmetric: $A^{s\mathsf{T}} = -A^s$. Continue $A^s(x)$, which is at first only defined on Γ, to a skew-symmetric matrix $A^s \in C^1(\Omega)$. [Here we have the same practical difficulty as in step (ia).] The entries of A^s define

$$L_2 := \sum_{i,j=1}^{n} \left[-\frac{\partial}{\partial x_j} a^s_{ij} \frac{\partial}{\partial x_i} + \left(a^s_{ij}\right)_{x_j} \frac{\partial}{\partial x_i} \right].$$

Again, $L_2 u = 0$ holds for all $u \in C^2(\Omega)$, since

$$-\left(a_{ij}^s u_{x_i}\right)_{x_j} - \left(a_{ji}^s u_{x_j}\right)_{x_i} + \left(a_{ij}^s\right)_{x_j} u_{x_i} + \left(a_{ji}^s\right)_{x_i} u_{x_j} = -a_{ij}^s u_{x_i x_j} - a_{ji}^s u_{x_j x_i} = 0.$$

Thus L can be replaced by $L + L_2$ without changing the boundary-value problem. The coefficients belonging to $L + L_2$ result in a boundary operator B whose derivative terms read

$$\sum_{i,j=1}^{n} n_j (a_{ij} + a_{ij}^s) \frac{\partial}{\partial x_i} = [\,(A + A^s)\mathbf{n}\,]^{\mathsf{T}} \nabla.$$

By the construction of A^s we have

$$A^s \mathbf{n} = (\mathbf{d}\mathbf{n}^{\mathsf{T}} - \mathbf{n}\mathbf{d}^{\mathsf{T}})\mathbf{n} = \mathbf{d} = \mathbf{b} - \tilde{\mathbf{b}},$$

since $\langle \mathbf{n}, \mathbf{n} \rangle = 1$ and $\langle \mathbf{d}, \mathbf{n} \rangle = 0$. Since we also have $A\mathbf{n} = \tilde{\mathbf{b}}$, the derivative term in B gives $\mathbf{b}^{\mathsf{T}} \nabla$ as desired. The transition $L \to L + L_2$ does not change the absolute term in B, so that from (7.29) follows $B = \mathbf{b}^{\mathsf{T}} \nabla + b_0$.

(iib) Let $B = \mathbf{b}^{\mathsf{T}} \nabla + b_0$ be given (cf. (7.29)) such that $\mathbf{d} = \mathbf{b} - \tilde{\mathbf{b}}$ is orthogonal to $\mathbf{n}$. From this the boundary operator

$$T := \mathbf{d}^{\mathsf{T}} \nabla$$

is the derivative in a tangential direction if $n = 2$ [resp. in the tangent hyperplane if $n \geq 3$]. If f is sufficiently smooth then the restriction $v|_\Gamma$ of $v \in H^1(\Omega)$ is an element of $H^{1/2}(\Gamma)$. By Remark 6.76a, one can show that

$$T \in L(H^{1/2}(\Gamma), H^{-1/2}(\Gamma)).$$

Since $T(u|_\Gamma) \in H^{-1/2}(\Gamma)$, $\int_\Gamma \psi \, T(u|_\Gamma) \, \mathrm{d}\Gamma$ is well defined for $\psi \in H^{1/2}(\Gamma)$, in particular, for $\psi = v|_\Gamma$ with $v \in H^1(\Omega)$. Thus

$$b(u, v) := \int_\Gamma (v|_\Gamma) \, T(u|_\Gamma) \, \mathrm{d}\Gamma$$

is a bilinear form bounded on $H^1(\Omega) \times H^1(\Omega)$. We add $b(\cdot, \cdot)$ to $a(\cdot, \cdot)$ in (7.28). Integration by parts yields the boundary operator

$$\tilde{B} + T = \tilde{\mathbf{b}}^{\mathsf{T}} \nabla + b_0 + (\mathbf{b} - \tilde{\mathbf{b}})^{\mathsf{T}} \nabla = \mathbf{b}^{\mathsf{T}} \nabla + b_0 = B. \qquad \blacksquare$$

Theorem 7.34. *Let the bilinear form (7.7) be $H^1(\Omega)$-coercive (cf. Theorem 7.25). Let its coefficients, as well as the boundary Γ, be sufficiently smooth ($\in C^1$). Then the constructions of the preceding proof again result in an $H^1(\Omega)$-coercive form.*

Proof. We go through the steps (ia) to (iib) of the proof of Theorem 7.33.

(α) In step (ia) only terms of lower order are added so that Lemma 7.12 is applicable. As for step (ib), see step (β_2).

(β_1) In step (iia) one adds

$$b(u,v) := \int_\Omega \sum_{i,j} \left[a_{ij}^s u_{x_i} v_{x_j} + (a_{ij}^s)_{x_j} v u_{x_i} \right] \mathrm{d}\mathbf{x}.$$

Here Lemma 7.12 is also applicable, since the skew symmetry of A^s results in

$$b(u,u) = \int_\Omega \sum (a_{ij}^s)_{x_j} u u_{x_i} \mathrm{d}\mathbf{x}.$$

(β_2) In the construction in step (iib) Lemma 7.12 is applicable analogously. This is easiest to understand in the case $n = 2$. Let Γ be described by $\{(x_1(s), x_2(s)) : 0 \le s \le 1\}$. If $x_1, x_2 \in C^1([0,1])$ and $x_i'(0) = x_i'(1)$, and if we have $\mathbf{d} \in C^1(\Gamma)$ for the $\mathbf{d}$ in $T = \mathbf{d}^\mathsf{T}\nabla$, then $b(\cdot,\cdot)$ has the representation

$$b(u,v) = \int_0^1 \bar{v}(s)\tau(s)\bar{u}'(s)\mathrm{d}s, \text{ where } \tau \in C^1([0,1]) \text{ and } \begin{cases} \bar{u}(s) = u(x_1(s), x_2(s)), \\ \bar{v}(s) = v(x_1(s), x_2(s)). \end{cases}$$

Thanks to periodicity, integration by parts yields in $b(u,v) = -\int_0^1 (\tau\bar{v})'\bar{u}\,\mathrm{d}s$ without boundary terms, so that

$$b(u,u) = \frac{1}{2}\left[\int_0^1 \bar{u}\tau\bar{u}'\mathrm{d}s - \int_0^1 (\tau\bar{u})'\bar{u}\mathrm{d}s \right] = -\frac{1}{2}\int_0^1 \tau'\bar{u}^2\mathrm{d}s.$$

This implies

$$|b(u,u)| \le \frac{1}{2}\|\tau\|_{C^1([0,1])} \|u|_\Gamma\|_{L^2(\Gamma)}^2 \le C\|\tau\|_{C^1([0,1])} \|u\|_{H^s(\Omega)}^2$$

for $\frac{1}{2} < s < 1$ (cf. Theorem 6.58a). Since $|u|_s^2 \le \varepsilon |u|_1^2 - C|u|_0^2$, one can apply Lemma 6.110c. ∎

7.4.4 Boundary Conditions for $m \ge 2$

The case $m \ge 2$ has been excluded in this section (except for Theorem 7.23). Boundary-value problems of order $2m$ require m boundary conditions $B_j u = \varphi_j$ on Γ ($j = 1, \ldots, m$; cf. Section 5.3). For $m \ge 2$ the proof of $H^m(\Omega)$-coercivity becomes more complicated. In order to carry over Theorem 7.13, one needs in addition the *condition of Agmon* [3] (cf. Wloka [308, §19], Lions–Magenes [194, page 210]).

The resulting complications can be seen with the aid of the biharmonic equation.

Example 7.35. (a) To the variational problem: find $u \in H^2(\Omega)$ with

$$\begin{aligned} a(u,v) := \textstyle\int_\Omega \Delta u \Delta v \mathrm{d}\mathbf{x} = f(v) := \textstyle\int_\Omega g(\mathbf{x}) v(\mathbf{x}) \mathrm{d}\mathbf{x} \\ + \textstyle\int_\Gamma \left(\varphi_1(\mathbf{x}) \frac{\partial v(\mathbf{x})}{\partial n} - \varphi_2(\mathbf{x}) v(\mathbf{x}) \right) \mathrm{d}\Gamma \quad \text{for all } v \in H^2(\Omega) \end{aligned} \tag{7.30a}$$

corresponds the classical formulation

$$\Delta^2 u = g \quad \text{in } \Omega, \qquad \Delta u = \varphi_1 \ \text{ and } \ \frac{\partial}{\partial n} \Delta u = \varphi_2 \quad \text{on } \Gamma. \tag{7.30b}$$

But the bilinear form $a(\cdot,\cdot)$ is not $H^2(\Omega)$-coercive.

(b) To the variational problem: find $u \in H^2(\Omega) \cap H_0^1(\Omega)$ with

$$a(u,v) = f(v) := \int_\Omega g(\mathbf{x}) v(\mathbf{x}) \mathrm{d}\mathbf{x} + \int_\Gamma \varphi \frac{\partial v}{\partial n} \mathrm{d}\Gamma \qquad \text{for all } v \in H^2(\Omega) \cap H_0^1(\Omega) \tag{7.30c}$$

($a(\cdot,\cdot)$ as in (7.30a)) corresponds the classical formulation

$$\Delta^2 u = g \quad \text{in } \Omega, \qquad u = 0 \ \text{ and } \ \Delta u = \varphi \quad \text{on } \Gamma. \tag{7.30d}$$

The bilinear form is $H^2(\Omega) \cap H_0^1(\Omega)$-coercive.

(c) The boundary conditions in

$$\Delta^2 u = g \quad \text{in } \Omega, \qquad u = 0 \ \text{ and } \ \frac{\partial}{\partial n} \Delta u = \varphi \quad \text{on } \Gamma. \tag{7.30e}$$

are admissible. Nevertheless this boundary-value problem cannot be written in the present form as a variational problem. A variational formulation for (7.30e) reads:

$$\text{find } u \in H^2(\Omega) \cap H_0^1(\Omega) \text{ such that } a(u,v) = f(v) \text{ for all } v \in H^2(\Omega) \text{ with } \frac{\partial v}{\partial n} = 0$$

with $f(v) := \int_\Omega g v \mathrm{d}\mathbf{x} + \int_\Gamma \varphi v \mathrm{d}\Gamma$ ($a(\cdot,\cdot)$ as in (7.30a)). But this does not agree with the present concept since u and v belong to different spaces. A possible remedy is mentioned in §8.9.1.

Proof. The equivalence of the variational and the classical formulation can be shown via integration by parts:

$$a(u,v) = \int_\Omega v \Delta^2 u \mathrm{d}\mathbf{x} + \int_\Gamma \left[\Delta u \frac{\partial v}{\partial n} - \frac{\partial \Delta u}{\partial n} v \right] \mathrm{d}\Gamma.$$

The noncoercivity in part (a) results as follows. Let $\Omega \subset \mathbb{R}^n$ be bounded. For all $\alpha \in \mathbb{R}$, $u_\alpha(x_1, \ldots, x_n) = \sin(\alpha x_1) \exp(\alpha x_2)$ lies in $H^2(\Omega)$ and satisfies $\Delta u_\alpha = 0$, hence also $a(u_\alpha, u_\alpha) = 0$. If $a(\cdot,\cdot)$ were coercive, there would exist a C with $0 = a(u_\alpha, u_\alpha) \geq \varepsilon |u_\alpha|_2 - C |u_\alpha|_0$ for all α, i.e., $|u_\alpha|_2 \leq (C/\varepsilon) |u_\alpha|_0$. The contradiction results from $|u_\alpha|_2 \geq \left| \partial^2 u_\alpha / \partial x_1^2 \right|_0 = \alpha^2 |u_\alpha|_0$ for sufficiently large α. ∎

7.4.5 Further Boundary Conditions

Natural and Dirichlet conditions can occur together. In Example 7.35b, $u = 0$ is a Dirichlet condition and $\partial \Delta u/\partial n = \varphi$ a natural boundary condition. Even in the case $m = 1$ both sorts of boundary conditions can occur.

Example 7.36. Let γ be a subset of Γ with positive boundary measure. The boundary-value problem

$$-\Delta u = g \quad \text{in } \Omega, \qquad u = 0 \ \text{ on } \gamma, \quad \frac{\partial}{\partial n} u = \varphi \ \text{ on } \Gamma \backslash \gamma \tag{7.31}$$

in the variational formulation reads as follows: find $u \in H^1_\gamma(\Omega)$ such that

$$a(u,v) := \int_\Omega \langle \nabla u, \nabla v \rangle \, \mathrm{d}\mathbf{x} = f(v) := \int_\Omega g\, v \, \mathrm{d}\mathbf{x} + \int_{\Gamma \backslash \gamma} \varphi \, v \, \mathrm{d}\Gamma \ \text{ for all } v \in H^1_\gamma(\Omega),$$

where

$$H^1_\gamma(\Omega) := \{u \in H^1(\Omega) : u = 0 \ \text{ on } \ \gamma\}.$$

Equation (7.31) is occasionally termed a *Robin problem*, *mixed boundary-value problem*, or *third boundary-value problem*.

Exercise 7.37. Let $a(u,v) := \int_\Omega [\, \langle \nabla u, \nabla v \rangle + auv \,] \, \mathrm{d}\mathbf{x}$ with $a(x) \geq \alpha > 0$ on $V \times V$ with

$$V := \{u \in H^1(\Omega) : u \ \text{ constant on } \ \Gamma\}$$

be defined. Show that

(a) $a(\cdot,\cdot)$ is V-elliptic.

(b) The weak formulation: $u \in V$, $a(u,v) = \int_\Omega g v \mathrm{d}\mathbf{x} + \int_\Gamma \varphi v \mathrm{d}\Gamma$ for all $v \in V$ corresponds to the problem

$$-\Delta u + au = g \quad \text{in } \Omega, \qquad u \text{ constant on } \Gamma \text{ with } \int_\Gamma \frac{\partial u}{\partial n} \mathrm{d}\Gamma = \int_\Gamma \varphi \, \mathrm{d}\Gamma,$$

which is also called an *Adler problem*.[7]

Finally, we want to point out the difficulty of classically interpreting a weak solution. In the variational formulation (7.21) the right-hand sides g and φ of the differential equation and the boundary condition are combined in the functional f. In the variational formulation the components g and φ are indistinguishable! $u \in H^1(\Omega)$ has first derivatives in $L^2(\Omega)$, whose restrictions to Γ do not have to make sense. That is why Bu cannot be defined in general; $Bu = \varphi$ cannot be viewed as an equality in the space $H^{-1/2}(\Gamma)$ although $\varphi \in H^{-1/2}(\Gamma)$ (cf. Corollary 7.26b).

[7] The physical example, as described by Walter [301, Kap. IV, §31.XVI], is a parabolic differential equation for the heat conduction in a body embedded in a fluid of *constant temperature*. In [301] one finds citations of Adler (1956, 1959).

But even if there is a classical solution, the following paradox arises. Let u be a classical solution of $Lu = 0$ in Ω, $Bu = \varphi$ on Γ. According to (7.20a) define $f_\varphi \in (H^1(\Omega))'$ by $f_\varphi(v) := \int_\Gamma \varphi v \mathrm{d}\Gamma$. One may also view u as a solution of $Lu = f_\varphi$ in Ω, $Bu = 0$ on Γ. These equations may even be interpreted classically in the following way: there exist $f_\nu \in C^\infty(\Omega)$ ($\nu \in \mathbb{N}$) with $f_\nu \to f_\varphi$ in $(H^1(\Omega))'$. Let u_ν be the classical solution of $Lu_\nu = f_\nu$, $Bu_\nu = 0$. Then u_ν converges in $H^1(\Omega)$ to the above-mentioned classical solution u.

Incorporating the boundary values $Bu = \varphi$ in the differential equation $Lu = f_\varphi$, corresponds to a modification of the discretised problem as used in Chapter 4. The difference equations $D_h u_h = f_h$ in Ω_h and the boundary conditions $u_h = \varphi$ on Γ_h resulted in the system of equations $L_h u_h = q_h := f_h + \varphi_h$ (cf. (4.13b)). If one defines $\bar{u}_h$ by $\bar{u}_h = u_h$ in Ω_h, $\bar{u}_h = 0$ on Γ_h, then $\bar{u}_h$ satisfies the equations $D_h \bar{u}_h = q_h$ in Ω_h, $\bar{u}_h = 0$ on Γ_h. Just as the functional f cannot be uniquely separated into g and φ, f_h and φ_h cannot be reconstructed from q_h. In contrast to the discrete case, the separation of f into g and φ is possible, however, provided that stronger conditions than $g \in (H^1(\Omega))'$ are imposed on g (for example, $g \in L^2(\Omega)$).

7.5 Pseudo-Differential Equations

In §3.6 the integral equation method is mentioned. Let $\Omega = \Omega_-$ be a bounded domain (now called the *interior domain*) with the boundary Γ. Then $\Omega_+ := \mathbb{R}^n \backslash \overline{\Omega_-}$ is called the *exterior domain*, which is unbounded and has the same boundary Γ. Since finite-element methods are less suited for unbounded domains (cf. Thatcher [283]), the integral equation method is the method of choice. The Neumann problem

$$\Delta u = 0 \text{ in } \Omega_\pm\,, \qquad \partial u/\partial n = \varphi \text{ on } \Gamma$$

can be solved in both the interior and exterior domains if $\int_\Gamma \varphi \,\mathrm{d}\Gamma = 0$ (cf. (3.17) with $f = 0$). As ansatz for u we take the double-layer potential $u(\mathbf{x}) = -\int_\Gamma \frac{\partial s(\boldsymbol{\xi},\mathbf{x})}{\partial n_\xi} g(\boldsymbol{\xi}) \mathrm{d}\Gamma_\xi$ with a function g still to be determined and the singularity function s in (2.4a). The characteristic equation for g is

$$\int_\Gamma \frac{\partial^2 s(\mathbf{x},\mathbf{y})}{\partial n_x \partial n_y}\, g(\mathbf{y})\, \mathrm{d}\Gamma_\mathbf{y} = \varphi(\mathbf{x}),$$

where the integral must be interpreted in the sense of Hadamard since the integrand has a nonintegrable singularity. Integral equations with such integrands are called hypersingular integral equations. Multiplying by a test function ψ and integration over Γ gives

$$a(g,\psi) := \int_\Gamma \int_\Gamma \frac{\partial^2 s(\mathbf{x},\mathbf{y})}{\partial n_x \partial n_y}\, \psi(\mathbf{x})\, g(\mathbf{y})\, \mathrm{d}\Gamma_\mathbf{y}\, \mathrm{d}\Gamma_\mathbf{x} = \varphi(\psi) := \int_\Gamma \varphi\, \psi \,\mathrm{d}\Gamma.$$

This is the variational formulation $a(g,\psi) = \varphi(\psi)$. We omit the definition of Hadamard integrals since we can turn it in an improper integral (cf. [136, §7.5]):

$$a(g,\psi) = \frac{1}{2}\int_\Gamma \int_\Gamma \frac{\partial^2 s(\mathbf{x},\mathbf{y})}{\partial n_x \partial n_y} \left[\psi(\mathbf{x}) - \psi(\mathbf{y})\right] \left[g(\mathbf{x}) - g(\mathbf{y})\right] \mathrm{d}\Gamma_{\mathbf{y}}\, \mathrm{d}\Gamma_{\mathbf{x}}. \quad (7.32a)$$

Note that $\left|\frac{\partial^2 s(\mathbf{x},\mathbf{y})}{\partial n_x \partial n_y}\right| \le C\,|\mathbf{x}-\mathbf{y}|^{-n}$ (cf. (2.4a) and (9.26a)). Schwarz' inequality yields the estimate

$$|a(g,\psi)| \le \frac{C}{2} |||g||| \, |||\psi||| \qquad \text{with}$$
$$|||f||| := \sqrt{\iint\limits_{\Gamma\times\Gamma} |f(\mathbf{x}) - f(\mathbf{y})|^2 \, |\mathbf{x}-\mathbf{y}|^{-n} \, \mathrm{d}\Gamma_{\mathbf{x}} \, \mathrm{d}\Gamma_{\mathbf{y}}}\,.$$

$|||f|||^2$ corresponds to the second term in (6.22a). Since Γ is an $(n-1)$-dimensional surface, the exponent n in $|\mathbf{x}-\mathbf{y}|^{-n}$ should be interpreted as $(n-1)+2\lambda$ with $\lambda = 1/2$. Thus $|||\cdot|||$ is equivalent to the Sobolev norm $\|\cdot\|_{H^{1/2}(\Gamma)}$ restricted to the subspace

$$H^{1/2}(\Gamma)/\mathbb{R} := \left\{ f \in H^{1/2}(\Gamma) : \int_\Gamma f \, \mathrm{d}\Gamma = 0 \right\}.$$

Hence,

$$a(\cdot,\cdot) : V \times V \to \mathbb{R} \qquad \text{for } V := H^{1/2}(\Gamma)/\mathbb{R} \quad (7.32b)$$

is a bounded bilinear form on $V \times V$.

The unique solvability of $a(g,\psi) = \int_\Gamma \varphi\,\psi\,\mathrm{d}\Gamma$ ($\psi \in V$) follows from the next statement.

Theorem 7.38. *The bilinear form (7.32a,b) is V-elliptic.*

Proof. Compare [136, Theorem 8.3.2] and Giroire–Nédélec [118]. ∎

According to Lemma 6.91, there is an operator $A : H^{1/2}(\Gamma)/\mathbb{R} \to H^{-1/2}(\Gamma)/\mathbb{R}$ associated to $a(\cdot,\cdot)$. It behaves like a differential operator of first order. However, while differential operators are local, i.e., $\operatorname{supp}(D^\alpha f) \subset \operatorname{supp}(f)$, A is a nonlocal operator. Such mappings are called *pseudo-differential operators*.

Chapter 8
The Finite-Element Method

Abstract In Chapter 7 the variational formulation has been introduced to prove the existence of a (weak) solution. Now it will turn out that the variational formulation is extremely important for numerical purposes. It establishes a new, very flexible discretisation method. After historical remarks in **Section 8.1** we introduce the Ritz–Galerkin method in **Section 8.2**. The basic principle is the replacement of the function space V in the variational formulation by an N-dimensional space. This leads to a system of N linear equations (§8.2.1). As described in §8.2.2, the theory from Chapter 7 can be applied. In §8.2.3 two criteria, the inf-sup condition and V-ellipticity are described which are sufficient for solvability. §8.2.4 contains numerical examples. Error estimates are discussed in **Section 8.3**. The quasi-optimality of the Ritz–Galerkin method proved in §8.3.1 shifts the discussion to the approximation properties of the subspace (§8.3.2). The finite elements introduced in **Section 8.4** form a special finite-dimensional subspace offering many practical advantages. The corresponding error estimates are given in **Section 8.5**. Generalisations to differential equations of higher order and to non-polygonal domains are investigated in **Section 8.6**. An important practical subject are a-posteriori error estimates discussed in **Section 8.7**. When solving the arising system of linear equations, the properties of the system matrix is of interest which are investigated in **Section 8.8**. Several other topics are sketched in the final **Section 8.9**.

8.1 Historical Remarks

The finite-element method is based on two independent concepts: the variational formulation involving a subspace and the special subspace of finite elements. The variational principle was first applied to eigenvalue problems by Rayleigh[1]. In the case of a symmetric and V-elliptic bilinear form the smallest eigenvalue λ and the eigenvector u of $Lu = \lambda u$ in Ω and, e.g., $u = 0$ on Γ can be obtained by minimising

[1] In 1873 John William Strutt inherited the barony of the small town Rayleigh in Essex. Afterwards he published under the name "John William Strutt, Baron Rayleigh".

W. Hackbusch, *Elliptic Differential Equations*, Springer Series
in Computational Mathematics 18, DOI 10.1007/978-3-662-54961-2_8

the Rayleigh quotient $a(u,u)/(u,u)_{L^2(\Omega)}$ over $0 \neq u \in V$. Rayleigh [234] (1877) minimised this quotient over certain subspaces.

On May 16, 1908, Walter Ritz presented the paper [241] to the mathematical-physical class of the Göttingen Academy of Sciences and Humanities. A similar article [242] of Ritz appeared shortly later in the Crelle journal. In both papers the biharmonic problem (5.26) with homogeneous boundary values (5.27) ($\varphi_1 = \varphi_2 = 0$) is treated. The variational problem $\int_\Omega \{\frac{1}{2}(\Delta u)^2 - uf\}\mathrm{d}\mathbf{x}$ is minimised over a subspace spanned by functions $\psi_1, \ldots, \psi_m$. Ritz generates the system of equations and emphasises that the matrix coefficients can be defined by[2] "purely numerical quadrature: hence the practical implementation of these quadratures with given accuracy poses no difficulties." The linearly independent ansatz functions together with their derivatives are required to be continuous and must satisfy the Dirichlet conditions $\psi_i = \partial\psi_i/\partial n = 0$ on the boundary. Concerning their construction, Ritz proposes the interpolation by a piecewise analytic function with support in a rectangle $\rho \subset\subset \Omega$. In [242, page 5] he emphasises in italic[2] "that in different regions the (ansatz functions) ψ_i may be given by different analytic[3] functions respectively by different expressions, provided that on the boundary between these regions certain continuity conditions are fulfilled. Herein lies a great facilitation for applications to experimental results." Today, these lines remind us of finite elements. However, Ritz still had a strong solution in mind; he required continuous derivatives up to third order (cf. [241, page 241]), while a continuous first derivative is sufficient for the weak formulation. He needed these assumptions for his proof of convergence. Ritz also discusses the Poisson equation ([242, page 45]) and differential equations with variable coefficients ([242, page 52]).

In the same papers Ritz also applied his method to the minimisation of the Rayleigh quotient in order to compute the string vibration. In [242, §16] he gives numerical results emphasising the "surprisingly good convergence"[2], but could not prove the convergence for the eigenvalue problem.[4] His article [240] in the same year treats the plate vibrations with explicit reference to Rayleigh. In this respect the name *Rayleigh–Ritz method* applies. More historical details can be found, e.g., in Parlett [216], Gaul [112], and Leissa [190].

In 1915 Galerkin published the article [106]. There he approaches the problem $Lu = f$ by the ansatz $u = \sum A_n \varphi_n$ and determines the coefficients A_n by $\int (Lu)\varphi_n \mathrm{d}\mathbf{x} = \int f\varphi_n \mathrm{d}\mathbf{x}$. In contrast to Ritz he can also treat nonsymmetric L (cf. [106, (13)]). The functions φ_n he used are trigonometric functions and polynomials. This explains the name *Ritz–Galerkin method*. In the Russian literature the usual term is *Bubnov–Galerkin method*, since even earlier I.G. Bubnov[5] [59] has mentioned this method (cf. Afendikova [2]).

[2] Translation of the original German text.

[3] Here he considered polynomials and trigonometric functions.

[4] Final remark in [242, page 61]: "In light of the presented example a physicist applying the new computational method to cases with still missing theoretical convergence proof will not be worried because of this deficiency."

[5] In 1913 Bubnov reported on a work of Timoshenko. Bubnov proposed orthogonal ansatz functions (in particular trigonometric functions).

Finite elements in the explicit form of triangular elements have first been used in 1851 by Schellbach [259] and 1943 by Courant[6,7] [76]. But only about 1960 when powerful computers were available, this method was rediscovered in engineering sciences (details in Stein [268] and Zienkiewicz [320]). The mathematical error analysis started with Zlámal [322]. An early, but still actual standard reference for the finite-element method is the monograph of Ciarlet [67]. Although the finite-element method has Russian roots (Galerkin, Bubnov), is was not promoted in the Soviet Union. Instead – in particular due to Samarskii (cf. [248]) – the difference methods became the standard. Variational techniques were published, e.g., under a title as 'variational difference method' (cf. Oganesjan–Ruchovec [214]). The connection between difference methods and variational problems was already stated in Courant–Friedrichs–Lewy [77] (concerning the common history we refer to Thomée [288]).

8.2 The Ritz–Galerkin Method

8.2.1 Basics

8.2.1.1 Subspace Discretisation

Suppose we have a boundary-value problem in its variational formulation as described in Chapter 7:

$$\text{find } u \in V, \quad \text{so that} \quad a(u,v) = f(v) \qquad \text{for all } v \in V, \tag{8.1}$$

where we are thinking, in particular, of $V = H_0^m(\Omega)$ and $V = H^1(\Omega)$ (cf. Sections 7.2 and 7.4). Of course, it is assumed that $a(\cdot,\cdot)$ is a bounded bilinear form defined on $V \times V$, and that $f \in V'$:

$$|a(u,v)| \le C_S \|u\|_V \|v\|_V \quad \text{for } u, v \in V, \qquad f \in V'. \tag{8.2}$$

Difference methods arise through discretising the differential operators. Now we wish to leave the differential operator hidden in $a(\cdot,\cdot)$ unchanged. The Ritz–Galerkin discretisation consists in replacing the infinite-dimensional space V with a finite-dimensional space[8] V_N:

$$V_N \subset V \qquad \text{with} \quad \dim V_N = N < \infty. \tag{8.3}$$

[6] In 1922 Hurwitz and Courant [159, page 338] mention piecewise linear finite elements on a triangulation as a theoretical example of a sequence approaching the minimiser in a footnote which, however, disappeared in the second edition of the book.

[7] See appendix of [76] on the pages 20–22. The figure on page 21 shows a triangulation.

[8] The condition $\dim V_N < \infty$ is necessary to obtain a practically solvable system of finitely many equations. However, the following theory also holds for a family of infinite-dimensional subspaces $V_n \subset V$. This fact will be used in the proof of Corollary 9.25.

V_N equipped with the norm $\|\cdot\|_V$ is still a Banach space. Since $V_N \subset V$, both $a(u,v)$ and $f(v)$ are defined for $u, v \in V_N$. Thus we may pose the problem (8.4):

$$\text{find } u^N \in V_N, \quad \text{so that} \quad a(u^N, v) = f(v) \qquad \text{for all } v \in V_N. \tag{8.4}$$

The transition from (8.1) to (8.4) characterises the *Ritz–Galerkin* or briefly *Galerkin discretisation*. More precisely, (8.4) defines a *conforming* Galerkin discretisation, since V_N is a subspace of V (cf. (8.3) and §8.9.2).

The solution u^N of (8.4), if it exists, is called the *Ritz–Galerkin solution* (belonging to V_N) of the boundary-value problem (8.1).

Remark 8.1. We have to distinguish two different scenarios:

(i) There is only one subspace V_N and the associated problem (8.4).

(ii) Let $\mathbb{N}' \subset \mathbb{N}$ be an infinite subset.[9] There is a sequence of Galerkin discretisations in V_N for each $N \in \mathbb{N}'$.

In case (i) we are interested in the solvability of (8.4) and an error estimate of $u - u^N$ for the special dimension N.

In case (ii), the limit $N \to \infty$ exists. For instance, solvability of (8.4) can be required for N sufficiently large. Asymptotic statements about error estimates of $\{u - u^N : N \in \mathbb{N}'\}$ are of interest.

8.2.1.2 Generation of the System of Equations

Notation 8.2. The variables $u, v, \ldots$ denote functions, capitals $L, \ldots$ are operators. Vectors $\mathbf{u}, \mathbf{v}, \ldots \in \mathbb{R}^N$ and matrices $\mathbf{L}, \mathbf{M}, \ldots \in \mathbb{R}^{N \times N}$ are written in boldface.

To calculate a solution one needs a basis of V_N. Let $\{b_1, \ldots, b_N\}$ be such a basis, i.e.,

$$V_N = \operatorname{span}\{b_1, \ldots, b_N\}. \tag{8.5}$$

For each *coefficient vector* $\mathbf{v} = (v_1, \ldots, v_N)^\mathsf{T} \in \mathbb{R}^N$ ($\mathbb{C}^N$ also possible) we define the mapping[10]

$$P : \mathbb{R}^N \to V_N \subset V, \qquad P\mathbf{v} = \sum_{i=1}^{N} v_i b_i \,. \tag{8.6}$$

Remark 8.3. P is an isomorphism between $\mathbb{R}^N$ and V_N. The inverse mapping $P^{-1} : V_N \to \mathbb{R}^N$ is thus well defined on V_N.

---

[9] In the case of the difference method in §4.2, $\mathbb{N}' := \{N = (n-1)^2 : n \in \mathbb{N}\}$ is the sequence of possible dimensions.

[10] The letter P abbreviates "prolongation" since in later applications the nodal values v_i (associated at certain points x_i) are extended to the function $P\mathbf{v} \in V_N$.

Lemma 8.4. *Assuming (8.5) the problem (8.4) is equivalent to*

$$\textit{find } u^N \in V_N, \quad \textit{so that} \quad a(u^N, b_i) = f(b_i) \qquad \textit{for all } i = 1, \dots, N. \tag{8.7}$$

Proof. (i) "(8.7) $\Rightarrow$ (8.4)" Suppose that $v = \sum_{i=1}^N v_i b_i \in V_N$ be arbitrary. Then (8.7) and the linearity of $a(u^N, \cdot)$ and f give (8.4):

$$\begin{aligned} a(u^N, v) - f(v) &= a\Big(u^N, \sum_{i=1}^N v_i b_i\Big) - f\Big(\sum_{i=1}^N v_i b_i\Big) = \sum_{i=1}^N v_i \left[a(u^N, b_i) - f(b_i)\right] \\ &= 0. \end{aligned}$$

(ii) "(8.4) $\Rightarrow$ (8.7)" Putting $v = b_i$ in (8.4) gives (8.7). ∎

We now seek $\mathbf{u} \in \mathbb{R}^N$ so that $u^N = P\mathbf{u}$ represents the Galerkin solution. The following theorem transforms the problem (8.4) [resp. (8.7)] into a system of linear equations.

Theorem 8.5. *Assume (8.5). The $N \times N$-matrix $\mathbf{L} = (L_{ij})_{i,j=1}^N$ and the N-vector $\mathbf{f} = (f_1, \dots, f_N)^\mathsf{T}$ are defined by*

$$L_{ij} := a(b_j, b_i) \qquad (i, j = 1, \dots, N)\,, \tag{8.8a}$$

$$f_i := f(b_i) \qquad (i = 1, \dots, N)\,. \tag{8.8b}$$

Then the problem (8.4) and the system

$$\mathbf{L}\mathbf{u} = \mathbf{f} \tag{8.9}$$

are equivalent. If $\mathbf{u}$ is a solution of (8.9), then $u^N := P\mathbf{u}$ solves the problem (8.4). In the opposite direction, if u^N is a solution of (8.4), then $\mathbf{u} := P^{-1}u^N$ is a solution of (8.9) (cf. Remark 8.3). Otherwise, both problems are not solvable.

Proof. (8.4) is equivalent to (8.7). In (8.7) put $u^N = P\mathbf{u} = \sum u_j b_j$:

$$a(u^N, b_i) = a\Big(\sum_{j=1}^N u_j b_j, b_i\Big) = \sum_{j=1}^N u_j\, a(b_j, b_i) = \sum_{j=1}^N L_{ij} u_j = f(b_i) = f_i\,,$$

and thus $\mathbf{L}\mathbf{u} = \mathbf{f}$. On the other hand the solution of $\mathbf{L}\mathbf{u} = \mathbf{f}$ gives the solution $u^N := P\mathbf{u}$ of $a(u^N, b_i) = f(b_i)$. ∎

In engineering applications where the boundary-value problems arise from continuum mechanics (cf. the first paragraph of §5.3.1), one calls $\mathbf{L}$ the *stiffness matrix*. In the following we call this matrix the *system matrix*. The connections between $\mathbf{L}$ and $a(\cdot, \cdot)$, on the one hand, and between $\mathbf{f}$ and $f(\cdot)$, on the other, are clear from the next remark.

Remark 8.6. If $\langle \mathbf{u}, \mathbf{v} \rangle := \sum_i u_i \, v_i$ is the usual scalar product, then

$$a(u, v) = \langle \mathbf{L}\mathbf{u}, \mathbf{v} \rangle \qquad \text{and} \qquad f(v) = \langle \mathbf{f}, \mathbf{v} \rangle$$

with $u = P\mathbf{u}$ and $v = P\mathbf{v}$.

Proof. $a(u,v) = a\,(P\mathbf{u}, P\mathbf{v}) = a\,(\sum_j u_j b_j, \sum_i v_i b_i) = \sum_{i,j} u_j \, v_i \, a(b_j, b_i) = \sum_{i,j} L_{ij} u_j v_i = \langle \mathbf{L}\mathbf{u}, \mathbf{v} \rangle$ and $f(v) = f(\sum_i v_i b_i) = \sum_i v_i f(b_i) = \sum_i v_i f_i = \langle \mathbf{f}, \mathbf{v} \rangle$. ∎

A trivial consequence of Theorem 8.5 is the following.

Corollary 8.7. The Ritz-Galerkin discretisation (8.4) has a unique solution u^N for each $f \in V'$ exactly when the matrix $\mathbf{L}$ in (8.8a) is nonsingular.

8.2.2 Analysis of the Discrete Equation

The analysis in Chapter 7 of the solvability of the variational problem holds for any Banach space V, also for finite-dimensional spaces. Therefore the statements in §7 hold also for V_N instead of V. However the terms must be transferred correctly. In §§8.2.2.1–3 we discuss the discrete spaces V_N, U_N, and V_N' of the discrete Gelfand triple. Sections 8.2.2.4–5 introduce the mapping P and the projection Q_N which connect the abstract space V_N with $\mathbb{R}^N$ and the system (8.9). The operator L_N from §8.2.2.6 is important for the application of Lemma 6.94, and Lemma 8.12 explains the relation between L_N and the matrix $\mathbf{L}$.

8.2.2.1 Spaces V_N and U_N

The variational problem (8.7) uses that $a : V \times V \to \mathbb{R}$ is also well defined on $V_N \times V_N$. For clarification we denote the restriction of $a(\cdot,\cdot)$ to $V_N \times V_N$ by a_N:

$$a_N : V_N \times V_N \to \mathbb{R} \quad \text{with} \quad a_N(u, v) := a(u, v) \ \text{ for all } u, v \in V_N. \tag{8.10}$$

The restriction of the norm $\|\cdot\|_V$ and of the scalar product $(\cdot,\cdot)_V$ to V_N defines the Hilbert space $(V_N, (\cdot,\cdot)_V)$.

Instead of the pivot space U in (6.36) we now have U_N which is the same set and vector space as V_N (i.e., $U_N = V_N$), but $U_N \subset U$ is equipped with the scalar product $(\cdot,\cdot)_U$ and the norm $\|\cdot\|_U$.

The Gelfand triple $V \subset U \subset V'$ in (6.36) is based on the continuous and dense embedding $V \subset U$, the equivalent dual statement $U' \subset V'$, and the identification $U = U'$. Correspondingly we identify U_N with the dual space U_N' of U_N.

8.2.2.2 Dual Space V_N'

V_N' coincides with $V_N = U_N$ as a set and vector space. However, it needs a proof that V_N' can be equipped with the (restriction of the) norm of V'. Lemma 6.63 states that $V \subset U$ implies $U' \subset V'$. This statement does not hold for $V_N \subset V$ since the embedding is not dense. In a certain sense the opposite is true: $V_N' \subset V'$, as shown in Lemma 6.71. This lemma can be applied because finite-dimensional spaces are closed. The $\|\cdot\|_{V_N'}$-norm is defined by[11]

$$\|v\|_{V_N'} := \max\left\{ |(v,u)_U| \,/\, \|u\|_V : 0 \neq u \in V_N \right\} \quad \text{for } v \in V_N' = V_N. \tag{8.11}$$

Lemma 6.71 yields the following statement.

Conclusion 8.8. *$\|f\|_{V_N'} = \|f\|_{V'}$ holds for $f \in V_N'$. $\|\cdot\|_{V_N'}$ defines a seminorm*[12] *on V' and satisfies $\|f\|_{V_N'} \leq \|f\|_{V'}$.*

8.2.2.3 The Discrete Gelfand Triple

We summarise the above considerations. In the discrete case, the Gelfand triple $V \subset U \subset V'$ becomes $V_N \subset U_N \subset V_N'$, where the spaces are the same as sets:

$$V_N = U_N = V_N' .$$

However, the three spaces are equipped with different norms:

$$V_N = (V_N, \|\cdot\|_V), \qquad U_N = (U_N, \|\cdot\|_U), \qquad V_N' = \big(V_N', \|\cdot\|_{V_N'}\big).$$

8.2.2.4 Maps P and P^*

The adjoint P^* is formed with respect to the $(\cdot,\cdot)_U$ scalar product of U and to the Euclidean scalar product $\langle\cdot,\cdot\rangle$ of $\mathbb{R}^N$:

$$(u, P\mathbf{x})_U = \langle P^* u, \mathbf{x}\rangle \quad \text{for all } \mathbf{x} \in \mathbb{R}^N,\ u \in U \text{ (in particular, } u \in U_N). \tag{8.12}$$

Selecting $\mathbf{x}$ as the i-th unit vector, the characterisation of[13] $P^* u \in \mathbb{R}^N$ follows from the definition (8.6) of P:

$$\mathbf{y} := P^* u \in \mathbb{R}^N \quad \text{has the components } y_i = (u, b_i)_U$$

[11] We recall that $v(u) = (v,u)_U$.

[12] A seminorm satisfies all norm axioms except the property $\|x\| = 0 \Rightarrow x = 0$.

[13] Formally, we conclude from $P : \mathbb{R}^N \to U_N$ that $P' : U_N' \to (\mathbb{R}^N)'$. We identify $(\mathbb{R}^N)'$ with $\mathbb{R}^N$. Occasionally, $\mathbf{y}^\mathsf{T}$ is written for elements of $(\mathbb{R}^N)'$; then $\langle P^* u, \mathbf{x}\rangle$ becomes $(P^* u)^\mathsf{T}\mathbf{x}$.

with b_i in (8.5). The usual definition of $(\cdot,\cdot)_U$ is $y_i = \int_\Omega u b_i \mathrm{d}\mathbf{x}$. Because of $V_N \subset V$, $P : \mathbb{R}^N \to V_N$ may be considered as a map $P : \mathbb{R}^N \to V$. Hence $P^\star$ can be extended to a map in $L(V', \mathbb{R}^N)$. From (8.8b) we obtain the following result.

Remark 8.9. $P^* : V' \to V_N'$ defines the right-hand side $\mathbf{f} = P^* f$ in (8.9).

8.2.2.5 Projection Q_N

Exercise 8.10. Show the following:
(a) The kernel of $P^\star$ is $V_N^\perp \subset U$ ($V_N^\perp$: orthogonal space relative to $(\cdot,\cdot)_U$). The map $P^\star|_{V_N} : V_N \to \mathbb{R}^N$ is an isomorphism.
(b) $(P^*)^{-1} = (P^{-1})^*$.

Since $P^{-1} : V_N \to \mathbb{R}^N$ exists (cf. Remark 8.3), we can define

$$C_p := \|P^{-1}\|_{\mathbb{R}^N \leftarrow U_N} = \max\left\{ \|P^{-1}u\|_{\mathbb{R}^N} / \|u\|_U : 0 \neq u \in V_N \right\}, \tag{8.13}$$

where $\|\cdot\| = \|\cdot\|_{\mathbb{R}^N}$ is the Euclidean norm of $\mathbb{R}^N$.

Lemma 8.11. *The matrix $P^\star P : \mathbb{R}^N \to \mathbb{R}^N$ has an inverse whose spectral norm is $\|(P^\star P)^{-1}\|_2 = C_P^2$ with C_P defined in (8.13). The mapping*

$$Q_N := P\,(P^\star P)^{-1} P^\star : U \to U \tag{8.14}$$

is the orthogonal projection *onto $U_N = V_N$ (with respect to $\|\cdot\|_U$). Besides $Q_N \in L(U,U)$ also $Q_N \in L(V',V)$ is valid.*

Proof. (a) The matrix $P^\star P$ is symmetric. Hence,

$$\|(P^\star P)^{-1}\|_2 = \max\left\{ \left|\left\langle (P^\star P)^{-1}\mathbf{x}, \mathbf{x}\right\rangle\right| : \|\mathbf{x}\|_{\mathbb{R}^N} = 1 \right\}.$$

Using $(P^*)^{-1} = (P^{-1})^*$ from Exercise 8.10b, we conclude that $\left\langle (P^\star P)^{-1}\mathbf{x}, \mathbf{x}\right\rangle = \left\langle P^{-1}(P^\star)^{-1}\mathbf{x}, \mathbf{x}\right\rangle = \left\langle (P^\star)^{-1}\mathbf{x}, (P^\star)^{-1}\mathbf{x}\right\rangle = \|(P^\star)^{-1}\mathbf{x}\|^2 = \|(P^{-1})^*\mathbf{x}\|^2$. Together we obtain $\|(P^\star P)^{-1}\|_2 = \|(P^{-1})^*\|^2_{U_N \leftarrow \mathbb{R}^N}$. By (6.32) the latter norm coincides with $\|P^{-1}\|^2_{\mathbb{R}^N \leftarrow U_N} = C_p^2$.

(b) Since

$$Q_N^2 = P\,(P^\star P)^{-1} P^\star P\,(P^\star P)^{-1} P^\star = P\,(P^\star P)^{-1} P^\star = Q_N,$$

Q_N is a projection. As $Q_N = Q_N^*$, it is an orthogonal projection. The image space is $\mathrm{range}(Q_N) = \mathrm{range}(P) = V_N$. Considered as mapping in $L(U,U)$, Q_N has the norm 1 as any nonvanishing orthogonal projection. Because of $P^\star \in L(V', \mathbb{R}^N)$ (see above), $(P^\star P)^{-1} : \mathbb{R}^N \to \mathbb{R}^N$ and $P : \mathbb{R}^N \to V$ are bounded so that $Q_N \in L(V', V)$. ∎

8.2.2.6 Operator L_N

According to Lemma 6.91 the bilinear form $a_N(\cdot,\cdot) : V_N \times V_N \to \mathbb{R}$ in (8.10) is associated with an operator

$$L_N : V_N \to V_N' \quad \text{so that} \quad a_N(u,v) = (L_N u, v)_U \quad \text{for all } u, v \in V_N. \tag{8.15}$$

Here we write $(\cdot,\cdot)_U$ instead of $\langle \cdot,\cdot \rangle_{V_N' \times V_N}$ (cf. Remark 6.74). This abstract operator is important since the statements of Lemma 6.94 refer to L_N^{-1}. The next lemma illustrates the link between L_N and the concrete system matrix $\mathbf{L}$.

Lemma 8.12. *Let $L_N \in L(V_N, V_N')$ be the operator corresponding to $a_N(\cdot,\cdot) : V_N \times V_N \to \mathbb{R}$, and $L \in L(V, V')$ that for $a(\cdot,\cdot) : V \times V \to \mathbb{R}$. The following relationships hold between the operators L_N, L, and the system matrix $\mathbf{L}$:*

$$\begin{aligned} &\mathbf{L} = P^* L\, P = P^* L_N P : V \to V', \\ &L_N = P^{*-1}\, \mathbf{L}\, P^{-1} : V_N \to V_N', \\ &L_N \text{ is the restriction of } Q_N L \text{ and } Q_N L Q_N \text{ to } V_N. \end{aligned} \tag{8.16}$$

The first two equations correspond to the following commuting diagram:

$$\begin{array}{ccc} V & \xrightarrow{L} & V' \\ P \uparrow & & \downarrow P^* \\ \mathbb{R}^N & \xrightarrow{\mathbf{L}} & \mathbb{R}^N \\ P \downarrow\uparrow P^{-1} & & P^* \downarrow\uparrow P^{*-1} \\ V_N & \xrightarrow{L_N} & V_N' \end{array} \qquad \begin{array}{l} \textit{(continuous level)} \\ \\ \textit{(map as matrix, discretisation level)} \\ \\ \textit{(abstract map, discretisation level)} \end{array}$$

Proof. (a) The identities

$$\langle \mathbf{L}\mathbf{u}, \mathbf{v} \rangle \underset{\text{Remark 8.6}}{=} \begin{cases} a_N(P\mathbf{u}, P\mathbf{v}) = (L_N P\mathbf{u}, P\mathbf{v})_U = \langle P^* L_N P\mathbf{u}, \mathbf{v} \rangle \\ a(P\mathbf{u}, P\mathbf{v}) \quad = (L\, P\, \mathbf{u}, P\mathbf{v})_U \quad = \langle P^* L\, P\, \mathbf{u}, \mathbf{v} \rangle \end{cases}$$

for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^N$ prove (8.16).

(b) Let $u \in V_N$. Since the image of $Q_N L u$ lies in $V_N' = U_N$, we have to prove $(L_N u, v)_U = (Q_N L u, v)_U$ for all $v \in U_N$. For this use $(L_N u, v)_U = a(u,v)$ and $(Q_N L u, v)_U \underset{Q_N = Q_N^*}{=} (L u, Q_N v)_U \underset{v \in U_N}{=} (Lu, v)_U = a(u,v)$. ∎

By Remark 6.74 we can write $(f,v)_U$ for $f(v)$. If $v = P\mathbf{v}$, it follows from (8.12) that $f(v) = (f, P\mathbf{v})_U = \langle P^\star f, \mathbf{v} \rangle$. Using Remark 8.6 we then have

$$\mathbf{f} = P^* f$$

for the $\mathbf{f}$ in the right-hand side of equation (8.9) (cf. Exercise 8.10a).

8.2.3 Solvability of the Discrete Problem

Below the statements and criteria on the solvability of the variational problem are applied to the discrete case.

8.2.3.1 The inf-sup Condition

Since, according to (8.16), $\mathbf{L}^{-1} = P^{-1}L_N^{-1}(P^*)^{-1}$, the discrete problem $\mathbf{Lu} = \mathbf{f}$ is solvable if and only if L_N^{-1} exists. As stated in Lemma 6.94, the existence of L_N^{-1} is equivalent to the inf-sup conditions. Because of the finite dimension of V_N, Exercise 6.95 simplifies the criteria. For better overview we repeat Lemma 6.94 in consideration of Exercise 6.95.

Lemma 8.13 (inf-sup condition). *Let $L_N \in L(V_N, V_N')$ be the operator corresponding to the continuous bilinear form $a_N(\cdot,\cdot)$. Then the statements (i) and (ii) are equivalent:*

(i) $L_N^{-1} \in L(V_N', V_N)$ *exists;*

(ii) there is $\varepsilon_N > 0$ *so that*

$$\inf_{x\in V_N,\, \|x\|_V=1} \sup_{y\in V_N,\, \|y\|_V=1} |a_N(x,y)| = \varepsilon_N > 0\,. \tag{8.17a}$$

If one of the statements (i) or (ii) is valid, then

$$\|L_N^{-1}\|_{V_N \leftarrow V_N'} = 1/\varepsilon_N. \tag{8.17b}$$

The operator norm (8.17b) yields estimates with respect to the appropriate norms. The following lemma uses Conclusion 8.8.

Lemma 8.14. *Let (8.4) be the Ritz–Galerkin discretisation of (8.1). If the problem is solvable then*

$$\|u^N\|_V \le \|L_N^{-1}\|_{V_N \leftarrow V_N'} \|f\|_{V_N'} \le \|L_N^{-1}\|_{V_N \leftarrow V_N'} \|f\|_{V'}. \tag{8.18}$$

The solvability of the continuous and of the discrete problem are independent as stated in the next remark and illustrated by Example 8.20.

Remark 8.15. The inf-sup condition (6.43a,b) of the continuous variational problem is neither sufficient nor necessary for the discrete inf-sup condition (8.17a). But note the asymptotic statement in Theorem 8.29.

8.2.3.2 *V*-Ellipticity

In the case of a continuous variational problem the V-ellipticity guarantees its unique solvability. The same condition is also sufficient in the discrete case.

Theorem 8.16. *Assume (8.3). Suppose the bilinear form be V-elliptic:*

$$a(u,u) \geq C_E \|u\|_V^2 \quad \textit{for all } u \in V \textit{ with } C_E > 0.$$

Then also the Ritz–Galerkin discretisation is solvable for any subspace V_N. The estimate

$$\|L_N^{-1}\|_{V_N \leftarrow V_N'} \leq \frac{1}{C_E} \tag{8.19}$$

is independent of N, and the Ritz–Galerkin solution $u^N \in V_N$ satisfies (8.18).

Proof. Restricting $a(u,u) \geq C_E \|u\|_V^2$ to $u \in V_N$ does not change the constant C_E (cf. Exercise 6.98a). Theorem 6.100 yields $\|L_N^{-1}\|_{V_N' \leftarrow V_N} \leq 1/C_E$. ∎

Exercise 8.17. Show the following:

(a) If $a(\cdot,\cdot)$ is symmetric then so is $\mathbf{L}$.

(b) If $\mathbf{L}$ is symmetric and V-elliptic then $\mathbf{L}$ is positive definite. Under the same assumptions the Ritz-Galerkin solution u^N solves the following minimisation problem (cf. Theorem 6.104):

$$J(u^N) \leq J(u) := a(u,u) - 2f(u) \qquad \text{for all } u \in V_N.$$

8.2.3.3 *V*-Coercivity

The coercivity condition (6.48) holds for the Ritz–Galerkin discretisation with the same constant. However, this condition loses its significance for finite-dimensional spaces.[14] Theorem 6.107 ($\lambda = 0$) states that either the problem is solvable or $\lambda = 0$ is an eigenvalue. For finite-dimensional spaces this statement is trivial since it is always true.

In the case (ii) of Remark 8.1 there is an infinite sequence of subspaces $\{V_N : N \to \infty\}$. In this case we shall show in Chapter 11 that under suitable conditions the sequence of discrete eigenvalues converges to a continuous eigenvalue. If the continuous problem is solvable (and therefore $\lambda = 0$ no eigenvalue), also the discrete eigenvalues must be different from zero for sufficiently large N and the discrete problems are solvable. The precise statement is in Theorem 8.29.

[14] The conclusions from the coercivity condition (6.48) make use of the fact that $\|\cdot\|_U$ is a strictly weaker norm than $\|\cdot\|_V$, i.e., $\sup_{0 \neq v \in V} \|v\|_V / \|v\|_U = \infty$. However, all finite-dimensional spaces have equivalent norms (cf. Hackbusch [142, §B.1.2]).

8.2.4 Examples

The following examples of low dimension will illustrate the Galerkin discretisation.

Example 8.18 (Dirichlet problem). The boundary-value problem is

$$-\Delta u = 1 \quad \text{in } \Omega = (0,1)\times(0,1)\,, \qquad u = 0 \quad \text{on } \Gamma.$$

The weak formulation is given by (8.1) with $V = H_0^1(\Omega)$,

$$f(v) = \int_0^1\int_0^1 v\,\mathrm{d}x\mathrm{d}y, \quad a(u,v) = \int_\Omega \langle\nabla u, \nabla v\rangle\,\mathrm{d}\mathbf{x} = \int_0^1\int_0^1 (u_x v_x + u_y v_y)\,\mathrm{d}x\mathrm{d}y.$$

The functions

$$\begin{aligned} b_1(x,y) &= \sin(\pi x)\sin(\pi y), \qquad b_2(x,y) = \sin(3\pi x)\sin(\pi y),\\ b_3(x,y) &= \sin(\pi x)\sin(3\pi y), \qquad b_4(x,y) = \sin(3\pi x)\sin(3\pi y) \end{aligned}$$

fulfil the boundary conditions $u = 0$ on Γ and so belong to the space $V = H_0^1(\Omega)$. They form a basis of $V_N = V_4 := \operatorname{span}\{b_1,\dots,b_4\} \subset V$. The matrix elements $L_{ii} = a(b_i,b_i)$ can be worked out to be $L_{11} = \pi^2/2$, $L_{22} = L_{33} = 5\pi^2/2$, $L_{44} = 9\pi^2/2$. In addition the chosen basis is $a(\cdot,\cdot)$-orthogonal: $L_{ij} = 0$ for $i \neq j$, so that the system matrix $\mathbf{L}$ is diagonal. Furthermore one may calculate the values $f_i = f(b_i) = \int_\Omega b_i(x,y)\mathrm{d}x\mathrm{d}y$ getting

$$f_1 = 4/\pi^2, \qquad f_2 = f_3 = 4/(3\pi^2), \qquad f_4 = 4/(9\pi^2).$$

Hence $\mathbf{u} = \mathbf{L}^{-1}\mathbf{f}$ has the components $u_1 = \dfrac{8}{\pi^4}$, $u_2 = u_3 = \dfrac{8}{15\pi^4}$, $u_4 = \dfrac{8}{81\pi^4}$, and the Ritz–Galerkin solution is then

$$u^N(x,y) = \frac{8}{\pi^4}\left[\sin(\pi x)\sin(\pi y) + \frac{1}{15}\Big(\sin(3\pi x)\sin(\pi y) + \sin(\pi x)\sin(3\pi y)\Big) + \frac{1}{81}\sin(3\pi x)\sin(3\pi y)\right].$$

The Ritz–Galerkin solution and the exact solution evaluated at $x = y = \frac{1}{2}$ are

$$\begin{aligned} u^N\left(\frac{1}{2},\frac{1}{2}\right) &= \frac{2848}{405}\pi^{-4} = 0.072\,191\,4\ldots,\\ u\left(\frac{1}{2},\frac{1}{2}\right) &= \sum_{\nu,\mu=0}^{\infty}\frac{16}{\pi^4}\frac{(-1)^{\nu+\mu}}{(1+2\nu)(1+2\mu)\left[(1+2\nu)^2+(1+2\mu)^2\right]}\\ &= 0.073\,671\,3\ldots \end{aligned}$$

Example 8.19 (natural boundary conditions). Let the boundary-value problem be

$$-\Delta u = \pi^2 \cos(\pi x) \quad \text{in } \Omega = (0,1)\times(0,1)\,, \qquad \partial u/\partial n = 0 \quad \text{on } \Gamma\,.$$

The solution is given by $u = \cos \pi x + \text{const}$. The weak formulation is in terms of (8.1) with $V = H^1(\Omega)$,

$$a(u,v) = \int_0^1\int_0^1 (u_x v_x + u_y v_y)\,\mathrm{d}x\mathrm{d}y, \quad f(v) = \pi^2 \int_0^1\int_0^1 v(x,y)\cos(\pi x)\mathrm{d}x\mathrm{d}y.$$

The boundary-value problem has a unique solution in $W := \{v \in V\colon \int_\Omega v\,\mathrm{d}x\mathrm{d}y = 0\}$. The basis functions $b_1(x,y) := x - \frac{1}{2}$, $b_2(x,y) := (x-\frac{1}{2})^3$ are in W. The system matrix $\mathbf{L}$ and the vector $\mathbf{f}$ are then

$$\mathbf{L} = \begin{bmatrix} 1 & 1/4 \\ 1/4 & 9/80 \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} -2 \\ -\frac{3}{2} + \frac{12}{\pi^2} \end{bmatrix}, \quad \text{so that } \mathbf{u} = \mathbf{L}^{-1}\mathbf{f} = \begin{bmatrix} 3 - 60/\pi^2 \\ -20 + 240/\pi^2 \end{bmatrix}.$$

The solution is $u^N(x,y) = (3 - \frac{60}{\pi^2})(x - \frac{1}{2}) - (20 - \frac{240}{\pi^2})(x-\frac{1}{2})^3$. The Ritz–Galerkin solution satisfies the boundary condition $\partial u/\partial n = 0$ and the differential equation only approximately:

$$\partial u^N/\partial n = 12 - 120/\pi^2 = -0.158\,54\ldots$$

For $x = 1/4$ the approximation has the value $u^N(1/4, y) = -7/16 + 45/(4\pi^2) = 0.702\,36\ldots$, whereas $u(1/4,y) = \cos\frac{\pi}{4} = 0.707\,1\ldots$ is the exact value.

In the following we shall consider the case in which $a(\cdot,\cdot)$ is no longer V-elliptic, though it is V-coercive. That $a(\cdot,\cdot)$ is V-coercive guarantees that either problem (8.1) is solvable or $\lambda = 0$ is an eigenvalue. Even if one assumes V-coercivity and the solvability of the problem (8.1) one cannot deduce the solvability of the discrete problem (8.4). In the next example, either the discrete or the continuous problem are not solvable depending on the choice of the coefficient.

Example 8.20. (a) $a(u,v) := \int_0^1 (u'v' - 10uv)\mathrm{d}x$ is $H_0^1(0,1)$-coercive and the variational problem $a(u,v) = f(v) := \int_0^1 gv\mathrm{d}x$ $(v \in H_0^1(0,1))$ has a unique solution. Let V^N be spanned by $b_1(x) = x(1-x) \in V = H_0^1(0,1)$ (i.e., $N = 1$). Then the discrete problem (8.4) is not solvable since $\mathbf{L} = 0$.
(b) The continuous problem $a(u,v) := \int_0^1 (u'v' - \pi^2 uv)\mathrm{d}x$ has no solution since $\lambda = \pi^2$ is an eigenvalue. Nevertheless, the discrete problem with V_1 as in (a) is solvable since $\mathbf{L} = \frac{1}{3} - \frac{1}{30}\pi^2 \neq 0$.

The requirement (8.19) guarantees the existence of $\mathbf{L}^{-1}$, but it does not say anything about its condition $\mathrm{cond}(\mathbf{L}) = \|\mathbf{L}\|\|\mathbf{L}^{-1}\|$, which is the deciding factor for the sensitivity of the system of equations $\mathbf{L}\mathbf{u} = \mathbf{f}$. For example, choosing for $a(u,v) := \int_0^1 u'v'\mathrm{d}x$ the basis $b_i = x^i$ $(i = 1,\ldots,N)$ of monomials, one obtains the very ill-conditioned matrix[15] $L_{ij} = ij/(i+j-1)$. The conditioning of $\mathbf{L}$ is optimal ($\mathrm{cond}(\mathbf{L}) = 1$) is optimal if one chooses the basis to be $a(\cdot,\cdot)$-orthogonal: $a(b_i, b_j) = \delta_{ij}$, as is the case in Example 8.18, up to a scaling factor. More details about the matrix condition will follow in §8.8.3.

[15] Up to a scaling by a diagonal matrix, $\mathbf{L}$ is the Hilbert matrix. The huge norm of the inverse follows from $(\mathbf{L}^{-1})_{ij} = (-1)^{i+j}\frac{i+j-1}{i\,j}\binom{n+i-1}{n-j}\binom{n+j-1}{n-i}\binom{i+j-2}{i-1}^2$.

8.3 Error Estimates

8.3.1 Quasi-Optimality

For difference methods the solution u and the grid function u_h are defined on different sets. The Ritz–Galerkin solution, u^N, is, on the other hand, directly comparable with u. One can measure the error due to discretisation with $\|u - u^N\|_V$ or with $\|u - u^N\|_U$. First, we only use the V-norm. Error estimates with respect to $\|\cdot\|_U$ will follow in §8.5.4.

Let u be the solution of (8.1): $a(u, v) = f(v)$ for $v \in V$. Suppose—by chance or because of a clever choice of V_N—that u also belongs to V_N; then $u^N := u$ also satisfies (8.4). That means: The discretisation error is zero if $u \in V_N$. The following theorem, usually called the *Lemma of Céa*, shows: The 'closer' that u is to V_N the smaller is the discretisation error. The proof in Céa [65, Prop. 3.1, page 365] treats the symmetric, V-elliptic case and gives the better factor $\sqrt{C_S/\varepsilon_N}$ in (8.20). The general case is analysed by Birkhoff–Schulz–Varga [39, Theorem 13].

Theorem 8.21. *Assume (8.2), (8.3), and (8.17a) hold. Let $u \in V$ be a solution of the problem (8.1), and let $u^N \in V_N$ be the Ritz–Galerkin solution of (8.4). Then the following estimate holds:*

$$\|u - u^N\|_V \le \left(1 + \frac{C_S}{\varepsilon_N}\right) \inf_{w \in V_N} \|u - w\|_V \quad \begin{cases} C_S \text{ from (8.2),} \\ \varepsilon_N \text{ from (8.17a)} \end{cases} . \tag{8.20}$$

Inequality (8.20) states that the discrete solution u^N is *quasi-optimal*, since its error coincides with the best possible approximation up to a factor. On the one hand, this is a rather strong statement (since the error can only be improved by a factor), on the other hand, the statement is also weak since the *best-approximation error* $\inf_{w \in V_N} \|u - w\|_V$ remains unknown. $\inf_{w \in V_N} \|u - w\|_V$ is the distance of the function u from V_N and is abbreviated in the following by

$$d(u, V_N) := \inf_{w \in V_N} \|u - w\|_V . \tag{8.21}$$

Proof of Theorem 8.21. If u satisfies $a(u, v) = f(v)$ for all $v \in V$ then it does in particular for all $v \in V_N \subset V$. Since we also have $a(u^N, v) = f(v)$ for $v \in V_N$, we can form the difference:

$$a(u - u^N, v) = 0 \qquad \text{for all } v \in V_N. \tag{8.22}$$

For arbitrary $v, w \in V_N$ with $\|v\|_V = 1$ we can therefore conclude

$$\begin{aligned} & a(u^N - w, v) = a\left(\left[u^N - u\right] + \left[u - w\right], v\right) = a\left(u - w, v\right) \\ \text{and} \quad & \left|a(u^N - w, v)\right| \le C_S \|u - w\|_V \|v\|_V = C_S \|u - w\|_V . \end{aligned} \tag{*}$$

From (8.17a) we then obtain

$$\left\|u^N - w\right\|_V \le \frac{1}{\varepsilon_N} \sup_{v \in V_N,\, \|v\|_V = 1} \left|a(u^N - w, v)\right| \underset{(*)}{\le} \frac{C_S}{\varepsilon_N} \|u - w\|_V .$$

The triangle inequality then gives

$$\|u-u^N\|_V \;\le\; \|u-w\|_V + \|w-u^N\|_V \;\le\; \left(1+\frac{C_S}{\varepsilon_N}\right)\|u-w\|_V\ .$$

Since $w \in V_N$ is arbitrary we deduce the assertion (8.20). ∎

Remark 8.22. The property (8.22) is characteristic for the Galerkin method. Often it is expressed as follows: *The Galerkin error* $u-u^N$ *is orthogonal*[16] *to the subspace* V_N.

Note that in Theorem 8.21 the unique solvability of the problem (8.1) was not assumed, but only the existence of at least one solution. In Theorem 8.24 the (unique) existence of a solution of $Lu=f$ will follow from suitable assumptions on the discrete problems.

Since the underlying spaces are Hilbert (and not general Banach) spaces, the infimum in (8.21) is taken as a minimum.

Remark 8.23. $V_N^\perp := \{v \in V : (v,w)_V = 0 \text{ for all } w \in V_N\}$ is the orthogonal space of V_N with respect to the $(\cdot,\cdot)_V$ scalar product. Since V_N is closed because of $\dim(V_N)<\infty$, we have $V = V_N \oplus V_N^\perp$ (cf. Lemma 6.15). Let $P_{V_N} : V \to V_N$ be the V-orthogonal projection onto V_N. Then for all $v \in V$, $v_N^* := P_{V_N} v \in V_N$ is the unique element with

$$d(v,V_N) = \|v-v_N^*\|_V = \|v^\perp\|_V,$$

where $v^\perp := (I-P_{V_N})v = P_{V_N^\perp} v$ is the orthogonal part.

8.3.2 Convergence of the Ritz–Galerkin Solutions

Speaking about convergence means that we have a sequence $N \to \infty$ ($N \in \mathbb{N}'$) of dimensions with corresponding subspaces V_N (see case (ii) in Remark 8.1). For all $N \in \mathbb{N}'$ assume the inf-sup condition (8.17a) with $\varepsilon_N > 0$. We require in addition that this inequality is uniform: There is some $\underline{\varepsilon} > 0$ so that

$$\varepsilon_N \ge \underline{\varepsilon} > 0 \qquad \text{for all } N \in \mathbb{N}'. \tag{8.23}$$

An equivalent statement is $\inf_{N\in\mathbb{N}'} \varepsilon_N > 0$. Even $\limsup_{N\in\mathbb{N}'} \varepsilon_N > 0$ is sufficient, since then one can select a subsequence $\mathbb{N}'' \subset \mathbb{N}'$ with $\inf_{N\in\mathbb{N}''} \varepsilon_N = \limsup_{N\in\mathbb{N}'} \varepsilon_N$.

If the discretisation error is supposed to converge to zero one makes use of a sequence of subspaces $V_N \subset V$ that converge to V in the following pointwise sense.

[16] In the strict sense this is only correct if $a(\cdot,\cdot)$ defines a scalar product (cf. Exercise 6.98c).

Theorem 8.24. *(a) Let $V_N \subset V$ $(N \in \mathbb{N}')$ be a sequence of subspaces with*[17]

$$\lim_{N\to\infty} d(u, V_N) = 0 \qquad \textit{for all } u \in V. \tag{8.24a}$$

In addition assume (8.23) and (8.2) (continuity of the bilinear form). Then there exists a unique solution u of the problem (8.1), and the Ritz–Galerkin solution u^N converges to u:

$$\|u - u^N\|_V \to 0 \qquad \textit{for } N \to \infty\,.$$

Sufficient to ensure (8.24a) is

$$V_N \subset V_M \subset V \textit{ for all } N < M \in \mathbb{N}' \quad \textit{and} \quad \bigcup\nolimits_{N\in\mathbb{N}'} V_N \textit{ dense in } V. \tag{8.24b}$$

Proof. (i) First we show that (8.24a) follows from (8.24b). The inclusion $V_N \subset V_M$ implies $d(u, V_M) \le d(u, V_N)$ for $N < M$. Hence $d(u, V_N)$ is monotonously decreasing as $N \to \infty$. To prove $d(u, V_N) \to 0$ we have to show that for each $\varepsilon > 0$ there exists an N such that $d(u, V_N) \le \varepsilon$. From the assumption (8.24b) there is, for each $u \in V$ and $\varepsilon > 0$, a $w \in \bigcup_{N\in\mathbb{N}'} V_N$ with $\|u - w\|_V \le \varepsilon$. Therefore we have $w \in V_N$ for an $N \in \mathbb{N}'$. That $d(u, V_N) \le \|u - w\|_V \le \varepsilon$ then proves (8.24a).

(ii) Let $Y := \{Lv : v \in V\} \subset V'$ be the image of the operator $L : V \to V'$ corresponding to $a(\cdot,\cdot)$. For each $f \in Y \subset V'$ there is at least one solution $u \in V$ of $a(u,\cdot) = f$. Inequality (8.20) proves convergence:

$$\|u - u^N\|_V \le \left(1 + \frac{C_S}{\varepsilon_N}\right) d(u, V_N) \le \left(1 + \frac{C_S}{\underline{\varepsilon}}\right) d(u, V_N) \to 0.$$

Since the discrete solutions corresponding to f are unique, the convergence $u^N \to u$ proves the uniqueness of the solution u.

(iii) The discrete solutions are uniformly bounded: $\|u^N\|_V \le \|f\|_{V'}/\varepsilon_N \le \|f\|_{V'}/\underline{\varepsilon}$. By part (ii) also the continuous solution u is bounded by $\|u\|_V \le \|f\|_{V'}/\underline{\varepsilon}$ for all $f \in Y$. Thus the inverse $L^{-1} : Y \to V$ exists, is bounded and therefore continuous. This prove that Y is closed.

(iv) It remains to show $Y = V'$. Otherwise we can decompose the Hilbert space V' into $Y \oplus Y^\perp$ with $Y^\perp \ne \{0\}$ (cf. Conclusion 6.69b). Select some $f \in Y^\perp$ with $\|f\|_{V'} = 1$. The Riesz isomorphism $J_V : V \to V'$ from Conclusion 6.69 shows that $\langle \cdot, J_V^{-1} f\rangle_{V'\times V} = (\cdot, f)_{V'}$. For $v_f := -J_V^{-1} f \in V$ we obtain

$$\begin{aligned} f(v_f) &= \langle f, v_f\rangle_{V'\times V} = -(f,f)_{V'} = -\|f\|_{V'}^2 = -1 \quad \text{and} \\ a(u, v_f) &= \langle Lu, v_f\rangle_{V'\times V} = -(Lu, f)_{V'} \underset{Lu\in Y,\, f\in Y^\perp}{=} 0 \quad \text{for all } u \in V. \end{aligned}$$

Let $u^N \in V_N$ be the Ritz–Galerkin solution to f. It is bounded by $\|u^N\|_V \le \|f\|_{V'}/\underline{\varepsilon} = 1/\underline{\varepsilon}$. Since $V_N \subset V$, we have $a(u^N, v_f) = 0$, i.e.,

[17] More precisely, $d(u, V_N) \to 0$ is only needed for the solution u. In the case of (8.24b), u must lie in the closure of $\bigcup_{N\in\mathbb{N}'} V_N$. If V is not separable, (8.24a) cannot be valid. However, if the solution belongs to a separable subspace $V_0 \subset V$ (e.g., with better smoothness properties), it is sufficient to assume $\lim_{N\to\infty} d(u, V_N) = 0$ for all $u \in V_0$.

$$a(u^N, v_f) - f(v_f) = 1.$$

Decompose v_f into $v^N + w^N$ with $v^N \in V_N$, $w^N \perp V_N$ (orthogonality with respect to the V-scalar product) and $d(v_f, V_N) = \|w^N\|_V$ (cf. Remark 8.23). From (8.24a) we infer $\|w^N\|_V \to 0$. Since the discrete solution fulfils $a(u^N, v^N) - f(v^N) = 0$, the contradiction follows from

$$\begin{aligned} 1 &= a(u^N, v_f) - f(v_f) = a(u^N, v^N) - f(v^N) + a(u^N, w^N) - f(w^N) \\ &= a(u^N, w^N) - f(w^N) \to 0. \end{aligned}$$

Hence $Y^\perp = \{0\}$ and $Y = V'$ must be valid, i.e., for any $f \in V'$ there exists a unique solution u of problem (8.1). ∎

The uniform inequality (8.23) follows immediately from V-ellipticity.

Corollary 8.25. The requirement (8.23) is satisfied with $\underline{\varepsilon} := C_E > 0$ if $a(\cdot,\cdot)$ is V-elliptic with the constant $C_E > 0$, i.e., $a(u,u) \geq C_E \|u\|_V^2$.

8.3.3 Ritz Projection

Let Q_N be the orthogonal projection onto V_N defined in (8.14). The factors $L : V \to V'$, $Q_N : V' \to V_N' = V_N$, and $L_N^{-1} : V_N' = V_N \to V_N \subset V$ can be composed to give

$$S_N := L_N^{-1} Q_N L : \; V \to V_N \subset V. \tag{8.25}$$

Exercise 8.26. Show there is also the representation

$$S_N = P\,\mathbf{L}^{-1} P^* L\,.$$

Lemma 8.27 (Ritz projection). *S_N is a projection onto V_N and is called the* Ritz projection*. It sends the solution u of the problem (8.1) to the Ritz-Galerkin solution $u^N \in V_N$: $u^N = S_N u$. Assuming (8.2) and (8.17a) we have*

$$\|S_N\|_{V \leftarrow V} \leq C_S/\varepsilon_N \qquad (C_S \text{ from (8.2), } \varepsilon_N \text{ from (8.17a)}). \tag{8.26}$$

A definition of S_N equivalent to that in (8.25) is

$$S_N u \in V_N \qquad \text{and} \qquad a(S_N u, v) = a(u, v) \quad \text{for all } v \in V_N,\ u \in V. \tag{8.27}$$

Proof. Since $a(u,v) = \langle Lu, v\rangle_{V' \times V} = (Lu, v)_U = (Lu, Q_N v)_U = (Q_N Lu, v)_U$ for all $v \in V_N$, it follows that $S_N u$ in (8.27) is the Ritz–Galerkin solution for the right-hand side $f := Q_N Lu \in V'$, i.e., $S_N u = L_N^{-1} Q_N Lu$. Conversely one may argue similarly, and so show the equivalence of the definitions (8.25) and (8.27).

(8.27) shows that $u \in V_N$ leads to $S_N u = u$. Thus $S_N^2 = S_N$, i.e., S_N is a projection. Inequality (8.26) follows from $\|S_N u\|_V \leq \|Q_N Lu\|_{V_N'}/\varepsilon_N$ (cf. the proof of (8.19)) and $\|Q_N Lu\|_{V_N'} = \|Lu\|_{V_N'} \leq \|Lu\|_{V'} \leq C_S \|u\|_V$ with C_S from (8.2) (cf. Conclusion 8.8). ∎

Remark 8.28. Let $a(\cdot,\cdot)$ be V-elliptic and symmetric. $|||v|||_V := \sqrt{a(v,v)}$ is a norm equivalent to $\|\cdot\|_V$. The Ritz projection S_N is, with respect to $|||\cdot|||_V$, an orthogonal projection onto V_N. Thus, in particular, we have for $N > 0$

$$|||S_N|||_{V \leftarrow V} = 1. \tag{8.28}$$

Proof. The equivalence of $\|\cdot\|_V$ and $|||\cdot|||_V$ is stated in Exercise 6.98c. The scalar product associated to $|||\cdot|||_V$ is $a(\cdot,\cdot)$, so that it is to be shown that $a(S_N v, w) = a(v, S_N w)$. (8.27) implies $a(S_N v, S_N w) = a(v, S_N w)$, since $S_N w \in V_N$. The symmetry of $a(\cdot,\cdot)$ and exchanging v and w give

$$a(S_N v, w) = a(w, S_N v) = a(S_N w, S_N v) = a(S_N v, S_N w),$$

so that $a(S_N v, w) = a(v, S_N w)$. (8.28) results from Remark 6.70. ∎

Remark 8.28 shows once again that the Ritz–Galerkin solution $u^N = S_N u$ is the best approximation to u in V_N in the sense of the norm $|||\cdot|||_V$. This assertion is equivalent to the minimisation problem $J(u^N) \le J(v)$ for all $v \in V_N$ (cf. Exercise 8.17b).

The stability condition (8.23), $\varepsilon_N \ge \underline{\varepsilon} > 0$, may be difficult to prove, except for V-elliptic bilinear forms. However, the following theorem shows that this condition does hold for subspaces approximating V well enough. The proof of this will be postponed to a supplement to Lemma 11.12 (page 335).

Theorem 8.29. *Let the bilinear form $a(\cdot,\cdot)$ be V-coercive, where $V \subset U \subset V'$ is a continuous, dense, and compact embedding. Let the continuous problem (8.1) be solvable for all $f \in V'$. Assume that (8.24a) holds for the subspaces $V_N \subset V$ ($N \in \mathbb{N}'$). For large enough N the stability condition (8.23) is then satisfied.*[18]

8.3.4 Further Stability and Error Estimates

In the variational formulation (8.1) the right-hand side f and the bilinear form a can be perturbed:

$$\text{find } \tilde{u}^N \in V_N \quad \text{so that} \quad \tilde{a}(\tilde{u}^N, v^N) = \tilde{f}(v^N) \qquad \text{for all } v^N \in V_N\,, \tag{8.29}$$

where

$$\begin{aligned} \tilde{a}(z^N, v^N) &= a(z^N, v^N) + \delta a(z^N, v^N), \\ \tilde{f}(v^N) &= f(v^N) + \delta f(v^N) \end{aligned} \qquad \text{for all } z^N, v^N \in V_N\,.$$

It is sufficient that the perturbations δa and δf are defined on $V_N \times V_N$ and V_N, respectively.

[18] One may select a suitable subsequence $\mathbb{N}'' \subset \mathbb{N}$ so that (8.23) holds for all $N \in \mathbb{N}''$.

The usual source of a perturbation is numerical quadrature: The integral $\int_\Omega fv d\mathbf{x}$ is replaced by a quadrature formula (cf. Exercise 8.34) based on point evaluations of the integrand fv. Obviously this requires that fv be continuous. Since $v \in V$ for $d \geq 2$ does not guarantee continuity (cf. Example 6.24), the perturbation δf may be unbounded with respect to the norm $\|\cdot\|_{-1}$. However, ansatz functions $v^N \in V_N$ have better smoothness properties. Correspondingly, the norm $\|\cdot\|_{-1}$ must be replaced by $\|\cdot\|_{V_N'}$ in (8.11). The treatment of δa is similar.

The following statement is called the *first Lemma von Strang* (cf. Strang [275]).

Theorem 8.30. *Besides (8.2) and (8.3) assume that*

$$\delta a : V_N \times V_N \to \mathbb{R} \qquad \textit{and} \qquad \delta f : V_N \to \mathbb{R}$$

are continuous. Let $\tilde{a} = a + \delta a$ be V_N-elliptic. Let u and u^N be the (exact) solutions of (8.1) and (8.4). Then the solution $\tilde{u}^N$ of the perturbed problem (8.29) satisfies

$$\|u - \tilde{u}^N\|_V \leq \text{const}\Big[\|\delta f\|_{V_N'} + \inf_{z^N \in V_N} \big\{\|u - z^N\|_V + \|\delta a(z^N, \cdot)\|_{V_N'}\big\}\Big]. \quad (8.30)$$

Proof. For arbitrary $z^N, v^N \in V_N$ we have

$$\begin{aligned}
\tilde{a}(\tilde{u}^N - z^N, v^N) &= \tilde{a}(\tilde{u}^N, v^N) - a(z^N, v^N) - \delta a(z^N, v^N) \\
&\underset{(8.29)}{=} \tilde{f}(v^N) - a(z^N, v^N) - \delta a(z^N, v^N) \\
&= \tilde{f}(v^N) + \underbrace{a(u, v^N) - f(v^N)}_{=0 \text{ because of (8.1)}} - a(z^N, v^N) - \delta a(z^N, v^N) \\
&= \delta f(v^N) + a(u - z^N, v^N) - \delta a(z^N, v^N)
\end{aligned}$$

because of $\tilde{f} = f + \delta f$. We estimate $\tilde{u}^N - z^N$ by

$$\begin{aligned}
\tilde{C}_E \|\tilde{u}^N - z^N\|_V^2 &\leq \tilde{a}(\tilde{u}^N - z^N, \tilde{u}^N - z^N) \\
&= \delta f(\tilde{u}^N - z^N) + a(\tilde{u}^N - z^N, \tilde{u}^N - z^N) - \delta a(z^N, \tilde{u}^N - z^N) \\
&= \|\delta f\|_{V_N'} \|\tilde{u}^N - z^N\|_V + C_S \|u - z^N\|_V \|\tilde{u}^N - z^N\|_V \\
&\quad + \|\delta a(z^N, \cdot)\|_{V_N'} \|\tilde{u}^N - z^N\|_V.
\end{aligned}$$

Dividing both sides by $\|\tilde{u}^N - z^N\|_V$ gives

$$\tilde{C}_E \|\tilde{u}^N - z^N\|_V \leq \|\delta f\|_{V_N'} + C_S \|u - z^N\|_V + \|\delta a(z^N, \cdot)\|_{V_N'}.$$

Since $z^N \in V_N$ is arbitrary, (8.30) follows from $\|u - \tilde{u}^N\|_V \leq \|u - z^N\|_V + \|z^N - \tilde{u}^N\|_V$ and the previous inequality. ∎

8.4 Finite Elements

The finite-element method (abbreviated by FEM) is a special case of the Ritz–Galerkin method. For an introduction we first treat the one-dimensional case.

8.4.1 Introduction: Linear Elements for $\Omega = (a, b)$

As soon as the dimension $N = \dim V_N$ becomes larger the essential disadvantage of the general Ritz–Galerkin method becomes apparent. The matrix $\mathbf{L}$ is in general fully populated, i.e., $L_{ij} \neq 0$ for all $i, j = 1, \ldots, N$. Therefore one needs N^2 integrations to obtain the values of $L_{ij} = a(b_j, b_i) = \int_\Omega \ldots$, whether exactly or approximately. The final solution of the system of equations $\mathbf{Lu} = \mathbf{f}$ requires up to $\mathcal{O}(N^3)$ operations. As soon as N is no longer small the general Ritz–Galerkin method therefore turns out to be unusable.

A glance at the difference method shows that the matrices L_h which occur there are sparse. Thus it is natural to wonder if it is possible to choose the basis $\{b_1, \ldots, b_N\}$ of V_N so that the system matrix $L_{ij} = a(b_j, b_i)$ is also sparse. The best situation would be that the b_i were orthogonal with respect to $a(\cdot, \cdot)$: $a(b_j, b_i) = 0$ for $i \neq j$. However, such a basis can be found only for special model problems such as the one in Example 8.18. Instead we shall base our further considerations on the following remark.

Remark 8.31. Let the bilinear form $a(\cdot, \cdot)$ be given by (7.7). Let B_i be the interior of the support $\mathrm{supp}(b_i)$ of the basis function b_i, i.e., $B_i := \mathrm{supp}(b_i) \backslash \partial \mathrm{supp}(b_i)$. A sufficient condition that ensures $L_{ij} = a(b_j, b_i) = 0$ is $B_i \cap B_j = \emptyset$.

Proof. The integration $a(b_j, b_i) = \int_\Omega \ldots$ can be restricted to $B_i \cap B_j$. ∎

To be able to apply Remark 8.31 the basis functions should have as small supports as possible. In constructing them in general one goes about it from the desired goal: one defines partitions of Ω into small pieces, the so-called *finite elements*, from which the supports of b_i are pieced together.

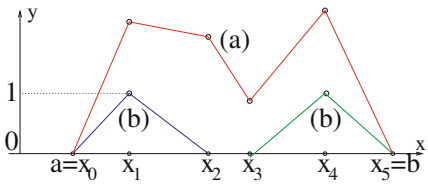


Fig. 8.1 (a) Piecewise linear function, (b) Basis functions corresponding to x_1 and x_4.

As an introduction let us investigate the one-dimensional boundary-value problem

$$-u''(x) = g(x) \qquad \text{for } a < x < b, \qquad u(a) = u(b) = 0. \tag{8.31}$$

Assume we have a partition of the interval $[a, b]$ given by $a = x_0 < x_1 < \ldots < x_{N+1} = b$. Denote the interval pieces by $I_i := (x_{i-1}, x_i)$ $(1 \leq i \leq N + 1)$. For the subspace $V_N \subset H_0^1(a, b)$ let us choose the *piecewise linear*[19] *functions*:

[19] Although they are called *linear* functions, affine functions of the form $y = a + bx$ are meant. We shall follow the standard terminology and use 'linear' instead of 'affine'.

$$V_N = \left\{ u \in C^0([a,b]) : u|_{I_i} \text{ linear for } 1 \le i \le N+1 \text{ and } u(a) = u(b) = 0 \right\} \tag{8.32}$$

(cf. Figure 8.1). The continuity $u \in C^0([a,b])$ is equivalent to continuity at the nodes x_i $(1 \le i \le N)$: $u(x_i + 0) = u(x_i - 0)$.

Remark 8.32. Assume (8.32). (a) $u \in V_N$ is uniquely determined by its nodal values $u(x_i)$ $(1 \le i \le N)$:

$$u(x) = \frac{u(x_i)\,(x_{i+1} - x) + u(x_{i+1})\,(x - x_i)}{x_{i+1} - x_i} \quad \text{for } x \in I_{i+1},\ 0 \le i \le N,$$

where $u(x_0) = u(x_{N+1}) = 0$. From Theorem 6.60 we conclude that $V_N \subset H_0^1(a,b)$. The (weak) derivative $u' \in L^2(a,b)$ is piecewise constant:

$$u'(x) = [\,u(x_{i+1}) - u(x_i)\,] \,/\, [\,x_{i+1} - x_i\,] \qquad \text{for } x \in I_{i+1},\ 0 \le i \le N.$$

The basis functions can be defined by

$$b_i(x) = \left\{ \begin{array}{ll} (x - x_{i-1}) \,/\, (x_i - x_{i-1}) & \text{for } x_{i-1} < x \le x_i \\ (x_{i+1} - x) \,/\, (x_{i+1} - x_i) & \text{for } x_i < x < x_{i+1} \\ 0 & \text{otherwise} \end{array} \right\} \qquad (1 \le i \le N)$$

(see Figure 8.1b). In the engineering sciences the basis functions are also called the *shape functions*.

(b) One has the representation $u = \sum_{i=1}^{N} u(x_i)\, b_i$. The supports of the functions b_i are $\operatorname{supp}(b_i) = \bar{I}_{i-1} \cup \bar{I}_i = [x_{i-1}, x_{i+1}]$. The set B_i from Remark 8.31 is now $B_i = (x_{i-1}, x_{i+1})$.

The weak formulation of the boundary-value problem (8.31) is $a(u,v) = f(v)$ with

$$a(u,v) := \int_a^b u'v' \mathrm{d}x, \qquad f(v) := \int_a^b gv \mathrm{d}x \qquad \text{for } u, v \in H_0^1(a,b).$$

By Remark 8.31 we have, for the matrix elements, $L_{ij} = 0$ as soon as $|i - j| \ge 2$, since then $B_i \cap B_j = \emptyset$. For $|i - j| \le 1$ we obtain

$$L_{i,i-1} = a(b_{i-1}, b_i) = \int_{x_{i-1}}^{x_i} \frac{-1}{x_i - x_{i-1}} \frac{1}{x_i - x_{i-1}} \mathrm{d}x = \frac{-1}{x_i - x_{i-1}}, \tag{8.33a}$$

$$\begin{aligned} L_{i,i} = a(b_i, b_i) &= \int_{x_{i-1}}^{x_i} (x_i - x_{i-1})^{-2}\, \mathrm{d}x + \int_{x_i}^{x_{i+1}} (x_{i+1} - x_i)^{-2}\, \mathrm{d}x \\ &= \frac{1}{x_i - x_{i-1}} + \frac{1}{x_{i+1} - x_i}, \end{aligned}$$

$$L_{i,i+1} = a(b_{i+1}, b_i) = L_{i+1,i} = \frac{-1}{x_{i+1} - x_i}.$$

The right-hand side $\mathbf{f} = (f_1, \ldots, f_N)^{\mathsf{T}}$ is given by (8.33b):

$$f_i = f(b_i) = \int_{x_{i-1}}^{x_{i+1}} g b_i \mathrm{d}x = \frac{1}{x_i - x_{i-1}} \int_{x_{i-1}}^{x_i} g(x)\,(x - x_{i-1})\,\mathrm{d}x \qquad (8.33b)$$
$$+ \frac{1}{x_{i+1} - x_i} \int_{x_i}^{x_{i+1}} g(x)\,(x_{i+1} - x)\,\mathrm{d}x.$$

Remark 8.33. The system of equations $Lu = f$, given in (8.33a,b), is tridiagonal, thus, in particular, sparse. For an equidistant partitioning $x_i := a + ih$ with $h := (b-a)/(N+1)$ the coefficients are

$$L_{i,i\pm1} = -1/h, \quad L_{ii} = 2/h, \quad f_i = h \int_0^1 [\,g(x_i + th) + g(x_i - th)\,](1-t)\mathrm{d}t.$$

If $g \in C^0(a,b)$, then by the intermediate value theorem $f_i = hg(x_i + \Theta_i h)$, $|\Theta_i| < 1$. Therefore the system of equations $\mathbf{Lu} = \mathbf{f}$ is, up to a scaling, identical with the difference equation $L_h u_h = f_h$ in Section 4.1, if one defines f_h by $f_h(x_i) := f_i = g(x_i + \Theta_i h)$ instead of $f_h(x_i) = g(x_i)$; then $\mathbf{L} = hL_h$, $\mathbf{f} = hf_h$.

This example shows that it is possible to find basis functions with small supports so that $\mathbf{L}$ is relatively easy to calculate. In addition, the similarity is apparent between the resulting discretisation method and difference methods.

Finally, we consider the numerical evaluation of the integral (8.33b).

Exercise 8.34. Show that the Gaussian quadrature formula for $\int_{x_{i-1}}^{x_{i+1}} g b_i \mathrm{d}x$ with b_i as weight function and one support point is:

$$f_{i,\mathrm{GQuad}} := \frac{x_{i+1} - x_{i-1}}{2}\, g\left(\frac{x_{i+1} + x_i + x_{i-1}}{3}\right).$$

What is the formula for a partition of equal intervals?

If one replaces the Dirichlet boundary conditions in (8.31) by the Neumann condition then $V = H^1(a,b)$. The subspace $V_N \subset H^1(a,b)$ results from (8.32) after removal of the condition $u(a) = u(b) = 0$. In order that $\dim V_N = N$ the numbering has to be changed: The partition of (a,b) is given by $a = x_1 < x_2 < \ldots < x_N = b$. The support of the basis functions b_1 and b_N consists of only one interval.

Exercise 8.35. Let $\frac{\partial u}{\partial n}(a) = \varphi_a$, $\frac{\partial u}{\partial n}(b) = \varphi_b$, with $a(\cdot,\cdot)$ as before. Suppose the partition of (a,b) is equidistant: $x_i = a + (i-1)h$, $h = (b-a)/(N-1)$. What is the form of the equation $\mathbf{Lu} = \mathbf{f}$? Show that $\mathbf{Lu} = \mathbf{f}$ has at least one solution if

$$\int_a^b g(x)\mathrm{d}x + \varphi_a + \varphi_b = 0.$$

In §8.4.2–8.4.4 we discuss typical examples for finite-element functions in two spatial dimensions. More can be found in Braess [45, Table 2 on page 66] and Brenner–Scott [52, §3.2].

8.4.2 Linear Elements for $\Omega \subset \mathbb{R}^2$

We shall assume:

$$\Omega \subset \mathbb{R}^2 \text{ is a polygon.} \quad (8.34)$$

As shown in Figure 8.2 we divide Ω into triangles ('finite elements'). The decomposition is called a *triangulation*.[20]

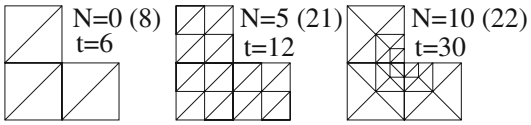


Fig. 8.2 Triangulation of Ω; t: number of triangles, N: number of inner nodes (number of all nodes).

Definition 8.36. $\mathcal{T} := \{T_1, \ldots, T_t\}$ is called an *admissible triangulation* of Ω if the following conditions are fulfilled:

$$T_i \ (1 \le i \le t) \text{ are open triangles,} \quad (8.35a)$$

$$T_i \text{ are disjoint, i.e., } T_i \cap T_j = \emptyset \text{ for } i \ne j, \quad (8.35b)$$

$$\bigcup_{i=1}^{t} \overline{T}_i = \overline{\Omega}\,, \quad (8.35c)$$

$$\text{for } i \ne j,\ \overline{T}_i \cap \overline{T}_j \text{ is } \begin{cases} \text{(i)} & \text{either empty, or} \\ \text{(ii)} & \text{a common side of the elements } T_i, T_j \text{ or} \\ \text{(iii)} & \text{a common edge of the elements } T_i, T_j. \end{cases} \quad (8.35d)$$

Remark 8.37. (a) The conditions (8.35a) and (8.35c) imply the polygonal shape of Ω, i.e., the assumption (8.34).

(b) Figure 8.2 shows only admissible triangulations. An example of an inadmissible triangulation would be, e.g., a square partitioned as follows: ⊠.

Let $\mathcal{T}$ be an admissible triangulation. The point $\mathbf{x}$ is called a *node* (of $\mathcal{T}$) if $\mathbf{x}$ is a vertex of one of the $T \in \mathcal{T}$. One distinguishes *inner* and *boundary* nodes, according to whether $\mathbf{x} \in \Omega$ or $\mathbf{x} \in \partial\Omega$. Let N be the total number of inner nodes. We define V_N as the subspace of the piecewise linear functions:

$$V_N = \left\{ u \in C^0(\overline{\Omega}) : \begin{array}{l} u = 0 \text{ on } \partial\Omega, \text{ on each } T \in \mathcal{T} \text{ the function} \\ u|_T \text{ agrees with a linear function,} \\ \text{i.e., } u(x,y) = a_{T,1} + a_{T,2}x + a_{T,3}y \text{ on } T. \end{array} \right\} \quad (8.36)$$

Remark 8.32 can be applied to the two-dimensional case under consideration:

Remark 8.38. (a) It is true that $V_N \subset H_0^1(\Omega)$.

(b) Each function $u \in V_N$ is uniquely determined by its node values $u(\mathbf{x}^i)$ at the inner nodes $\mathbf{x}^i$, $1 \le i \le N$.

[20] We will also use this term if other elements (quadrangles, tetrahedra, etc.) occur. A more general term is 'tessellation'.

Proof. (a) Example 6.22 shows that $V_N \subset H^1(\Omega)$. Since $u = 0$ on $\partial\Omega$, Theorem 6.60 proves that $V_N \subset H_0^1(\Omega)$.

(b) Let $\mathbf{x}$, $\mathbf{x}'$, $\mathbf{x}''$ be the three vertices of $T \in \mathcal{T}$. The linear function $u(x,y) = a_{i1}+a_{i2}x+a_{i3}y$ is uniquely determined on T by the values $u(\mathbf{x})$, $u(\mathbf{x}')$, $u(\mathbf{x}'')$. ■

The converse to Remark 8.38b is as follows.

Remark 8.39. Let $\mathbf{x}^i$ $(1 \le i \le N)$ be the inner nodes $\mathcal{T}$. For arbitrary values u_i $(1 \le i \le N)$ there exists exactly one $u \in V_N$ with $u(\mathbf{x}^i) = u_i$. It may be written as $u = \sum_{i=1}^N u_i\, b_i$, where the basis functions b_i are characterised by

$$b_i(\mathbf{x}^i) = 1, \qquad b_i(\mathbf{x}^j) = 0 \quad \text{for all } i \ne j \qquad (1 \le i,j \le N). \tag{8.37a}$$

If $T \in \mathcal{T}$ is a triangle with the vertices $\mathbf{x}^i = (x_i, y_i)$ [$\mathbf{x}^i$ as in (8.37a)] and $\mathbf{x}' = (x', y')$, $\mathbf{x}'' = (x'', y'')$, then

$$b_i(x,y) = \frac{(x-x')(y''-y') - (y-y')(x''-x')}{(x_i-x')(y''-y') - (y_i-y')(x''-x')} \qquad \text{on } T\,. \tag{8.37b}$$

On all $T \in \mathcal{T}$ that do not have $\mathbf{x}^i$ as a vertex, $b_i|_T = 0$.

Proof. Obviously, using (8.37b) we get a linear function defined on all $T \in \mathcal{T}$, which satisfies (8.37a). In addition, (8.37a) forces continuity at all nodes. If the vertices $\mathbf{x}^j$ and $\mathbf{x}^k$ are directly connected by the side of a triangle, then (8.37b) provides the representation $b_i(\mathbf{x}^j + s(\mathbf{x}^k - \mathbf{x}^j)) = s b_i(\mathbf{x}^k) + (1-s) b_i(\mathbf{x}^j)$ with $s \in [0,1]$ for both triangles that have this side in common. Thus b_i is also continuous on the common edges, so that $b_i \in C^0(\Omega)$. Applying this consideration to two boundary nodes $\mathbf{x}', \mathbf{x}''$ with $b_i(\mathbf{x}') = b_i(\mathbf{x}'') = 0$ we derive $b_i = 0$ on $\partial\Omega$. Thus b_i belongs to V_N. ■

Remark 8.40. (a) The dimension of the subspace determined by (8.36) is the number of inner nodes of $\mathcal{T}$.

(b) The support of the basis function b_i (belonging to the node $\mathbf{x}^i$) is

$$\operatorname{supp}(b_i) = \bigcup \left\{ \overline{T} : T \in \mathcal{T} \text{ has } \mathbf{x}^i \text{ as a corner} \right\}.$$

(c) Let B_i be the interior of $\operatorname{supp}(b_i)$. There holds $B_i \cap B_j = \emptyset$ if and only if the nodes $\mathbf{x}^i \ne \mathbf{x}^j$ are not directly connected by a side.

Conclusion 8.41. *$a(u,v) = \int_\Omega \langle \nabla u, \nabla v \rangle \mathrm{d}\mathbf{x}$ is the bilinear form associated to the Poisson equation. The integrals $L_{ij} = a(b_j, b_i) = \sum_k \int_{T_k} \langle \nabla b_j, \nabla b_i \rangle \mathrm{d}\mathbf{x}$ are to be taken over the following $T_k \in \mathcal{T}$:*

$$\begin{aligned} &\text{all } T_k \text{ with } \mathbf{x}^i \text{ as a vertex,} && \text{if } i = j,\\ &\text{all } T_k \text{ with } \mathbf{x}^i \text{ and } \mathbf{x}^j \text{ as vertices,} && \text{if } i \ne j. \end{aligned}$$

We have $L_{ij} = 0$, if $\mathbf{x}^i$ and $\mathbf{x}^j$ are not directly connected by the side of a triangle.

We get an especially regular triangulation when we first divide Ω into squares with sides of length h, and then divide these into two triangles ($\boxslash$). The first and second triangulations in Figure 8.2 are of this sort. We call them 'square-grid triangulations'. The corresponding basis function is depicted in Figure 8.3. Its support consists of six triangles. One therefore expects that the matrix $\mathbf{L}$ corresponds to a 7-point formula. For the Laplace operator, however, one finds the well-known 5-point formula (4.19) from Section 4.2.

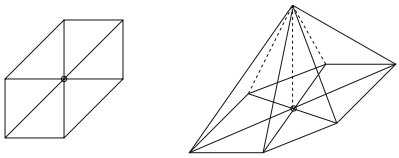


Fig. 8.3 Piecewise linear basis functions for the node (x_0, y_0). Left: triangles of the support; right: sketch of the function.

Exercise 8.42. Let $\mathcal{T}$ be a square-grid triangulation, and consider the bilinear form $a(u,v) = \int_\Omega \langle \nabla u, \nabla v \rangle \mathrm{d}\mathbf{x}$. The basis functions are described by Figure 8.3. Show that one has for the entries of the system matrix $\mathbf{L}$

$$L_{ii} = 4, \qquad L_{ij} = -1, \quad \text{if } \mathbf{x}^i - \mathbf{x}^j \in \{(0,h), (0,-h), (h,0), (-h,0)\},$$

and $L_{ij} = 0$ otherwise, i.e., $\mathbf{L}$ agrees with $h^2 L_h$ from (4.16).

Although $\mathbf{L} = h^2 L_h$ holds in the case of the Poisson equation $-\Delta u = g$, the finite-element discretisation and the difference method still do not coincide, since $h^2 f_h$ has $h^2 g(x^i)$ as components and so differs from $f_i = \int_\Omega g\, b_i \, \mathrm{d}\mathbf{x}$.

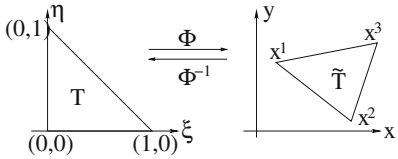


Fig. 8.4 Reference triangle T.

The integration $\int_{T_i} \ldots \mathrm{d}\mathbf{x}$ over the triangle $T_i \in \mathcal{T}$ seems at first difficult. However, for each i one can express $\int_{T_i} \ldots \mathrm{d}\mathbf{x}$ as an integral over the reference triangle T in Figure 8.4. The details are given below.

Exercise 8.43. Let $\mathbf{x}^i = (x_i, y_i)$ $(i = 1,2,3)$ be the vertices of $\tilde{T} \in \mathcal{T}$, and let T be the unit triangle in Figure 8.4 (left). Show:

(a) $\Phi : (\xi, \eta) \mapsto \mathbf{x}^1 + \xi\left(\mathbf{x}^2 - \mathbf{x}^1\right) + \eta\left(\mathbf{x}^3 - \mathbf{x}^1\right)$ maps T onto $\tilde{T}$.

(b) $\det \Phi'(\xi,\eta) = (x_2 - x_1)(y_3 - y_1) - (y_2 - y_1)(x_3 - x_1)$ is a constant.

(c) The substitution rule gives

$$\int_{\tilde{T}} v(x,y)\mathrm{d}x\mathrm{d}y = \left|\, (x_2 - x_1)(y_3 - y_1) - (y_2 - y_1)(x_3 - x_1) \,\right| \int_T v(\Phi(\xi,\eta))\, \mathrm{d}\xi \mathrm{d}\eta\,.$$

In general one evaluates the integral $\int_T \ldots \mathrm{d}\xi \mathrm{d}\eta$ over the unit triangle numerically. Examples of integration formulae can be found in Schwarz [262, §2.4.3] and Ciarlet [67, §4.1]. The necessary order of the quadrature formulae is discussed in the same place by Ciarlet and by Witsch [307]. See also Scott [263].

In contrast to the difference methods finite-element discretisation offers one the possibility of changing the size of the triangles locally. The third triangulation in Figure 8.2 contains triangles that become smaller as they are nearer to the intruding corner. This flexibility of the finite-element method is an essential advantage. On the other hand, one does obtain systems of equations $\mathbf{Lu} = \mathbf{f}$ with more complex structures, since (a) $\mathbf{u}$ can no longer be stored in a two-dimensional array, (b) $\mathbf{L}$ cannot be characterised by a star as in (4.20).

Remark 8.44. If one replaces the Dirichlet condition $u = 0$ on $\partial\Omega$ by natural boundary conditions, then the following changes take place:

(a) $N = \dim V_N$ is the number of all nodes (inner *and* boundary nodes).

(b) $V_N \subset H^1(\Omega)$ is given by (8.36) without the restriction "$u = 0$ on $\partial\Omega$".

(c) In Remarks 8.38–8.40 it should read "nodes" instead of "inner nodes".

(d) The matrix elements L_{ij} calculated in Exercise 8.42 are valid only for inner nodes $\mathbf{x}^i$. For a Neumann boundary-value problem in the square $\Omega = (0,1)\times(0,1)$, $\mathbf{L}$ coincides with $h^2 D_h L_h$ from Exercise 4.63b.

(e) In calculating $f_i = f(b_i)$ with $\varphi \neq 0$ in (7.20a) one has to take account of possible boundary integrals over $\partial\Omega \cap \mathrm{supp}(b_i)$, if $\mathrm{supp}(b_i) \cap \partial\Omega \neq \emptyset$.

8.4.3 Bilinear Elements for $\Omega \subset \mathbb{R}^2$

The difference methods in a square grid suggest partitioning Ω into squares of side h (cf. Figure 8.5, left). If, more generally, one replaces the squares by parallelograms one obtains partitions like those in Figure 8.5. An admissible partition by parallelograms is described by the conditions (8.35a–d), if in (8.35a) the expression "triangle" is replaced by "parallelogram".

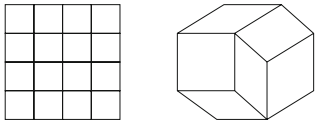


Fig. 8.5 Partition of Ω into parallelograms.

If one were to define the subspace V_N by the condition that u must be a linear function in each parallelogram, then there would be only three of the four vertex values that could be arbitrarily assigned. In the case of the partition in Figure 8.5 (right) one can see that the only piecewise linear function u with $u = 0$ on $\partial\Omega$ is the null function. Thus in each parallelogram u must be a function that involves four free parameters. We next consider the case of a rectangle parallel to the axes, $P = (x_1, x_2) \times (y_1, y_4)$, and define a *bilinear function* on P by

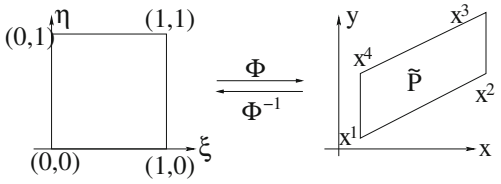


Fig. 8.6 Unit square as reference parallelogram for $\tilde{P}$.

$$u(x,y) := a_1 + a_2 x + a_3 y + a_4 xy \qquad \text{in } P\,. \tag{8.38a}$$

Then u is linear in each direction parallel to the axes—thus, in particular, along the sides of the rectangle. For an arbitrary parallelogram $\tilde{P}$ such as in Figure 8.6, the restriction of the function (8.38a) to the side of a parallelogram is in general a quadratic function. Therefore one generalises the definition as follows. Let

$$\Phi : (\xi, \eta) \mapsto \mathbf{x}^1 + \xi \left(\mathbf{x}^2 - \mathbf{x}^1\right) + \eta \left(\mathbf{x}^4 - \mathbf{x}^1\right) \in \mathbb{R}^2 \tag{8.38b}$$

be the mapping taking the unit square $(0,1) \times (0,1)$ onto the parallelogram $\tilde{P}$ (cf. Figure 8.6). A bilinear function is defined on $\tilde{P}$ by

$$u(x,y) := v(\Phi^{-1}(x,y)), \quad v(\xi,\eta) := \alpha + \beta\xi + \gamma\eta + \delta\xi\eta.$$

It is not necessary to calculate $v(\Phi^{-1}(\cdot,\cdot))$ explicitly, since all the integrations can be carried out over $(0,1) \times (0,1)$ as the reference parallelogram (cf. Exercise 8.43).

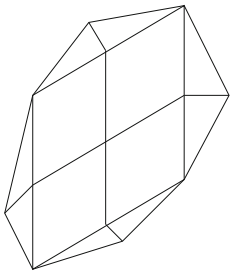


Fig. 8.7 A combination of triangles and parallelograms.

If $\pi = \{P_1, \ldots, P_t\}$ is an admissible partition into parallelograms one defines $V_N \subset H^1(\Omega)$ [resp. $V_N \subset H_0^1(\Omega)$] by

$$V_N = \left\{ u \in C^0(\overline{\Omega}) : \begin{array}{l} \text{on all } P \in \pi\,, \ u \text{ coincides} \\ \text{with a bilinear function} \end{array} \right\} \tag{8.39a}$$

respectively

$$V_N = \left\{ u \in C^0(\overline{\Omega}) : \begin{array}{l} u = 0 \text{ on } \Gamma = \partial\Omega, \\ \text{on all } P \in \pi\,, \ u \text{ coincides} \\ \text{with a bilinear function.} \end{array} \right\} \tag{8.39b}$$

Here $N = \dim V_N$ is the number of nodes (case (8.39a)) [resp. inner nodes, case (8.39b)]. A bilinear function on $P \in \pi$ is linear along each side of P. Continuity in the node points thus already implies continuity in Ω.

Remark 8.45. The Remarks 8.37a and 8.38–8.40 hold with appropriate changes.

Exercise 8.46. Let the bilinear form associated to $-\Delta$ be $a(u,v) = \int_\Omega \langle \nabla u, \nabla v \rangle \mathrm{d}\mathbf{x}$. Assume $\Omega = (0,1) \times (0,1)$ is divided into squares of side h as in Figure 8.5. What are the basis functions characterised by (8.37a)? Show that the matrix $\mathbf{L}$ coincides with the difference star $\frac{1}{3}\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$.

Remark 8.47. Triangle and parallelogram divisions can also be combined. A polygonal domain Ω can be divided up into both triangles and parallelograms (cf. Figure 8.7). In this case V_N is defined as

$$\{u \in C^0(\Omega) : \ u \text{ linear on the triangles, bilinear on the parallelograms}\}.$$

8.4.4 Quadratic Elements for $\Omega \subset \mathbb{R}^2$

Let $\mathcal{T}$ be an admissible triangulation of a polygonal domain Ω. We wish to increase the dimension of the finite-element subspace by allowing, instead of linear functions, quadratics

$$u(x,y) = a_1 + a_2 x + a_3 y + a_4 x^2 + a_5 xy + a_6 y^2 \qquad \text{on } T \in \mathcal{T} \tag{8.40}$$

so that

$$\begin{aligned} V_N = \{u \in C^0(\overline{\Omega}) : u = 0 \text{ on } \Gamma = \partial\Omega, \\ \text{on each } T_i \in \mathcal{T},\ u \text{ coincides with a quadratic function}\}. \end{aligned} \tag{8.41}$$

Lemma 8.48. *(a) Let* $\mathbf{x}^1$, $\mathbf{x}^2$, $\mathbf{x}^3$ *be the vertices of a triangle* $T \in \mathcal{T}$, *while* $\mathbf{x}^4$, $\mathbf{x}^5$, $\mathbf{x}^6$ *are the midpoints of the sides*[21] *(cf. Figure 8.8a). Each function quadratic on* T *is determined by the values* $\{u(\mathbf{x}^j) : j = 1, \ldots, 6\}$.

(b) The restriction of the function (8.40) to a side of $T \in \mathcal{T}$ *gives a one-dimensional quadratic function, which is uniquely determined by three of the nodes lying on this side (e.g.,* $u(\mathbf{x}^1)$, $u(\mathbf{x}^4)$, $u(\mathbf{x}^2)$ *in Figure 8.8a).*

(c) If u *is quadratic on each* $T \in \mathcal{T}$ *and continuous at all nodes (i.e., vertices of triangles and midpoints of sides), then* u *is continuous on* Ω.

Proof. (a) The quadratic function can be obtained as the uniquely defined interpolating polynomial of the form (8.40). Part (b) is also elementary.

(c) Let T and $\hat{T}$ be neighbouring triangles in Figure 8.8a. Since $u|_T$ and $u|_{\hat{T}}$ coincide on $\mathbf{x}^1$, $\mathbf{x}^4$, $\mathbf{x}^2$, by part (b) they represent the same quadratic function on the common side of T and $\hat{T}$. ∎

We call all [inner] vertices of triangles and midpoints of sides [inner] nodes. By Lemma 8.48a we can find for each inner node $\mathbf{x}^i$ a basis function b_i which is quadratic on each $T \in \mathcal{T}$ and satisfies (8.37a):

$$b_i(\mathbf{x}^i) = 1, \quad b_i(\mathbf{x}^j) = 0 \quad \text{for } j \neq i.$$

By Lemma 8.48c b_i belongs to V_N. This proves the next statement.

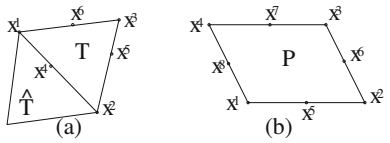


Fig. 8.8 Nodes for (a) a quadratic ansatz on a triangle, (b) a quadratic ansatz of the serendipity class on a parallelogram.

Remark 8.49. The number of inner nodes is the dimension of V_N in (8.41). Each $u \in V_N$ admits the expression $u = \sum_{i=1}^{N} u(\mathbf{x}^i)\, b_i$, where the basis function belonging to the node $\mathbf{x}^i$ is characterised by (8.37a).

[21] Instead of the midpoints any other point on the interior of the side can be chosen. However, the choice must be the same for both triangles attached to this side.

In the sequel we wish to assume that, as in Figure 8.7, both triangles and parallelograms are used in the partition. On the triangles the function $u \in V_N$ are quadratic. The functions on the parallelograms P must satisfy the following conditions:

(α) $u(x,y)|_P$ must be uniquely determined by the values at the 4 vertices and 4 midpoints of the sides (cf. Figure 8.8b).

(β) The restriction to a side of P gives a (one-dimensional) quadratic function.

By condition (α) the ansatz must contain exactly 8 coefficients. The quadratic (8.40) has only 6 coefficients, while the biquadratic

$$\sum_{0 \le i,j \le 2} a_{ij} \xi^i \eta^j$$

has one parameter too many. If one omits the term $\xi^2\eta^2$ in the biquadratic ansatz one obtains for the unit square the function

$$v(\xi,\eta) := a_1 + a_2\xi + a_3\eta + a_4\xi^2 + a_5\xi\eta + a_6\eta^2 + a_7\xi^2\eta + a_7\xi\eta^2, \qquad (8.42)$$

the so-called quadratic ansatz of the *serendipity class*. The restrictions to the sides $\xi = 0, 1$ [resp. $\eta = 0, 1$] give a quadratic function in η [resp. ξ]. The mapping Φ in (8.38b) (cf. Figure 8.6a,b) yields the function

$$u(x,y) = v(\Phi^{-1}(x,y)),$$

defined on the parallelogram P, which still satisfies the conditions (α) and (β).

If one were to use the full biquadratic ansatz one would need one additional node which one could choose as the barycentre of the parallelogram. Cubic ansatzes can be carried out in the same way in triangles and parallelograms (cf. Schwarz [262]).

8.4.5 Elements for $\Omega \subset \mathbb{R}^3$

In the three-dimensional case assume

$$\Omega \subset \mathbb{R}^3 \qquad \text{is a polyhedron.}$$

The triangulation from §8.4.2 corresponds now to a division of Ω into tetrahedra. It is called admissible, if (8.35a–d) hold in the appropriate sense: (8.35a) becomes "T_i $(1 \le i \le t)$ are open tetrahedra"; in (8.35d) it now should read "For $i \ne j$, $\bar{T}_i \cap \bar{T}_j$ is either empty, or a common vertex, side, or face of T_i and T_j".

Each linear function $a_1 + a_2x + a_3y + a_4z$ is uniquely determined by its values at the 4 vertices of the tetrahedron. As a basis for the space

$$V_N = \left\{ u \in C^0(\overline{\Omega}) : \; u = 0 \text{ on } \partial\Omega, \; u \text{ linear on each tetrahedron } T_i \; (1 \le i \le t) \right\}$$

one chooses b_i with the property (8.37a): $b_i(\mathbf{x}^i) = 1$, $b_i(\mathbf{x}^j) = 0$ $(j \neq i)$. The support of b_i consists of all tetrahedra that share $\mathbf{x}^i$ as a vertex. The dimension $N = \dim V_N$ is again the number of inner nodes (i.e., vertices of tetrahedra).

As in the two-dimensional case the linear ansatz may be replaced by a quadratic one. Instead of a tetrahedron one can use a parallelepiped or a triangular prism with corresponding ansatzes for the functions (cf. Schwarz [262, §2.7]).

8.4.6 Handling of Side Conditions

The space V which lies at the foundation of the whole matter may be a subspace of a simply discretisable space $W \supset V$. A given function $w \in W$ belongs to V if certain side conditions are satisfied. Before we describe this situation in full generality, two examples will be provided as illustrations.

Example 8.50. If one wishes to make the Neumann boundary-value problem $-\Delta u = g$ in Ω and $\frac{\partial u}{\partial n} = \varphi$ on Γ uniquely solvable by the addition of the side condition $\int_\Omega u(x)\mathrm{d}\mathbf{x} = 0$, one can choose the space V to be

$$V = \left\{ u \in H^1(\Omega) : \int_\Omega u(\mathbf{x})\mathrm{d}\mathbf{x} = 0 \right\}. \tag{8.43a}$$

For a bounded domain Ω the bilinear form $a(u,v) := \int_\Omega \langle \nabla u, \nabla v \rangle \mathrm{d}\mathbf{x}$ is V-elliptic. The weak formulation (8.1) with $f(v) := \int_\Omega gv\mathrm{d}x + \int_\Gamma \varphi v \mathrm{d}\Gamma$ corresponds to the equation

$$-\Delta u = g, \;\; \partial u/\partial n = \varphi,$$

if g and φ satisfy the integrability condition $f(1) = 0$ (cf. (3.17)). But even when $f(1) \neq 0$ there exists a weak solution of the corrected equation

$$-\Delta u = \tilde{g}, \qquad \tilde{g}(\mathbf{x}) := g(\mathbf{x}) - \left[\int_\Omega g\,\mathrm{d}\mathbf{x} + \int_\Gamma \varphi\,\mathrm{d}\Gamma \right] / \int_\Omega \mathrm{d}\mathbf{x}\,.$$

Example 8.51. The Adler problem of Exercise 7.37 uses

$$V = \left\{ u \in H^1(\Omega) : u \text{ constant on } \Gamma \right\}. \tag{8.43b}$$

In both cases $W = H^1(\Omega)$ is a proper superset of V. Let $\mathcal{T}$ be a triangulation of Ω with I_{in} inner nodes and I_{bd} boundary nodes. Let $W_h \subset W$ be the space of linear triangular elements[22] (cf. Remark 8.44). Its dimension is

$$N_h := \dim W_h = I_{\mathrm{in}} + I_{\mathrm{bd}}. \tag{8.44a}$$

[22] Here we use the (maximal) grid size h as index of V_h, W_h, and M_h, since the dimension N differs in the cases with and without side condition.

The basis functions $\{b_i : 1 \le i \le N_h\} \subset W_h$ will be described as usual by $b_i(\mathbf{x}^j) = \delta_{ij}$ (cf. (8.37a)). As the finite-element subspace of V we define $V_h := W_h \cap V$, which is of smaller dimension. The difference M_h corresponds to the number of side conditions:

$$V_h := W_h \cap V, \qquad M_h := N_h - \dim V_h. \tag{8.44b}$$

In the Examples 8.50 and 8.51 we have $M_h = 1$, resp. $M_h = I_{\mathrm{bd}} - 1$. The difficulty in the numerical solution of the discrete problem (8.45),

$$\text{find } u^h \in V_h \quad \text{with} \quad a(u^h, v) = f(v) \quad \text{for all } v \in V_h \tag{8.45}$$

begins with the choice of a basis for V_h. It is not possible to use a subset of the functions $b_1, \ldots, b_{N_h}$. Thus, for example, in Example 8.50 no b_i belongs to V, and thus also none to V_h, since $\int_\Omega b_i \, \mathrm{d}\mathbf{x} > 0$.

In principle, it is possible in the case of (8.43a) to construct as new basis functions linear combinations $b_i := \alpha b_{i_1} + \beta b_{i_2} \in V_h$, which have again localised (but a bit larger) supports. In the case of (8.43b) it is still relatively simple to find a practical basis for V_h: The basis functions $b_i \in W_h$ which belong to inner nodes are also in V_h, since $b_i \in H_0^1(\Omega) \subset V$. As further elements one uses $b_0 := \sum' b_j$, where $\sum'$ denotes the sum over all boundary nodes. In spite of this, even in this case, it would simply be easier if one could work with the standard basis of W_h.

In order to treat the problem (8.45) with the aid of $b_i \in W_h$, we reintroduce the notation $w^h = P\mathbf{w}$ ($\mathbf{w} \in \mathbb{R}^{N_h}$ a coefficient vector, $w^h \in W_h$, P from (8.6)). Each $v \in V_h \subset W_h$ can be written as $P\mathbf{v}$ with $\mathbf{v}$ in $\mathbf{V} := \{\mathbf{v} \in \mathbb{R}^{N_h} : P\mathbf{v} \in V_h\} = P^{-1}V_h$. Thus problem (8.45) is equivalent to:

$$\text{find } \mathbf{u} \in \mathbf{V} \quad \text{with} \quad a(P\mathbf{u}, P\mathbf{v}) = f(P\mathbf{v}) \quad \text{for all } \mathbf{v} \in \mathbf{V}. \tag{8.45$'$}$$

From (8.44a,b) we have $\dim \mathbf{V} = \dim V_h = N_h - M_h = \dim W_h - M_h$. The spaces V_h and $\mathbf{V}$ are described by M_h linear conditions

$$\mathbf{V} = \ker(\mathbf{C}) = \left\{\mathbf{w} \in \mathbb{R}^{N_h} : \mathbf{C}\mathbf{w} = \mathbf{0}\right\}, \tag{8.46}$$

where $\mathbf{C} = (c_{ij})$ is the $M_h \times N_h$ matrix of the coefficients of the M_h linear side conditions

$$\sum_{j=1}^{N_h} c_{ij} w_j = 0 \qquad (1 \le i \le M_h). \tag{8.47}$$

In the case of Example 8.50 we have

$$M_h = 1, \qquad c_{1j} = \int_\Omega b_j(\mathbf{x}) \mathrm{d}\mathbf{x} \qquad (1 \le j \le N_h).$$

For Example 8.51 let $\mathbf{x}^0, \ldots, \mathbf{x}^{M_h}$, with $M_h = I_{\mathrm{bd}} - 1$, be the boundary nodes. Then $\mathbf{C}$ can be defined as follows:

$$M_h = I_{\mathrm{bd}} - 1,\ c_{ii} = 1\ (1 \le i \le M_h),\ c_{i,i+1} = -1\ (1 \le i < M_h),\ c_{M_h,0} = -1$$

and $c_{ij} = 0$ otherwise.

The variation of $\mathbf{v}$ over $\mathbf{V}$ in problem (8.45′) can be replaced by the variation of $\mathbf{w}$ over $\mathbb{R}^{N_h}$, if one couples the conditions (8.47) by using Lagrange multipliers λ_i $(1 \le i \le M_h)$, which can be put together to form a vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{M_h})$. The resulting formulation of the problem is:

$$\text{find } \mathbf{u} \in \mathbb{R}^{N_h} \quad \text{and} \quad \boldsymbol{\lambda} \in \mathbb{R}^{M_h} \quad \text{with } \mathbf{C}\mathbf{u} = \mathbf{0} \quad \text{and} \tag{8.48a}$$

$$a(P\mathbf{u}, P\mathbf{w}) + \langle \boldsymbol{\lambda}, \mathbf{C}\mathbf{w} \rangle = f(P\mathbf{w}) \qquad \text{for all } \mathbf{w} \in \mathbb{R}^{M_h}, \tag{8.48b}$$

where $\langle \boldsymbol{\lambda}, \boldsymbol{\mu} \rangle = \sum \lambda_i \mu_i$ is the scalar product in $\mathbb{R}^{M_h}$.

Theorem 8.52. *The problems (8.45) and (8.48a,b) are equivalent in the following sense. If* $\mathbf{u}$, $\boldsymbol{\lambda}$ *is a solution pair for (8.48a,b) then* $\mathbf{u}$ *is a solution of (8.45′) and* $P\mathbf{u}$ *is a solution of (8.45). Conversely, if* $u^h = P\mathbf{u}$ *is a solution of (8.45) then there exists precisely one* $\boldsymbol{\lambda} \in \mathbb{R}^{M_h}$ *such that* $\mathbf{u}$ *and* $\boldsymbol{\lambda}$ *solve (8.48a,b).*

Before we prove this theorem, we give a matrix formulation which is equivalent to (8.48a,b).

Remark 8.53. Let $\mathbf{L}$ and $\mathbf{f}$ be defined as in (8.8a,b). Further, suppose that $\mathbf{B} := \mathbf{C}^{\mathsf{T}}$ with $\mathbf{C}$ from (8.46). Then (8.48a,b) is equivalent to the system of equations

$$\begin{bmatrix} \mathbf{L} & \mathbf{B} \\ \mathbf{B}^{\mathsf{T}} & \mathbf{O} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{0} \end{bmatrix}. \tag{8.48′}$$

Proof. $\mathbf{L}\mathbf{u} + \mathbf{B}\boldsymbol{\lambda} = \mathbf{f}$ is equivalent to

$$a(P\mathbf{u}, P\mathbf{w}) + \langle \boldsymbol{\lambda}, \mathbf{C}\mathbf{w} \rangle = \langle \mathbf{L}\mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{B}\boldsymbol{\lambda}, \mathbf{w} \rangle = \langle \mathbf{f}, \mathbf{w} \rangle = f(P\mathbf{w})$$

for all $\mathbf{w} \in \mathbb{R}^{N_h}$. Note $\mathbf{B}^{\mathsf{T}}\mathbf{u} = \mathbf{0}$ is the same as $\mathbf{C}\mathbf{u} = \mathbf{0}$. ∎

Proof of Theorem 8.52. (a) Assume a solution of (8.48a,b) is given in terms of $\mathbf{u}, \boldsymbol{\lambda}$. Then $\mathbf{C}\mathbf{u} = \mathbf{0}$ implies $\mathbf{u} \in \mathbf{V}$ and $u^h := P\mathbf{u} \in V_h$. Equation (8.48b) holds in particular for all $\mathbf{w} \in \mathbf{V}$, so that (8.45′) follows since $\mathbf{C}\mathbf{w} = \mathbf{0}$.

(b) (8.45′) implies $\mathbf{f} - \mathbf{L}\mathbf{u} \in \mathbf{V}^{\perp}$. Since $\mathbf{V}^{\perp} = (\ker(\mathbf{C}))^{\perp} = (\ker(\mathbf{B}^{\mathsf{T}}))^{\perp} = \mathrm{range}(\mathbf{B})$, there exists a $\boldsymbol{\lambda} \in \mathbb{R}^{M_h}$ with $\mathbf{f} - \mathbf{L}\mathbf{u} = \mathbf{B}\boldsymbol{\lambda}$, so that (8.48′), and thus (8.48a,b) are satisfied. For the $N_h \times M_h$ matrix $\mathbf{B}$ with rank M_h we have $\ker(\mathbf{B}) = \{0\}$, so that the solution is unique. ∎

Note that in the case of Example 8.50 the system of equations (8.48′) is essentially identical to (4.68a,b).

8.5 Error Estimates for Finite-Element Methods

In this section we shall restrict ourselves to the consideration of the linear elements from §8.4.2 and therefore assume:

$$\mathcal{T} \quad \text{is an admissible triangulation of } \Omega \subset \mathbb{R}^2, \tag{8.49a}$$

$$V_N \quad \text{is defined by (8.36), if } V = H_0^1(\Omega), \tag{8.49b}$$

$$V_N \quad \text{is as in Remark 8.44b, if } V = H^1(\Omega). \tag{8.49c}$$

8.5.1 Preparations

By Theorem 8.21 one has to determine $d(u, V_N) = \inf\{|u - w|_1 : w \in V_N\}$. To do this one begins by looking at the reference triangle T from Figure 8.4. The right-hand side in (8.50) consists of values of u at the corners of T and the L^2 norm of the second partial derivatives.

Lemma 8.54. *Let* $T = \{(\xi, \eta) : \xi, \eta \geq 0,\ \xi + \eta \leq 1\}$. *For all* $u \in H^2(T)$ *there holds*

$$\|u\|^2_{H^2(T)} \leq C\Big[\, |u(0,0)|^2 + |u(1,0)|^2 + |u(0,1)|^2 + \sum_{|\alpha|=2} \|D^\alpha u\|^2_{L^2(T)} \Big]\,. \tag{8.50}$$

Proof. (i) First it has to be shown that the bilinear form

$$a(u,v) := u(0,0)v(0,0) + u(1,0)v(1,0) + u(0,1)v(0,1) + \sum_{|\alpha|=2} (D^\alpha u, D^\alpha v)_{L^2(T)}$$

is continuous on $H^2(T) \times H^2(T)$ and $H^2(T)$-coercive. The continuity follows from the continuous embedding $H^2(T) \subset C^0(T)$, which implies $|u(\mathbf{x})| \leq \hat{C}\,|u|_2$ for all $\mathbf{x} \in T$ (cf. Theorem 6.48). Thus we have

$$|a(u,v)| \leq (1 + 3\hat{C})\,|u|_2\,|v|_2\,.$$

Since $T \in C^{0,1}$, one may apply Theorem 6.86b: $H^2(T)$ is compactly embedded in $H^1(T)$. By Lemma 6.90, there exists a constant $C_{1/2}$ such that

$$|u|_1^2 \leq \Big(\tfrac{1}{2}\,|u|_2 + C_{1/2}\,|u|_0\Big)^2 \leq \frac{1}{2}\,|u|_2^2 + 2C_{1/2}^2\,|u|_0^2 \quad \text{for all } u \in H^2(T). \quad (*)$$

Since $\sum_{|\alpha|=2} |D^\alpha u|_0^2 = |u|_2^2 - |u|_1^2$, the estimate

$$a(u,u) \geq |u|_2^2 - |u|_1^2 \underset{(*)}{\geq} \frac{1}{2}\,|u|_2^2 - 2\,C_{1/2}^2\,|u|_0^2$$

shows that $a(\cdot,\cdot)$ is $H^2(T)$-coercive.

(ii) Since the embedding $H^2(T) \subset L^2(T)$ is also compact one may apply Theorem 6.107. The operator $A \in L(H^2(T), H^2(T)')$ which is associated with $a(\cdot,\cdot)$ either has an inverse $A^{-1} \in L(H^2(T)', H^2(T))$ or has $\lambda = 0$ as an eigenvalue with an eigenfunction $0 \neq e \in H^2(T)$. In the latter case e must, in particular, satisfy the equation $a(e,e) = 0$. From this follows that $D^\alpha e = 0$ for all $|\alpha| = 2$, so that e must be linear: $e(x,y) = \alpha + \beta x + \gamma y$. Furthermore, from $a(e,e) = 0$ one concludes that $e(0,0) = e(1,0) = e(0,1) = 0$, so that $e = 0$ in contradiction to what was just assumed. From Lemma 6.94 and Exercise 6.98c one may show the $H^2(T)$-ellipticity: $a(u,u) \geq C_E\,|u|_2^2$ with

$$C_E := \frac{1}{(1+3\hat{C})\,\|A^{-1}\|^2_{H^2(T)\leftarrow H^2(T)'}} > 0$$

and $\hat{C}$ from part (i). Thus we have assertion (8.50) with $C := 1/C_E$. ∎

Lemma 8.54 is tailored to the case of linear elements, since the right-hand side in (8.50) vanishes for linear interpolants at $(0,0)$, $(1,0)$, $(0,1)$. The generalisation to higher order elements is as follows : Let $\mathbf{x}^i \in T$ $(1 \leq i \leq q)$ be the nodes which uniquely determine an interpolating polynomial of degree $\leq t-1$. Then the norms $\|u\|_{H^t(T)}$ and

$$\sqrt{\sum_{i=1}^{q} |u(\mathbf{x}^i)|^2 + \sum_{|\alpha|=t} \|D^\alpha u\|^2_{L^2(T)}}$$

are equivalent.

The estimate (8.50) is also valid for other triangles T, but the constant C depends on T. First we study the case of a scaled unit triangle. Note that derivatives of different order scale differently. Therefore we do not estimate $\|u\|^2_{H^2(T_h)}$, but $\|D^\beta u\|^2_{L^2(T)}$ for $|\beta| \leq 2$.

Lemma 8.55. *Let $h > 0$ and $T_h := hT = \{(x,y) : x,y \geq 0,\ x+y \leq h\}$. For each $u \in H^2(T_h)$ there holds with $\beta \in \mathbb{N}_0^2$, $|\beta| \leq 2$:*

$$\left\|D^\beta u\right\|^2_{L^2(T)} \leq C \left\{ h^{2-2|\beta|} \left[|u(0,0)|^2 + |u(1,0)|^2 + |u(0,1)|^2 \right] + h^{4-2|\beta|} \sum_{|\alpha|=2} \|D^\alpha u\|^2_{L^2(T)} \right\}. \tag{8.51}$$

C is the constant in (8.50) and therefore independent of u, β, and h.

Proof. Let $v(\xi,\eta) := u(\xi h, \eta h) \in H^2(T)$. The derivatives with respect to (x,y) and to $(\xi,\eta) = (x,y)/h$ are related by $D^\beta_{x,y} = h^{-|\beta|} D^\beta_{\xi,\eta}$. For each $|\beta| \leq 2$ the substitution rule gives

$$\left\|D^\beta u\right\|^2_{L^2(T_h)} = \iint_{T_h} \left|D^\beta_{x,y} u\right|^2 \mathrm{d}x\mathrm{d}y = \iint_T \left|h^{-|\beta|} D^\beta_{\xi,\eta} v\right|^2 h^2 \mathrm{d}\xi\mathrm{d}\eta$$

$$= h^{2-2|\beta|} \left\|D^\beta v\right\|^2_{L^2(T)} \le h^{2-2|\beta|} \left\|v\right\|^2_{H^2(T)} \qquad (**)$$

$$\underset{\text{Lemma 8.54}}{\le} Ch^{2-2|\beta|} \left[|v(0,0)|^2 + |v(1,0)|^2 + |v(0,1)|^2 + \sum_{|\alpha|=2} \|D^\alpha v\|^2_{L^2(T)} \right].$$

From $v(0,0) = u(0,0)$, $v(1,0) = u(h,0)$, $v(0,1) = u(0,h)$ and (**) with $\beta := \alpha$ and $|\alpha| = 2$ we obtain

$$\|D^\alpha_{\xi,\eta} v\|^2_{L^2(T)} = h^2 \|D^\alpha_{x,y} u\|^2_{L^2(T_h)}$$

and thus the assertion (8.51). ∎

Lemma 8.56. *Let $\tilde{T}$ be an arbitrary triangle with*

$$\textit{side lengths} \le h_{\max}, \qquad \textit{interior angles} \ge \alpha_0 > 0. \tag{8.52}$$

For each $u \in H^2(\tilde{T})$ and $|\beta| \le 2$, there holds

$$\left\|D^\beta u\right\|^2_{L^2(\tilde{T})} \le C(\alpha_0) \left\{ h^{2-2|\beta|}_{\max} \sum_{\substack{\mathbf{x} \text{ vertex} \\ \text{of } \tilde{T}}} |u(\mathbf{x})|^2 \;+\; h^{4-2|\beta|}_{\max} \sum_{|\alpha|=2} \|D^\alpha u\|^2_{L^2(\tilde{T})} \right\}, \tag{8.53}$$

where $C(\alpha_0)$ depends only on α_0 and not on u, β or h.

Proof. Let $h \le h_{\max}$ be one of the lengths of the sides of $\tilde{T}$, while T_h is the triangle from Lemma 8.55. In a way similar to that in Exercise 8.43a, let $\Phi : T_h \to \tilde{T}$ be the linear transformation that maps T_h onto $\tilde{T}$. Then $v(\xi,\eta) := u(\Phi(\xi,\eta))$ belongs to $H^2(T_h)$. Under the condition (8.52) we know that the determinant $|\det \Phi| \in [1/K(\alpha_0), K(\alpha_0)]$ is bounded both above and below. From

$$\|D^\beta_{x,y} u\|^2_{L^2(\tilde{T})} \le C_1(\alpha_0) \sum_{|\beta'|=|\beta|} \|D^{\beta'}_{\xi,\eta} v\|^2_{L^2(T_h)},$$

(8.51) and

$$\sum_{|\alpha|=2} \|D^\alpha_{\xi,\eta} v\|^2_{L^2(T_h)} \le C_2(\alpha_0) \sum_{|\alpha|=2} \|D^\alpha_{x,y} u\|^2_{L^2(\tilde{T})},$$

there then follows (8.53). ∎

From Lemma 8.56 follows Theorem 8.57 as the main result. In (8.54) we use $H^2(\Omega) \cap V$. For $V = H^1(\Omega)$ this is $H^2(\Omega)$, only if $V = H^1_0(\Omega)$ the space $H^2(\Omega) \cap H^1_0(\Omega)$ carries additional zero boundary conditions.

Theorem 8.57. *Assume that conditions (8.49a–c) hold for $\mathcal{T}$, V_N, and V. Let $\alpha_0 > 0$ be the smallest interior angle of all $T \in \mathcal{T}$, while h is the maximum length of the sides of all $T \in \mathcal{T}$. Then*

$$\inf_{v \in V_N} \|u - v\|_{H^k(\Omega)} \le C'(\alpha_0) h^{2-k} \|u\|_{H^2(\Omega)} \quad \begin{cases} \text{for } k \in \{0,1,2\} \text{ and} \\ \text{all } u \in H^2(\Omega) \cap V. \end{cases} \tag{8.54}$$

Proof. For a $u \in H^2(\Omega)$ [resp. $u \in H^2(\Omega) \cap H_0^1(\Omega)$, if $V = H_0^1(\Omega)$] one chooses $v := \sum_i u(\mathbf{x}^i) b_i \in V_N$, i.e., $v \in V_N$ with $v(\mathbf{x}^i) = u(\mathbf{x}^i)$ at the [inner] nodes $\mathbf{x}^i$. Since $w := u - v$ vanishes at the vertices of each $T_i \in \mathcal{T}$ and $D^\alpha w = D^\alpha u$ for $|\alpha| = 2$ (because $D^\alpha v = 0$ for all linear(!) functions $v \in V_N$), inequality (8.53) implies the estimate

$$\begin{aligned} \left\|D^\beta w\right\|^2_{L^2(\Omega)} &= \sum_{T_i \in \mathcal{T}} \left\|D^\beta w\right\|^2_{L^2(T_i)} \underset{(8.53)}{\le} \sum_{T_i \in \mathcal{T}} C(\alpha_0) h^{4-2|\beta|} \sum_{|\alpha|=2} \|D^\alpha u\|^2_{L^2(T_i)} \\ &= C(\alpha_0) h^{4-2|\beta|} \sum_{|\alpha|=2} \|D^\alpha u\|^2_{L^2(\Omega)} \le C(\alpha_0) h^{4-2|\beta|} \|u\|^2_{H^2(\Omega)} . \end{aligned}$$

Summation over $|\beta| \le k$ now proves the inequality (8.54). ∎

The interpolation $v(\mathbf{x}^i) = u(\mathbf{x}^i)$ in the proof above only makes sense for continuous functions. Since the Sobolev embedding $H^2(\Omega) \subset C^0(\overline{\Omega})$ holds for the (physically realistic) dimensions $d \le 3$ (cf. Theorem 6.48), the statement of Theorem 8.57 can also be extended to piecewise linear functions on tetrahedra. However, for inequality (8.54) in $d \ge 4$ dimensions or a weaker norm than $\|\cdot\|_{H^2(\Omega)}$ on the right-hand side of (8.54), one may apply the interpolation by the Clément operator (cf. Clément [72], Brezzi–Fortin [55, §III.2, Proposition 2.2], and Braess [45, §II.6.9]).

8.5.2 Properties of Sequences of Finite-Element Spaces

8.5.2.1 Uniform Triangulations, Regular Triangulations, and K-Grids

For a discussion of the convergence of finite-element solutions one needs a sequence of finite-element spaces. In the case of difference schemes the step size h is the essential parameter which approaches zero. In the case of finite-element spaces the dimension N is only a partial information. The smallest interior angle is another important quantity.

Let $\{\mathcal{T}_\nu : \nu \in \mathbb{N}\}$ be a sequence of triangulations (here used in the true sense: $\mathcal{T}_\nu$ contains triangles $T \subset \Omega \subset \mathbb{R}^2$). The corresponding finite-element spaces V_{N_ν} are indexed by their dimension $N_\nu \to \infty$. For each $T \in \mathcal{T}_\nu$ let

$$\rho_{T,\text{outer}} := \inf\{\rho : T \subset K_\rho(\mathbf{x}),\ \rho > 0,\ \mathbf{x} \in \mathbb{R}^2\}$$

be the radius of the outer circle of T. Correspondingly let

$$\rho_{T,\text{inner}} := \sup\{\rho : T \supset K_\rho(\mathbf{x}),\ \rho > 0,\ \mathbf{x} \in \mathbb{R}^2\}$$

be the radius of the inner circle. Further let h_T be the longest side of $T \in \mathcal{T}_\nu$, then $h_{\max}(\mathcal{T}_\nu)$ is the longest side-length which occurs in $\mathcal{T}_\nu$:

$$h_T := \operatorname{diam}(T) \quad \text{and} \quad \begin{cases} h_{\max,\nu} := h_{\max}(\mathcal{T}_\nu) := \max\{h_T : T \in \mathcal{T}_\nu\}, \\ h_{\min,\nu} := h_{\min}(\mathcal{T}_\nu) := \min\{h_T : T \in \mathcal{T}_\nu\}. \end{cases} \tag{8.55}$$

Obviously we have $\rho_{T,\text{inner}} \le h_T \le \rho_{T,\text{outer}}$ for all $T \in \mathcal{T}_\nu$.

We define the following properties of the sequence $\{\mathcal{T}_\nu\}_{\nu\in\mathbb{N}}$:

- $\{\mathcal{T}_\nu\}_{\nu\in\mathbb{N}}$ is called *uniform*, if $\sup\{h_{\max,\nu}/h_{\min,\nu} : \nu \in \mathbb{N}\} < \infty$.
- $\{\mathcal{T}_\nu\}_{\nu\in\mathbb{N}}$ is called *shape regular* (or, simply, *regular*) if

$$\sup\left\{\frac{\rho_{T,\text{outer}}}{\rho_{T,\text{inner}}} : T \in \mathcal{T}_\nu,\ \nu \in \mathbb{N}\right\} < \infty.$$

- A triangulation $\mathcal{T}$ has the *K-grid property for some $K > 0$*, if the inequality

$$h_{\max,T_1} \le K h_{\max,T_2}$$

holds for all pairs of neighboured triangles $T_1, T_2 \in \mathcal{T}$ ($T_1 \cap T_2 \neq \emptyset$). A sequence $\{\mathcal{T}_\nu\}_{\nu\in\mathbb{N}}$ satisfies the K-grid property, if all $\mathcal{T}_\nu$ are K-grids with the same constant K. If $\{\mathcal{T}_\nu\}_{\nu\in\mathbb{N}}$ has the K-grid property for some K, we call it *quasi-uniform*.[23]

Remark 8.58. (a) A uniform triangulation has the K-grid property with $K := \sup\{h_{\max,\nu}/h_{\min,\nu} : \nu \in \mathbb{N}\}$.
(b) A shape regular triangulation is quasi-uniform.[24]

In the case of a uniform sequence $\{\mathcal{T}_\nu\}_{\nu\in\mathbb{N}}$ all triangles $T \in \mathcal{T}_\nu$ have a diameter deviating from h_ν only by a fixed factor $C := \sup\{h_{\max,\nu}/h_{\min,\nu} : \nu \in \mathbb{N}\}$. Local refinements as in Figure 8.2 (right) are only possible to a limited extent.

On the other hand, local refinements do not necessarily deteriorate the shape regularity. In the right part of Figure 8.2 the inequality $\frac{\rho_{T,\text{outer}}}{\rho_{T,\text{inner}}} \le 2$ is also valid after further refinements. In §8.7.3.3 we discuss the connection between shape regularity and K-grid property.

We can also characterise the shape regularity by the smallest inner angle α_T of the triangles T instead by the inner and outer circles: $\inf\{\alpha_T : T \in \mathcal{T}_\nu,\ \nu \in \mathbb{N}\} > 0$. The latter condition is used in (8.52).

[23] In the *local* view, i.e., considering a triangle and a fixed number of neighbouring triangles, the triangulation is uniform.

[24] However, in the one-dimensional case of intervals instead of triangles, shape regularity always holds, but does not imply the K-grid property and therefore not quasi-uniformity.

The estimate in (8.54) uses the maximum grid size h. This approach is appropriate for uniform grids. However, in the quasi-uniform case with a large ratio $h_{\max}/h_{\min}$ the inequality (8.54) cannot express the better local behaviour of the small elements. The estimates (8.56) use the individual size $h_{\tilde{T}}$ of each triangle $\tilde{T}$. Instead enlarging $h_{\tilde{T}}$ to $h_{\max}$ and then summing over all triangles, one can use $h_{\tilde{T}}$ as local factor and replace $h^s \|u\|_{H^t(\Omega)}$ by $\|h^s u\|_{H^t(\Omega)}$. If $t \leq 1$ let h be the piecewise linear function whose value at a node $\mathbf{x}$ equals $\max\{h_{\tilde{T}} : \tilde{T} \text{ contains the vertex } \mathbf{x}\}$. In the K-grid case we have $\|\nabla h\|_\infty \leq K$. Estimates of the weighted expressions $h^s u$ are, e.g., studied in Dahmen et al. [81].

8.5.2.2 Maximum Angle Condition

The condition "$\alpha_T \geq \alpha_0 > 0$" in (8.52) is stronger than necessary. Since the sum of all inner angles is π, (8.52) implies that inner angles are bounded below $\pi - \alpha_0$:

$$\alpha_{T,\max} := \text{maximum inner angle of } T \;\leq\; \bar{\alpha} \leq \pi - \alpha_0 < \pi. \tag{8.56}$$

The reverse implication is not true: The angles $\alpha_1 = \pi/2$, $\alpha_2 = \pi/2 - \varepsilon$, $\alpha_3 = \varepsilon$ of a flat, right-angled triangle satisfy (8.56) with $\alpha_{T,\max} = \pi/2 < \pi$ although α_3 can be arbitrarily small. A more detailed estimate of the transformations in Lemma 8.56 shows that already the weaker maximum angle condition (8.56) is sufficient to prove an estimate (8.54) in which $C'(\alpha_0)$ is replaced by a function $C(\bar{\alpha})$ depending on $\bar{\alpha}$ only (cf. Knabner–Angermann [172, §3.4.2]).

8.5.3 H^1-Estimates for Linear Elements

Let h_T be the longest side of $T \in \mathcal{T}$ according to (8.55). $h := \max\{h_T : T \in \mathcal{T}\}$ is the longest side-length which occurs in $\mathcal{T}$. The parameter h is the essential one and will be used as an index in the sequel: $\mathcal{T} = \mathcal{T}_h$.

When constructing the triangulations one must take care that the shape regularity of $\mathcal{T}_{h_\nu}$ is preserved while $h_\nu \to 0$. Strategies for the systematic construction of quasi-uniform sequences will be discussed in §8.7.3.3.

Theorem 8.57 yields the following result for $k = 1$.

Theorem 8.59. *Assume conditions (8.49a–c) hold for a sequence of quasi-uniform triangulations $\mathcal{T}_\nu$. Then there exists a constant C, such that for all $h = h_\nu$ and $V_h = V_{h_\nu}$ there holds the following estimate:*

$$\inf_{v \in V_h} |u - v|_1 \leq C h \, |u|_2 \qquad \textit{for all } u \in H^2(\Omega) \cap V. \tag{8.54$'$}$$

Combining this theorem with Theorem 8.21 gives the next statement.

Theorem 8.60. *Assume conditions (8.49a–c) hold for a sequence of quasi-uniform triangulations $\mathcal{T}_\nu$. Let the bilinear form fulfil conditions (8.2) and (8.17a). Suppose the constants $\varepsilon_{h_\nu} > 0$ from (8.17a) are bounded below by $\varepsilon_{h_\nu} \geq \underline{\varepsilon} > 0$ (cf. Corollary 8.25). Let problem (8.1) have the solution $u \in H^2(\Omega) \cap V$. Let $u^h \in V_h$ be the finite-element solution. Then there exists a constant C, which does not depend on u, $h = h_\nu$, or ν, such that*

$$\left|u^h - u\right|_1 \leq Ch\,|u|_2\,. \tag{8.57}$$

Combining the Theorems 8.59 and 8.24 yields the following.

Theorem 8.61. *Assume conditions (8.49a–c) hold for a sequence of quasi-uniform triangulations $\mathcal{T}_\nu$ with $h_\nu \to 0$. Let the bilinear form fulfil (8.2) and (8.17a) with $\varepsilon_{h_\nu} \geq \underline{\varepsilon} > 0$. Then problem (8.1) has a unique solution $u \in V$, and the finite-element solution $u^{h_\nu} \in V_h$ converges to u:*

$$\left|u^{h_\nu} - u\right|_1 \to 0 \qquad (\nu \to \infty).$$

Proof. Let $u \in V$ and $\varepsilon > 0$ be arbitrary. Since $H^2(\Omega) \cap V$ is dense in V, there exists $u_\varepsilon \in H^2(\Omega) \cap V$ with $|u - u_\varepsilon|_1 \leq \varepsilon/2$. From (8.54) and $h_\nu \to 0$ we see there are ν and $v_\varepsilon \in V_{h_\nu}$ with $|u_\varepsilon - v_\varepsilon|_1 \leq \varepsilon/2$; thus $|u - v_\varepsilon|_1 \leq \varepsilon$. This proves (8.24a). ∎

Theorem 8.61 proves convergence without restrictive conditions on u. On the other hand the order of convergence $\mathcal{O}(h)$ in the estimate (8.57) requires the assumption that the solution $u \in V$ lies in $H^2(\Omega)$. As will later become clear, this assumption is hardly to be taken as fulfilled in every situation. A weaker assumption is $u \in V \cap H^s(\Omega)$ with $s \in (1, 2)$. The corresponding result is as follows.

Theorem 8.62. *Assume that in the assumptions of Theorem 8.60 $u \in V \cap H^2(\Omega)$ is replaced by $u \in V \cap H^s(\Omega)$ with $s \in [1, 2]$. Let Ω be sufficiently smooth. Then there holds*

$$\left|u^h - u\right|_1 \leq Ch^{s-1}\,|u|_s\,. \tag{8.58}$$

The proof uses a generalisation of Theorem 8.57 with $k = 1$.

Lemma 8.63. *Under the assumptions of Theorem 8.57 and suitable conditions on Ω, there holds*

$$\inf_{v \in V_h} |u - v|_1 \leq C''(\alpha_0) h^{s-1} \|u\|_{H^s(\Omega)} \qquad \begin{cases} \text{for all } s \in [1, 2] \\ \text{and } u \in H^s(\Omega) \cap V. \end{cases} \tag{8.59}$$

The proof is based on an interpolation argument that will not be further explained here. Equation (8.59) holds for $s = 2$ (cf. (8.57)) and for $s = 1$, since $\inf |u - v|_1 \leq |u - 0|_1 = |u|_1$. From this follows (8.59) for $s \in (1, 2)$ with the norm $|\cdot|_s$ of the interpolating space $[H^1(\Omega) \cap V, H^2(\Omega) \cap V]_{s-1}$ (cf. Lions–Magenes [194]), which under suitable assumptions on Ω coincides with $H^s(\Omega) \cap V$.

For $s=1$ the right-hand side of (8.59) becomes $\text{const} \cdot |u|_1$. In fact the estimate $\inf\{|u-v|_1 : v \in V_h\} \leq |u|_1$ is the best possible. To prove this, choose $u \perp V_h$ (orthogonal with respect to $|\cdot|_1$). On the other side $s=2$ is the maximal value for which the estimates (8.58) and (8.59) can hold. Even $u \in C^\infty(\Omega)$ permits no better order of approximation than $\mathcal{O}(h)$!

The Ritz projections $S_N : V \to V_N$ introduced in (8.25) will now be written $S_h : V \to V_h$. The inequality (8.58) becomes

$$|u - S_h u|_1 / |u|_s \leq C h^{s-1}$$

and this proves the Conclusion 8.64.

Conclusion 8.64. *Assume conditions (8.49a–c) for $\mathcal{T}$, V_h, and V. Let the bilinear form fulfil (8.2) and (8.17a). The Ritz projection S_h satisfies the estimate*

$$\|I - S_h\|_{H^1(\Omega) \leftarrow H^2(\Omega) \cap V} \leq Ch. \tag{8.60}$$

Under the assumptions of Lemma 8.63 there holds

$$\|I - S_h\|_{H^1(\Omega) \leftarrow H^s(\Omega) \cap V} \leq Ch^{s-1} \qquad \textit{for all } s \in [1,2]\,. \tag{8.60$'$}$$

8.5.4 L^2 Estimates for Linear Elements

According to Theorem 8.60, $\mathcal{O}(h)$ is the optimal order of convergence. This result seems to contradict the $\mathcal{O}(h^2)$ convergence of the five-point formula (cf. Section 4.5), since the finite-element method with a particular triangulation is almost identical to the five-point formula (of. Exercise 8.42). The reason for this is that the finite-element error $u^h - u$ is measured in the $|\cdot|_1$ norm. The estimate

$$\inf\{|u-v|_0 : v \in V_h\} \leq Ch^2 |u|_2$$

in Theorem 8.57 suggests the conjecture that the $|\cdot|_0$ norm of the error is of the order $\mathcal{O}(h^2)$: $|u-u_h|_0 \leq Ch^2\,|u|_2$. However, this statement is false without further assumptions.

Up to this point only the existence of a weak solution $u \in V$ (e.g., $V = H_0^1(\Omega)$ or $V = H^1(\Omega)$) has been guaranteed. But in Theorem 8.60 we needed the stronger assumption $u \in H^2(\Omega) \cap V$. A similar regularity condition will also be imposed on the *adjoint problem* to (8.1):

$$\text{find} \quad u \in V \quad \text{with} \quad a^\star(u,v) = f(v) \qquad \text{for all } v \in V, \tag{8.61}$$

which uses the adjoint bilinear form $a^\star(u,v) := a(v,u)$. For $f \in L^2(\Omega) \subset V'$, the value $f(v)$ becomes $(f,v)_{L^2(\Omega)}$. The H^2 *regularity condition* is:

$$\begin{aligned}&\text{For each } f \in L^2(\Omega) \text{ the problem (8.61)}\\ &\text{has a solution } u \in H^2(\Omega) \cap V \text{ with } |u|_2 \le C_R\,|f|_0\,.\end{aligned} \tag{8.62}$$

In Chapter 9 we shall see that this statement holds for sufficiently smooth domains.

The following statement is called the *Lemma of Aubin–Nitsche* and is independently described in the articles of Aubin [11], Nitsche [212], and Oganesjan–Ruchovets [213].

Theorem 8.65. *Assume (8.62), (8.2), (8.17a) with* $\varepsilon_N = \varepsilon_h \ge \underline{\varepsilon} > 0$, *and*

$$\inf_{v \in V_h} |u - v|_1 \le C_0 h\,|u|_2 \qquad \textit{for all } u \in H^2(\Omega) \cap V. \tag{8.63}$$

Let problem (8.1) have the solution $u \in V$. *Let* $u^h \in V_h \subset V$ *be the finite-element solution. Then, with a constant* C_1 *independent of* u *and* h,

$$\left|u^h - u\right|_0 \le C_1 h\,|u|_1\,. \tag{8.64a}$$

If the solution u *also belongs to* $H^2(\Omega) \cap V$, *then there is a constant* C_2, *independent of* u *and* h, *such that*

$$\left|u^h - u\right|_0 \le C_2 h^2\,|u|_2\,. \tag{8.64b}$$

From Theorem 8.57, it is sufficient to ensure (8.63) that V_h *be the space of finite elements of an admissible and quasi-uniform triangulation.*

Proof. For each $e := u^h - u \in L^2(\Omega)$ define $w \in H^2(\Omega) \cap V$ as the solution of (8.61) for $f := e$:

$$a(v, w) = (e, v)_{L^2(\Omega)} \qquad \text{for all } v \in V. \tag{8.65a}$$

Corresponding to w there is, by (8.63), a $w^h \in V_h$ with

$$\left|w^h - w\right|_1 \le C_0 h\,|w|_2 \underset{(8.62)}{\le} C_0 C_R h\,|e|_0\,. \tag{8.65b}$$

Equation (8.22) is $a(e, v) = 0$ for all $v \in V_h$; therefore, in particular, we have

$$a(e, w^h) = 0. \tag{8.65c}$$

From (8.65a–c) we obtain

$$\begin{aligned}|e|_0^2 = (e,e)_{L^2(\Omega)} &\underset{(8.65a)}{=} a(e,w) \underset{(8.65c)}{=} a(e, w - w^h) \underset{(8.2)}{\le} C_S\,|e|_1\left|w^h - w\right|_1\\ &\underset{(8.65b)}{\le} C_S\,|e|_1\,C_0 C_R h\,|e|_0\end{aligned}$$

and after division by $|e|_0$:

$$|e|_0 \le C_S C_0 C_R h\,|e|_1\,.$$

From (8.20) one deduces

$$|e|_1 = \left\|u^h - u\right\|_V \le \left(1 + \frac{C_S}{\varepsilon_N}\right) \inf_{v \in V_N} \|u - v\|_V \le \left(1 + \frac{C_S}{\underline{\varepsilon}}\right) \|u\|_V$$

so that (8.64a) follows with $C_1 := C_S C_0 C_R (1 + C_S/\underline{\varepsilon})$. If $u \in H^2(\Omega) \cap V$, one may use the inequality (8.57), $|u^h - u|_1 \le Ch|u|_2$, and deduce (8.64b) with $C_2 := C_S C_0 C_R C$. ∎

Corollary 8.66. The inequalities (8.64a) and (8.64b) are equivalent to the respective properties (8.66) of the Ritz projection S_h:

$$\|I - S_h\|_{L^2(\Omega) \leftarrow V} \le C_1 h, \qquad \|I - S_h\|_{L^2(\Omega) \leftarrow H^2(\Omega) \cap V} \le C_2 h^2. \tag{8.66}$$

The estimates (8.66) may also be proved directly. The definition of $\hat{S}_h$ and the connection between S_h and $\hat{S}_h$ may be found in the next exercise.

Exercise 8.67. To $a(\cdot,\cdot)$ let the operator $L : V \to V'$, the Ritz projection S_h, and the system matrix $\mathbf{L}$ be associated. Show:
(a) To the adjoint bilinear form $a^\star(\cdot,\cdot)$ belong the system matrix $\mathbf{L}^\mathsf{T}$ and the Ritz projection

$$\hat{S}_h = P \left(\mathbf{L}^\mathsf{T}\right)^{-1} P^* L^*.$$

(b) There holds

$$S_h = L^{-1} \hat{S}_h^* L\,. \tag{8.67}$$

Second proof of Theorem 8.65. Let

$$H^2_\star(\Omega) := \{u \in H^2(\Omega) : u \text{ is a solution of (8.61) for an } f \in L^2(\Omega)\} \subset V\}$$

be the range of $L^{\star -1} \in L(L^2(\Omega), H^2(\Omega))$. Equip the space $H^2_\star(\Omega)$ with the norm $|\cdot|_2$. Since $L^{-1} \in L(H^2_\star(\Omega)', L^2(\Omega))$ we have that $L^{\star -1} \in L(L^2(\Omega), H^2_\star(\Omega))$ is equivalent to $L^{-1} \in L(H^2_\star(\Omega)', L^2(\Omega))$ (cf. Lemma 6.64), so that there is a C_α with

$$\|L^{-1}\|_{L^2(\Omega) \leftarrow H^2_\star(\Omega)'} \le C_\alpha. \tag{8.68a}$$

The assumptions (8.2) and (8.17a) of Theorem 8.65 can be brought over without change of the constants to the adjoint problem, so that the statement of Theorem 8.21 can be written in the form $\|I - \hat{S}_h\|_{V \leftarrow H^2_\star(\Omega)} \le C_\beta h$. For the adjoint operator we have the estimate

$$\|(I - \hat{S}_h)^*\|_{H^2_\star(\Omega)' \leftarrow V'} \le C_\beta h. \tag{8.68b}$$

The basic assumption (8.2) may be written

$$\|L\|_{V' \leftarrow V} \le C_S. \tag{8.68c}$$

From equation (8.67) one may read the expression

$$I - S_h = I - L^{-1} \hat{S}_h^\star L = L^{-1}(I - \hat{S}_h^\star) L = L^{-1}(I - \hat{S}_h)^\star L.$$

From (8.68a–c) one has

$$\|I - S_h\|_{L^2(\Omega)\leftarrow V} \leq \|L^{-1}\|_{L^2(\Omega)\leftarrow H^2_\star(\Omega)'} \|(I - \hat{S}_h)^*\|_{H^2_\star(\Omega)'\leftarrow V'} \|L\|_{V'\leftarrow V}$$
$$\leq C_\alpha C_\beta h C_S \,. \tag{8.68d}$$

Therefore we have proved the first inequality in (8.66) with $C_1 = C_\alpha C_\beta C_S$. Since S_h is a projection, so is $I - S_h$, so that

$$\|I - S_h\|_{L^2(\Omega)\leftarrow H^2(\Omega)\cap V} = \| \left(I - S_h\right)^2 \|_{L^2(\Omega)\leftarrow H^2(\Omega)\cap V}$$
$$\leq \|I - S_h\|_{L^2(\Omega)\leftarrow V} \|I - S_h\|_{V\leftarrow H^2(\Omega)\cap V}.$$

From the first inequality in (8.66) and (8.60) we deduce the second inequality in (8.66). ∎

The regularity condition (8.62) is weakened in the following theorem.

Theorem 8.68. *Assume (8.2), and (8.17a) with $\varepsilon_N = \varepsilon_h \geq \tilde{\varepsilon} > 0$, and inequality (8.59) for all $1 \leq s \leq 2$. In the place of (8.62) we assume H^{2-t}-regularity for some $t \in [0,1]$. If $V = H^1(\Omega)$ then assume $L^{\star -1} \in L(H^t(\Omega)', H^{2-t}(\Omega))$; if $V = H^1_0(\Omega)$, then assume $L^{\star -1} \in L(H^{-t}(\Omega), H^{2-t}(\Omega) \cap V)$. Then the finite-element solution satisfies the inequality*

$$\left|u^h - u\right|_t \leq C_{t,s} h^{s-t} \left|u\right|_s \qquad (0 \leq t \leq 1 \leq s \leq 2)\,. \tag{8.69}$$

The Ritz projection satisfies

$$\|I - S_h\|_{H^t(\Omega)\leftarrow H^s(\Omega)\cap V} \leq C_{t,s} h^{s-t} \qquad (0 \leq t \leq 1 \leq s \leq 2)\,. \tag{8.70}$$

Proof. For simplicity we only consider the case $V = H^1(\Omega)$. In (8.68a–c) replace $L^2(\Omega)$ by $H^t(\Omega)$, and $H^2_\star(\Omega)'$ by $H^{2-t}_\star(\Omega)'$ with

$$H^{2-t}_\star(\Omega) := \{u \in H^{2-t}(\Omega) :\ u \text{ is the solution to } f \in H^t(\Omega)'\}.$$

Because of (8.60′), (8.68b) becomes [with s replaced by $2 - t$]

$$\|(I - \hat{S}_h)^*\|_{H^{2-t}_\star(\Omega)'\leftarrow V'} = \|I - S_h\|_{V\leftarrow H^{2-t}_\star(\Omega)} \leq C_\beta h^{1-t}.$$

As in (8.68d) one shows that $\|I - S_h\|_{H^t(\Omega)\leftarrow V} \leq C h^{1-t}$. Combining this with $\|I - S_h\|_{V\leftarrow H^s(\Omega)\cap V} \leq C h^{s-1}$ (cf. (8.60′)) there follows (8.70), and thus (8.69). ∎

For nonlinear boundary-value problems, estimates of the errors in terms of other norms (e.g., $L^\infty(\Omega)$, $L^p(\Omega)$) are of interest. For this we refer to Ciarlet [67, §3.3], Schatz [253], and Schatz–Wahlbin [254, 255, 256].

8.6 Generalisations

8.6.1 Error Estimates for Other Elements

The estimates of the errors for linear functions on triangular elements proved in Section 8.5 are also true for bilinear functions on parallelograms and for combinations of both sorts of elements (cf. Figure 8.7). The proofs are along analogous lines. Even for tetrahedral elements in a three-dimensional region $\Omega \subset \mathbb{R}^3$ the same results can be carried over. For the proof one notices that $u \in H^2(\Omega)$ has well-defined nodal values $u(\mathbf{x}^i)$ even for $\Omega \subset \mathbb{R}^3$, since $2 > n/2$ ($n = 3$, dimension of Ω; cf. Theorem 6.48).

For quadratic elements (cf. Section 8.4.4) one expects a correspondingly improved order of convergence. In general one can show the following: If the ansatz function is, in each $T \in \mathcal{T}$, a polynomial of degree $k \geq 1$ (i.e., $u(\mathbf{x}) = \sum_{|\nu|\leq k} \alpha_\nu \mathbf{x}^\nu$), then

$$d(u, V_h) := \inf \{|u - v|_1 : v \in V_h\} \leq Ch^k \, |u|_{k+1} \quad \text{for all } u \in H^{k+1}(\Omega) \cap V \tag{8.71}$$

(cf. Ciarlet [67, Theorem 3.2.1]). If one uses parallelograms as a basis and uses ansatz functions that are polynomials of degree at least k, then (8.71) also holds. For example, biquadratic elements and quadratic ansatz functions of the serendipity class fulfil this requirement for $k = 2$. The result corresponding to Theorem 8.60 follows from Theorem 8.21.

Theorem 8.69. *Let V_h fulfil (8.71), and let the bilinear form satisfy (8.2) and (8.17a) with $\varepsilon_N =: \varepsilon_h \geq \underline{\varepsilon} > 0$. Assume problem (8.1) has a solution $u \in V \cap H^{k+1}(\Omega)$. Then the finite-element solution $u^h \in V_h$ satisfies the inequality*

$$\left|u^h - u\right|_1 \leq C \, h^k \, |u|_{k+1} \, .$$

The Ritz projection satisfies $\|I - S_h\|_{V \leftarrow H^{k+1}(\Omega)\cap V} \leq Ch^k$.

The estimate (8.58) now holds for $s \in [1, k]$ if $u \in H^s(\Omega) \cap V$. With suitable regularity conditions, there are, as in Theorem 8.65, the error estimates

$$\left|u^h - u\right|_0 \leq Ch^{k+1} \, |u|_{k+1} \, , \quad \left\|u^h - u\right\|_{H^{k+1}(\Omega)'} \leq Ch^{2k} \, |u|_{k+1} \, . \tag{8.72}$$

For (8.72) one needs, for instance, H^{k+1}-regularity: For each $f \in H^{k-1}(\Omega)$ the adjoint problem (8.61) has a solution $u \in H^{k+1}(\Omega)$.

Remark 8.70. Under suitable assumptions the inequality (8.69) holds for $1 - k \leq t \leq 1 \leq s \leq k + 1$. For negative t the norm $|\cdot|_t$ should be understood as the dual norm of $H^{-t}(\Omega)$.

8.6.2 Finite Elements for Equations of Higher Order

8.6.2.1 Introduction: The One-Dimensional Biharmonic Equation

All the spaces of finite elements, V_h, constructed so far are useless for equations of order $2m > 2$, since $V_h \not\subset H^m(\Omega)$. According to Example 6.22, in order to have $V_h \subset H^2(\Omega)$ it is necessary that not only the function u but also its derivatives u_{x_i} $(1 \le i \le n)$ change continuously between elements. The ansatz functions must therefore be piecewise smooth and globally in $C^1(\overline{\Omega})$.

As a model problem we introduce the one-dimensional biharmonic equation

$$u''''(x) = g(x) \quad \text{for } 0 < x < 1, \qquad u(0) = u'(0) = u(1) = u'(1) = 0$$

which becomes in its weak formulation

$$\begin{aligned} &a(u,v) = f(v) \qquad \text{for all } v \in H_0^2(0,1), \quad \text{where} \\ &a(u,v) := \int_0^1 u''v''\,\mathrm{d}x, \quad f(v) := \int_0^1 gv\,\mathrm{d}x. \end{aligned} \tag{8.73}$$

Divide the interval $\Omega = (0,1)$ into equal subintervals of length h. Piecewise linear functions (cf. Figure 8.1) can be viewed as linear spline functions, so that it is natural to define V_h as the space of *cubic splines* (with $u = u' = 0$ for $x = 0$ and $x = 1$) (cf. Stoer [274, §2.4] and Quarteroni et al. [230, §8.7.1]). We can take as basis functions B-splines, whose supports in general consist of four subintervals (cf. Figure 8.9a). Since cubic spline functions belong to $C^2(0,1)$, they are not just in $H_0^2(0,1)$, but even in $H^3(0,1) \cap H_0^2(0,1)$.

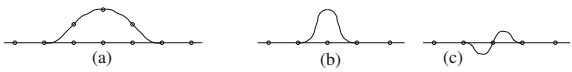


Fig. 8.9 (a) B-spline, (b,c) Hermite basis functions.

Simpler yet than working with spline functions is to use *cubic Hermite interpolation*:

$$V_h := \left\{ u \in C^1(0,1) : \begin{array}{l} u \text{ is cubic on each subinterval,} \\ u(0) = u'(0) = u(1) = u'(1) = 0 \end{array} \right\}. \tag{8.74}$$

To each of the inner nodes $x_j = jh$ there are two basis functions b_{1i} and b_{2i} with $b_{1i}(x_i) = 1$, $b'_{1i}(x_i) = 0$, $b_{2i}(x_i) = 0$, $b'_{2i}(x_i) = 1$, $b_{ki}(x_j) = b'_{ki}(x_j) = 0$ for $k = 1, 2$, $j \ne i$ (cf. Figure 8.9b,c). The support consist of only two subintervals. The expressions are

$$\begin{aligned} b_{1i}(x) &= (h - |x - x_i|)^2 (h + 2|x - x_i|)/h^3 && \text{for } x_{i-1} \le x \le x_{i+1}, \\ b_{2i}(x) &= (h - |x - x_i|)^2 (x - x_i)/h^2 && \text{for } x_{i-1} \le x \le x_{i+1}, \\ b_{1i}(x) &= b_{2i}(x) = 0 && \text{for } x \notin [x_{i-1}, x_{i+1}]. \end{aligned}$$

Exercise 8.71. Let u_{1i} and u_{2i} $(0 < i < 1/h)$ be the coefficients of the expression

$$u^h = \sum_{i=1}^{1/h-1} [u_{1i} b_{1i} + u_{2i} b_{2i}] \in V_h .$$

Show that the coefficients of the finite-element solution of the problem (8.73) are given by the equations

$$\begin{aligned} \frac{-12u_{1,i-1} + 24u_{1,i} - 12u_{1,i+1}}{h^3} + \frac{-6u_{2,i-1} + 6u_{2,i+1}}{h^2} &= f_{1,i}, \\ \frac{6u_{1,i-1} - 6u_{1,i+1}}{h^2} + \frac{2u_{2,i-1} + 8u_{2,i} + 2u_{2,i+1}}{h} &= f_{2,i} \end{aligned} \tag{8.75}$$

where $f_{1i} := \int_{x_{i-1}}^{x_{i+1}} b_{1i}\, g \,\mathrm{d}x$, $f_{2i} := \int_{x_{i-1}}^{x_{i+1}} b_{2i}\, g \,\mathrm{d}x$, $x_i = ih$.

The system of equations (8.75) differs completely from the difference equations (cf. §5.3.3), since in equation (8.75) there appear values u_{1i} of the functions at the nodes together with the values u_{2i} of the derivatives at the nodes.

8.6.2.2 The Two-Dimensional Case

The ansatz (8.74) can be carried over to $\Omega \subset \mathbb{R}^2$ if one starts with a partition into rectangular elements (as, e.g., in the left part of Figure 8.5 on page 206). The ansatz function is *bicubic*:

$$V_h := \left\{ u \in C^1(\overline{\Omega}) : \begin{array}{l} u = \frac{\partial u}{\partial n} = 0 \text{ on } \Gamma, \\ u = \sum\limits_{\nu,\mu=0}^{3} \alpha_{\nu\mu} x^\nu y^\mu \text{ on each rectangle of the partition} \end{array} \right\} .$$

At each inner node (corner $\mathbf{x} \in \Omega$ of the rectangular elements) one may prescribe the four values u, u_x, u_y, u_{xy}. Consequently there are four unknowns and four basis functions $b_{1i}(x,y), \ldots, b_{4i}(x,y)$ belonging to each node. The latter are products $b_{ji}(x) b_{ki}(y)$ $(j,k = 1,2)$ of the one-dimensional basis functions described in §8.6.2.1 (cf. Meis–Marcowitz [201, pages 312ff], Schwarz [262, §2.6.1]).

If one starts from a triangulation $\mathcal{T}$, a *fifth-order ansatz* gives what is wanted: $u(\mathbf{x}) = \sum_{|\nu|\le 5} \alpha_\nu \mathbf{x}^\nu$ on $T \in \mathcal{T}$. The number of degrees of freedom is 21 (the number of $\nu \in \mathbb{N}_0^2$ with $|\nu| \le 5$). For the nodes one chooses the points $\mathbf{x}^1, \ldots, \mathbf{x}^6$ in Figure 8.8a. At each vertex $\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3$ one prescribes 6 values $\{D^\mu u(\mathbf{x}^j) : |\mu| \le 2\}$. The 3 remaining degrees of freedom result from giving the normal derivatives $\partial u(\mathbf{x}^j)/\partial n$ at the midpoints of the sides $\mathbf{x}^4, \mathbf{x}^5, \mathbf{x}^6$. Notice that if two neighbouring triangles (T and $\hat{T}$ in Figure 8.8a) have the same vertex values $\{D^\mu u(\mathbf{x}^j) : |\mu| \le 2, j = 1,2\}$ and the same normal derivative at $\mathbf{x}^4$, then both u and ∇u are continuous on the shared side, i.e., $V_h \subset H^2(\Omega)$. According to whether

$V = H_0^2(\Omega)$, $V = H^2(\Omega) \cap H_0^1(\Omega)$, or $V = H^2(\Omega)$ one has to set u, or the first derivatives, to be zero at the boundary nodes $\mathbf{x}^i$.

Remark 8.72. The finite-element space $V_h \subset H^2(\Omega)$ described here can of course be used for differential equations of the order $2m = 2$.

8.6.2.3 Estimating Errors

Instead of (8.71) one obtains

$$\inf \{|u - v|_m : v \in V_h\} \leq Ch^{k+1-m} |u|_{k+1} \quad \text{for all } u \in H^{k+1}(\Omega) \cap V, \quad (8.76)$$

where $k \geq m$ depends on the order of the polynomial ansatz (e.g., $k = 3$ for cubic splines, cubic [resp. bicubic] Hermite interpolation). Here $m = 2$ for the biharmonic equation. As in Theorem 8.69, there follows from (8.76) the error estimate

$$|u^h - u|_m \leq Ch^{k+1-m} |u|_{k+1}$$

for the finite-element solution $u^h \in V_h$. Under suitable regularity assumptions one gets

$$\left|u^h - u\right|_t \leq C_{t,s} h^{s-t} |u|_s \qquad (2m - k - 1 \leq t \leq m \leq s \leq k + 1) \quad (8.77)$$

(cf. Remark 8.70). The maximal order of convergence $2(k - m) + 2$ results for $s = k + 1$, $t = 2m - k - 1$ and requires $u \in H^{k+1}(\Omega) \cap V$. In addition each solution of the adjoint problem (8.61) with $f \in H^{k+1-2m}(\Omega)$ must belong to $H^{k+1}(\Omega)$.

8.6.3 Finite Elements for Non-Polygonal Regions

Since the union of triangles and parallelograms only generates polygonal regions a polygonal shape was assumed in (8.34). The finite-element method is, however, in no way restricted to just such regions. On the contrary finite elements can readily be adapted to curved boundaries.

We discuss two approaches. The first one tries to exhaust Ω by usual finite elements. However, to obtain the usual error estimate, one has to use the *isoparametric elements*.

Let $V = H^1(\Omega)$ and let Ω be arbitrary. The triangulation $\mathcal{T}$ can be so chosen that (8.35a,b,d) hold and the 'outer' triangles have two vertices on $\Gamma = \partial\Omega$ as in Figure 8.10. In the convex case of Figure 8.10a one extends the linear functions defined on T onto $\tilde{T} := T \backslash B$. In the case that the boundary is concave (see Figure 8.10b) one should replace T by $\tilde{T} := T \backslash B$. One copes correspondingly in the

situation of Figure 8.10c. The nodes and the expressions for the basis functions remain undisturbed by these changes. $\overline{\Omega}$ is the union of the closures of all the inner triangles $T_i \in \mathcal{T}$ and all modified triangles $\tilde{T}$ which lie on the boundary. All the properties and results of Sections 8.4–8.5 extend to the new situation. The only difficulty is of a practical sort: To calculate $\mathbf{L}$ and $\mathbf{f}$ one has to work out integrals over the triangle $\tilde{T}$ which involves arcs.

Assume now $V = H_0^1(\Omega)$. The previous construction will not be a success, since the extensions of the linear functions to $\tilde{T}$ will not vanish on the boundary piece $\Gamma \cap \partial\tilde{T}$. Thus $V_h \subset H_0^1(\Omega)$ will not be satisfied. As long as the region Ω is convex, only the situation shown in Figure 8.10a occurs, and u^h may be extended to $B \subset \tilde{T}$ by $u^h = 0$. In the case of Figure 8.10b, one must set the values at the inner nodes to zero, so that $u^h = 0$ on $\tilde{T} = T \backslash B$, and in particular $u^h = 0$ on $\Gamma \cap \partial(T \backslash B)$. All in all, what results is that the support of any $u^h \in V_h$ is in a polygonal region inscribed in Ω. One interpretation is the following: In the boundary-value problem, Ω should be replaced by an approximating region $\Omega_h \subset \Omega$ (cf. Theorem 2.29). The finite-element solution described agrees with that which would result from the smaller region. However, the error estimate in Theorem 8.57 only holds for the smaller region Ω_h:

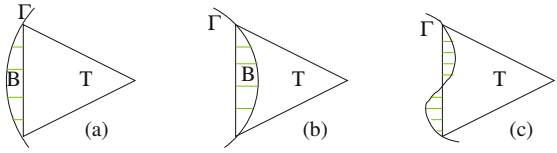


Fig. 8.10 Curved boundary.

$$\inf_{v \in V_h} \|u - v\|_{H^1(\Omega_h)} \leq Ch \, |u|_{H^2(\Omega)} .$$

Since $v = 0$ on $\Omega \backslash \Omega_h$, for any $v \in V_h$ one should also estimate $\|u\|_{H^1(\Omega \backslash \Omega_h)}$. Now $\Omega \backslash \Omega_h$ is contained in a strip

$$S_\delta = \{x \in \Omega : \operatorname{dist}(x, \Gamma) \leq \delta\}$$

of width

$$\delta := \max\{\operatorname{dist}(x, \Gamma) : x \in \Omega \backslash \Omega_h\}.$$

For $u \in H^2(\Omega) \cap H_0^1(\Omega)$ one can estimate as follows:

$$\inf_{v \in V_h} \|u - v\|_{H^1(\Omega \backslash \Omega_h)} = \|u\|_{H^1(\Omega \backslash \Omega_h)} \leq \|u\|_{H^1(S_\delta)} \leq C\sqrt{\delta} \, |u|_{H^2(\Omega)} .$$

If $\Omega \backslash \Omega_h$ consists only of arc segments B as in Figure 8.10a (for instance, in the case of a convex region), and if $\Omega \in C^{1,1}$, then $\delta = \mathcal{O}(h^2)$. From this follows the estimate $\inf_{v \in V_h} \|u - v\|_{H^1(\Omega)} \leq Ch^1 \, |u|_{H^2(\Omega)}$, as for a polygonal region. If, however, as in Figure 8.10b the whole triangle is part of $\Omega \backslash \Omega_h$, then δ becomes $\mathcal{O}(h)$ and the approximation worsens to $\inf_{v \in V_h} \|u - v\|_{H^1(\Omega)} \leq Ch^{1/2} \, |u|_{H^2(\Omega)}$ (cf. Strang–Fix [276, page 192]).

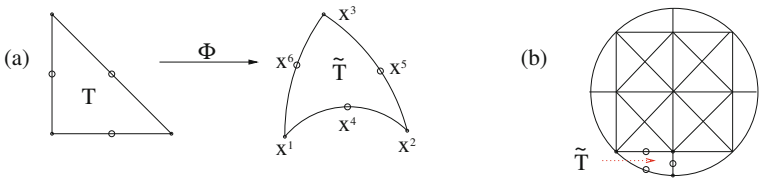


Fig. 8.11 (a) Mapping of the reference triangle T to the curvilinear triangle $\tilde{T}$; (b) isoparametric triangulation.

In order to adapt the triangular or parallelogram elements better to a curved boundary one can use the technique of *isoparametric finite elements*. From Figure 8.4 (page 205) we see that the reference triangle T may be mapped into an arbitrary triangle by an affine transformation $\Phi : (\xi, \eta) \mapsto \mathbf{x}^1 + \xi(\mathbf{x}^2 - \mathbf{x}^1) + \eta(\mathbf{x}^3 - \mathbf{x}^1)$. The linear [resp. in the case of §8.4.4, quadratic] function $u(\mathbf{x})$ can be expressed as the image, $v(\xi, \eta) = u(\Phi(\xi, \eta))$, of a linear [resp. quadratic] function $u(\mathbf{x})$ on $\tilde{T}$. We now replace the affine transformation of triangles Φ by a more general quadratic mapping

$$\Phi(\xi, \eta) := \begin{bmatrix} a_1 + a_2\xi + a_3\eta + a_4\xi^2 + a_5\xi\eta + a_6\eta^2 \\ a_7 + a_8\xi + a_9\eta + a_{10}\xi^2 + a_{11}\xi\eta + a_{12}\eta^2 \end{bmatrix} : \; T \longrightarrow \tilde{T} \subset \mathbb{R}^2.$$

The image $\tilde{T}$ is a curvilinear triangle. The coefficients $a_1, \ldots, a_{12}$ are uniquely determined by the nodes $\mathbf{x}^1, \ldots, \mathbf{x}^6$ of $\tilde{T}$ which are the images of the vertices and midpoints of the sides of the reference triangle T (cf. Figure 8.11a). The triangulation $\mathcal{T}$ used so far can be replaced by a "triangulation" by triangles with curved sides, if neighbouring elements have the same arcs as common boundaries and the midpoints also coincide. The ansatz functions on $\tilde{T} \in \tilde{\mathcal{T}}$ have the form $u(\mathbf{x}) = v(\Phi^{-1}(\mathbf{x}))$, where $v(\xi, \eta)$ is linear [resp. quadratic] in ξ and η. The resulting finite-element space is the space of isoparametric linear [resp. quadratic] elements (cf. Jung–Langer [164, §4.5.6], Knabner–Angermann [172, §3.8], Strang–Fix [276, page 192], Ciarlet [67, §4.3], Schwarz [262, §2.4], Wahlbin [298], Zienkiewicz [319]).

In general, there is no reason for using curvilinear triangles in the interior of Ω. As in Figure 8.11b, one chooses ordinary triangles in the interior (i.e., the quadratic transformation Φ is again taken to be linear). A boundary triangle, like $\tilde{T}$ in Figure 8.11b, is, on the other hand, defined as follows: $\mathbf{x}^1$ and $\mathbf{x}^2$ are the vertices of $\tilde{T}$ which lie on Γ. One chooses another boundary point $\mathbf{x}^4$ between $\mathbf{x}^1$ and $\mathbf{x}^2$ and requires that the boundary $\partial\tilde{T}$ of the triangle cuts the boundary Γ of the region at $\mathbf{x}^1$, $\mathbf{x}^4$, and $\mathbf{x}^2$.

Remark 8.73. Section 5.2.2 shows that, in the case of other than Dirichlet boundary conditions, the construction of difference schemes is increasingly complicated. Instead one may restrict the usual difference methods to the *inner* grid points, and near the boundary discretise by using (e.g., isoparametric) finite elements.

8.7 A-posteriori Error Estimates, Adaptivity

The following considerations are concerned with two different questions. First, having computed a finite-element solution, one wants to guess the quality (accuracy) of this approximation. This question will be discussed in §8.7.1. Second, the efficiency of the finite-element discretisation is of interest. This topic will be treated in §8.7.2. It will turn out that both tasks are connected; at least one tries to obtain efficiency with the help of a-posteriori error estimates.

8.7.1 A-posteriori Error Estimates

8.7.1.1 Criticism of the Previous Error Estimates

The previous error estimates for finite-element solutions (but also for the solutions of the difference methods) are determined *a priori*, i.e., without considering the information obtainable by the computed solution. These estimates are of the form $\|\text{error}\| \leq \alpha(h) \cdot |||u|||$ and describe the asymptotic behaviour as h tends to zero: $\|\text{error}\| \leq \mathcal{O}(\alpha(h))$. Moreover, this inequality (even for fixed h) indicates how the error depends on the smoothness of u expressed by the norm $||| \cdot |||$. Although these are interesting statements, other questions remain open.

• Usually the error bound is only described qualitatively, i.e., constants remain unknown. Therefore a concrete question like

$$\text{does} \quad \|\text{error}\| \leq 0.001 \quad \text{hold?} \tag{8.78}$$

cannot be answered. Even is the factor $\alpha(h)$ in $\|\text{error}\| \leq \alpha(h) \cdot |||u|||$ is known quantitatively, the following problem remains.

• The factor $|||u|||$ is unknown since it is the norm of a (at least before the computation, i.e., *a priori*) unknown function.

Generally all estimates from above, even (8.78), may be far too pessimistic (it would be unsatisfactory if we ensure that (8.78) holds, while $\|\text{error}\| = 10^{-6}$ is the true value). Finally, there is the question concerning the

• choice of the norms.

The norms used above are chosen since then the estimates are easy to prove. There are reasons why the Sobolev norms used so far are not optimal: The regularity analysis in §9.2 will show that solutions of $Lu = f$ (f smooth) are smoother in the interior of the domain (distant from the boundary) than close to the boundary. Therefore it would be appropriate to use norms of u which reflect this behaviour.

One can go a step further and ask whether at all norms are the right tool. In many applications one is not interested in the solution u as a function, but in certain point values or in local mean values, etc. In these cases the general formulation uses a (linear) functional $\phi(u)$. One wants to find the value $\phi(u)$ or a tuple of functionals $\phi_i(u)$. Correspondingly, the errors $\phi(u-u^h)$ or $\phi_i(u-u^h)$ are of interest (cf. Becker–Rannacher [34]).

8.7.1.2 Concept of A-posteriori Estimates

The term 'a posteriori' means that we study the error $e := u - u^h$ *after* the computation of u^h. In particular one can use the computed approximation u^h to get concrete information about the error.

A possible solution could be as follows. Let u be the solution of $Lu = f$ (e.g., with zero Dirichlet condition) and u^h the finite-element solution (with the same Dirichlet condition). The error satisfies the boundary-value problem

$$Le = L(u-u^h) = Lu - Lu^h = f - Lu^h =: r. \tag{8.79}$$

The right-hand side r (called the *defect* or *residual*) belongs to the dual space $H^{-1}(\Omega)$. If we succeed in obtaining an estimate $\|r\|_{-1} \le \varepsilon$ with a concrete value of ε, we can derive $\|e\|_1 \le \|L^{-1}\|_{1\leftarrow -1}\,\varepsilon$. Assuming that the constant C in $\|L^{-1}\|_{1\leftarrow -1} \le C$ is explicitly known, $\|e\|_1 \le C\varepsilon$ is the desired error estimate with a known bound on the right-hand side.

Unfortunately the above estimation of r by $\|r\|_{-1} \le \varepsilon$ is impossible in the strict sense: $\|r\|_{-1} = \sup_{v\in H_0^1(\Omega)} |\,(r,v)\,| / \|v\|_1$ cannot be obtained with finite computational work. The simple attempt to restrict the test functions $v \in H_0^1(\Omega)$ to the finite-element space V_h fails completely since r is perpendicular to V_h (cf. (8.22)). Even the computation (or estimation) of the L^2-norm of a right-hand side r by finitely many data is impossible. Any estimate of $\|r\|_{-1}$ requires a-priori assumptions about f and the coefficients of L. However, there may be means[25] to obtain $\|L^{-1}\|_{1\leftarrow -1} \le C$.

The next section explains the direct estimate of $\|e\|_1$ via r.

8.7.1.3 Example of a Residual Based Error Estimator

Assume the model case

$$-\Delta u = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \Gamma, \tag{8.80}$$

so that the variational formulation is given by $a(u,v) = f(v)$ $(v \in V := H_0^1(\Omega))$ from Example 7.10. Let $\mathcal{T}$ be an admissible triangulation (cf. Definition 8.36).

[25] The value $\|L^{-1}\|_{1\leftarrow -1}$ is the eigenvalue of a suitable eigenvalue problem. Its approximative solution is not necessarily an upper bound, but also a good approximation of the right-hand side in $\|L^{-1}\|_{1\leftarrow -1} \le C$ is useful.

We introduce the following notation: The three sides of the triangle $T \in \mathcal{T}$ form the set $\mathcal{E}(T)$, while the three vertices of $T \in \mathcal{T}$ define the vertex set $\mathcal{N}(T)$. Their unions lead to

$$\mathcal{E} := \bigcup_{T\in\mathcal{T}} \mathcal{E}(T)\,, \qquad \mathcal{N} := \bigcup_{T\in\mathcal{T}} \mathcal{N}(T)\,.$$

The parts belonging to the Dirichlet boundary is $\mathcal{E}_D$, its complement is $\mathcal{E}_\Omega$:

$$\mathcal{E}_D := \{E \in \mathcal{E} : E \subset \Gamma\}\,, \qquad \mathcal{E}_\Omega := \{E \in \mathcal{E} : E \subset \Omega\}\,.$$

Each edge $E \in \mathcal{E}$ has two vertices as endpoints; they form the set $\mathcal{N}(E)$.

The sets of triangles neighboured to a triangle $T \in \mathcal{T}$, to an edge $E \in \mathcal{E}$, or to a node $\mathbf{x} \in \mathcal{N}$, respectively, are denoted by

$$\omega_T := \bigcup_{T':\mathcal{E}(T)\cap\mathcal{E}(T')\neq\emptyset} T', \qquad \omega_E := \bigcup_{T:E\in\mathcal{E}(T)} T, \qquad \omega_{\mathbf{x}} := \bigcup_{T:\mathbf{x}\in\mathcal{N}(T)} T$$

(cf. Figure 8.12). To each edge $E \in \mathcal{E}$ we associate a normal direction $\mathbf{n}_E$ (the sign can be chosen arbitrarily; i.e., if T and T' share the edge $E \in \mathcal{E}(T) \cap \mathcal{E}(T')$, $\mathbf{n}_E$ is the usual outer normal direction of one of the triangles). Since the derivative φ of a finite-element function may be discontinuous across the edges $E \in \mathcal{E}_\Omega$, there are two one-sided limits $\varphi(x \pm t\mathbf{n}_E)$ for $t \searrow 0$. Their difference is denoted by the symbol $[\,\cdot\,]$:

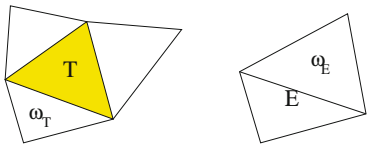


Fig. 8.12 ω_T and ω_E.

$$[\varphi]_E(x) := \lim_{t\searrow 0} \varphi(x + t\mathbf{n}_E) - \lim_{t\searrow 0} \varphi(x - t\mathbf{n}_E) \qquad \text{for } x \in E \in \mathcal{E}_\Omega.$$

Subtracting $a(u^h, v)$ from both sides of the continuous equation $a(u, v) = f(v)$, we obtain

$$\int_\Omega \langle \nabla\left(u - u^h\right), \nabla v\rangle \mathrm{d}\mathbf{x} = \int_\Omega f v \mathrm{d}\mathbf{x} - \int_\Omega \langle \nabla u^h, \nabla v\rangle \mathrm{d}\mathbf{x} \text{ for all } v \in V = H_0^1(\Omega). \tag{8.81}$$

The right-hand side represents the residual r in (8.79) via

$$r(v) := \int_\Omega f v \,\mathrm{d}\mathbf{x} - \int_\Omega \left\langle \nabla u^h, \nabla v\right\rangle \mathrm{d}\mathbf{x}.$$

Lemma 8.74. *Let c_Ω be the Poincaré–Friedrichs constant in $\|v\|_{L^2(\Omega)} \le c_\Omega\, |v|_{1,0}$ for $v \in H_0^1(\Omega)$ (cf. Lemma 6.29). Then the discretisation error $\|u - u^h\|_{H^1(\Omega)}$ can be bounded from both sides by r:*

$$\|r\|_{-1} \le \|u-u^h\|_{H^1(\Omega)} \le \left(1+c_\Omega^2\right)\|r\|_{-1}\,, \tag{8.82}$$
$$\|r\|_{-1} = \sup_{\substack{v\in V=H_0^1(\Omega)\\ \|v\|_{H^1(\Omega)}=1}} \left| \int_\Omega f v\,\mathrm{d}\mathbf{x} - \int_\Omega \left\langle \nabla u^h, \nabla v\right\rangle \mathrm{d}\mathbf{x}\right|.$$

Proof. By assumption $w \in H_0^1(\Omega)$ satisfies $\|w\|_{L^2(\Omega)} \le c_\Omega\,|w|_{1,0}$ so that

$$|w|_{1,0}^2 \le \|w\|_{H^1(\Omega)}^2 \le \left(1+c_\Omega^2\right)|w|_{1,0}^2\,. \tag{8.83}$$

From $|u-u^h|_{1,0} = \|\nabla\left(u-u^h\right)\|_{L^2(\Omega)} = \sup_{\|v\|_{H^1(\Omega)}=1} \left|\int_\Omega \left\langle \nabla\left(u-u^h\right), \nabla v\right\rangle \mathrm{d}\mathbf{x}\right|$ we deduce

$$\begin{aligned}
\|r\|_{-1} &= \sup_{0\neq v\in H_0^1(\Omega)} \frac{|r(v)|}{\|v\|_{H^1(\Omega)}} \underset{(8.81)}{=} \sup_{0\neq v\in H_0^1(\Omega)} \frac{\left|\int_\Omega \left\langle \nabla\left(u-u^h\right), \nabla v\right\rangle \mathrm{d}\mathbf{x}\right|}{\|v\|_{H^1(\Omega)}} \\
&\underset{(8.83)}{\le} \sup_{0\neq v\in H_0^1(\Omega)} \frac{\left|\int_\Omega \left\langle \nabla\left(u-u^h\right), \nabla v\right\rangle \mathrm{d}\mathbf{x}\right|}{\|\nabla v\|_{L^2(\Omega)}} \\
&= \|\nabla\left(u-u^h\right)\|_{L^2(\Omega)} \le \|u-u^h\|_{H^1(\Omega)}
\end{aligned}$$

as well as

$$\begin{aligned}
\|u-u^h\|_{H^1(\Omega)} &\underset{(8.83)}{\le} \sqrt{1+c_\Omega^2}\,|u-u^h|_{1,0} \\
&= \sqrt{1+c_\Omega^2} \sup_{0\neq v\in H_0^1(\Omega)} \frac{\left|\int_\Omega \left\langle \nabla\left(u-u^h\right), \nabla v\right\rangle \mathrm{d}\mathbf{x}\right|}{\|\nabla v\|_{L^2(\Omega)}} \\
&\underset{(8.81)}{=} \sqrt{1+c_\Omega^2} \sup_{0\neq v\in H_0^1(\Omega)} \frac{|r(v)|}{|v|_{1,0}} \\
&\underset{(8.83)}{\le} \left(1+c_\Omega^2\right) \sup_{0\neq v\in H_0^1(\Omega)} \frac{|r(v)|}{\|v\|_{H^1(\Omega)}} = \left(1+c_\Omega^2\right)\|r\|_{-1}\,.
\end{aligned}$$

Together the assertion follows. ∎

In (8.82) we may integrate $\int_\Omega f v\mathrm{d}\mathbf{x} - \int_\Omega \left\langle \nabla u^h, \nabla v\right\rangle \mathrm{d}\mathbf{x}$ by parts. For each T we have

$$\int_T \left\langle \nabla u^h, \nabla v\right\rangle \mathrm{d}\mathbf{x} = -\int_T \left(\Delta u^h\right) v\mathrm{d}\mathbf{x} - \sum_{E\in\mathcal{E}(T)} \int_E \frac{\partial u^h}{\partial n} v\mathrm{d}\mathbf{\Gamma} \quad \text{for all } v\in H^1(T).$$

Let $v \in H_0^1(T)$ be arbitrary. Summation over all $T \in \mathcal{T}$ yields

$$\int_\Omega f v \mathrm{d}\mathbf{x} - \int_\Omega \langle \nabla u^h, \nabla v \rangle \, \mathrm{d}\mathbf{x} \tag{8.84}$$
$$= \sum_{T\in\mathcal{T}} \left[\int_T \left(f + \Delta u^h\right) v \mathrm{d}\mathbf{x} + \sum_{E\in\mathcal{E}_\Omega(T)} \int_E \left[\langle \nabla u^h, \mathbf{n}_E\rangle\right]_E v \mathrm{d}\Gamma \right]$$

(note that both the scalar product $\left[\langle \nabla u^h, \mathbf{n}_E\rangle\right]_E$ and the difference $[\,\cdot\,]_E$ depend on the sign of $\mathbf{n}_E$, so that the expression $\left[\langle \nabla u^h, \mathbf{n}_E\rangle\right]_E$ is independent of the choice of $\mathbf{n}_E$). The edges $E \subset \Gamma$ do not appear in the last sum of (8.84) because of $v|_\Gamma = 0$.

In the case of piecewise finite elements we have $\Delta u^h = 0$ on each T. In the following estimate we replace f on each triangle by a constant function f_T which in the optimal case is chosen as the mean value

$$f_T := \frac{1}{\mathrm{area}(T)} \int_T f(\mathbf{x}) \mathrm{d}\mathbf{x} \qquad \text{(constant function on } T\text{)}.$$

The local quantity

$$\eta_T := \sqrt{h_T^2 \left\| f_T \right\|^2_{L^2(T)} + \frac{1}{2} \sum_{E\in\mathcal{E}_\Omega(T)} h_E \left\| \left[\langle \nabla u^h, \mathbf{n}_E\rangle\right]_E \right\|^2_{L^2(E)}} \tag{8.85}$$

with $h_T = \mathrm{diam}(T)$ and $h_E = \mathrm{length}(E)$ will be crucial in the following.

Theorem 8.75. *Let u be the solution of (8.80) and u^h the finite-element solution for piecewise linear elements of a triangulation $\mathcal{T}$. Then there are constants $\bar{c}$ and $\underline{c}$ only depending on Ω and the shape regularity of the triangulation $\mathcal{T}$ such that*

$$\|u - u^h\|_{H^1(\Omega)} \le \bar{c} \sqrt{\sum_{T\in\mathcal{T}} \eta_T^2 + \sum_{T\in\mathcal{T}} h_T^2 \left\| f - f_T \right\|^2_{L^2(T)}} \tag{8.86a}$$

and
$$\eta_T \le \underline{c} \sqrt{\|u - u^h\|^2_{H^1(\omega_T)} + \sum_{T'\subset\omega_T} h_{T'}^2 \left\| f - f_{T'} \right\|^2_{L^2(T')}}\,. \tag{8.86b}$$

Proof. We use the representation (8.84) of $r(v)$ and the Galerkin equation $r(v^h) = 0$ for all $v^h \in V_h$.

Inserting $-\Delta u^h = 0$ we obtain for all $v^h \in V_h$ that

$$r(v) = r(v - v^h) = \sum_{T\in\mathcal{T}} \left[\int_T f\left(v - v^h\right) \mathrm{d}\mathbf{x} - \sum_{E\in\mathcal{E}_\Omega(T)} \int_E \left[\langle \nabla u^h, \mathbf{n}_E\rangle\right]_E \left(v - v^h\right) \mathrm{d}\Gamma \right].$$

Choose v^h as the Clément interpolant of v (see page 216). It follows that

$$\left\| v - v^h \right\|_{L^2(T)} \le C_1 h_T \|u\|_{H^1(\widetilde{\omega_T})}, \quad \left\| v - v^h \right\|_{L^2(E)} \le C_2 \sqrt{h_E} \|u\|_{H^1(\widetilde{\omega_E})}$$

(see Clément [72]), where

$$\widetilde{\omega_T} := \bigcup_{T':\mathcal{N}(T)\cap\mathcal{N}(T')\neq\emptyset} T' \supset \omega_T \quad \text{and} \quad \widetilde{\omega_E} := \bigcup_{T':\mathcal{N}(E)\cap\mathcal{N}(T')\neq\emptyset} T' \supset \omega_E\,.$$

The constants C_1, C_2 only depend on the shape regularity (i.e., the smallest inner angle). This yields the estimate

$$\begin{aligned}|r(v)| &\le C_1 \sum_{T\in\mathcal{T}} h_T \,\|f\|_{L^2(T)}\,\|v\|_{H^1(\widetilde{\omega_T})} \\ &\quad + C_2 \sum_{E\in\mathcal{E}_\Omega(T)} \sqrt{h_E}\,\big\|\big[\langle\nabla u^h, \mathbf{n}_E\rangle\big]_E\big\|_{L^2(E)}\,\|v\|_{H^1(\widetilde{\omega_E})} \\ &\le C_3\|v\|_{H^1(\Omega)} \sqrt{\sum_{T\in\mathcal{T}} h_T^2\,\|f\|^2_{L^2(T)} + \sum_{E\in\mathcal{E}_\Omega(T)} h_E\,\big\|\big[\langle\nabla u^h, \mathbf{n}_E\rangle\big]_E\big\|^2_{L^2(E)}}\end{aligned}$$

for all $v \in H^1_0(\Omega)$, where the last line uses the Cauchy–Schwarz inequality. The term $\|f\|^2_{L^2(T)}$ can be bounded by $2\,\|f_T\|^2_{L^2(T)} + 2\,\|f - f_T\|^2_{L^2(T)}$. Together with (8.82) and (8.85) we prove the assertion (8.86a).

The second inequality (8.86b) uses a special choice of v. Details can be found in Verfürth [297, pages 15–17]. ∎

Inequality (8.86a) offers an error estimate using essentially the local quantities η_T. Here it is crucial that η_T can be determined *a posteriori*, i.e., after the computation of u^h. It remains to discuss $f - f_T$, whose bound $\|f - f_T\|_{L^2(T)}$ must be known *a priori*. Therefore one needs assumptions about f as discussed in §8.7.1.2.

The second inequality (8.86b) shows that η_T can be bounded from above by the error $\|u - u^h\|^2_{H^1(\omega_T)} + \mathcal{O}(h^2\,\|f - f_T\|^2_{L^2(\omega_T)})$ based on a somewhat enlarged neighbourhood $\omega_T \supset T$. Summation of all η_T^2 yields

$$\sum_{T\in\mathcal{T}} \eta_T^2 \le C\left[\|u - u^h\|^2_{H^1(\Omega)} + \mathrm{osc}^2(\mathcal{T})\right],$$

where

$$\mathrm{osc}(\mathcal{T}) := \sqrt{\textstyle\sum_{T\in\mathcal{T}} h_T^2\,\|f - f_T\|^2_{L^2(T)}}$$

is called the *oscillation term*. Therefore the error estimator $\eta := \sqrt{\sum_{T\in\mathcal{T}} \eta_T^2}$ and the error $\|u - u^h\|_{H^1(\Omega)}$ are equivalent modulo $\mathrm{osc}(\mathcal{T})$.

An error estimator η is called *reliable* if

$$\|u - u^h\|_{H^1(\Omega)} \le \mathrm{const}\cdot\eta + \mathcal{O}(\mathrm{osc}(\mathcal{T})).$$

To avoid an overestimation because of $\|u - u^h\|_{H^1(\Omega)} \ll \eta$ one needs the reverse inequality. An estimator η with the property

$$\eta \le \mathrm{const} \cdot \|u - u^h\|_{H^1(\Omega)} + \mathcal{O}(\mathrm{osc}(\mathcal{T}))$$

is called *efficient*. To answer a question as (8.78) the size of const in the above inequalities should be known approximately. The optimal situation is characterised by $\mathrm{const} = 1$ in both inequalities or at least by $\|u - u^h\|_{H^1(\Omega)}/\eta \to 1$ as $h \to 0$. In the latter case η is called *asymptotically exact*.

More about error estimators can be found in Verfürth [297]. We also refer to Eriksson et al. [96], Großmann–Roos–Stynes [124, §4.7], and Knabner–Angermann [172, §4.2], and Stein [269].

The reliability condition and the efficiency contain constants which usually are unknown. Another approach avoiding unknown constants is described by Braess–Schöberl [47] based on an idea of Prager–Synge [228]. Here one makes use of the fact that the original minimisation problem (6.47) has the same solution as a dual maximisation problem (see the example in §8.9.4).

At the end of §8.7.1.1 we mentioned the case of functionals instead of norms. Error estimators for functionals $\phi_i(u - u^h)$ are discussed in Becker–Rannacher [34] and Giles–Süli [116].

8.7.2 Efficiency of the Finite-Element Method

8.7.2.1 Measuring the Quality of a Discretisation

In one way or another each (consistent and convergent) discretisation determines an approximation u^h of u. An obvious question is how to compare two discretisations.

Two different goals can serve as criterion: (a) smallest possible computational cost (e.g., measured as run time of the computer program) or (b) smallest possible error (measured by some norm). Since both goals are opposite, one has to combine cost and accuracy. This leads us to two questions:

(i) Given some $\varepsilon > 0$. Which method yields an approximation $\tilde{u}$ with $\|u - \tilde{u}\| \le \varepsilon$ consuming the smallest computational cost?

(ii) Let u^I and u^{II} be the results of two methods with same computational cost. Which approximation is more accurate?

Taking the computational cost as criterion leads us to a new difficulty, since the computational cost refers to a certain algorithm which is not uniquely determined by the discretisation method. The latter yields the system of equation which is to be solved. For the solution there exist many methods which may require quite different computational cost (cf. Hackbusch [142]). Since there are solvers whose cost is proportional to the dimension N of the system, the remedy is using the dimension N as a measure of the computational cost.[26]

[26] This is a strong simplification. The computational cost depends not only on the dimension N, but also on the structure (e.g., sparsity of the matrix). Another measure would be the amount of storage.

8.7.2.2 Order of Consistency

Considering the above criterion (ii), we have to discuss $\varepsilon = \varepsilon(N)$ as a function of the dimension N. Piecewise linear finite elements lead to an error $\varepsilon = \mathcal{O}(h^2)$ with respect to the L^2-norm. Assuming a uniform triangulation and the spatial dimension n ($\Omega \subset \mathbb{R}^n$), we obtain $h = \mathcal{O}(N^{-1/n})$ so that $\varepsilon = \mathcal{O}(N^{-2/n})$. Using piecewise polynomials[27] of degree p, the error is $\varepsilon = \mathcal{O}(N^{-(p+1)/n})$. Here we assume that $u \in H^{p+1}(\Omega) \cap V$ (e.g., with $V = H_0^1(\Omega)$). Then question (ii) can be answered asymptotically (for large N): methods of higher order of consistency are better than those of lower order.

8.7.2.3 Choice of the Triangulation

Fixing the local polynomial degree p, the question remains how to choose the triangulation $\mathcal{T}$. Since the squared error $\|e\|^2_{L^2(\Omega)}$ ($e := u - u^h$) is the sum of local contributions $\|e\|^2_{L^2(T)}$ ($T \in \mathcal{T}$), one can apply the heuristic *principle of error equilibration*: All error terms $\|e\|_{L^2(T)}$ should be of comparable size. This principle will be applied in §8.7.3 as a constructive criterion for generating the adaptive triangulation.

8.7.3 Adaptive Finite-Element Method

8.7.3.1 Posing the Problem

We use the goal (i) from §8.7.2.1: Given a desired error tolerance $\varepsilon > 0$ with respect to some norm $\|\cdot\|$, the finite-element discretisation with the smallest possible dimension N is sought satisfying the error bound ε.

8.7.3.2 Adaptive Grid Refinement

Let $\eta(T)$ be defined, e.g., by the error estimator in §8.7.1.3 which associates each triangle of $\mathcal{T}$ with a local error. The heuristic expectation is that a grid refinement

---

Note that, for the same N, a difference method requires much less storage than a finite-element method. Using N as measure is completely unrealistic for spatially high-dimensional problems (vgl. Hackbusch [138, §16], [141]).

[27] By numerical reasons one has to use suitable, e.g., orthogonal polynomial bases. Possible difficulties are demonstrated by the polynomials $b_\nu(x) = x^\nu/\nu$ ($\nu = 1, \ldots, N$) for the one-dimensional boundary-value problem $u'' = f$ in $\Omega = (0,1)$, $u(0) = 0$, $u'(1) = 0$ (natural boundary condition). The system matrix is the *Hilbert matrix* $A_{ij} = 1/(i+j-1)$ which is an example of an extremely large condition (cf. Maeß [198, S. 108] and Footnote 15 on page 193).

in triangles T with a large value $\eta(T)$ reduces the total error better than a refinement of triangles with smaller $\eta(T)$. According to the definition of efficiency, one wants to reduce the error without increasing N too much. Therefore one chooses the following subset of elements to be refined:

Step I: Let $\vartheta \in (0,1)$ be fixed. Determine $\eta_{\max} := \max\{\eta(T) : T \in \mathcal{T}\}$ and set $\mathcal{T}_{\text{refine}} := \{T \in \mathcal{T} : \eta(T) \geq \vartheta\, \eta_{\max}\}$.

Step II: Refine all triangles in $\mathcal{T}_{\text{refine}}$ according to the method in §8.7.3.3.

Step III: Determine the solution u^h and $\eta(T)$ for the new triangulation $\mathcal{T}_{\text{new}}$.

Step IV: If the a-posteriori error estimator indicates a sufficient accuracy accept u^h as result, otherwise repeat the process at Step I.

Step I is called the *maximum marking strategy*. If the estimator $\eta(T)$ is already equilibrated, i.e., all $\eta(T)$ are of similar size, the full set $\mathcal{T}_{\text{refine}} = \mathcal{T}$ may result and the refinement is uniform. Otherwise, $\#\mathcal{T}_{\text{refine}} \ll \#\mathcal{T}$ may occur.

A variant of Step I is the *bulk chasing marking* or *Dörfler marking* (cf. Dörfler [90]): Given some $\vartheta \in (0,1)$, order $\eta(T)$ $(T \in \mathcal{T})$ by size and collect as many of the largest contributions in $\mathcal{T}_{\text{refine}}$ such that $\mathcal{T}_{\text{refine}} \subset \mathcal{T}$ is the subset of minimal cardinality with the property

$$\vartheta \sum_{T \in \mathcal{T}} \eta(T) \leq \sum_{T \in \mathcal{T}_{\text{refine}}} \eta(T).$$

The schemes described above determines for each T whether it should be refined or not. Alternatively, one may decide how often a triangle is to be refined (cf. Bank–Deotte [24]).

8.7.3.3 Techniques for Grid Refinement

The basic task reads as follows. Starting from a subset $\mathcal{T}_v$ of $\mathcal{T}$ one wants to create a new triangulation $\mathcal{T}_{\text{new}}$ which replaces all $T \in \mathcal{T}_v$ by smaller elements, while the changes of $\mathcal{T}$ should be minimal. The following remarks are to be considered.

(a) The more element are in $\mathcal{T} \cap \mathcal{T}_{\text{new}}$ the less new computations of matrix entries are required.

(b) If hanging nodes are not allowed, the refinement of a triangle $T \in \mathcal{T}_v$ may force a refinement of its neighbours.

(c) The shape regularity requires constructions which prevents too small inner angles.

(d) If the K-grid property does not follow from the shape regularity, one has to take care that K is not increasing.

First we comment on (d). In the one-dimensional case, where the elements are intervals, shape regularity does not apply. Figure 8.13 shows a sequence of refinements in $\Omega = (0, 1)$ where all subintervals in $(0, 1/4)$ should be refined. The first refinement leads to a K-grid with $K = 2$. The second, quite local halving of the subintervals in $(0, 1/4)$ yields $K = 4$, since the lengths of the neighbouring intervals $[3/16, 4/16]$ and $[1/4, 1/2]$ form the ratio 4. To keep a K-grid with $K = 2$ one has to halve the interval $[1/4, 1/2]$.

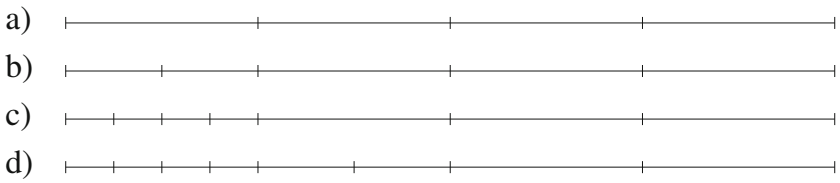


Fig. 8.13 K-grid with $K = 2$, grid (c) is not accepted.

There is a simple sufficient criterion for controlling the shape regularity. If the refinement produces only triangles that are congruent to the initial triangle, the inner angles are not changed and in particular the minimal angle is constant. Therefore the standard refinement ("type A") is a partition of $T \in \mathcal{T}_v$ in four congruent subtriangles of half side length (cf. Figure 8.14a).

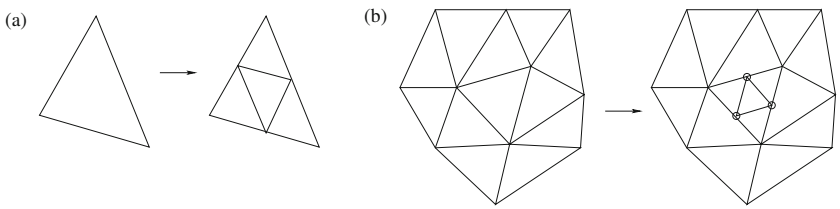


Fig. 8.14 (a) Type A: regular refinement of a triangle, (b) hanging node after a local refinement.

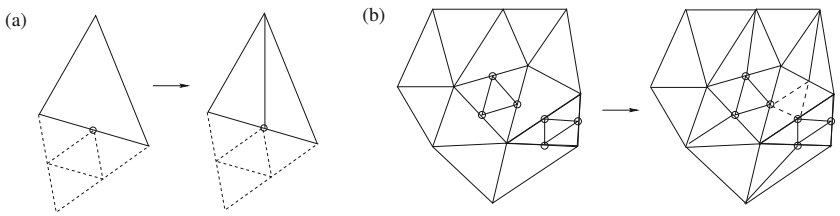


Fig. 8.15 (a) Type B: refinement of a triangle by a median, (b) result of the refinement.

A triangle $T \in \mathcal{T}_v$ with a neighbouring triangle $T' \in \mathcal{T} \backslash \mathcal{T}_v$ requires special care (cf. Figure 8.14b), since a hanging node occurs if T is refined according to type A but T' not (cf. §8.9.3). On the other hand, if also T' is refined according to type A, the refinement spreads all over the domain and we obtain a uniform refinement. Therefore $T' \in \mathcal{T} \backslash \mathcal{T}_v$ is only refined according to type A if it has two neighbours in $\mathcal{T}_v$ (see dashed triangle in Figure 8.15b). Otherwise the refinement strategy of Bank [22] requires that T' is halved according to Figure 8.15a (refinement of "type B"). This step may lead to a smaller inner angle. Therefore, in the next step, the refinements of type B are undone.

Concerning alternative strategies we refer to Rivara [243, 244] and Knabner–Angermann [172, §4.1].

Since the refinement strategy is derived from the heuristic principle of error equilibration, one may ask whether a sequence of refined triangulations converges and whether the convergence rate is optimal. This question is answered by Dörfler [90] (see also Kreuzer–Siebert [177]; the maximum marking strategy is discussed, e.g., by Diening–Kreuzer–Stevenson [85]). For three-dimensional problems with domains containing edges an optimal order of convergence requires the use of anisotropic refinements (cf. Apel–Nicaise [6]), hence regular refinements cannot be optimal.

8.7.3.4 Adaptive Grid Coarsening

So far we have considered a triangulation which is too coarse and requires refinement. However, there are also situations in which the starting triangulation has too small triangles, i.e., the triangles are smaller than necessary. Similar to the above refinement method one can try to determine a subset $\mathcal{T}_{\text{coarse}} \subset \mathcal{T}$ (e.g., via $\eta(T) \leq \eta_{\min}$) and to coarsen the grid locally. In this case it is helpful if the previous refinement steps are stored in a tree structure: the subtriangles are the sons of the unrefined triangle (father vertex). Then a refinement means that sons are added to a leaf vertex while coarsening corresponds to an erasing of the sons. As for the refinement one has to take care that no hanging nodes occur.

8.7.3.5 *hp*-Method

The previously discussed finite elements has been piecewise polynomials of a fixed degree p, while the element size h is the variable quantity approaching zero. In the hp approach we allow different polynomial degrees in different elements. In the 'refinement step' we may either refine the element as discussed above or increase the (local) polynomial degree p. According to §8.7.2.2 an enlargement of p should be advantageous at least if the solution is sufficiently smooth. Step II in §8.7.3.2 has to be modified: one must decide for each triangle whether h or p or both should be changed (cf. Melenk [204], Schwab [261], Giles–Süli [116, §9], Großmann–Roos–Stynes [124, §4.9.4], Le Borne–Ovall [187], Bank–Deotte [24], and Canuto et al. [62, 63]). The implementation of hp-finite elements becomes easier in combination with DGFEM discussed in §8.9.10.2 (cf. Houston–Schwab–Süli [156]).

8.8 Properties of the System Matrix

8.8.1 Connection of $\mathbf{L}$ and L_h

The properties of the matrix L_h have been carefully studied for difference schemes since the solvability of the difference equations and the analysis of convergence properties depend on them. In the case of finite-element discretisation we obtain the corresponding assertions in another way. The factors that control the solvability and convergence are the subspace V_h and the operator $L_h : V_h \to V_h$, which is associated to the bilinear form $a(\cdot,\cdot) : V_h \times V_h \to \mathbb{R}$ and which was introduced as L_N in (8.15).

The system matrix (stiffness matrix) $\mathbf{L}$, for a given V_h, depends on the basis $\{b_1, \ldots, b_{N_h}\}$ chosen. Let

$$N_h := \dim(V_h).$$

The isomorphism $P : \mathbb{R}^{N_h} \to V_h$ defined in (8.6) maps the coefficient vector $\mathbf{v}$ to $P\mathbf{v} = \sum_{i=1}^{N_h} v_i b_i \in V_h$. For finite elements, in general, the coefficients coincide with the nodal values: $v_i = (P\mathbf{v})(\mathbf{x}^i)$, $1 \le i \le N_h$. The connection between the matrix $\mathbf{L}$ and the operator $L_h : V_h \to V_h$ is, according to Lemma 8.12,

$$\mathbf{L} = P^* L_h P, \qquad L_h = P^{*-1} \mathbf{L} P^{-1}.$$

The definition of $P^\star$ in (8.12), $\langle P^\star u, v\rangle = (u, Pv)_{L^2(\Omega)}$, also depends on the choice of the scalar product $\langle\cdot,\cdot\rangle$ in $\mathbb{R}^{N_h}$. Let us choose here the usual

$$\langle \mathbf{u}, \mathbf{v}\rangle = \sum_{i=1}^{N_h} u_i v_i \,. \tag{8.87a}$$

In Exercise 8.17 we have already established the following properties of $\mathbf{L}$: If $a(\cdot,\cdot)$ is symmetric [and V-elliptic] then $\mathbf{L}$ is symmetric [and positive definite]. Notice that the difference methods do not always have these properties. For the Poisson problem there is a symmetric form, but nonetheless the matrix L_h in Section 4.8.1 is not symmetric (cf. Theorem 4.77).

8.8.2 Equivalent Norms and Mass Matrix

The condition $\mathrm{cond}(\mathbf{L})$ plays an important role in the solution of the equation $\mathbf{Lu} = \mathbf{f}$. We would like to show that, under standard conditions, we have $\mathrm{cond}(\mathbf{L}) = \mathcal{O}(h^{-2m})$ (cf. Remark 5.45). First we have to decide on the fundamental norm of $\mathbb{R}^{N_h}$. Up to a scalar factor h^n

$$\|\mathbf{u}\|_h := \sqrt{h^n \sum_{i=1}^{N_h} |u_i|^2} \qquad (n\text{: dimension of } \Omega \subset \mathbb{R}^n) \tag{8.87b}$$

is the Euclidean norm coming from the scalar product (8.87a). The corresponding matrix norm

$$\|\mathbf{L}\| := \sup\{\|\mathbf{L}\mathbf{v}\|_h / \|\mathbf{v}\|_h : \mathbf{0} \neq \mathbf{v} \in \mathbb{R}^{N_h}\}$$

is the usual spectral norm of $\mathbf{L}$ (it is independent of the scaling factor in $\|\cdot\|_h$).

As an alternative to $\|\cdot\|_h$ let us introduce

$$\|\mathbf{u}\|_P := \|P\mathbf{u}\|_{L^2(\Omega)} \qquad \text{for } \mathbf{u} \in \mathbb{R}^{N_h}.$$

In some important cases $\|\cdot\|_P$ and $\|\cdot\|_h$ are equivalent (uniformly in h). As an example consider the linear elements.

Theorem 8.76. *Let $\{\mathcal{T}_h\}$ be a sequence of shape-regular triangulations. Let V_h be the space of linear elements defined in (8.36) with the usual basis (see (8.37a)). Then there is a constant C_P, which does not depend on h, such that*

$$\frac{1}{C_P} \|\mathbf{u}\|_P \leq \|\mathbf{u}\|_h \leq C_P \|\mathbf{u}\|_P . \tag{8.88}$$

For various finite elements concrete bounds in inequality (8.88) are described by Wathen [303]. The basis for the proof of Theorem 8.76 is the following statement.

Lemma 8.77. *Let $T = \{(\xi, \eta) : \xi, \eta \geq 0, \xi + \eta \leq 1\}$ be the unit triangle (see Figure 8.4). If u is linear on T, then*

$$\begin{aligned} &\frac{1}{24}\left[u(0,0)^2 + u(1,0)^2 + u(0,1)^2\right] \\ &\leq \iint_T u(\xi,\eta)^2 \mathrm{d}\xi\mathrm{d}\eta \leq \frac{1}{6}\left[u(0,0)^2 + u(1,0)^2 + u(0,1)^2\right]. \end{aligned}$$

Proof. From $u(\xi,\eta) = u_1 + (u_2 - u_1)\xi + (u_3 - u_1)\eta$ with $u_1 := u(0,0)$, $u_2 := u(1,0)$, $u_3 := u(0,1)$ it follows that

$$\iint_T u(\xi,\eta)^2 \mathrm{d}\xi\mathrm{d}\eta = \frac{1}{24} [u_1\ u_2\ u_3] \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}.$$

The eigenvalues of the symmetric 3×3 matrix are $\lambda_1 = \lambda_2 = 1$ and $\lambda_3 = 4$. Therefore the right-hand side is $\geq \sum_{i=1}^3 u_i^2/24$ and $\leq 4 \sum u_i^2/24$. ∎

Proof of Theorem 8.76. Write $\|\mathbf{u}\|_P^2$ as $\int_\Omega (P\mathbf{u})^2 \mathrm{d}\mathbf{x} = \sum_{T_i \in \mathcal{T}_h} \int_{T_i} (P\mathbf{u})^2 \mathrm{d}\mathbf{x}$. Let $\Phi_i : T \to T_i$ be the linear maps from T to T_i as in Exercise 8.43a. Let u_1, u_2, u_3 be the values of $u = P\mathbf{u}$ at the vertices of T_i. The inequality $0 < C_1 h^2 \leq |\det \Phi_i'| \leq C_2 h^2$ and Lemma 8.77 imply

$$\frac{C_1 h^2}{24}\left(u_1^2+u_2^2+u_3^2\right) \le \int_{T_i}(P\mathbf{u})^2 \mathrm{d}\mathbf{x} = |\det \Phi_i'| \iint_T \left[(P\mathbf{u})(\Phi_i(\xi,\eta))\right]^2 \mathrm{d}\xi \mathrm{d}\eta$$
$$\le \frac{C_2 h^2}{6}\left(u_1^2+u_2^2+u_3^2\right). \tag{8.89}$$

By definition of a shape-regular triangulation all the angles of the triangles are $\ge \alpha_0 > 0$. Each node belongs to at least $M_{\min}$ triangles, and to at most $M_{\max} \le \frac{2\pi}{\alpha_0}$ triangles. Since u_1, u_2, u_3 are the components of the vector $\mathbf{u}$, the summation in equation (8.89) over all $T_i \in \mathcal{T}_h$ gives

$$\frac{M_{\min} C_1}{24}\|\mathbf{u}\|_h^2 \le \|\mathbf{u}\|_P^2 = \sum_{T_i \in \mathcal{T}_h} \int_{T_i}(P\mathbf{u})^2 \mathrm{d}\mathbf{x} \le \frac{M_{\max} C_2}{6}\|\mathbf{u}\|_h^2 , \tag{8.90}$$

so that inequality (8.88) holds with $C_P := \sqrt{\max(M_{\max}C_2/6, 24/(M_{\min}C_1))}$. ■

The matrix

$$\mathbf{M} := P^* P, \qquad \text{i.e.,} \quad M_{ij} = \int_\Omega b_i(\mathbf{x}) b_j(\mathbf{x}) \mathrm{d}\mathbf{x}, \tag{8.91}$$

is called the *mass matrix* when it occurs in engineering applications.

Exercise 8.78. Let $V_h \subset H_0^1(\Omega)$ be derived from the square-grid triangulation (cf. Exercise 8.42). Show the following:
(a) (8.89) and (8.90) hold with $C_1 = C_2 = 1$ and $M_{\min} = M_{\max} = 6$ so that

$$\tfrac{1}{2}\|\mathbf{u}\|_h \le \|\mathbf{u}\|_P \le \|\mathbf{u}\|_h .$$

(b) The mass matrix yields the star $\frac{h^2}{12}\begin{bmatrix}0&1&1\\1&6&1\\1&1&0\end{bmatrix}$ (cf. (4.20)).

Remark 8.79. (a) Inequality (8.88) is equivalent to

$$\|\mathbf{M}\| \le C_P^2 h^n, \qquad \|\mathbf{M}^{-1}\| \le C_P^2 h^{-n}.$$

(b) We have $\|\mathbf{u}\|_P \le \|h^{-n}\mathbf{M}\|^{1/2}\|\mathbf{u}\|_h$, $\|\mathbf{u}\|_h \le \|h^n\mathbf{M}^{-1}\|^{1/2}\|\mathbf{u}\|_P$.

Proof. (i) The calculation $\|\mathbf{u}\|_P^2 = \int_\Omega (P\mathbf{u})^2 \mathrm{d}\mathbf{x} = (P\mathbf{u}, P\mathbf{u})_0 = \langle P^\star P\mathbf{u}, \mathbf{u}\rangle = \langle \mathbf{M}\mathbf{u}, \mathbf{u}\rangle \le \|\mathbf{M}\|\langle \mathbf{u}, \mathbf{u}\rangle = \|\mathbf{M}\| h^{-n}\|\mathbf{u}\|_h^2$ proves the first inequality in (b). The next one follows from

$$\|\mathbf{u}\|_h^2 = h^n\langle \mathbf{u}, \mathbf{u}\rangle = h^n\langle \mathbf{M}^{1/2}\mathbf{u}, \mathbf{M}^{-1}\mathbf{M}^{1/2}u\rangle$$
$$\le \|h^n\mathbf{M}^{-1}\|\langle \mathbf{M}^{1/2}\mathbf{u}, \mathbf{M}^{1/2}\mathbf{u}\rangle = \|h^n\mathbf{M}^{-1}\|\langle \mathbf{M}u, u\rangle = \|h^n\mathbf{M}^{-1}\|\,\|\mathbf{u}\|_P^2 ,$$

since $\mathbf{M}$ is positive definite.

(ii) From $h^{-n}\|\mathbf{M}\| \le C_P^2$ and $h^n\|\mathbf{M}^{-1}\| \le C_P^2$, using (b), we deduce the inequality (8.88). Conversely $\langle \mathbf{M}\mathbf{u}, \mathbf{u}\rangle^{1/2} = \|\mathbf{u}\|_P \le C_P\|\mathbf{u}\|_h = C_P h^{n/2}\langle \mathbf{u}, \mathbf{u}\rangle^{1/2}$ implies that $h^{-n}\|\mathbf{M}\| \le C^2$, since $\mathbf{M}$ is positive definite (cf. Lemma 4.33). Similarly, $h^n\|\mathbf{M}^{-1}\| \le C_P^2$ follows from $\|\mathbf{u}\|_h \le C_P\|\mathbf{u}\|_P$. ■

8.8.3 *Inverse Estimate and Condition of* L

The H^1-norm cannot be bounded by the L^2-norm since the supremum of $\frac{|u|_1}{|u|_0}$ over $0 \neq u \in V$ is infinite. However, in the case of a finite-dimensional space V_h, the supremum $C_h := \sup\{|u|_1 / |u|_0 : 0 \neq u \in V_h\}$ is finite. One says that V_h satisfies the *inverse estimate* if $C_h = \mathcal{O}(h^{-1})$, that is,

$$|u|_1 \leq C_I \, h^{-1} \, |u|_0 \qquad \text{for all } u \in V_h, \tag{8.92}$$

where C_I does not depend on h as $h \to 0$.

Theorem 8.80. *Let $\{\mathcal{T}_h\}$ be a sequence of shape-regular triangulations. Then the space of finite elements introduced in (8.36) satisfies the inverse estimate.*

Proof. As in Theorem 8.76, the proof follows by transforming the integrals $\int_{T_i} |D^\alpha(Pu)|^2 \, \mathrm{d}\mathbf{x}$ for $T_i \in \mathcal{T}_h$ and $|\alpha| \leq 1$ into integrals over T. In addition we have to transform $D = D_x$ into $D_{\xi,\eta}$ (see the proof of Lemma 8.55). ∎

Theorem 8.81. *Suppose $\Omega \subset \mathbb{R}^n$. Assume that (8.2), (8.92), and $\|\mathbf{u}\|_P \leq C_P \, \|\mathbf{u}\|_h$ hold. Then we have*

$$\|\mathbf{L}\| \leq h^{n-2} C_S C_P^2 C_I^2 \, . \tag{8.93}$$

Proof. Multiply the inequality

$$\begin{aligned} \langle \mathbf{u}, \mathbf{L}\mathbf{v} \rangle = a(P\mathbf{u}, P\mathbf{v}) &\leq C_S \, |P\mathbf{u}|_1 \, |P\mathbf{v}|_1 \leq C_S C_I^2 h^{-2} \, |P\mathbf{u}|_0 \, |P\mathbf{v}|_0 \\ &\leq C_S C_I^2 h^{-2} C_P^2 \, \|\mathbf{u}\|_h \, \|\mathbf{v}\|_h \end{aligned}$$

by h^n and set $\mathbf{u} := \mathbf{L}\mathbf{v}$:

$$\|\mathbf{L}\mathbf{v}\|_h^2 = h^n \, \langle \mathbf{L}\mathbf{v}, \mathbf{L}\mathbf{v} \rangle \leq C_S C_I^2 h^{n-2} C_P^2 \, \|\mathbf{L}\mathbf{v}\|_h \, \|\mathbf{v}\|_h \, ,$$

and thus $\|\mathbf{L}\mathbf{v}\|_h \leq C_S C_I^2 h^{-2} C_P^2 \, \|\mathbf{v}\|_h$ for all $\mathbf{v}$, that is (8.93). ∎

Theorem 8.82. *Assume $\Omega \subset \mathbb{R}^n$. Let (8.17a) hold with $\varepsilon_h \geq \varepsilon > 0$ and $\|\cdot\|_h \leq C_P \, \|\cdot\|_P$. Then there holds*

$$\|\mathbf{L}^{-1}\| \leq C_P^2 h^{-n} / \varepsilon. \tag{8.94}$$

Proof. We know $\mathbf{L}^{-1} = P^{-1} L_h^{-1} P^{\star-1}$ and $\|L_h^{-1}\|_{V_h \leftarrow V_h'} \leq \frac{1}{\varepsilon_h} \leq \frac{1}{\varepsilon}$ (cf. Section 8.2), so that

$$\begin{aligned} \|\mathbf{L}^{-1}\mathbf{f}\|_h \leq C_P \|\mathbf{L}^{-1}\mathbf{f}\|_P = C_P \left|P\mathbf{L}^{-1}\mathbf{f}\right|_0 &= C_P \left|L_h^{-1} P^{\star-1}\mathbf{f}\right|_0 \\ \leq C_P \|L_h^{-1} P^{\star-1}\mathbf{f}\|_{V_h} &\leq (C_P/\varepsilon) \|P^{\star-1}\mathbf{f}\|_{V_h'}. \end{aligned}$$

Since

$$\begin{aligned}\|P^{\star-1}\mathbf{f}\|_{V_h'} &= \sup_{0\neq v\in V_h} \left|\left(v, P^{\star-1}\mathbf{f}\right)_0\right| / |v|_1 = \sup_{0\neq v\in V_h} \left|\left\langle P^{-1}v, \mathbf{f}\right\rangle\right| / |v|_1 \\ &= \sup_{\mathbf{v}\neq\mathbf{0}} |\langle \mathbf{v}, \mathbf{f}\rangle| / |P\mathbf{v}|_1 \underset{|P\mathbf{v}|_1 \geq |P\mathbf{v}|_0}{\leq} \sup_{\mathbf{v}\neq\mathbf{0}} h^{-n} \|\mathbf{v}\|_h \|\mathbf{f}\|_h / |P\mathbf{v}|_0 \\ &\leq C_P \sup_{\mathbf{v}\neq\mathbf{0}} h^{-n} \|\mathbf{v}\|_h \|\mathbf{f}\|_h / \|\mathbf{v}\|_h = C_P h^{-n} \|\mathbf{f}\|_h ,\end{aligned}$$

we have that $\|\mathbf{L}^{-1}\mathbf{f}\|_h \leq C_P^2 h^{-n} \|\mathbf{f}\|_h / \varepsilon$ for all $\mathbf{f}$; thus we conclude (8.94). ∎

The next theorem is the combination of Theorems 8.81 and 8.82.

Theorem 8.83. *Assume (8.2), (8.17a) with $\varepsilon_h \geq \varepsilon > 0$, (8.88), $m = 1$, and (8.92). Then we have*

$$\operatorname{cond}(\mathbf{L}) \leq h^{-2} C_S C_P^4 C_I^2 / \varepsilon .$$

The ideas involved in the proof can, in principle, be carried over to the case $2m > 2$, i.e., to boundary-value problems of higher order. The inverse estimate becomes

$$|u|_m \leq C_I h^{-m} |u|_0 \qquad \text{for all } u \in V_h.$$

In order that inequality (8.88) hold one must be careful in defining the norm $\|\cdot\|_h$. If all the components u_i of $\mathbf{u}$ are nodal values $(P\mathbf{u})(\mathbf{x}^i)$ (as for instance is the case for the spline ansatz for $V_h \subset H^2(\Omega)$, then $\|\cdot\|_h$ can be defined as in (8.87b). However as soon as the components u_i of $\mathbf{u}$ involve derivatives $(D^\alpha P\mathbf{u})(\mathbf{x}^i)$, the u_i^2 in (8.87b) must be replaced by $(h^{|\alpha|}u_i)^2$. For example, when the Hermite functions of (8.74) are used then $\mathbf{u}$ has the components u_{1i} and u_{2i} (cf. Exercise 8.71), where $u_{1i} = (P\mathbf{u})(x^i)$ and $u_{2i} = \frac{\mathrm{d}}{\mathrm{d}x}(P\mathbf{u})(x^i)$. The appropriate definition of $\|\cdot\|_h$ is $\|u\|_h^2 := h^n \sum_i (u_{1i}^2 + h^2 u_{2i}^2)$ with $n = 1$, since $(0,1) \subset \mathbb{R}^1$.

Exercise 8.84. Let $V \subset U \subset V'$ be a Gelfand triple and assume $V_h \subset V$ with $\dim(V_h) < \infty$. Show that the inverse estimate

$$\|u\|_V \leq C_I h^{-m} \|u\|_U \qquad \text{for all } u \in V_h$$

implies

$$\|u\|_U \leq C_I h^{-m} \|u\|_{V'} \qquad \text{and} \qquad \|u\|_V \leq C_I^2 h^{-2m} \|u\|_{V'} \quad \text{for all } u \in V_h .$$

The previous considerations require a uniform grid. A quasi-uniform grid enlarges the condition, since $\|\mathbf{L}\|$ is determined by the minimal and $\|\mathbf{L}^{-1}\|$ by the maximal element size. Nevertheless, the result can be recovered by a suitable scaling. First we introduce the dimension $N = \dim V_h$ as actual parameter. In the uniform case h is proportional to $N^{-1/n}$, where n is the spatial dimension ($\Omega \subset \mathbb{R}^n$). The condition $\mathcal{O}(h^{-2})$ in Theorem 8.83 corresponds to $\mathcal{O}(N^{2/n})$. Bank–Scott [26] prove for quasi-uniform grids in dimension $n \geq 3$ that a suitable scaling of the basis functions also gives $\operatorname{cond}(\mathbf{L}) = \mathcal{O}(N^{2/n})$ (for $n = 2$ an additional logarithmic factor appears).

8.8.4 Element Matrices

The element matrices, which we are going to define, allow the generation of the system matrix $\mathbf{L}$. Furthermore, the element matrices are useful for domain decomposition methods.

Let $T \in \mathcal{T}_h$ be an element with nodal points $\mathbf{x}^{T,1}, \mathbf{x}^{T,2}, \ldots, \mathbf{x}^{T,m} \in T$ (the number m of nodal points may depend on T, e.g., $m = 3$ for linear elements on triangles). Let the integrands $\varphi_{\Omega/\Gamma}(b_j, b_i)$ in $L_{ij} = a(b_j, b_i) = \int_\Omega \varphi_\Omega(b_j, b_i)\mathrm{d}\mathbf{x} + \int_\Gamma \varphi_\Gamma(b_j, b_i)\mathrm{d}\Gamma$ be defined by (7.28). The integrals can be written as sums of $\int_T$ for all T in the support of the integrand:

$$L_{ij} = \sum_{T\in\mathcal{T}_h} \left\{ \int_T \varphi_\Omega(b_j, b_i)\,\mathrm{d}\mathbf{x} + \int_{\Gamma\cap\overline{T}} \varphi_\Gamma(b_j, b_i)\,\mathrm{d}\Gamma \right\} \quad (1 \le i, j \le n)\,. \tag{8.95}$$

The boundary integral only occurs if $T \in \mathcal{T}_h$ touches the boundary.

In the following we assume that the nodal values of b_i are the function values $b_i(\mathbf{x}^{T,\nu})$, $1 \le \nu \le m$ (for more complicated finite-element ansatzes also derivatives $D^\alpha b_i(\mathbf{x}^{T,\nu})$ or directional derivatives may appear). By definition of the support of finite-element functions b_i, $\mathrm{supp}(b_i) \cap T \neq \emptyset$ holds if and only if one of the nodal values $b_i(\mathbf{x}^{T,\nu})$ does not vanish. This explains the following procedure.

Let the space V_T be the restriction of the finite-element functions from V_h to T (e.g., consisting of linear functions in (8.36), bilinear ones in (8.39a,b), quadratic ones in (8.40), or the serendipity functions in (8.42)). Interpolation at the nodal points $\mathbf{x}^{T,\nu}$ must be unique (necessary: $\dim V_T = m$). For $T \in \mathcal{T}_h$ and $1 \le \nu \le m$ let ϕ_ν^T be the Lagrange basis function, i.e., $\phi_\nu^T \in V_T$, $\phi_\nu^T(\mathbf{x}^{T,\mu}) = \delta_{\nu,\mu}$. Then the basis functions b_i of V_h satisfy

$$b_i|_T = \sum\nolimits_{\nu=1}^{m} b_i(\mathbf{x}^{T,\nu})\,\phi_\nu^T\,. \tag{8.96}$$

For each $T \in \mathcal{T}_h$ one defines the $m \times m$ matrix $\mathbf{E}_T$ by the coefficients

$$\mathbf{E}_{T,\nu\mu} = \int_T \varphi_\Omega(\phi_\mu^T, \phi_\nu^T)\,\mathrm{d}\mathbf{x} + \int_{\Gamma\cap\overline{T}} \varphi_\Gamma(\phi_\mu^T, \phi_\nu^T)\,\mathrm{d}\Gamma \quad (1 \le \nu, \mu \le m)\,. \tag{8.97}$$

The practical computation uses the map onto the unit element (cf. Exercise 8.43). Note that the term $\int_{\Gamma\cap\overline{T}} \ldots$ vanishes for homogeneous Dirichlet boundary condition and for $\Gamma \cap \overline{T} = \emptyset$.

Combining (8.96), (8.97), and the definition (8.95) of $\mathbf{L}$, we obtain the representation

$$\mathbf{L}_{ij} = \sum_{T\in\mathcal{T}_h} \sum_{\nu=1}^{m} \sum_{\mu=1}^{m} b_i(\mathbf{x}^{T,\nu})\, b_j(\mathbf{x}^{T,\mu})\, \mathbf{E}_{T,\nu\mu}. \tag{8.98}$$

In spite of the three-fold sum, the evaluation of (8.98) is cheap. The sum over $T \in \mathcal{T}_h$ can be restricted to $T \subset \mathrm{supp}(b_i) \cap \mathrm{supp}(b_j)$. The ν- and μ-sums often reduce to a single term since usually $b_i(\mathbf{x}^{T,\nu}) = 1$ holds for only one point and vanishes otherwise.[28]

[28] Modifications may occur if hanging nodes are present (cf. §8.9.3).

Remark 8.85. (a) Instead of using the matrix $\mathbf{L}$ for a matrix-vector multiplication $\mathbf{v} \in \mathbb{R}^n \mapsto \mathbf{L}\mathbf{v}$ one can directly use the element matrices.

(b) The element matrices $\{\mathbf{E}_T : T \in \mathcal{T}_h\}$ require more storage than $\mathbf{L}$. The square-grid triangulation of $\Omega = (0,1)^2$ (cf. Figure 8.2) consists of about $2N$ triangles. Piecewise linear elements require $m = 3$ nodal points so that the element matrices take $18N$ storage units while $\mathbf{L}$ needs $5N$ units (5 entries per row).

(c) While $\mathbf{L}$ can be computed from $(\mathbf{E}_T)_{T\in\mathcal{T}_h}$, one cannot regain the element entries $(\mathbf{E}_T)_{T\in\mathcal{T}_h}$ from $\mathbf{L}$.

There are different reasons (e.g., the domain decomposition method, cf. [142, §12]) why the domain Ω is described as the disjoint union of two (or more) subdomains Ω_1 and Ω_2 (cf. Figure 10.1 on page 312) and the corresponding system matrices $\mathbf{L}_1$ and $\mathbf{L}_2$ are used. The latter are defined by replacing $\int_\Omega$ in the definition of the bilinear form by $\int_{\Omega_1}$ and $\int_{\Omega_2}$. Under the tacit assumption that each element $T \in \mathcal{T}_h$ belongs to only one subdomain, $\mathbf{L}_1$ can be obtained from (8.98) by replacing $\sum_{T\in\mathcal{T}_h}$ with $\sum_{T\in\mathcal{T}_h,\, T\subset\Omega_1}$. According to Remark 8.85c it is impossible to extract the matrices $\mathbf{L}_1$ and $\mathbf{L}_2$ from $\mathbf{L}$.

8.8.5 Positivity, Maximum Principle

Assume that the maximum principle holds for the (continuous) differential equation. In the construction of difference methods one often tries to satisfy the sign conditions (4.21a) to obtain the M-matrix property for the purpose of stability and to ensure the (discrete) maximum principle. So far the signs of the finite-element entries $\mathbf{L}_{ij}$ have not been discussed. First there is a negative statement: the continuous maximum principle does not necessarily imply the discrete one. In particular the off-diagonal entries $\mathbf{L}_{ij}$ $(i \neq j)$ may take the "wrong" sign.

In the case of finite-element discretisation of $L = -\Delta$ with piecewise linear elements one can get the following statement. Let the indices i, j belong to two neighboured corner points $\mathbf{x}^i$ and $\mathbf{x}^j$. The intersection of the supports of the basis functions b_i and b_j consists of the two triangles T' and T'' in Figure 8.16. Let α' and α'' be the angles in T' and T'' opposite to the side $\overline{\mathbf{x}^i\mathbf{x}^j}$. Then the identity

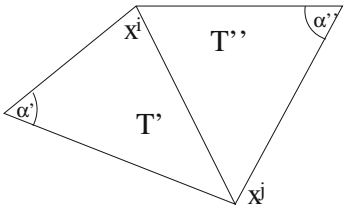


Fig. 8.16 Angles α' and α''.

$$\mathbf{L}_{ij} = \int_{T'\cup T''} \langle \nabla b_i, \nabla b_j \rangle \, \mathrm{d}\mathbf{x} = -\frac{1}{2}\frac{\sin(\alpha' + \alpha'')}{\sin(\alpha')\sin(\alpha'')},$$

is proved in Knabner–Angermann [172, §3.9]. Hence the desired sign condition $\mathbf{L}_{ij} \leq 0$ holds if and only if $\alpha' + \alpha'' \leq \pi$. For triangles touching the boundary additional conditions can be found in [172, page 173].

8.9 Further Remarks

There are many more details and modifications of the finite-element method which are not discussed here. In the following we mention some topics.

8.9.1 Mixed and Hybrid Finite-Element Methods

Although the name 'mixed/hybrid finite elements' suggests special finite elements (i.e., a special choice of the subspace), this term denotes a particular reformulation of the boundary-value problem.

The biharmonic equation $\Delta^2 u = f$ with Dirichlet conditions $u = \frac{\partial u}{\partial n} = 0$ (cf. (5.26), (5.27)) has the variational formulation with $a(u,v) = \int_\Omega (\Delta u)(\Delta v) \mathrm{d}\mathbf{x}$. However, a conforming finite-element discretisation $V_N \subset H_0^2(\Omega)$ requires elements which are continuously differentiable across the edges of triangles. Such finite elements can be constructed, but their use is costly. On the other hand, the equation of fourth order can be reformulated by two equations of second order. We introduce the auxiliary variable $v := -\Delta u$. The system $-\Delta v = f,\ -\Delta u = v$ with $u = \frac{\partial u}{\partial n} = 0$ on Γ has the variational form

$$a_1(v,\varphi) := \int_\Omega \langle \nabla v, \nabla \varphi \rangle \, \mathrm{d}\mathbf{x} = \int_\Omega f\varphi \mathrm{d}\mathbf{x}, \quad a_2(u,\psi) := \int_\Omega \langle \nabla v, \nabla \psi \rangle \, \mathrm{d}\mathbf{x} = \int_\Omega v\psi \mathrm{d}\mathbf{x}$$

for $u, \varphi \in H_0^1(\Omega),\ v, \psi \in H^1(\Omega)$. Note that the arguments of a_1 and a_2 belong to different spaces. We obtain a more familiar form after introducing the spaces $X := H_0^1(\Omega) \times H^1(\Omega)$ and $x = \binom{u}{v},\ y = \binom{\varphi}{\psi} \in X$:

$$c(x,y) = c\left(\binom{u}{v}, \binom{\varphi}{\psi}\right) := \int_\Omega \Big\{ \langle \nabla v, \nabla \varphi \rangle + \langle \nabla v, \nabla \psi \rangle - v\psi \Big\} \mathrm{d}\mathbf{x}$$

is a bilinear form on $X \times X$. The system from above becomes $c\big(\binom{u}{v}, \binom{\varphi}{\psi}\big) = \int_\Omega f\varphi \mathrm{d}\mathbf{x}$. The conforming finite-element discretisation may use simple piecewise linear elements. Concerning the analysis of the problem we refer to Ciarlet–Raviart [71] and Monk [207]. The uniform treatment of the components u and v in the mixed variational formulation cannot hide the fact that the reformulation does not change the smoothness properties. From $u \in H_0^2(\Omega)$ we only obtain $v \in L^2(\Omega)$. If the extra smoothness $v \in H^1(\Omega)$ does not hold, it cannot be the solution of the new formulation.

Also the Poisson problem $-\Delta u = f$ in Ω and $u = g$ on Γ can be split. The Laplace operator is the product $\Delta = \operatorname{div} \nabla$. As above we introduce a vector-valued variable $v := \nabla u$. The equation $-\Delta u = f$ becomes $-\operatorname{div} v = f$ and $\nabla u = v$. The variation formulation is

$$c\left(\binom{u}{v}, \binom{\varphi}{\psi}\right) := \int_\Omega \big\{ \langle v, \psi \rangle + u \operatorname{div} \psi - (\operatorname{div} v)\, \varphi \big\} \mathrm{d}\mathbf{x} = \int_\Omega f\varphi \mathrm{d}\mathbf{x} + \int_\Gamma g \, \langle \psi, \mathbf{n} \rangle \, \mathrm{d}\mathbf{x}$$

with $v, \psi \in H(\mathrm{div})$ and $u, \varphi \in L^2(\Omega)$ (cf. Raviart–Thomas [232, 233]), where

$$H(\mathrm{div}) = \big\{ u \in [L^2(\Omega)]^n : \operatorname{div} u \in L^2(\Omega) \big\} .$$

In this case the ansatz for u may be discontinuous. In $H(\mathrm{div})$ discontinuous functions may have a continuous divergence. Concrete conforming finite-element functions for $H(\mathrm{div})$ and for the analogous space $H(\mathrm{curl})$ are described by Nédélec [208, 209].

The book of Brezzi–Fortin [55] is a standard literature for mixed finite elements. Concerning the use of Raviart–Thomas elements for the Maxwell system we refer to Hiptmair [153, 154].

The variational problems constructed above are saddle-point problems which are more closely studied in Chapter 12.

8.9.2 Nonconforming Elements

Condition (8.3), $V_h \subset V$, characterises the conforming finite-element methods. Discretisations based on $V_h \not\subset V$ are called *nonconforming*. An example of a nonconforming element is the following 'Wilson rectangle'. We consider a partition into rectangles $R_i = (x_{i1}, x_{i2}) \times (y_{1i}, y_{2i})$ and define quadratic functions on R_i:

$$V_h := \{u \text{ quadratic on each } R_i,\ u \text{ continuous at all corners of } R_i\}\,.$$

Note that $u \in V_h$ is only continuous at the nodal points (corners of R_i), but may be discontinuous across the edges. Therefore these functions cannot belong to $H^1(\Omega)$: $V_h \not\subset V = H^1(\Omega)$. A possible basis for V_h is the following. For each of the four nodes $\mathbf{x}^j$ let b_j be the basis function of the bilinear elements, $b_i^{(1)}(x,y) := \frac{(x_{2i}-x)(y_{2i}-y)}{(x_{2i}-x_{1i})(y_{2i}-y_{1i})}$ (cf. §8.4.3). The ansatz chosen above has two more degrees of freedom. Therefore for each rectangle $R_i = (x_{1i}, x_{2i}) \times (y_{1i}, y_{2i})$ we add the two basis functions

$$b_i^{(5)}(x,y) := (x - x_{1i})\,(x_{2i} - x)\,, \qquad b_i^{(6)}(x,y) := (y - y_{1i})\,(y_{2i} - y)\,,$$

and extent these functions by zero outside of R_i. Since $b_i^{(5)}$ and $b_i^{(6)}$ vanish at all four corners of R_i, the required continuity is ensured.

The first difficulty arising from nonconforming elements is the definition of the system matrix $\mathbf{L}$. It may happen that the definition by $L_{ij} = a(b_j, b_i)$ as in (8.8a) does not make sense since $a(\cdot,\cdot)$ need not be defined for $b_j, b_i \notin V$! Therefore the replacement of V by V_h must be go along with a replacement of $a(\cdot,\cdot) : V \to V$ by a bilinear form $a_h(\cdot,\cdot) : V_h \times V_h \to \mathbb{R}$. Then we can set $L_{ij} := a_h(b_j, b_i)$. As a kind of consistency we require $a(u,v) = a_h(u,v)$ for all $u, v \in V$. Note that the new bilinear form is dependent of V_h. In the case of Wilson's rectangle and the decomposition of Ω into rectangles R_i we define $a_h(\cdot,\cdot)$ by

$$a_h(u,v) := \sum_{|\alpha|\le m} \sum_{|\beta|\le m} \sum_i \int_{R_i} a_{\alpha\beta}(x)\,[D^\alpha u(x)]\,[D^\beta v(x)]\mathrm{d}\mathbf{x}$$

using the coefficients from $a(\cdot,\cdot)$ in (7.7), i.e., the integration $\int_\Omega$ is replaced by $\sum_i \int_{R_i}$.

Conforming finite-element discretisations always yield consistent discretisations. In the case of nonconforming methods consistency is not guaranteed. Concerning this problem and in particular the so-called 'patch test' we refer to Strang–Fix [276, page 174], Ciarlet [67, §4.2], Stummel [278], Gladwell–Wait [119, pages 83–92], Thomasset [284]. The next exercise shows that seemingly reasonable choices of V_h may produce completely wrong discretisations.

Exercise 8.86. Let $\mathcal{T}$ be an admissible triangulation and define $V_h := \{u \text{ constant on all } T \in \mathcal{T}\}$, where h is the longest side of the triangles. Prove the following:
(a) $V_h \not\subset V := H^1(\Omega)$.
(b) $\inf\{|u-v|_0 : v \in V_h\} \le \frac{h}{\pi}|u|_1$ for all $u \in V$ looks promising, but does not help concerning consistency.
(c) Let $a(u,v) := \int_\Omega [\langle \nabla u, \nabla v\rangle + uv]\,\mathrm{d}\mathbf{x}$, $a_h(u,v) := \sum_{T\in\mathcal{T}} \int_T [\langle \nabla u, \nabla v\rangle + uv]\,\mathrm{d}\mathbf{x}$. What are the entries $L_{ij} = a_h(b_j, b_i)$?

Nonconforming elements are of particular interest for equations of higher order. Conforming finite elements in $V = H^2(\Omega)$ must be continuously differentiable across the elements which requires complicated constructions. Any simplification necessarily leads to $V_h \not\subset V$.

Exactly speaking, in the case of $V = H_0^1(\Omega)$, also the isoparametric finite elements in Section 8.6.3 belong to the nonconforming methods since the elements of V_h approximate zero on Γ but are not exactly zero, so that $V_h \not\subset V = H_0^1(\Omega)$. However, still $V_h \subset H^1(\Omega)$ holds so that the bilinear form $a(\cdot,\cdot)$ is well defined and need not replaced by another bilinear form $a_h(\cdot,\cdot)$.

The following statement about the error estimate of nonconforming methods is called the *second Lemma of Strang* (cf. Strang [275]). Here, a_h must be a continuous bilinear form on $Z_h \times Z_h$ where $Z_h := V + V_h = \{v + v_h : v \in V, v^h \in V_h\}$.

Theorem 8.87. *Let the bilinear form a_h be continuous, symmetric and positive definite on $Z_h \times Z_h$ so that $|||z|||_h := \sqrt{a_h(z,z)}$ defines a norm on Z_h. Let $u^h \in V_h$ and $u \in V$ be the respective solutions of $a_h(u^h, v) = f_h(v)$ $(v \in V_h)$ and (8.1). Then there holds*

$$|||u-u^h|||_h \le \text{const}\left\{\|f_h - a_h(u,\cdot)\|_{V_N'} + \inf_{z^h\in V_h} |||u-z^h|||_h\right\}.$$

Proof. For all $z^h, v^h \in V_h$ we have

$$\begin{aligned} a_h(u^h - z^h, v^h) &= a_h(u - z^h, v^h) + a_h(u^h, v^h) - a_h(u, v^h) \\ &= a_h(u - z^h, v^h) + f_h(v^h) - a_h(u, v^h). \end{aligned}$$

Taking $v^h := u^h - z^h$, we obtain

$$\begin{aligned} |||u^h - z^h|||_h^2 &= a_h(u^h - z^h, u^h - z^h) \\ &= a_h(u - z^h, u^h - z^h) + f_h(u^h - z^h) - a_h(u, u^h - z^h) \\ &\le C_h |||u - z^h|||_h |||u^h - z^h|||_h + \|f_h - a_h(u,\cdot)\|_{V_N'} |||u^h - z^h|||_h, \end{aligned}$$

where C_h is the bound of a_h. Division by $|||u^h - z^h|||_h$ yields the inequality $|||u^h - z^h|||_h \le C_h |||u - z^h|||_h + \|f_h - a_h(u,\cdot)\|_{V_N'}$. In

$$|||u-u^h|||_h \le |||u-z^h|||_h + |||z^h-u^h|||_h \le (C_h+1)\,|||u-z^h|||_h + \|f_h-a_h(u,\cdot)\|_{V_N'}$$

we use the triangle inequality and the previous estimate and prove the theorem. ■

8.9.3 Inadmissible Triangulations

We discuss inadmissible triangulations using the example of Figure 8.17. The shown triangulation arises from a regular and admissible triangulation after refining the middle triangles. The admissibility conditions (8.35a–d) are violated since $\overline{T_0} \cap \overline{T_1}$ is a side of T_1, but not of T_0. Another inadmissibility concerns the point $P = \overline{T_0} \cap \overline{T_2}$ which is a vertex of T_2, but not of T_0. The points P, Q, R, S violating the admissibility conditions are called 'hanging nodes'. The other vertices $\mathbf{x}^i$ in Figure 8.17 are denoted by $i = 1, \dots, 17$.

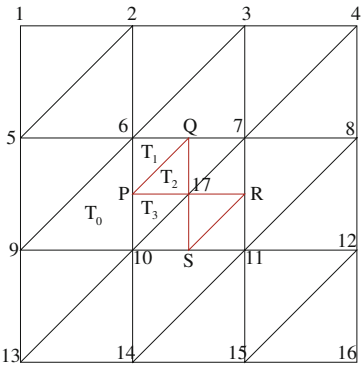


Fig. 8.17 Inadmissible triangulation.

The finite-element space (here without zero boundary condition) can be defined by (8.36) as before:

$$V_h = \left\{ u \in C^0(\overline{\Omega}) : u|_T \text{ linear on } T \in \mathcal{T}_h \right\}.$$

The difference between admissible and inadmissible triangulations becomes obvious when we ask for the dimension $\dim V_h$. Since, in particular, $u|_{T_0}$ is linear, the nodal value $u(P)$ is determined by $u(\mathbf{x}^6)$ and $u(\mathbf{x}^{10})$:

$$u(P) = \frac{1}{2}\left(u(\mathbf{x}^6) + u(\mathbf{x}^{10})\right). \qquad (8.99)$$

Therefore, different from Remark 8.39, one cannot prescribe arbitrary values of u at all nodes of the triangulation, but only at the regular, i.e., non-hanging nodes $\mathbf{x}^1, \dots, \mathbf{x}^{17}$. Correspondingly, the basis of V_h is given by $\{b_1, \dots, b_{17}\} \subset V_h$, where $b_i(\mathbf{x}^j) = \delta_{ij}$ $(1 \le i, j \le 17)$. If one computes the system matrix via the element matrices one has to note that the values $b_i(\mathbf{x}^{T,\nu})$ in (8.98) may have a value different from 0 or 1.

Remark 8.88. (a) Also for an inadmissible triangulation $\mathcal{T}_h$ one can construct the (conforming) finite-element space V_h. When defining the basis functions one has to distinguish between the regular and the hanging nodes.
(b) In the case of a finite-element space V_h that do not require continuity between elements[29] (for example piecewise constant functions) one need not distinguish between regular and the hanging nodes, i.e., in this case the admissibility condition can be omitted.

[29] Piecewise constant elements often appear in finite-element discretisations of integral equations. They may also be used for mixed formulations (see the second example in §8.9.1).

In the case of admissible triangulations the support of a basis function is easy to characterise. Remark 8.40 states that the support of b_i consists of all triangles attached to $\mathbf{x}^i$. This does not hold in the presence of hanging nodes. Figure 8.18 shows a part of a triangulation with hanging nodes P, Q, R. The basis function b_1 in the regular node $\mathbf{x}^1$ has the value $b_1(P) = 1/2$ at P because of (8.99). This induced the value $b_1(Q) = 1/4$ at Q and $b_1(R) = 1/8$ at R. Hence the support of b_1 consists of the triangles T_0, T_1, T_2, T_3 where the latter three triangles are only indirectly connected with $\mathbf{x}^1$. Obviously, this fact makes the computation of the finite-element matrix more complicated. A possible modification is the restriction to hanging nodes of first degree, i.e., triangle T_1 is not allowed to have a further hanging node Q (which would be of degree two).

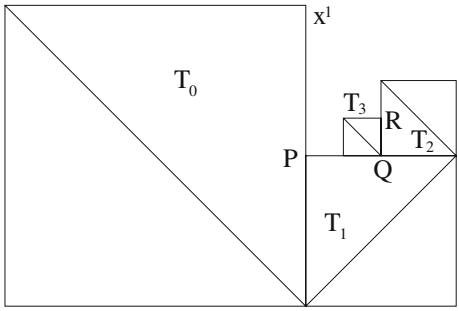


Fig. 8.18 Hanging nodes P, Q, R.

8.9.4 Trefftz' Method

Theorem 6.104 states that for symmetric and V-elliptic bilinear forms the weak formulation (8.1) of the boundary-value problem is equivalent to the minimisation problem

$$\text{find} \quad u \in V \quad \text{with} \quad J(u) := \min\{J(v) : v \in V\} \tag{8.100}$$

(see also Theorem 7.9 and Exercise 7.22). The maximisation problem

$$\text{find} \quad w \in W \quad \text{with} \quad K(w) := \max\{K(\hat{w}) : \hat{w} \in W\}$$

is called the *dual* (or *complementary*) *variational problem* of (8.100) if both have the same solution $u = w \in V \cap W$.

For instance, the Poisson problem $-\Delta u = f \in L^2(\Omega)$ in Ω, $u = 0$ on Γ, leads to

$$J(v) = \int_\Omega |\nabla v|^2 \, \mathrm{d}\mathbf{x} \qquad \text{for } v \in V := H_0^1(\Omega).$$

A particular dual variational problem is due to Trefftz[30] [292, 293]:

$$K(w) := -\int_\Omega |\nabla w|^2 \, \mathrm{d}\mathbf{x} \qquad \text{for } w \in W := \{v \in H^1(\Omega) : -\Delta v = f\}$$

(cf. Velte [296, pages 91–101]). Note that the functions $v \in V$ satisfy the boundary condition $v = 0$ on Γ, whereas $w \in W$ requires no boundary condition, but the differential equation $Lw = f$.

[30] An acknowledgement of Erich Trefftz is given by Stein [267].

8.9.5 Finite-Element Methods for Singular Solutions

The previous error estimates as, e.g., (8.57) are based on the assumption $u \in H^2(\Omega) \cap V$, which will be discussed more closely in Section 9.1. The following example shows that this assumption may be wrong.

Example 8.89. (a) The Laplace equation $\Delta u = 0$ in the L-shaped domain Ω of Example 2.4 has the solution[31] $u = r^{2/3}\sin((2\varphi - \pi)/3)$ which only belongs to $H^s(\Omega)$ with $s < 1 + 2/3$.
(b) The solution $u = r^{1/2}\sin(\varphi/2)$ of Example 5.28 only belongs to $H^s(\Omega)$ with $s < 3/2$.

Consider again the L-shaped domain Ω and replace the Laplace equation by the more general Poisson equation $-\Delta u = f$ in Ω, $u = 0$ on Γ. Then one can show that for $f \in L^2(\Omega)$ the solution u can be split into

$$u = u_0 + \alpha\,\chi(r)\,r^{2/3}\sin\left(\tfrac{2\varphi-\pi}{3}\right) \quad \text{with } u_0 \in H^2(\Omega) \cap H_0^1(\Omega) \text{ and } \alpha \in \mathbb{R}$$

(cf. Strang–Fix [276, pages 257ff]). Here the radial function $\chi(r) \in C^\infty(\Omega)$ is an arbitrary cut-off function with $\chi(r) = 1$ in $0 \le r \le \frac{1}{4}$ and $\chi(r) = 0$ for $r \ge \frac{1}{2}$, so that $\chi(r)\,r^{2/3}\sin\left(\frac{2\varphi-\pi}{3}\right) \in H_0^1(\Omega)$ is guaranteed. See also Wahlbin [299].

In the cases described above the usual discretisation with linear elements but without local refinement only yields $|u - u^h|_1 = \mathcal{O}(h^s)$, $s < 2/3$. However, if one enriches the space V_h by the function $\chi(r)r^{2/3}\sin(\ldots)$ or similar functions, the approximation of u in V_h can be improved: $d(u, V_h) = \mathcal{O}(h)$, so that $|u - u^h|_1 = \mathcal{O}(h)$. Concerning this approach we refer to Babuška–Rosenzweig [20], Blum–Dobrowolski [41], as well as Gladwell–Wait [119, p. 119], Hackbusch [132]. Zenger–Gietl [317] describe quite a different approach in the case of difference methods: The difference scheme is suitably modified only at the corner so that the singular part of the solution is avoided. Egger–Rüde–Wohlmuth [94] describe and analyse an analogous finite-element approach.

The standard approach are traditional finite elements together with local grid refinement. If the location of the singularity is known, there are explicit constructions of the refined triangulation. Otherwise one uses the technique from §8.7.3.3. Concerning the error analysis we refer to Schatz–Wahlbin [255].

8.9.6 Hierarchical Bases

We illustrate the underlying idea by the one-dimensional case. Let $V_h \subset H_0^1(0,1)$ be the space of piecewise linear functions and equidistant partition of the interval $\Omega = (0,1)$ in subintervals of length h (cf. (8.32)). For $h = 1/2$ we have $\dim(V_{1/2}) = 1$ and b_1 in Figure 8.19a is the only basis function. Obviously $V_{1/4} \subset V_{1/2}$ are nested spaces. As basis functions in $V_{1/4}$ we take $b_1 \in V_{1/2}$

[31] In general, a domain $\Omega \subset \mathbb{R}^2$ with an inner angle $\omega \in (0, 2\pi)$ leads to a solution with a factor $r^{\pi/\omega}$. Such solutions belong to $H^s(\Omega)$ for all $s < 1 + \pi/\omega$.

and add the functions b_2, b_3 shown in Figure 8.19b. $\{b_1, b_2, b_3\}$ is a basis of $V_{1/4}$. Similarly we have $V_{1/8} \subset V_{1/4}$. b_1, b_2, b_3 are supplemented by $b_4, \ldots, b_7$ (cf. Figure 8.19c) resulting in a basis of $V_{1/8}$, the so-called *hierarchical basis*. In contrast to the previously used basis functions the supports of b_i need not be of length $\mathcal{O}(h)$, e.g., $\mathrm{supp}(b_1) = [0, 1]$. Therefore the system matrix $\mathbf{L}$ is not as sparse as usual. Nevertheless the choice of this basis has important advantages.

(i) The previously computed matrix $\mathbf{L} = \mathbf{L}^{(2h)}$ corresponding to V_{2h} is a principal submatrix of the system matrix $\mathbf{L} = \mathbf{L}^{(h)}$ corresponding to V_h, so that a (global or local) grid refinement does not require a new computation.

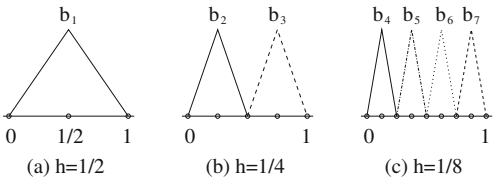


Fig. 8.19 Hierarchical basis.

(ii) The condition $\mathrm{cond}(\mathbf{L}) = \|\mathbf{L}\| \, \|\mathbf{L}^{-1}\|$ of the system matrix is essentially smaller than in the standard case (e.g., $\mathrm{cond}(\mathbf{L}) = \mathcal{O}(\log h)$ instead of $\mathcal{O}(h^{-2})$ in the two-dimensional case $\Omega \subset \mathbb{R}^2$; cf. Section 8.8).

(iii) For solving the system $\mathbf{Lu} = \mathbf{f}$ there exist suitable iterations. More details can be found in Yserentant [313], Bank–Dupont–Yserentant [25], Bank [23], and Hackbusch [142, §12.9.4].

8.9.7 Superconvergence

In the case of difference methods the error is given by the grid function $u_h - R_h u$ (u_h: difference solution, u: continuous solution). That means that the error is only defined at the grid points, whereas in the finite-element case the solution u^h and the error $u^h - u$ are defined in the complete domain Ω.

In principle the grid function u_h could be extended to a function u^h defined in Ω by interpolating the values $u_h(\mathbf{x}^i)$ ($\mathbf{x}^i$: grid point of Ω_h) by piecewise linear or bilinear finite elements: $u^h := Pu_h$ (P defined in (8.6)). Then the error $u^h - u$ contains an additional interpolation error $u - PR_h u$:

$$u^h - u = Pu_h - u = P(u_h - R_h u) - (u - PR_h u).$$

If the interpolation error is at least as small as $u_h - R_h u$, one obtains the same convergence behaviour; otherwise $Pu_h - u$ converges slower than $u_h - R_h u$.

Vice versa, in the finite-element case one can ask whether there is a better convergence behaviour at the nodal values $u_h = \{u^h(\mathbf{x}^i) : 1 \le i \le N_h\}$ or for suitable mean values. If there is a better behaviour, this is called the *superconvergence*. This also applies to difference quotients of the nodal values if these are more precise than $|D^\alpha u^h - D^\alpha u|_0 \le |u^h - u|_1$ ($|\alpha| = 1$). For this topic we refer to Wahlbin [300] and, e.g., Lesaint–Zlámal [191], Zlámal [324, 323], Bramble–Schatz [51], Thomée [286], Louis [197], Großmann–Roos–Stynes [124, §4.9.3]. Some error estimators are based on superconvergence results for gradients (cf. Verfürth [297, pages 36ff]).

Occasionally, the error estimate (8.77) for a negative t is also called superconvergence, since the estimate is by a factor h^{-t} better than the statement for $t = 0$.

8.9.8 Mortar Finite Elements

8.9.8.1 Introduction

So far we assumed that the whole domain Ω is covered by one connected finite-element grid satisfying the admissibility condition (8.35a–d). Even in the case of hanging nodes there are still common nodal point. Figure 8.20 shows a domain Ω split into two subdomains Ω_1 and Ω_2. The common interior boundary is denoted by $\Gamma_{12} := \overline{\Omega_1} \cap \overline{\Omega_2}$. In both subdomains the triangulations $\mathcal{T}_1$ and $\mathcal{T}_2$ are constructed independently so that *non-matching* grids result. Since $\mathcal{T}_i$ are assumed to be admissible triangulations of Ω_i it follows from (8.35c) that $\bigcup_{T\in\mathcal{T}_i} \overline{T} = \overline{\Omega_i}$ (here we assume that Ω_i is a polygon). Hence Γ_{12} lies in the union of the edges of $T \in \mathcal{T}_i$ for both $i = 1$ and $i = 2$. Hence both $\mathcal{T}_1$ and $\mathcal{T}_2$ cover Γ_{12}, but there is no connection between the vertices $\mathcal{T}_i$ $(i = 1, 2)$ lying on Γ_{12}.

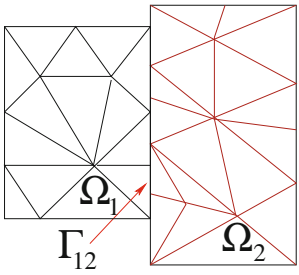


Fig. 8.20 Non-matching triangulations in two subdomains.

There may be several reasons to give up a global, admissible triangulation:

• Ω_1 and Ω_2 are modelled and discretised independently before—in a later step of the construction—the pieces are jointed together.

• Ω_1 and Ω_2 are subdomains whose data are associated to different processors. Possible grid refinements of a global, admissible triangulation would need communication which is avoided in the case of non-matching grids.

• In the case of instationary problems the domain Ω may change in time. For instance, the subdomain Ω_2 may move vertically while Ω_1 is fixed. To get a global, admissible triangulation one has to reconstruct a new grid in each time step and to transfer the data. These complications vanish if admissibility is required only locally in Ω_1 and Ω_2 (cf. Flemisch–Melenk–Wohlmuth [101]).

• *Contact problems* are characterised by two domains which in a certain way get into contact and form a common boundary (example: a sphere on an elastic body under gravitational force). We refer, e.g., to Hüeber–Wohlmuth [158].

To simplify the notation, consider the Poisson equation $-\Delta u = f$ in $\Omega \subset \mathbb{R}^2$ with Dirichlet condition $u = 0$ on $\partial\Omega$. This boundary-value problem in Ω is equivalent to the corresponding Poisson problems $-\Delta u^{(i)} = f|_{\Omega_i}$ in Ω_i, $u^{(i)}|_{\partial\Omega_i \setminus \Gamma_{12}} = 0$ $(i = 1, 2)$ together with the *matching conditions*

$$u^{(1)}|_{\Gamma_{12}} = u^{(2)}|_{\Gamma_{12}} \quad \text{and} \quad \partial u^{(1)}/\partial n|_{\Gamma_{12}} = \partial u^{(2)}/\partial n|_{\Gamma_{12}} .$$

Then $u^{(i)} = u|_{\Omega_i}$ $(i = 1, 2)$ are the restrictions of the global solution u to Ω_i.

On the finite-element level the matching conditions are expressed by a further variational formulation described below.

8.9.8.2 Variational Formulation in the Continuous Case

Set

$$X := \{v \in L^2(\Omega) \colon v|_{\Omega_i} \in H^1(\Omega_i),\ i=1,2,\ v|_{\partial\Omega}=0\}, \quad M = (H^{\frac{1}{2}}(\Gamma_{12}))'.$$

Functions $v \in X$ have two different boundary values on Γ_{12}. The restriction $v^{(1)} := v|_{\Omega_1}$ belongs to $H^1(\Omega_1)$ and defines the boundary value $v^{(1)}|_{\Gamma_{12}} \in H^{\frac{1}{2}}(\Gamma_{12})$. Similarly there is the boundary value $v^{(2)}|_{\Gamma_{12}} \in H^{1/2}(\Gamma_{12})$ of $v^{(2)} := v|_{\Omega_2}$. The jump

$$[v] := v^{(1)}|_{\Gamma_{12}} - v^{(2)}|_{\Gamma_{12}} \in H^{1/2}(\Gamma_{12})$$

appears in the second of the following bilinear forms:

$$\begin{aligned} a(u,v) &:= \sum_{i=1}^{2} \int_{\Omega_i} \langle \nabla u(x), \nabla v(x) \rangle \, \mathrm{d}\mathbf{x}, && \text{for } u, v \in X, \\ b(v,\mu) &:= (\mu, [v])_{L^2(\Gamma_{12})} = \langle \mu, [v] \rangle_{(H^{\frac{1}{2}}(\Gamma_{12}))' \times H^{\frac{1}{2}}(\Gamma_{12})} && \text{for } v \in X,\ \mu \in M. \end{aligned}$$

We are looking for the pair $(u, \lambda) \in X \times M$ with

$$\begin{aligned} a(u,v) + b(v,\lambda) &= (f,v)_{L^2(\Omega)} && \text{for all } v \in X, \\ b(u,\mu) &= 0 && \text{for all } \mu \in M. \end{aligned} \tag{8.101}$$

Assuming $u \in H^2(\Omega) \cap H_0^1(\Omega)$, integration by parts applied to the first part of (8.101) yields

$$-\int_\Omega v\, \Delta u \, \mathrm{d}\mathbf{x} + \int_{\Gamma_{12}} \left(v^{(1)} \frac{\partial u^{(1)}}{\partial n} - v^{(2)} \frac{\partial u^{(2)}}{\partial n} + [v]\, \lambda \right) \mathrm{d}\Gamma = \int_\Omega f\, v \, \mathrm{d}\mathbf{x}$$

($u^{(i)}$ is defined by the restriction $u|_{\Omega_i}$; $\mathbf{n}$ is the normal direction of Ω_1). Variation over $v \in H_0^1(\Omega) \subset X$ yields $-\Delta u = f$ in Ω. The remaining Γ_{12} integral gives

$$\begin{aligned} 0 &= \int_{\Gamma_{12}} \left(v^{(1)} \frac{\partial u^{(1)}}{\partial n} - v^{(2)} \frac{\partial u^{(2)}}{\partial n} + \left(v^{(1)} - v^{(2)} \right) \lambda \right) \mathrm{d}\Gamma \\ &= \int_{\Gamma_{12}} \left\{ v^{(1)} \left(\frac{\partial u^{(1)}}{\partial n} - \frac{\partial u^{(2)}}{\partial n} \right) + \left(v^{(1)} - v^{(2)} \right) \left(\frac{\partial u^{(2)}}{\partial n} + \lambda \right) \right\} \mathrm{d}\Gamma. \end{aligned}$$

Choosing $v^{(1)}$ and $v^{(2)}$ independently, we conclude that

$$\frac{\partial u^{(1)}}{\partial n} = \frac{\partial u^{(2)}}{\partial n} \qquad \text{and} \quad \lambda = -\frac{\partial u^{(2)}}{\partial n} \qquad \text{on } \Gamma_{12}.$$

The second part in (8.101) gives $[u] = 0$, i.e., $u^{(1)} = u^{(2)}$ on Γ_{12}.

8.9.8.3 Discrete Variational Formulation

Let X_h be the subspace of X, so that $u^h|_{\Omega_i}$ ($u^h \in X_h$) is the usual finite-element space of piecewise linear elements on the triangulation $\mathcal{T}_i$.

Let $\mathcal{T}_{12}$ be the restriction of the triangulation $\mathcal{T}_1$ on Γ_{12}. The one-dimensional grid $\mathcal{T}_{12}$ consists of edges $E \cap \Gamma_{12}$ (E: edge of $T \in \mathcal{T}_1$) and all vertices of $\mathcal{T}_1$ lying on Γ_{12}. The discrete space M_h consists of all piecewise linear functions on $\mathcal{T}_{12}$ vanishing at the endpoints of Γ_{12}. (Note that the construction is not symmetric: $\mathcal{T}_{21}$ as restriction of $\mathcal{T}_2$ on Γ_{12} yields another grid on Γ_{12}).

The variational formulation (8.101) with $X_h,\ M_h$ instead of $X,\ M$ defines the discrete problem.

8.9.8.4 Generalisation

The simple Poisson equation can be replaced by general differential equations and more other boundary conditions. In particular a decomposition of Ω in $k > 2$ subdomains is possible as require in the domain decomposition method. We refer to Bernardi–Maday–Patera [37], Wohlmuth [309], Kim–Lazarov–Pasciak–Vassilevski [171], and Großmann–Roos–Stynes [124, pp. 567–569]. Concerning the stability of the method see, e.g., Braess–Dahmen [46] and Dahmen–Faermann–Graham–Hackbusch–Sauter [81].

8.9.9 Composite Finite Elements

One of the advantages of the finite-element method is the fact that the triangulation can be adapted to the domain Ω (or at least to a polygonal approximation of the domain; cf. §8.6.3). However, a problem arises if the boundary contains small geometric details. Here we can consider two different cases. Let Ω be a domain of genus zero, so that $\partial\Omega$ is connected. Nevertheless, $\partial\Omega$ may be of tortuous shape like the coast line of an ocean. A second case is a domain with many small holes like a sieve. In the latter case, $\partial\Omega$ consists of many larger and smaller pieces. To resolve the geometry we need very small finite elements close to the boundary. Let h_{geom} be the size of these finite elements resolving the shape of the boundary. Inside of Ω there is an element size h_ε which is necessary to obtain an error of about ε. A conflict occurs if $h_{\mathrm{geom}} \ll h_\varepsilon$ since then the dimension of the finite-element problem is by far larger than for a smooth simple domain Ω. The additional degrees of freedom do not improve the accuracy.

The approach of the composite finite elements proposed by Hackbusch–Sauter [144, 145, 146, 147] avoids the unnecessary degrees of freedom. The method uses basis functions of size h_ε which are built by finite elements up to size h_{geom}. This allows an accurate approximation of the boundary, but does not create a higher

dimension of the composite finite-element space. The latter fact is in particular helpful for the solution of the linear system.

The upper Figure 8.21 shows a basis function satisfying homogeneous Dirichlet boundary conditions on the fine grid. The corresponding H^1-basis function is depicted below. For further details we refer to Sauter–Warnke [251, 252], Peterseim [220], Peterseim–Sauter [221, 222, 223], Preusser et al. [229].

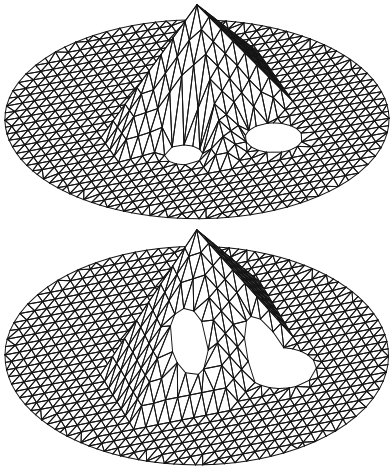


Fig. 8.21 Basis functions of the composite finite-element method.

A generalisation for handling geometric details in the domain (complicated coefficients, interior boundaries) is the adaptive local basis (*AL basis*) method (cf. Grasedyck–Greff–Sauter [120], Weymuth–Sauter [305]). The approximation of small details (e.g., oscillatory behaviour of the coefficients) within a coarser grid is called *homogenisation*. Here 'coarser' means that the finite element size is (much) larger than the size of the details. Such problems are called 'two-scale' (or *multiscale*) problems. In the irregular case one still needs a fine grid to describe the details. In a second step one tries to construct special coarse-grid finite elements. A very elegant approach in this direction with simple error estimates is described by Kornhuber–Yserentant [175], Kornhuber–Peterseim–Yserentant [174].

The approach described in Hackbusch [140, §12] is an approximate solution based of the original fine-grid data.

8.9.10 Related Discretisations

Besides the standard Galerkin method there are further variants which are briefly mentioned below.

8.9.10.1 Galerkin Method for the Least-Squares Formulation

The standard Galerkin discretisation minimises the energy norm if the problem is symmetric and V-elliptic. In the least-squares case one starts from $Lu - f = 0$ where u is restricted to an ansatz space V_N (e.g., with $V_N \subset H_0^1(\Omega)$), and tries to minimise the residual:

$$\text{minimise} \quad \|Lu - f\|_{L^2(\Omega)} \quad \text{over all} \quad u \in V_N$$

(cf. Bochev–Gunzburger [44] and Roos–Stynes–Tobiska [246, page 249]).

8.9.10.2 Discontinuous Galerkin Method

Standard finite-element functions are continuous. Also in the case of nonconforming elements continuity holds in a restricted sense. Elements which are completely discontinuous across the edges are more flexible and lead to the *discontinuous Galerkin finite element method* (DGFEM). In the meantime this variant is very popular. We refer to the proceedings Cockburn–Karniadakis–Shu [73] and the monographs Dolejší–Feistauer [88], Rivière [245], Di Pietro–Ern [84], Kanschat [166], Roos–Stynes–Tobiska [246, page 255], and Brenner–Scott [52, §10.5].

8.9.10.3 Petrov–Galerkin and Finite-Volume Methods

The Petrov–Galerkin method is characterised by different subspaces for the test and ansatz functions:

$$\text{find} \quad u^N \in V_N \quad \text{so that} \quad a(u^N, w) = f(w) \quad \text{for all } w \in W_N.$$

An obvious connection of the ansatz space V_N and the test space W_N is

$$\dim V_N = \dim W_N.$$

In particular in the case of unsymmetric bilinear forms there may be reasons to choose both spaces differently. This approach is due to Georgij Ivanovič Petrov [224] (1940, cf. [28]). Note that this discretisation is only reasonable if it is stable.

Also the bilinear form $a(\cdot,\cdot)$ of the original problem can be generalised to a bilinear form on $V \times W$ with different spaces $V \neq W$ (then the right-hand side must satisfy $f \in W'$). For finite-dimensional subspaces $V_N \subset V$ and $W_N \subset W$ the stability is equivalent to the inf-sup condition

$$\inf_{u \in V_N, \|u\|_V = 1} \sup_{w \in W_N, \|w\|_W = 1} |a(u,w)| \geq \varepsilon_N > 0.$$

In the following we consider a space W containing characteristic functions. Let $\omega \subset \Omega$ be a subset of Ω and consider the differential operator $Lu = \operatorname{div} a(\mathbf{x}) \nabla u$. Integration by parts yields

$$\int_\omega w\, Lu \,\mathrm{d}\mathbf{x} = \int_{\partial\omega} a(\mathbf{x}) w(\mathbf{x}) \frac{\partial u}{\partial n} \mathrm{d}\Gamma - \int_\omega a(\mathbf{x}) \left\langle \nabla w, \nabla u \right\rangle \mathrm{d}\mathbf{x}\,.$$

Choose $w(\mathbf{x})$ as the characteristic function of ω, i.e., $w = 1$ on ω and $w = 0$ outside:

$$\int_{\partial\omega} a(\mathbf{x}) \frac{\partial u}{\partial n} \mathrm{d}\Gamma = \int_\omega f \,\mathrm{d}\mathbf{x} \qquad \text{for all } \omega \in \mathcal{T}_{fV},$$

where the set $\mathcal{T}_{\mathrm{fV}}$ of *finite volumes* is still to be defined (traditionally even in the two-dimensional case ω is called the volume).

Let V_N be the usual space of piecewise linear elements on triangles of a triangulation $\mathcal{T}$. The dimension $N = \dim V_N$ is given by the number of vertices of $\mathcal{T}$. According to $\dim V_N = \dim W_N$ we choose a *dual grid* $\mathcal{T}_{\mathrm{fV}}$ which for each vertex $\mathbf{x} \in \mathcal{T}$ contains a 'finite volume' $\omega_{\mathbf{x}}$ (in general, this is a polygon; cf. Süli [279], Knabner–Angermann [172, §6 and §9.3] and Großmann–Roos–Stynes [124, §2.5]).

If $\Omega \subset \mathbb{R}^2$, the term 'box method' is a synonym for 'finite-volume method'. For differential operators $L = -\operatorname{div} a(\mathbf{x})\nabla$ consisting of the principal part only and under suitable assumption on the coefficient a, the box method and the standard Galerkin method with piecewise linear elements have identical system matrices (cf. Hackbusch [133]).

8.9.11 Sparse Grids

Since the physical space is three-dimensional, the standard problems are boundary-value problems in two or three spatial variables. However, there are also applications in higher dimensions $d > 3$. Then we confront the following problem. The accuracy of a discretisation depends on the grid size h which typically is $\mathcal{O}(h^\kappa)$ (κ: consistency order). Let $\Omega \subset \mathbb{R}^d$ be a domain discretised by a (finite element) grid of size h with N nodal points. Since N is proportional to h^{-d}, the discretisation error as a function of N becomes $\varepsilon_d(N) = \mathcal{O}(N^{-\kappa/d})$. As a consequence, for fixed N, $\varepsilon_d(N)$ tends to $\mathcal{O}(1)$ as $d \to \infty$. Vice versa, fixing ε, the problem size N must increase exponentially as $\mathcal{O}(\varepsilon^{-d/\kappa})$. This difficulty is often termed as the 'curse of dimensionality' (cf. Bellman [35, page 94]).

The following sparse-grid approach avoids the exponential behaviour $N = \mathcal{O}(\varepsilon^{-d/\kappa})$. We illustrate this method for the d-dimensional cube $\Omega = (0,1)^d$. By obvious reasons, the pictures refer to $d = 2$.

The bilinear finite elements explained in §8.4.3 can immediately be generalised to the d-dimensional case.

Let

$$b_i(x) = \max\left\{1 - \frac{x - ih}{h}, 0\right\}$$

be the one-dimensional piecewise linear basis function for the regular grid

$$0 < h < 2h < \ldots < 1 - h < 1$$

($h = 1/n$). Note that $b_i(jh) = \delta_{ij}$. The support of b_i is

$$[(i-1)\,h, (i+1)\,h] \cap [0,1].$$

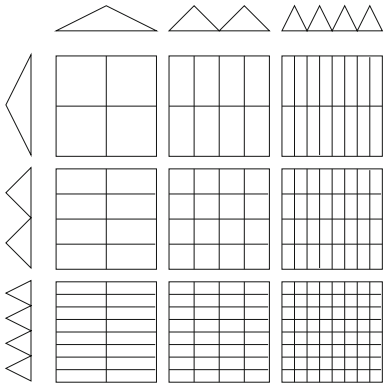


Fig. 8.22 All grids generated by the hierarchical basis

The d-linear basis functions on $\Omega = (0,1)^d$ are the products

$$B_{i_1 i_2 \ldots i_d}(x_1, x_2, \ldots, x_d) := b_{i_1}(x_1) \cdot b_{i_2}(x_2) \cdot \ldots \cdot b_{i_d}(x_d) \tag{8.102}$$

corresponding to the nodal point $\mathbf{x}_{i_1 i_2 \ldots i_d} = (i_1 h, i_2 h, \ldots, i_d h)$. Since $0 \le i_\nu \le n$ (or $1 \le i_\nu \le n-1$ for homogeneous Dirichlet values), we have $N := (n+1)^d$ (respectively $(n-1)^d$) grid points.

Now we recall the hierarchical basis in Section 8.9.6. Assume that $n = 2^L$ and introduce the step sizes $h_\ell = 2^{-\ell}$. On level[32] $\ell > 0$ we introduce the basis functions $b_i^{(\ell)}(x) = \max\{1 - \frac{x - i h_\ell}{h_\ell}, 0\}$ for odd $i = 1, 3, \ldots, 2^\ell - 1$. The hierarchical basis $\{b_i^{(\ell)} : 0 < \ell = L, 1 \le i \le 2^\ell - 1, i \text{ odd}\}$ spans the same space as the standard basis $\{b_i : 1 \le i \le n\}$. As in (8.102) we can form the products

$$B_{i_1 i_2 \ldots i_d}^{\ell_1 \ell_2 \ldots \ell_d}(x_1, x_2, \ldots, x_d) := b_{i_1}^{\ell_1}(x_1) \cdot b_{i_2}^{\ell_2}(x_2) \cdot \ldots \cdot b_{i_d}^{\ell_d}(x_d)$$

and the span of all $B_{i_1 i_2 \ldots i_d}^{\ell_1 \ell_2 \ldots \ell_d}$ (i_ν odd) coincides with the span of all $B_{i_1 i_2 \ldots i_d}$. The support of $B_{i_1 i_2 \ldots i_d}^{\ell_1 \ell_2 \ldots \ell_d}$ is anisotropic: $\overline{\Omega} \cap \times_{\nu=1}^{L} [(i_\nu - 1)\, h_{\ell_\nu}, (i_\nu + 1)\, h_{\ell_\nu}]$.

Figure 8.22 indicates the situation for $d = 2$. Combining the bases $\{b_i^{(\ell')}\}$ and $\{b_i^{(\ell'')}\}$ depicted at the top and at the left side of Figure 8.22 yields the corresponding grid with $\left(2^{\ell'} - 1\right)\left(2^{\ell''} - 1\right) \approx 2^{\ell' + \ell''}$ nodal points. The grids with large $\ell' + \ell''$ are the costly ones. However, it turns out that the costly grids are not necessary for obtaining the standard accuracy.

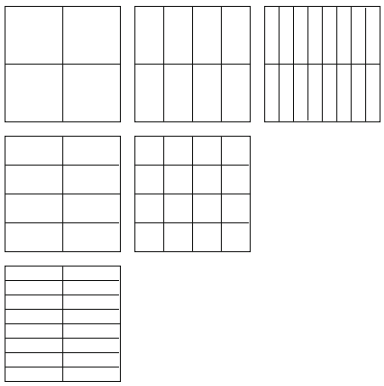


Fig. 8.23 Grids used in the sparse-grid method

The sparse-grid approximation uses the finite-element space spanned by all $B_{i_1 i_2 \ldots i_d}^{\ell_1 \ell_2 \ldots \ell_d}$ with

$$\ell_1 + \ell_2 + \ldots + \ell_d \le L + d - 1 \tag{8.103}$$

as depicted in Figure 8.23. The analysis presented in Bungartz–Griebel [60] shows that the grids omitted in Figure 8.23 yield contributions that can be neglected (a shorter description is given by Garcke [109]).

We have to assume that the function u has bounded second *mixed* derivatives, i.e., $\|D^\alpha u\|_{L^2(\Omega)}$ must be bounded for all multi-indices α with $|\alpha_j| \le 2$ (instead of the usual inequality $\alpha_1 + \ldots + \alpha_d \le d$). Then the finite-element space V_L spanned by

[32] On level $\ell = 0$ the basis functions are $b_i^{(0)}$ for $i = 0$ and $i = 1$, if inhomogeous boundary values are needed. In the homogeous case, there are no basis function at level 0.

$B_{i_1 i_2 \dots i_d}^{\ell_1 \ell_2 \dots \ell_d}$ with (8.103) leads to the solution u_L with the error estimate[33]

$$\|u - u_L\|_{L^2(\Omega)} \leq \mathcal{O}(h_L^2 L^{d-1}) = \mathcal{O}(h_L^2 \, |\log h_L|^{d-1})$$

(cf. [60, Theorem 3.8], [109, Theorem 1]). This is the standard estimate of the full grid except the logarithmic factor. On the other hand the dimension of the sparse-grid space is

$$\dim(V_L) = \mathcal{O}(h_L^{-1} \, |\log h_L|^{d-1})$$

(cf. [60, Lemma 3.6], [109, Lemma 6]). Up to the logarithmic factor, this is the dimension of the space of piecewise linear functions in *one* spatial dimension.

The sparse-grid method is introduced by Zenger [316] in 1991. It is also called *hyperbolic cross method* or Smolyak method (cf. Smolyak [265]).

[33] In [60, 109] these estimates are proved for the interpolation error. Therefore they are also an estimate of the finite-element method. The error estimate with respect to the H^1 seminorm holds without logarithmic factor; cf. [60, Theorem 3.8].

Chapter 9
Regularity

Abstract The previous results can only guarantee the existence of weak solutions, i.e., in the case of $m = 1$ only first derivatives in $L^2(\Omega)$ can be proven. In the beginning we also asked for second derivatives satisfying the equation $Lu = f$ in the classical sense. Also the error estimate in §8.5 has shown that an error of size $\mathcal{O}(h^k)$ requires a solution in $H^{1+k}(\Omega)$. Therefore the crucial question is, under what conditions the weak solution also belongs to Sobolev spaces of higher order (cf. **Section 9.1**). **Section 9.2** characterises a specific property of elliptic solutions: In the interior of the domain the solution is smoother than close to the boundary. In the case of analytic coefficients the solution is also analytic in the interior and the bounds of the (higher) derivatives improve with the distance from the boundary. This behaviour also holds for the singularity and Green's function. In **Section 9.3** the regularity properties of solutions of difference schemes is studied. When comparing the error estimates for difference methods in §4.5 with those for finite-element estimates in §8.5 one observes that the latter require much weaker smoothness of the solution. However, one gets similar estimate for difference methods if one uses suitable discrete regularity properties (cf. §9.3.3). Unfortunately, the proof of these properties is rather technical, much more involved, and inflexible compared with the finite-element case.

9.1 Solutions of the Boundary-Value Problem in $H^s(\Omega)$, $s > m$

In §9.1.2 we start the analysis with the case of $\Omega = \mathbb{R}^n$ and prove that the solution is smooth if the coefficients of the differential operator are sufficiently smooth. The next case considered in §9.1.3 is the half-space $\Omega = \mathbb{R}^n_+$. Here a boundary $\Gamma = \partial\Omega$ occurs, and the essential tool of the analysis is the continuous extension of functions from $\mathbb{R}^n_+$ to $\mathbb{R}^n$. The last step leads to a general bounded domain Ω (cf. §9.1.4). As soon as the domain and the coefficients of the differential equation are sufficiently smooth, also the solution is so. Special results hold for convex domains as discussed in §9.1.5.

W. Hackbusch, *Elliptic Differential Equations*, Springer Series
in Computational Mathematics 18, DOI 10.1007/978-3-662-54961-2_9

9.1.1 The Regularity Problem

The weak formulation of a boundary-value problem

$$Lu = g \text{ in } \Omega, \quad Bu = \varphi \text{ on } \Gamma \quad (2m\text{: order of the differential operator } L) \tag{9.1}$$

as

$$u \in V, \qquad a(u,v) = f(v) \quad \text{for all } v \in V \tag{9.2}$$

was, in Chapter 7, the basis upon which we were able to answer the questions of existence and uniqueness of the solution. Here, by existence of a solution we understand the existence of a weak solution $u \in V$.

The error estimates in Section 8.5 made it clear that the statement $u \in V$ is not enough. Under this assumption we can show $\|u^h - u\|_V \to 0$, but the convergence may be arbitrarily slow. The more interesting quantitative estimate $\|u^h - u\|_V = \mathcal{O}(h)$ as in, for example, Theorem 8.60 for $V = H_0^1(\Omega)$ or $V = H^1(\Omega)$, requires the assumption $u \in H^2(\Omega) \cap V$. The assertion $u \in H^2(\Omega)$ or, more generally $u \in H^s(\Omega)$, is a statement of *regularity*, i.e., a statement about the smoothness of the solution, which will be examined in greater detail in this section. If $s - \frac{n}{2}$ is sufficiently large, Sobolev's embedding Theorem 6.48 proves that the solution is smooth in the sense of classical function spaces. In particular, for $s - \frac{n}{2} > 2m$ the solution u is a classical solution of the boundary-value problems.

The regularity proofs in the following sections are very technical. To make the proof ideas clearer, let us sketch the proof of inequality (9.4) below for the Helmholtz equation $-\Delta u + u = f$ in $\Omega \subset \mathbb{R}^2$, $u = 0$ on Γ.

Step 1: $\Omega = \mathbb{R}^2$. Since the bilinear form $a(u,v) = \int_\Omega \langle \nabla u, \nabla v \rangle \mathrm{d}\mathbf{x} + \int_\Omega uv \mathrm{d}\mathbf{x}$ is $H^1(\mathbb{R}^2)$-elliptic, (9.4) holds for $s = m = 1$: $|u|_1 \le C_1' |f|_{-1}$. This is the start of an induction proof of (9.4) for $s = 1, 2, 3, \ldots$. First we want to prove (9.4) for $s = 2$. To this end we take the derivative of the differential equation with respect to x, the component in $\mathbf{x} = (x, y)$:

$$-\Delta u_x + u_x = f_x.$$

If $f \in H^0(\Omega)$, then $f_x \in H^{-1}(\Omega)$ and the equation $-\Delta v + v = f_x$, by the induction assumption, has a unique solution $v \in H^1(\Omega)$ with $|v|_1 \le C_1' |f_x|_{-1} \le C_1' |f|_0$. If one sets up this inequality for $v = u_x$, and likewise for $v = u_y$, the result is $|u|_2 \le |u|_1 + |u_x|_1 + |u_y|_1 \le 3C_1' |f|_0$. Thus, (9.4) has been shown for $s = 2$. The further induction steps for $s = 3, 4, \ldots$ are analogous.

Step 2: $\Omega = \mathbb{R}^2_+ := \mathbb{R} \times (0, \infty)$. As in the previous case, we can obtain the estimate $|u_x|_1 \le C_1' |f|_0$, since u_x also satisfies $-\Delta u_x + u_x = f_x$ and the (tangentially differentiated) boundary condition $u_x = 0$ on Γ. This is not the case for u_y. But $u_x \in H^1(\mathbb{R}^2_+)$ implies $u_{xx} \in H^0(\mathbb{R}^2_+)$ and $u_{xy} \in H^0(\mathbb{R}^2_+)$. We would have $u \in H^s(\mathbb{R}^2_+)$ if we could also show $u_{yy} \in H^0(\mathbb{R}^2_+)$. This property, however, results from the differential equation, $u_{yy} = \Delta u - u_{xx} = u - f - u_{xx} \in H^0(\mathbb{R}^2_+)$.

Step 3: Let Ω be arbitrary, but sufficiently smooth. As in Section 6.1, Ω is decomposed into (overlapping) pieces Ω_i which can be mapped into $\mathbb{R}^2$ or $\mathbb{R}^2_+$. Correspondingly one splits the solution u into $\sum \chi_i u$ (χ_i is a partition of unity). Then the arguments from steps 1 and 2 prove inequalities for $|\chi_i u|_2$ which together result in (9.4).

Note that only a sketch of the proof was given. Some of the steps of the proof are incomplete. For example, might not the equation $-\Delta u_x + u_x = f_x$ have a solution[1] $u_x \in L^2(\mathbb{R}^2)$, which does not belong to $H^1(\mathbb{R}^2)$ and hence does not coincide with the solution $v \in H^1(\mathbb{R}^2)$ of $-\Delta v + v = f_x$?

In the following, always let $s \geq m$. The boundary-value problem (9.2) with $V = H_0^1(\Omega)$ is said to be H^s*-regular* if each solution $u \in H_0^1(\Omega)$ of problem (9.2) with $f \in H^{s-2m}(\Omega)$ belongs to $H^s(\Omega) \cap H_0^1(\Omega)$ and satisfies the estimate

$$|u|_s \leq C_s \left[|f|_{s-2m} + |u|_m \right]. \tag{9.3}$$

If L is the operator associated with $a(\cdot,\cdot)$, then it is also said that L is H^s-regular.

Remark 9.1. Let m be defined as in (9.1). (a) H^m-regularity always holds.

(b) Let the variational problem (9.2) have a unique solution $u \in H_0^m(\Omega)$ with $|u|_m \leq C_0 |f|_{-m}$ for all $f \in H^{-m}(\Omega)$. If the boundary-value problem is H^s-regular, then the weak solution of (9.2) with $f \in H^{s-2m}(\Omega)$ satisfies the inequality

$$|u|_s \leq C_s' \, |f|_{s-m} \,. \tag{9.4}$$

(c) Let L be the operator associated with $a(\cdot,\cdot)$. Inequality (9.4) is equivalent to $L^{-1} \in L(H^{s-2m}(\Omega), H^s(\Omega))$ and the following statement:

$$\|L^{-1}\|_{H^s(\Omega) \leftarrow H^{s-2m}(\Omega)} \leq C_s' \,.$$

Proof. (a) (9.3) holds with $C_m = 1$ for all $u \in H^m(\Omega)$ because $|u|_s = |u|_m$.

(b) In (9.3) estimate $|u|_m$ by $C_0 \, |f|_{-m}$ and use the fact that because of $s \geq m$ the embedding $H^{s-2m}(\Omega) \subset H^{-m}(\Omega)$ is continuous, i.e., $|f|_{-m} \leq C' \, |f|_{s-2m}$. Thus we obtain (9.4) with $C_s' = C_s(1 + C_0 C')$. ∎

The following remark shows that a perturbation of $a(\cdot,\cdot)$ by a smooth term of order $< 2m$ does not change the H^s-regularity.

Remark 9.2. Let $m \leq s$ and $m, s \in \mathbb{N}$. Let the operator L be H^t-regular for all $t \in \{m, m+1, \ldots, s\}$. Let δL be an operator of order $\leq 2m-1$, i.e.,

$$\delta L \in L(H^r(\Omega) \cap H_0^m(\Omega), H^{r+1-2m}(\Omega)) \quad \text{for all } r \in \{m, m+1, \ldots, s-1\}.$$

Then $L + \delta L$ is also H^t-regular for $t = m, m+1, \ldots, s$. The assumption on δL holds in particular if δL belongs to the bilinear form $a''(\cdot,\cdot)$ in Lemma 7.12 and its coefficients $a_{\alpha\beta}$ are sufficiently smooth: for example, $a_{\alpha\beta} \in C^{\max\{0, s-2m+|\beta|\}}$.

[1] Compare Lions–Magenes [194, Chap. 2, §6] for *very weak* solutions.

Proof. We start the proof by induction with $s = m$ since for $s = m$ the statement follows from Remark 9.1a. Assume the assertion for $s-1$. Let u be the weak solution of $(L + \delta L)u = f \in H^{s-2m}(\Omega)$, hence also the solution of $Lu = \tilde{f} := f - \delta L u$. By induction the inequality $|u|_{s-1} \le C'_{s-1}[\,|f|_{s-2m-1} + |u|_m\,]$ already holds and shows that

$$|\tilde{f}|_{s-2m} \le |f|_{s-2m} + \|\delta L\|_{s-2m\leftarrow s-1}\,|u|_{s-1} \le C''_{s-1}[\,|f|_{s-2m} + |u|_m\,].$$

Because of the H^s-regularity of L, the solution u belongs to $H^s(\Omega)$ and satisfies $|u|_s \le C'_s[\,|\tilde{f}|_{s-2m} + |u|_m\,]$. Together these inequalities result in statement (9.4). ■

9.1.2 Regularity Theorems for $\Omega = \mathbb{R}^n$

The domain $\Omega = \mathbb{R}^n$ is distinguished by the fact that it has no boundary, and hence no boundary conditions either. Furthermore, the shifted version $v(\cdot + \delta)$ of $v \in H^s(\mathbb{R}^n)$ again belongs to $H^s(\mathbb{R}^n)$. The proof of the following theorem will be given on page 270.

Theorem 9.3 (regularity theorem). *Let $m \in \mathbb{N}$, $\Omega = \mathbb{R}^n$. Let the bilinear form*

$$a(u,v) := \sum_{|\alpha|\le m}\sum_{|\beta|\le m}\int_\Omega a_{\alpha\beta}(\mathbf{x})\,[D^\alpha u(\mathbf{x})]\,[D^\beta v(\mathbf{x})]\,\mathrm{d}\mathbf{x} \tag{9.5}$$

be $H^m(\mathbb{R}^n)$-coercive. For some $k \in \mathbb{N}$ let the following hold:

$$D^\gamma a_{\alpha\beta} \in L^\infty(\Omega) \qquad \text{for all } \alpha,\beta,\gamma \text{ with } |\gamma| \le \max\{0, k + |\beta| - m\}. \tag{9.6}$$

Then every weak solution $u \in H^m(\mathbb{R}^n)$ of the problem

$$a(u,v) = (f,v)_0 \qquad \text{for all } v \in H^m(\mathbb{R}^n) \tag{9.7}$$

with $f \in H^{-m+k}(\mathbb{R}^n)$ belongs to $H^{m+k}(\mathbb{R}^n)$ and satisfies the estimate

$$|u|_{m+k} \le C_k\left[\,|f|_{-m+k} + |u|_m\,\right]. \tag{9.8}$$

Corollary 9.4. Instead of inequality (9.8) one can also write

$$|u|_{m+k} \le C_{k,k-1}\left[\,|f|_{-m+k} + |u|_{m+k-1}\,\right]. \tag{9.8$'$}$$

Proof. (a) (9.8) $\Rightarrow$ (9.8′) follows from $|u|_m \le |u|_{m+k-1}$ because $k \ge 1$.

(b) If Theorem 9.3 holds for k, then also for $1, ..., k-1$ instead of k. Combining the inequalities (9.8′) for $k, k-1, \ldots, 1$, we obtain

$$|u|_{m+k} \le C_{k,k-1}\left\{|f|_{-m+k} + |u|_{m+k-1}\right\}$$

$$\leq C_{k,k-1}\left\{|f|_{-m+k}+C_{k-1,k-2}\left[|f|_{-m+k-1}+|u|_{m+k-1}\right]\right\}$$
$$\leq \ldots \leq C'\sum_{\ell=0}^{k}|f|_{-m+k-\ell}+C''|u|_m$$

and thus (9.8) since $|f|_{-m+k-\ell}\leq |f|_{-m+k}$ for $\ell\geq 0$. ∎

Lemma 9.5. *Let $a(\cdot,\cdot)$ be the bilinear form (9.5) with coefficients $a_{\alpha\beta}$ satisfying (9.6). Let $\ell\in\{0,\ldots,k\}$ and let $u\in H^{m+\ell}(\mathbb{R}^n)$ be chosen fixed.*
(a) Then $F:=a(u,\cdot)$ is a functional in $H^{-m+\ell}(\mathbb{R}^n)$ bounded by

$$|F(v)|=|a(u,v)|\leq C\,|u|_{m+\ell}\,|v|_{m-\ell}\,.$$

(b) $a_0(u,v):=\sum_{|\alpha|+|\beta|<2m}\int_\Omega a_{\alpha\beta}D^\alpha uD^\beta v\,\mathrm{d}\mathbf{x}$ — this is the bilinear form a without principal part — satisfies the stronger inequality

$$|a_0(u,v)|\leq C_0\,|u|_{m+\ell}\,|v|_{m-\ell-1}\,.$$

(c) Let D^γ, $|\gamma|=1$, be a first derivative. Define a bilinear form $b_\gamma(\cdot,\cdot)$ and a functional $b_\gamma(u,\cdot)$ by $b_\gamma(u,v):=a(D^\gamma u,v)+a(u,D^\gamma v)$. It satisfies

$$|b_\gamma(u,v)|\leq C\,|u|_{m+\ell}\,|v|_{m-\ell}\,.$$

Proof. (a) Let $b(u,v):=\int_{\mathbb{R}^n}a_{\alpha\beta}(D^\alpha u)(D^\beta v)\mathrm{d}\mathbf{x}$ be one of the terms in $a(u,v)$. Let $u\in H^{m+\ell}(\mathbb{R}^n)$. We distinguish the following three cases (a_1–a_3).

(a_1) Let $|\beta|\leq m-\ell$. Obviously

$$|b(u,v)|\leq\|a_{\alpha\beta}\|_\infty|u|_{|\alpha|}|v|_{|\beta|}\underset{|\alpha|\leq m+\ell}{\leq}\|a_{\alpha\beta}\|_\infty|u|_{m+\ell}|v|_{m-\ell}$$

holds. It remains to discuss the case $m-|\beta|<\ell$, i.e., $m-|\beta|+1\leq\ell$.

(a_2) Let $m-|\beta|+1\leq\ell\leq\min\{m,k\}$. Let $\gamma\leq\beta$ be a multi-index of length $|\gamma|=|\beta|-m+\ell\in[\,1,|\beta|\,]$. Integration by parts gives

$$b(u,v)=(-1)^{|\gamma|}\int_{\mathbb{R}^n}\left[D^\gamma\left(a_{\alpha\beta}D^\alpha u\right)\right]\left[D^{\beta-\gamma}v\right]\mathrm{d}\mathbf{x}\,. \qquad (*)$$

By assumption (9.6) the derivatives of $a_{\alpha\beta}$ are bounded so that

$$|b(u,v)|\leq C|u|_{|\alpha+\gamma|}|v|_{|\beta-\gamma|}\leq C|u|_{m+\ell}|v|_{m-\ell}.$$

(a_3) Let $\ell>m$. Equation (*) with $\gamma:=\beta$ shows that

$$b(u,v)=(-1)^{|\beta|}\int_{\mathbb{R}^n}\left[D^\beta(a_{\alpha\beta}D^\alpha u)\right]\,v\,\mathrm{d}\mathbf{x}=(g,v)_0\quad\text{with}$$
$$g:=(-1)^{|\beta|}\,D^\beta(a_{\alpha\beta}D^\alpha u)\in H^{m+\ell-|\alpha|-|\beta|}(\mathbb{R}^n).$$

Since $|\alpha|\leq m$ and $\ell>m\geq|\beta|$, it follows that $m+\ell-|\alpha|-|\beta|\geq\ell-m$, so that b can be estimated by $|b(u,v)|\leq|g|_{-m+\ell}\,|v|_{m-\ell}$. Together with

$$|g|_{-m+\ell} \le \left|D^\beta(a_{\alpha\beta}D^\alpha u)\right|_{-m+\ell} \le C\,|u|_{|\alpha|+|\beta|-m+\ell} \underset{|\alpha|+|\beta|\le 2m}{\le} C\,|u|_{m+\ell}$$

also in this case $|b(u,v)| \le C|u|_{m+\ell}|v|_{m-\ell}$ is proved. Summation over all α,β yields the desired result $|a(u,v)| \le C|u|_{m+\ell}|v|_{m-\ell}$.

(b) Let, e.g., $|\alpha| \le m$ and $|\beta| \le m-1$ for $b(u,v) := \int_{\mathbb{R}^n} a_{\alpha\beta}[D^\alpha u]\big[D^\beta v\big]\,\mathrm{d}\mathbf{x}$. If $|\beta| \le m-\ell-1$, $|b(u,v)| \le C_0|u|_{m+\ell}|v|_{m-\ell-1}$ is immediate. If $m-|\beta| \le \ell \le m-1$, we apply $|\beta|-m+\ell+1 \in [\,1,|\beta|\,]$ partial integrations and obtain $|b(u,v)| \le C|u|_{m+\ell}|v|_{m-\ell}$ because $|\alpha|+|\beta|+1-m+\ell \le 2m-m+\ell = m+\ell$. The case $|\alpha| \le m-1$ and $|\beta| \le m$ is left to the reader.

(c) Let $b_{\alpha\beta}(u,v) := \int_{\mathbb{R}^n} a_{\alpha\beta}D^\alpha u D^\beta v\mathrm{d}\mathbf{x}$ be a term of the bilinear form $a(u,v)$. Integration by parts shows

$$b_{\alpha\beta,\gamma}(u,v) = -\int_{\mathbb{R}^n} \left(D^\gamma a_{\alpha\beta}\right)\left(D^\alpha u\right)\left(D^\beta v\right)\mathrm{d}\mathbf{x}$$

$$\text{for } b_{\alpha\beta,\gamma}(u,v) := b_{\alpha\beta}(D^\gamma u, v) + b_{\alpha\beta}(u, D^\gamma v)$$

since $|\gamma| = 1$. Summation over α,β and analogous considerations as in part (i) prove $|b_\gamma(u,v)| \le C|u|_{m+\ell}|v|_{m-\ell}$.

Concerning the smoothness of the coefficients we observe that the bilinear form $b_\gamma(\cdot,\cdot)$ contains the coefficients $a_{\alpha\beta,\gamma} := D^\gamma a_{\alpha\beta}$ which fulfil (9.6) for $k-1$ instead of k. ∎

In Theorem 9.3 we only assume that $a(\cdot,\cdot)$ is $H^m(\mathbb{R}^n)$-coercive. Therefore a solution (e.g., an eigenfunction) may not be unique. Uniqueness is guaranteed, e.g., by $H^m(\mathbb{R}^n)$-ellipticity, which is discussed in part (b) of the next lemma.

Lemma 9.6. *Assume (9.5) and (9.6). Let a_H be the principal part:*

$$a_H(u,v) := \sum_{|\alpha|,|\beta|=m} \int_{\mathbb{R}^n} a_{\alpha\beta}\,[D^\alpha u]\,\big[D^\beta v\big]\,\mathrm{d}\mathbf{x}.$$

(a) If Theorem 9.3 holds for $a_H(\cdot,\cdot)$ and $k=1$, then also for the general bilinear form (9.5) and $k=1$.

(b) Without loss of generality, we may replace $a(\cdot,\cdot)$ in Theorem 9.3 by

$$a_\Lambda(u,v) = a(u,v) + \Lambda\int_{\mathbb{R}^n} u\,v\,\mathrm{d}\mathbf{x}$$

with $\Lambda \in \mathbb{R}$. Theorem 9.3 with $H^m(\mathbb{R}^n)$-ellipticity instead of $H^m(\mathbb{R}^n)$-coercivity is equivalent to the original form of Theorem 9.3. In part (a) $a_H(u,v)$ can be replaced by $a_H(u,v) + \Lambda\cdot(u,v)_0$.

Proof. (b) Let $a(\cdot,\cdot)$ be a $H^m(\mathbb{R}^n)$-coercive bilinear form with the constant C_K in (6.48). Then $a_\Lambda(\cdot,\cdot)$ for $\Lambda \ge C_K$ is an $H^m(\mathbb{R}^n)$-elliptic form. Let Theorem 9.3 hold for $a_\Lambda(\cdot,\cdot)$ and let u be a weak solution of (9.7): $a(u,v) = (f,v)_0$. Because of the $H^m(\mathbb{R}^n)$-ellipticity u is the *unique* solution of

$$a_{\mathbf{\Lambda}}(u,v) = F_{\mathbf{\Lambda}}(v) := (f+\Lambda u, v)_0\,.$$

Because of the inequality $|F_\Lambda(v)| \le |f|_{-m+k}||v|_{m-k} + |\Lambda|\,|u|_{-m+k}|v|_{m-k} \le \big[\,|f|_{-m+k}| + |\Lambda|\,|u|_{-m+k}\big]\,|v|_{m-k}$, F_Λ also belongs to $H^{-m+k}(\mathbb{R}^n)$. The assumption of Theorem 9.3 for $a_\Lambda(\cdot,\cdot)$ implies the regularity inequality (9.8′) in the form $|u|_{m+k} \le C_{k,k-1}\big[|F_\Lambda|_{-m+k} + |u|_{m+k-1}\big]$. Using the inequality $|F_\Lambda|_{-m+k} \le |f|_{-m+k}| + |\Lambda|\,|u|_{-m+k}$, we obtain

$$\begin{aligned}|u|_{m+k} &\le C_{k,k-1}\big[|f|_{-m+k}| + |\Lambda|\,|u|_{-m+k} + |u|_{m+k-1}\big]\\ &\le C'_{k,k-1}\big[|f|_{-m+k}| + |u|_{m+k-1}\big]\end{aligned}$$

since $|u|_{-m+k} \le |u|_{m+k-1}$. This proves the assertion of Theorem 9.3 for $a(\cdot,\cdot)$.

The replacement of $a(\cdot,\cdot)$ by $a_H(\cdot,\cdot)$ shows that also in part (a) we may add $\Lambda\cdot(u,v)_0$ without loss of generality.

(a) Let Theorem 9.3 hold for $a_H(\cdot,\cdot)$ and $k=1$, where according to part (b) $a_H(\cdot,\cdot)$ is assumed to be $H^m(\mathbb{R}^n)$-elliptic. Let $a_0 := a - a_H$ be the lower-order part. A solution $u \in H^m(\mathbb{R}^n)$ of $a_H(u,v) = (f,v)_0$ is the unique solution of

$$a_H(u,v) = F_H(v) := (f,v)_0 + a_0(u,v) \qquad \text{for all } v \in H^m(\mathbb{R}^n).$$

Lemma 9.5b with $\ell = 0$ applied to a_0 together with $|\,(f,v)_0\,| \le |f|_{-m+1}||v|_{m-1}$ yields

$$|F_H(v)| \le \big[\,|f|_{-m+1}| + C_0|u|_m\,\big]\,|v|_{m-1}$$

and thus $F_\Lambda \in H^{-m+1}(\mathbb{R}^n)$ for $\Lambda = 0$. Using

$$\begin{aligned}|u|_{m+1} &\le C_1\big[\,|F_H|_{-m+1} + |u|_m\,\big] \le C_1\,[\,|f|_{-m+1}| + C_0|u|_m + |u|_m\,]\\ &\le C'_1\big[\,|f|_{-m+1}| + |u|_m\,\big],\end{aligned}$$

we have proved (9.8) with $k=1$ for $a(\cdot,\cdot)$ instead of $a_H(\cdot,\cdot)$. ∎

Lemma 9.7. *Assume (9.5) and (9.6). If (9.8) holds for some $\ell \in \{1,\dots,k-1\}$ instead of k, so also for $\ell+1$.*

Proof. Let u be the solution of (9.7). By assumption, $u \in H^{m+\ell}(\mathbb{R}^n)$ holds with the estimate $|u|_{m+\ell} \le C_\ell\big[|f|_{-m+\ell} + |u|_m\big]$. Let D^γ, $|\gamma| = 1$, be a first derivative and let b_γ be the bilinear form from Lemma 9.5. Obviously $w := D^\gamma u$ is the solution of

$$a(w,v) = a(D^\gamma u, v) = -a(u, D^\gamma v) + b_\gamma(u,v) = (f, D^\gamma v)_0 + b_\gamma(u,v) =: F_\gamma(v).$$

Since $|\,(f,D^\gamma v)_0\,| \le |f|_{-m+k}|D^\gamma v|_{m-k} \le |f|_{-m+k}|v|_{m-k+1} \le |f|_{-m+k}|v|_{m-\ell}$ and $|b_\gamma(u,v)| \le C|u|_{m+\ell}|v|_{m-\ell}$, we conclude that

$$F_\gamma \in H^{-m+\ell}(\mathbb{R}^n) \qquad \text{with} \quad |F_\gamma|_{-m+\ell} \le |f|_{-m+k} + C|u|_{m+\ell}\,.$$

Theorem 9.3 with ℓ and F_γ instead of k and f shows that $w \in H^{m+\ell}(\mathbb{R}^n)$ and

$$\begin{aligned}|w|_{m+\ell} &\le C_\ell\Big[\,|F_\gamma|_{-m+\ell} + |w|_m\Big] \le C_\ell\,[\,|f|_{-m+k} + C|u|_{m+\ell} + |D^\gamma u|_m\,]\\ &\le C_\ell\,[\,|f|_{-m+k} + C'|u|_{m+\ell}\,]\,.\end{aligned}$$

Since the multi-index γ with $|\gamma| = 1$ in $w := D^\gamma u$ is arbitrary, we obtain $u \in H^{m+\ell+1}(\mathbb{R}^n)$ and the inequality

$$|u|_{m+\ell+1} \le (n+1)\, C_\ell \,[\, |f|_{-m+k} + C' |u|_{m+\ell} \,] .$$

Together with the regularity assumption $|u|_{m+\ell} \le C_\ell \left[|f|_{-m+\ell} + |u|_m \right]$ it follows that $|u|_{m+\ell+1} \le C_{\ell+1} \left[|f|_{-m+k} + |u|_m \right]$, i.e., the regularity statement for $\ell + 1$. ∎

Proof of Theorem 9.3. (i) First we must investigate the start of the induction, $k = 1$. According to Lemma 9.6a, $a(\cdot,\cdot)$ may be replaced by the principal part $a_H(u,v) := \sum_{|\alpha|,|\beta|=m} \int_{\mathbb{R}^n} a_{\alpha\beta}\, D^\alpha u\, D^\beta v \,\mathrm{d}\mathbf{x}$. Let $\partial_{h,i}$ be the difference operator[2]

$$\partial_{h,i} u(\mathbf{x}) := \frac{1}{h} \left[u\left(\mathbf{x} + \tfrac{h}{2}\mathbf{e}_i\right) - u\left(\mathbf{x} - \tfrac{h}{2}\mathbf{e}_i\right) \right], \qquad \begin{cases} \mathbf{e}_i\colon\ i\text{-th unit vector,} \\ 1 \le i \le n. \end{cases}$$

Let u be the solution of (9.7). For $v \in H^m(\mathbb{R}^n)$ set

$$d_i(u,v) := d(u,v) := a(u, \partial_{h,i} v) + a(\partial_{h,i} u, v) \quad \text{for all } v \in H^m(\mathbb{R}^n). \tag{9.9a}$$

Note that $d_i(u,v)$ is well defined since $\partial_{h,i} v,\ \partial_{h,i} u \in H^m(\mathbb{R}^n)$. Given a function w we introduce the notation $w^\pm(\mathbf{x}) := w(\mathbf{x} \pm h\mathbf{e}_i/2)$ for the translates. From

$$\begin{aligned} \int_{\mathbb{R}^n} a\, u\, \partial_{h,i} v \,\mathrm{d}\mathbf{x} = \frac{1}{h} \int_{\mathbb{R}^n} a u \left[v^+ - v^- \right] \mathrm{d}\mathbf{x} \quad &= \quad \frac{1}{h} \int_{\mathbb{R}^n} \left[a^- u^- - a^+ u^+ \right] v \,\mathrm{d}\mathbf{x} \\ = -\int_{\mathbb{R}^n} a\, v\, \partial_{h,i} u \,\mathrm{d}\mathbf{x} + \frac{1}{h} \int_{\mathbb{R}^n} &\left[(a^- - a) u^- + (a - a^+) u^+ \right] v \,\mathrm{d}\mathbf{x} \end{aligned}$$

and the substitutions $a \rightsquigarrow a_{\alpha\beta},\ u \rightsquigarrow D^\alpha u,\ v \rightsquigarrow D^\beta v$ we obtain the representation

$$d(u,v) = -\sum_{|\alpha|=m} \sum_{|\beta|=m} \int_{\mathbb{R}^n} \left\{ \frac{a_{\alpha\beta} - a^-_{\alpha\beta}}{h} D^\alpha u^- + \frac{a^+_{\alpha\beta} - a_{\alpha\beta}}{h} D^\alpha u^+ \right\} D^\beta v \,\mathrm{d}\mathbf{x} .$$

Since $\frac{|a_{\alpha\beta} - a^\pm_{\alpha\beta}|}{h} \le \|\partial a_{\alpha\beta}/\partial x_i\|_{L^\infty} \le \text{const}$ for $\max\{0, k + |\beta| - m\} = k \ge 1$ (cf. (9.6)), we have

$$|d(u,v)| \le C\, |u|_m\, |v|_m \qquad \text{for all } v \in H^m(\mathbb{R}^n). \tag{9.9b}$$

The $H^m(\mathbb{R}^n)$-coercivity with constants $C_E > 0$ and C_K (cf. (6.48)) shows

$$\begin{aligned} C_E\, |\partial_{h,i} u|_m^2 &\le a(\partial_{h,i} u, \partial_{h,i} u) + C_K\, |\partial_{h,i} u|_0^2 \\ &= -a(u, \partial_{h,i}^2 u) + d(u, \partial_{h,i} u) + C_K\, |\partial_{h,i} u|_0^2 \end{aligned} \tag{9.10a}$$

[2] Actually, we would prefer the derivative $\partial u/\partial x_i$, but since we do not know whether it belongs to $V = H^m(\mathbb{R}^n)$, the difference $\partial_{h,i} u \in V$ is taken as a substitute.

(here we use (9.9a) with $v := \partial_{h,i}u \in H^m(\mathbb{R}^n)$). The first term can be transformed according to (9.7): $a(u,\partial^2_{h,i}u) = (f,\partial^2_{h,i}u)_0$ and bounded by $|f|_{-m+1}\,|\partial^2_{h,i}u|_{m-1}$. The inequality $|\partial_{h,i}v|_s \le |v|_{s+1}$ for $s = m-1$ and $v = \partial_{h,i}u$ yields

$$\left|a(u,\partial^2_{h,i}u)\right| \le |f|_{-m+1}\,|\partial_{h,i}u|_m\,. \tag{9.10b}$$

By (9.9b), the bound for the second term in (9.10a) reads:

$$|d(u,\partial_{h,i}u)| \le C_d|u|_m|\partial_{h,i}u|_m\,. \tag{9.10c}$$

The last term in (9.10a) is bounded by

$$|\partial_{h,i}u|_0^2 \le |\partial_{h,i}u|_{m-1}\,|\partial_{h,i}u|_m \le |u|_m\,|\partial_{h,i}u|_m\,. \tag{9.10d}$$

(9.10a–d) yields

$$|\partial_{h,i}u|_m \le \left[|f|_{-m+1} + (C_d + C_K)\,|u|_m\right]/C_E \qquad \text{for all } h>0,\ 1\le i\le n.$$

Lemma 6.42 shows that $u \in H^{m+1}(\mathbb{R}^n)$ and proves inequality (9.8) for $k=1$.

(ii) The induction step is given by Lemma 9.7. ∎

Corollary 9.8. The $H^m(\mathbb{R}^n)$-coercivity in Theorem 9.3 can be replaced

(a) by the sufficient conditions described in Theorem 7.11;

(b) by the assumption that for some λ the bilinear form $a_H(u,v)+\lambda(u,v)_0$ satisfies the inf-sup conditions (6.43a,b).

Corollary 9.9. If, in addition, $a(\cdot,\cdot)$ is $H^m(\mathbb{R}^n)$-elliptic or if $a(\cdot,\cdot)$ satisfies condition (6.43a,b), then in Theorem 9.3 the estimate (9.8) can be replaced by

$$|u|_{m+k} \le C_k\,|f|_{-m+k}\,.$$

Proof. The assertion follows by Remark 9.2. ∎

Corollary 9.10. Let $a(\cdot,\cdot)$ be $H^m(\mathbb{R}^n)$-coercive. Let the conditions (9.6) and $f \in H^{k-m}(\mathbb{R}^n)$ be satisfied for $k \in \mathbb{N}$ with $k > s + n/2 > n/2$. Then the weak solution u of (9.7) belongs to $C^s(\mathbb{R}^n)$. Hence for $s > 2m$, the weak solution is also a classical solution.

Proof. The statement results from Sobolev's embedding (Theorem 6.48). ∎

Corollary 9.11. Let $a(\cdot,\cdot)$ be $H^m(\mathbb{R}^n)$-coercive. Let the conditions (9.6) and $f \in H^{k-m}(\mathbb{R}^n)$ be satisfied for all $k \in \mathbb{N}$. Then the weak solution of problem (9.7) belongs to $C^\infty(\mathbb{R}^n)$. The conditions are satisfied in particular if f belongs to $C_0^\infty(\mathbb{R}^n)$ and the coefficients $a_{\alpha\beta}$ are constant.

The generalisation of $u \in H^{m+k}(\mathbb{R}^n)$ with $k\in\mathbb{N}$ to $u \in H^{m+s}(\mathbb{R}^n)$ with real $s>0$ reads as follows.

Theorem 9.12. *Let $a(\cdot,\cdot)$ in (9.5) be $H^m(\mathbb{R}^n)$-coercive. Let*

$$s = k + \Theta, \qquad k \in \mathbb{N}_0, \qquad 0 < \Theta < \vartheta < 1, \qquad t := k + \vartheta.$$

For the coefficients ($\Omega = \mathbb{R}^n$) let

$$a_{\alpha\beta} \in \begin{cases} C^{t+|\beta|-m}(\Omega) & \text{for } k + |\beta| \geq m, \\ L^\infty(\Omega) & \text{otherwise.} \end{cases} \tag{9.11}$$

Then each weak solution of (9.7) with $f \in H^{-m+s}(\mathbb{R}^n)$ belongs to $H^{m+s}(\mathbb{R}^n)$, and satisfies the estimate

$$|u|_{m+s} \leq C_s \left[|f|_{-m+s} + |u|_m \right]. \tag{9.12}$$

Proof. The proof is similar to that of Theorem 9.3, but the difference quotient $\partial = \partial_{h,i}$ is now replaced by an approximation R of its power ∂^Θ, $0 < \Theta < 1$:

$$Ru(\mathbf{x}) := R_{h,i}u(\mathbf{x}) := h^{-\Theta} \sum_{\mu=0}^{\infty} \mathrm{e}^{-\mu h} (-1)^\mu \binom{\Theta}{\mu} u(\mathbf{x} + \mu h \mathbf{e}_i) \qquad \begin{cases} \mathbf{e}_i\text{: } i\text{-th} \\ \text{unit vector.} \end{cases}$$

Here $\binom{\Theta}{0} = 1$, and $\binom{\Theta}{\mu} = (-1)^\mu (-\Theta)(1-\Theta)(2-\Theta) \cdot \ldots \cdot (\mu - 1 - \Theta)/\mu!$ are the binomial coefficients.

Exercise 9.13. Let $0 < \Theta < 1$. Show that

(a) The operator adjoint to $R_{h,i}$ is

$$R^* u(\mathbf{x}) = R^*_{h,i} u(\mathbf{x}) = h^{-\Theta} \sum_{\mu=0}^{\infty} \mathrm{e}^{-\mu h} (-1)^\mu \binom{\Theta}{\mu} u(\mathbf{x} - \mu h \mathbf{e}_i).$$

(b) For any $z \in \mathbb{C}$ with $|z| < 1$ we have $(1-z)^\Theta = \sum_{\mu=0}^{\infty} \binom{\Theta}{\mu} (-z)^\mu$.

(c) For the Fourier transform we have $\widehat{u(\cdot + \delta \mathbf{e}_i)}(\boldsymbol{\xi}) = \mathrm{e}^{\mathrm{i}\xi_i \delta} \hat{u}(\boldsymbol{\xi})$ and

$$\widehat{(R_{h,i}u)}(\boldsymbol{\xi}) = \left[\left(1 - \mathrm{e}^{-h + \mathrm{i}\xi_i h}\right)/h \right]^\Theta \hat{u}(\boldsymbol{\xi}),$$
$$\widehat{(R^*_{h,i}u)}(\boldsymbol{\xi}) = \left[\left(1 - \mathrm{e}^{-h - \mathrm{i}\xi_i h}\right)/h \right]^\Theta \hat{u}(\boldsymbol{\xi}).$$

(d) There is a constant C so that

$$\frac{1}{C} \left(1 + t^2\right)^{1/2} \leq \left| \left(1 - \mathrm{e}^{-h - \mathrm{i}th}\right)/h \right| \leq C \left(1 + t^2\right)^{1/2}.$$

holds for all $\tau \in \mathbb{R}$ and $0 < h \leq 1$ with $|th| \leq 1$.

(e) $|R_{h,i}u|_\tau \leq C_{\tau,\Theta} |u|_{\tau+\Theta}$ for all $\tau \in \mathbb{R}$, $h > 0$, $1 \leq i \leq n$, $u \in H^{\tau+\Theta}(\mathbb{R}^n)$; similarly $|R^*_{h,i}u|_\tau \leq C_{\tau,\Theta} |u|_{\tau+\Theta}$. *Hint:* Use the norms $|\cdot|^\wedge_\tau$, $|\cdot|^\wedge_{\tau+\Theta}$ (cf. (6.20)) and prove

$$|\widehat{(R_{h,i}u)}(\boldsymbol{\xi})| \leq (1 + |\boldsymbol{\xi}|^2)^{\Theta/2} |\hat{u}(\boldsymbol{\xi})|.$$

(f) $\sum_{i=1}^n |\widehat{(R_{h,i}u)}(\boldsymbol{\xi})|^2 \geq \frac{1}{C}(1 + |\boldsymbol{\xi}|^2)^{\Theta/2}$.

Now we continue the proof. The expression in (9.9a) becomes

$$d(u,v) := a(u,R^*_{h,i}v) - a(R_{h,i}u,v) = -\sum_{|\alpha|=|\beta|=m} h^{-\Theta}\sum_{\mu=1}^{\infty} \mathrm{e}^{-\mu h}\,(-1)^{\mu}\binom{\Theta}{\mu}\times$$

$$\times \int_{\mathbb{R}^n} \left[a_{\alpha\beta}(\mathbf{x}+\mu h\mathbf{e}_i) - a_{\alpha\beta}(\mathbf{x})\right]\left[D^{\alpha}u(\mathbf{x}+\mu h\mathbf{e}_i)\right]\left[D^{\beta}v(\mathbf{x})\right]\mathrm{d}\mathbf{x}\,.$$

Since we have $|a_{\alpha\beta}(\mathbf{x}+\mu h\mathbf{e}_i) - a_{\alpha\beta}(\mathbf{x})| \le C(\mu h)^t$, it follows that

$$|d(u,v)| \le C\,|u|_m\,|v|_m\left[1 + h^{t-\Theta}\sum_{\mu=1}^{\infty}\mathrm{e}^{-\mu h}\,(-1)^{\mu}\binom{\Theta}{\mu}\mu^t\right] \le C'\,|u|_m\,|v|_m\ ,$$

for it is true that $\binom{\Theta}{\mu} = \mathcal{O}(\mu^{-\Theta-1})$ and $\sum_{\mu=1}^{\infty}\mathrm{e}^{-\mu h}\mu^{t-\Theta-1} = \mathcal{O}(h^{\Theta-t})$. This proves the inequality (9.9b).

Instead of (9.10a) we obtain

$$C_E\,|Ru|_m \le a(Ru,Ru) + C_K|Ru|_0^2 = a(u,R^*Ru) - d(u,Ru) + C_K|Ru|_0^2\,.$$

Since $|R^*Ru|_{m-\Theta} \le C\,|Ru|_m$ (cf. Exercise 9.13e), (9.10b) becomes

$$|a(u,R^*Ru)| \le C\,|f|_{-m+\Theta}\,|Ru|_m\ ,$$

while (9.10c) yields

$$|d(u,Ru)| \le C\,|u|_m\,|Ru|_m\,.$$

The analogue $|Ru|_0^2 \le C\,|u|_m\,|Ru|_m$ of (9.10d) is trivial. The same considerations as in the proof of Theorem 9.3 result in

$$\left|R^*_{h,i}u\right|_m \le C\left[|R_{h,i}f|_{-m} + |u|_m\right] \le C'\left[|f|_{-m+\Theta} + |u|_m\right] \qquad \left\{\begin{array}{l}\text{for all } h>0,\\ 1\le i\le n\end{array}\right.$$

(cf. Exercise 9.13e). To obtain the analogous estimate of $|u|_{m+\Theta}$ we express $(|v|_m^{\wedge})^2$ with $v = R^*_{h,i}u$ from (6.21b) as

$$\int_{|\boldsymbol{\xi}|\le 1/h}\left(1+|\boldsymbol{\xi}|^2\right)^m|\hat{v}(\boldsymbol{\xi})|^2\,\mathrm{d}\boldsymbol{\xi} + \int_{|\boldsymbol{\xi}|\ge 1/h}\left(1+|\boldsymbol{\xi}|^2\right)^m|\hat{v}(\boldsymbol{\xi})|^2\,\mathrm{d}\boldsymbol{\xi}\,.$$

The second integral tends to zero as $h\to 0$. Since

$$\sum_{i=1}^{n}\left(|R^*_{h,i}u|_m^{\wedge}\right)^2 \ge \int_{|\boldsymbol{\xi}|\le 1/h}\left(1+|\boldsymbol{\xi}|^2\right)^m\sum_{j=1}^{n}\left|\widehat{R^*_{h,j}u}(\boldsymbol{\xi})\right|^2\mathrm{d}\boldsymbol{\xi}$$

$$\underset{\text{Exercise 9.13f}}{\ge} \frac{1}{C}\int_{|\boldsymbol{\xi}|\le 1/h}\left(1+|\boldsymbol{\xi}|^2\right)^{m+\Theta}|\widehat{u}(\boldsymbol{\xi})|^2\,\mathrm{d}\boldsymbol{\xi}$$

holds with same C for all h, we conclude $|u|_{m+\Theta} \le C\lim\limits_{h\to 0}\sum\limits_{i=1}^{n}\left(|R^*_{h,i}u|_m^{\wedge}\right)^2$ and the inequality (9.12). ∎

9.1.3 Regularity Theorems for $\Omega = \mathbb{R}^n_+$

The half-space $\mathbb{R}^m_+$ in (6.23) is characterised by $x_n > 0$. As in Chapter 7, we limit ourselves to the following two cases: either a Dirichlet problem is given for arbitrary $m \geq 1$, or the natural boundary condition is posed for $m = 1$.

Theorem 9.14 (homogeneous Dirichlet problem). *An analogue to Theorem 9.3 holds for the Dirichlet problem:*

$$u \in H^m_0(\mathbb{R}^n_+), \qquad a(u,v) = (f,v)_0 \quad \textit{for all } v \in H^m_0(\mathbb{R}^n_+).$$

Theorem 9.12 can also be carried over if we exclude the values $s = \frac{1}{2}, \frac{3}{2}, \ldots, m-\frac{1}{2}$.

For the proof on page 276 we need the following highly technical lemma.

Lemma 9.15. *Let* $s > 0$, $s \notin \{1/2, 3/2, \ldots, m-1/2\}$. *The norm* $|\cdot|_s$ *of* $H^s(\mathbb{R}^n_+)$ *is equivalent to*

$$|||u|||_s := \sqrt{|u|^2_0 + \sum_{|\alpha|=m} |D^\alpha u|^2_{s-m}}. \tag{9.13}$$

Proof. The relatively elementary case $s \geq m$ is left to the reader. For $0 < s < m$ too, the proof would be considerably simpler if in (9.13) one were to replace the dual norm $|\cdot|_{s-m}$ of $(H^{m-s}_0(\mathbb{R}^n_+))'$ by that of $(H^{m-s}(\mathbb{R}^n_+))'$.

(i) First we prove the statement for $\Omega = \mathbb{R}^n$ instead of $\Omega = \mathbb{R}^n_+$. According to Theorem 6.43a, for $\Omega = \mathbb{R}^n$ the norm $|||\cdot|||_s$ is equivalent to

$$|||u|||^\wedge_s := \sqrt{|u|^2_0 + \sum\nolimits_{|\alpha|=m} \big(|D^\alpha u|^\wedge_{s-m}\big)^2}.$$

Since

$$\big(|||u|||^\wedge_s\big)^2 = \int_{\mathbb{R}^n} \Big[1 + \Big(\sum_{|\alpha|=m} |\boldsymbol{\xi}^\alpha|^2\Big)(1+|\boldsymbol{\xi}|^2)^{s-m}\Big]\,|\hat{u}(\boldsymbol{\xi})|^2\,\mathrm{d}\boldsymbol{\xi}$$

and

$$0 < C_0\big(1+|\boldsymbol{\xi}|^2\big)^s \leq 1 + \sum_{|\alpha|=m} |\boldsymbol{\xi}^\alpha|^2\,\big(1+|\boldsymbol{\xi}|^2\big)^{s-m} \leq C_1\big(1+|\boldsymbol{\xi}|^2\big)^s,$$

$|||u|||^\wedge_s$ and $|\cdot|^\wedge_s$ are equivalent.

(ii) For the transition to $\Omega = \mathbb{R}^n_+$ the following extension $\phi : H^s(\mathbb{R}^n_+) \to H^s(\mathbb{R}^n)$ must be investigated, where $\mathbf{x} = (\mathbf{x}', x_n) \in \mathbb{R}^n$ with $\mathbf{x}' = (x_1, \ldots, x_{n-1})$:

$$(\phi u)(\mathbf{x}) := u(\mathbf{x}) \qquad \text{for } \mathbf{x} \in \mathbb{R}^n_+ \text{ (i.e., } x_n > 0)$$

$$(\phi u)(\mathbf{x}', x_n) := \sum_{\nu=1}^{L} a_\nu\,[u(\mathbf{x}', -\nu x_n) + u(\mathbf{x}', -x_n/\nu)] \qquad \text{for } x_n < 0,$$

where the coefficients a_ν are defined in the following exercise.

Exercise 9.16. Let the coefficients a_ν of ϕ be selected as the solution of the system of equations

$$\sum_{\nu=1}^{L} a_\nu(\nu^k+\nu^{-k}) = (-1)^k \qquad (0\le k\le L-1).$$

Show that

(a) $u\in C^{L-1}(\mathbb{R}^n_+)$ yields $\phi u\in C^{L-1}(\mathbb{R}^n)$.

(b) $\phi\in L(H^k(\mathbb{R}^n_+),H^k(\mathbb{R}^n))$ for $k=0,1,\dots,L$.

(c) The operator adjoint to ϕ reads

$$(\phi^* u)(\mathbf{x}',x_n) = u(\mathbf{x}',x_n)+\sum_{\nu=1}^{L} a_\nu\left[\frac{1}{\nu}u(\mathbf{x}',\frac{-x_n}{\nu})+\nu u(\mathbf{x}',-\nu x_n)\right] \text{ for } x_n>0.$$

(d) $(\partial/\partial x_n)^k(\phi^* u)(\mathbf{x}',0)=0$ and $u\in C^k(\mathbb{R}^n)$ for $k=0,1,\dots,L-2$.

(e) $\phi^*\in L(H^k(\mathbb{R}^n),H_0^k(\mathbb{R}^n_+))$ for $k=0,1,\dots,L-1$. *Hint*: Corollary 6.61.

(f) $\phi\in L(H^k(\mathbb{R}^n_+),H^k(\mathbb{R}^n))$ for $k=1-L,2-L,\dots,0,1,\dots,L$.

(iii) $|||\cdot|||_s\le C\,|\cdot|_s$ results from $D^\alpha\in L(H_0^s(\mathbb{R}^n_+),H^{s-m}(\mathbb{R}^n_+))$ (provable via continuation arguments from Remark 6.76b) so that $|\cdot|_s\le|||\cdot|||_s$ remains to be shown.

(iv) $|u|_s\le|\phi u|_s\le|||\phi u|||_s$ is true according to part (i) of the proof. The inequality $|||\phi u|||_s\le C|||u|||_s$, which would finish the proof, reduces to

$$|D^\alpha\phi u|_{s-m}\le C\,|D^\alpha u|_{s-m} \qquad \text{for } |\alpha|=m,\ u\in H^k(\mathbb{R}^n_+). \tag{9.14}$$

Let $|\alpha|=m$. For $\phi_\alpha:\mathbb{R}^n_+\to\mathbb{R}^n$ defined by

$$(\phi_\alpha u)(\mathbf{x}',x_n) := \left\{\begin{array}{ll} u(\mathbf{x}',x_n) & \text{for } x_n>0,\\ \sum\limits_{\nu=1}^{L} a_\nu\left[\begin{array}{l}(-\nu)^{\alpha_n}u(\mathbf{x}',-\nu x_n)\\ +\left(\frac{-1}{\nu}\right)^{\alpha_n}u(\mathbf{x}',\frac{-x_n}{\nu})\end{array}\right] & \text{otherwise}\end{array}\right\}$$

one verifies $D^\alpha(\phi u)=\phi_\alpha D^\alpha u$. As in Exercise 9.16f one shows that ϕ_α belongs to $L(H^s(\mathbb{R}^n_+),H^s(\mathbb{R}^n))$ for $s=1+m-L,2+m-L,\dots,L-m$. This result can be carried over to real $s\in[1+m-L,L-m]$ except for the cases $\frac12-s\in\mathbb{N}$ (i.e., $s=-\frac12,-\frac32,\dots$) (cf. Lions–Magenes [194, pages 54ff]). Let $L\ge 2m+1$ and $v\in H^{m-s}(\mathbb{R}^n)$. Since $\phi_\alpha^\star v\in H_0^{m-s}(\mathbb{R}^n_+)$, one infers from the identity $(D^\alpha\phi u,v)_{L^2(\mathbb{R}^n)}=(D^\alpha u,\phi_\alpha^\star v)_{L^2(\mathbb{R}^n_+)}$ the estimate

$$(D^\alpha\phi u,v)_0\le|D^\alpha u|_{s-m}\,\|\phi_\alpha^\star\|_{H_0^{m-s}(\mathbb{R}^n_+)\leftarrow H^{m-s}(\mathbb{R}^n)}\,|v|_{m-s}$$

for all $v\in H^{m-s}(\mathbb{R}^n)$, and therefore (9.14) with $C:=\|\phi_\alpha^\star\|_{H_0^{m-s}(\mathbb{R}^n_+)\leftarrow H^{m-s}(\mathbb{R}^n)}=\|\phi_\alpha\|_{H^{s-m}(\mathbb{R}^n)\leftarrow H^{s-m}(\mathbb{R}^n_+)}$. ∎

Proof of Theorem 9.14. (i) First let $k = s = 1$. The proof of Theorem 9.3 can be repeated for the differences $\partial_{h,j} u$ ($j = 1, \dots, n-1$) and implies the existence of the derivatives $\partial u / \partial x_j \in H_0^m(\mathbb{R}^n_+)$, $j \neq n$. Thus one has $D^\alpha u \in H^1(\mathbb{R}^n_+)$ for all multi-indices $|\alpha| = m$ except for $\alpha = (0, \dots, 0, m)$.

(ii) We set $\hat{\alpha} := (0, \dots, 0, m) \in \mathbb{Z}^n$ and

$$w := a_{\hat{\alpha}\hat{\alpha}} D^{\hat{\alpha}} u, \quad F_\alpha(v) = \int_{\mathbb{R}^n_+} w(\mathbf{x}) D^\alpha v(\mathbf{x}) \mathrm{d}\mathbf{x} \quad \text{for } |\alpha| = m,$$

where $a_{\hat{\alpha}\hat{\alpha}}$ is the coefficient from the bilinear form (9.5). The remainder of the proof runs as follows. In part (iii) we shall show that

$$|F_\alpha(v)| \leq C \, |v|_{m-1} \qquad \text{for } |\alpha| = m, \ v \in H^{m-s}(\mathbb{R}^n_+). \tag{9.15}$$

Since $F_\alpha(v) = (w, D^\alpha v)_0 = (-1)^m (D^\alpha w, v)_0$, inequality (9.15) means that $D^\alpha w \in H^{1-m}(\mathbb{R}^n_+)$ and $|D^\alpha w|_{1-m} \leq C$ hold for $|\alpha| = m$. According to Lemma 9.15 it follows that $w \in H^1(\mathbb{R}^n_+)$. The coercivity of $a(\cdot, \cdot)$ implies uniform ellipticity of $L = \sum_{|\alpha|=|\beta|=m} (-1)^m D^\beta a_{\alpha\beta} D^\alpha$, i.e., $\sum a_{\alpha\beta} \boldsymbol{\xi}^{\alpha+\beta} \geq \varepsilon \, |\boldsymbol{\xi}|^{2m}$ (cf. Theorem 7.13). For $\boldsymbol{\xi} = (0, \dots, 0, 1)$ one obtains $a_{\hat{\alpha}\hat{\alpha}}(\mathbf{x}) \geq \varepsilon$. Hence it follows from $w \in H^1(\mathbb{R}^n_+)$ and $D^\gamma a_{\hat{\alpha}\hat{\alpha}} \in L^\infty(\mathbb{R}^n_+)$, $|\gamma| = 1$, that $D^{\hat{\alpha}} u \in H^1(\mathbb{R}^n_+)$. According to part (i) all other derivatives $D^\alpha u$ ($|\alpha| \leq m$, $\alpha \neq \hat{\alpha}$) belong to $H^1(\mathbb{R}^n_+)$ anyway so that $u \in H^{m-1}(\mathbb{R}^n_+)$ has been proved.

(iii) *Proof of (9.15).* For each $\alpha \neq \hat{\alpha}$ there exists a γ with $|\gamma| = 1$, $\gamma_n = 0$, $0 \leq \gamma \leq \alpha$ (componentwise inequalities). Integration by parts yields

$$F_\alpha(v) = -\int_{\mathbb{R}^n_+} \left[\left(D^\gamma a_{\hat{\alpha}\hat{\alpha}}\right) \left(D^{\hat{\alpha}} u\right) \left(D^{\alpha-\gamma} v\right) + a_{\hat{\alpha}\hat{\alpha}} \left(D^{\hat{\alpha}+\gamma} u\right) \left(D^{\alpha-\gamma} v\right) \right] \mathrm{d}\mathbf{x}$$

for all $v \in C_0^\infty(\mathbb{R}^n_+)$; and thus

$$|F_\alpha(v)| \leq C_\alpha \, |v|_{m-1} \ \text{ with } \ C_\alpha := C \left[\, |u|_m + |D^\gamma u|_m \, \right] \text{ for all } \ v \in C_0^\infty(\mathbb{R}^n_+).$$

Here we used the fact that $D^\gamma u \in H^m(\mathbb{R}^n_+)$ according to part (i). Since $C_0^\infty(\mathbb{R}^n_+)$ is dense in $H_0^{m-1}(\mathbb{R}^n_+)$, (9.15) follows for $\alpha \neq \hat{\alpha}$.

There remains to investigate $\alpha = \hat{\alpha}$. We write

$$F_{\hat{\alpha}}(v) = a(u, v) - \hat{a}(u, v) \qquad \text{with} \quad \hat{a}(u, v) := \sum{}' \int_{\mathbb{R}^n_+} a_{\alpha\beta} \left(D^\alpha u\right) \left(D^\beta v\right) \mathrm{d}\mathbf{x},$$

where $\sum'$ represents the summation over all pairs $(\alpha, \beta) \neq (\hat{\alpha}, \hat{\alpha})$. For each $(\alpha, \beta) \neq (\hat{\alpha}, \hat{\alpha})$ there exists a γ with

$$|\gamma| = 1, \qquad 0 \leq \gamma \leq \beta, \qquad \alpha + \gamma \neq (0, \dots, 0, m+1) \, .$$

As above, one integrates each term with $|\beta| = m$ by parts:

$$\begin{aligned} &\int_{\mathbb{R}^n_+} a_{\alpha\beta} \left(D^\alpha u\right) \left(D^\beta v\right) \mathrm{d}\mathbf{x} \\ &\quad = -\int_{\mathbb{R}^n_+} \left(D^\gamma a_{\alpha\beta}\right) \left(D^\alpha u\right) \left(D^{\beta-\gamma} v\right) \mathrm{d}\mathbf{x} - \int_{\mathbb{R}^n_+} a_{\alpha\beta} \left(D^{\alpha+\gamma} u\right) \left(D^{\beta-\gamma} v\right) \mathrm{d}\mathbf{x} \end{aligned}$$

for $v\in C_0^\infty(\mathbb{R}^n_+)$, and estimates using $C\,|v|_{m-1}$ with $C=C(u)$. Altogether one obtains $|\hat\alpha(u,v)|\le C\,|v|_{m-1}$. Together with $|a(u,v)|=|(f,v)_0|\le|f|_{-m+1}\,|v|_{m-1}$, (9.15) also follows for $\alpha=\hat\alpha$.

(iv) By induction for $k=2,\dots$ one proves in the same way $|F_\alpha(v)|\le C\,|v|_{m-k}$, and from this $u\in H^{k+m}(\mathbb{R}^n_+)$. For real $s>0$, $s\notin\{1/2,\dots,m-1/2\}$, one proves correspondingly $|F_\alpha(v)|\le C\,|v|_{m-s}$ and hence $u\in H^{s+m}(\mathbb{R}^n_+)$. ∎

The generalisation of Theorem 9.14 to inhomogeneous boundary values reads as follows.

Theorem 9.17. *Let the bilinear form $a(\cdot,\cdot)$ from (9.5) be $H_0^m(\mathbb{R}^n_+)$-coercive. For an $s>0$, $s\notin\{1/2,\dots,m-1/2\}$ either let (9.6) hold if $s=k\in\mathbb{N}$, or (9.11) if $s\notin\mathbb{N}$. Let $u\in H^m(\mathbb{R}^n_+)$ be the weak solution of the inhomogeneous Dirichlet problem*

$$\begin{aligned} a(u,v)&=(f,v)_0 \qquad && \text{for all } v\in H_0^m(\mathbb{R}^n_+),\\ \partial^\ell u/\partial n^\ell&=\varphi_\ell && \text{on } \Gamma=\partial\mathbb{R}^n_+ \text{ for } \ell=0,1,\dots,m-1, \end{aligned} \tag{9.16}$$

where

$$f\in H^{-m+s}(\mathbb{R}^n_+),\qquad \varphi_\ell\in H^{m+s-\ell-1/2}(\Gamma)\qquad (0\le\ell\le m-1)\,.$$

Then u belongs to $H^{m+s}(\mathbb{R}^n_+)$ and satisfies the inequality

$$|u|_{m+s}\le C_s\left[|f|_{-m+s}+\sum_{\ell=0}^{m-1}|\varphi_\ell|_{m+s-\ell-1/2}+|u|_m\right]. \tag{9.17}$$

Proof. For $m=1$ Theorem 6.50 guarantees the existence of $u_0\in H^{m+s}(\mathbb{R}^n_+)$ which satisfies the boundary conditions (9.16) (for $m>1$ cf. Wloka [308, Theorem 8.8b]). $w:=u-u_0$ is the solution of the homogeneous problem

$$a(w,v)=F(v):=(f,v)_0-a(u_0,v)$$

(cf. Remark 7.16). Theorem 9.14 may also be carried over to the right-hand side $F(v)$ under discussion here (instead of $(f,v)_0$) and yields $w\in H^{m+s}(\mathbb{R}^n_+)$. ∎

By similar means one proves the next theorem.

Theorem 9.18 (natural boundary conditions). *Let the bilinear form $a(\cdot,\cdot)$ from (9.5) be $H^1(\mathbb{R}^n_+)$-coercive. For $s>0$ either let (9.6) hold if $s=k\in\mathbb{N}$, or (9.11) if $s\notin\mathbb{N}$. Let $u\in H^1(\mathbb{R}^n_+)$ be the weak solution of the problem*

$$a(u,v)=f(v):=\int_{\mathbb{R}^n_+}g(\mathbf{x})v(\mathbf{x})\mathrm{d}\mathbf{x}+\int_\Gamma\varphi(\mathbf{x})v(\mathbf{x})\mathrm{d}\Gamma\quad\text{for all } v\in H^1(\mathbb{R}^n_+),$$

where $g\in H^{s-1}:=\left\{\begin{matrix}H^{s-1}(\mathbb{R}^n_+)\\ \big(H^{1-s}(\mathbb{R}^n_+)\big)'\end{matrix}\right.\begin{matrix}\text{for } s\ge1\\ \text{for } s<1\end{matrix}\Big\}$, $\varphi\in H^{s-1/2}(\Gamma)$. Then it belongs to $H^{1+s}(\mathbb{R}^n_+)$ and satisfies the estimate

$$|u|_{1+s}\le C_s\left[\|g\|_{H^{s-1}}+|\varphi|_{s-1/2}+|u|_1\right].$$

If a boundary-value problem is in the form (9.1): $Lu = g$, $Bu = \varphi$ with $B = \mathbf{b}^{\mathsf{T}}\nabla + b_0$, $|b_n(\mathbf{x})| \geq \varepsilon > 0$ on $\Gamma = \partial\mathbb{R}^n_+$ (b_n is the n-th coefficient of $\mathbf{b}$), then according to Theorem 7.33 one can find an associated variational formulation and apply Theorem 9.18.

9.1.4 Regularity Theorems for General Domains $\Omega \subset \mathbb{R}^n$

The following theorems show that the above regularity statements also hold for $\Omega \subset \mathbb{R}^n$ if Ω is bounded sufficiently smoothly.

Theorem 9.19. *Let $\Omega \in C^{t+m}$ for some $t \geq 0$. Let the bilinear form (9.5) be $H^m_0(\Omega)$-coercive. Let $s \geq 0$ satisfy*

$$s + 1/2 \notin \{1, 2, \ldots, m\}, \quad 0 \leq s \leq t, \text{ if } t \in \mathbb{N}, \quad 0 \leq s < t, \text{ if } t \notin \mathbb{N}.$$

For the coefficients let the following hold:

$$\begin{aligned} &D^\gamma a_{\alpha\beta} \in L^\infty(\Omega) \quad \left\{ \begin{array}{l} \text{for all } \alpha, \beta, \gamma \text{ with} \\ |\gamma| \leq \max\{0, t + |\beta| - m\} \end{array} \right\}, \quad \text{if } t \notin \mathbb{N}, \\ &a_{\alpha\beta} \in \left\{ \begin{array}{ll} C^{t+|\beta|-m}(\overline{\Omega}) & \text{for } |\beta| > m - t \\ L^\infty(\Omega) & \text{otherwise} \end{array} \right\}, \quad \text{if } t \in \mathbb{N}. \end{aligned} \tag{9.18}$$

Then each weak solution $u \in H^m_0(\Omega)$ of the problem

$$a(u, v) = f(v) \qquad \text{for all } v \in H^m_0(\Omega)$$

with $f \in H^{-m+s}(\Omega)$ belongs to $H^{m+s}(\Omega) \cap H^m_0(\Omega)$ and satisfies the estimate

$$|u|_{m+s} \leq C_s \left[|f|_{-m+s} + |u|_m \right]. \tag{9.19}$$

For inhomogeneous boundary conditions

$$\partial^\ell u / \partial n^\ell = \varphi_\ell \quad \text{with } \varphi_\ell \in H^{m+s-\ell-1/2}(\Gamma) \quad \text{for } \ell = 0, 1, \ldots, m-1$$

instead of $u \in H^m_0(\Omega)$, the statement $u \in H^m(\Omega)$ implies $u \in H^{m+s}(\Omega)$ and the estimate (9.17).

Proof. (i) Let $\{U^i : i = 0, 1, \ldots, N\}$ with $U^i \subset \Omega$ be a covering of Ω as in Lemma 6.54. Let $\{\chi_i\}$, with $\chi_i = \sigma_i^2 \in C^\infty(\Omega)$ and $\operatorname{supp}(\chi_i) \subset U^i$, be the associated partition of unity from Lemma 6.55. There exist maps $\alpha^i \in C^t(U^i)$ which map U^i ($i \geq 1$) into $\mathbb{R}^n_+$ such that $\alpha^i(\partial U^i \cap \Gamma) \subset \partial\mathbb{R}^n_+$. By contrast, U^0 lies in the interior of Ω so that $\Gamma \cap \partial U^0 = \emptyset$. The solution u can be written as $\sum_i \chi_i u$. In part (ii) of the proof we shall treat $\chi_0 u$, and in part (iii) $\chi_i u$ for $i \geq 1$.

(ii) We set $d_i(u, v) := a(\chi_i u, v) - a(u, \chi_i v)$ $(i = 0, 1, \ldots, N)$ and wish to show the estimate

$$|d_0(u,v)| \le C_d\,|u|_m\,|v|_{m-s} \qquad (u\in H_0^m(\Omega),\ v\in H_0^{m-s}(\Omega),\ s\ge 1) \quad (9.20)$$

for $s=1$. Each term of $d_0(u,v)$ has the form

$$\int_{U^0} a_{\alpha\beta}(\mathbf{x})\big[\,(D^\alpha(\chi_0 u))\,(D^\beta v) - (D^\alpha u)\,(D^\beta\,(\chi_0 v))\,\big]\,\mathrm{d}\mathbf{x}\,.$$

Since $D^\alpha(\chi_0 u)=\chi_0 D^\alpha u+$ lower derivatives of u, one has

$$\big[\,(D^\alpha(\chi_0 u))\,(D^\beta v) - (D^\alpha u)\,(D^\beta\,(\chi_0 v))\,\big] = \sum_{\gamma,\delta} c_{\gamma\delta}\,D^\gamma u\,D^\delta v$$

with $|\gamma|\,,|\delta|\le m,\ |\gamma|+|\delta|\le 2m-1$. One integrates $\int_{U^0} a_{\alpha\beta}\,c_{\gamma\delta}\,D^\gamma u\,D^\delta v\,\mathrm{d}\mathbf{x}$ with $|\delta|=m$ by parts, and obtains a bound $C\,|u|_m\,|v|_{m-1}$, whence (9.20) follows.

The coefficients $a_{\alpha\beta}$ can be extended to $\mathbb{R}^n$ in such a way that the corresponding condition (9.18) is satisfied on $\mathbb{R}^n$. Denote the resulting bilinear form by $\bar a_0(u,v)$. Since $\chi_0\in C^\infty(\Omega)$ has a support $\mathrm{supp}(\chi_0)\subset U^0$, the extension of $\chi_0 u$ through $(\chi_0 u)\,(\mathbf{x})=0$ for $\mathbf{x}\in\mathbb{R}^n\setminus U^0$ poses no problems. We may formally define $d_0(u,v)$ for $v\in H^{m-s}(\mathbb{R}^n)$ since only the restriction of v to U^0 is of any consequence. By (9.20) $d_0(u,v)$ can be written, for a fixed $u\in H_0^m(\Omega)$, in the form

$$d_0(u,v)=(d_0,v)_{L^2(\mathbb{R}^n)} \qquad \text{with } d_0\in H^{s-m}(\mathbb{R}^n),\ |d_0|_{s-m}\le C_d\,|u|_m\,.$$

$\chi_0 u$ is the weak solution of

$$\begin{aligned}\bar a_0(\chi_0 u,v) &= a(\chi_0 u,v)=a(u,\chi_0 v)+d_0(u,v)=(f,\chi_0 v)_0+(d_0,v)_0\\ &=(\chi_0 f+d_0,v)_{L^2(\mathbb{R}^n)} \qquad (v\in H^m(\mathbb{R}^n))\,.\end{aligned}$$

Theorem 9.3 [resp. 9.12] proves $\chi_0 u\in H^{m+s}(\mathbb{R}^n)$ (thus too $\chi_0 u\in H^{m+s}(\Omega)$) and

$$\begin{aligned}|\chi_0 u|_{m+s} &\le C\,\big[\,|\chi_0 f|_{s-m}+|d_0|_{s-m}+|\chi_0 u|_m\,\big] \qquad (9.21a)\\ &\le C'\,\big[\,|f|_{s-m}+C_d\,|u|_m+C_0\,|u|_m\,\big]\le C''\,\big[\,|f|_{s-m}+|u|_m\,\big]\,,\end{aligned}$$

where s is still restricted to $s\le 1$.

(iii) The same reasoning as for $\chi_i u\ (i=1,\dots,N)$ shows

$$a(\chi_i u,v)=(\chi_i f+d_i,v)_0 \quad \text{for all } v\in H^m(\mathbb{R}^n) \text{ with } |d_i|_{s-m}\le C_d\,|u|_m\,.$$

By assumption the maps $\alpha^i: U^i\to\mathbb{R}^n_+$ and their inverses $(\alpha^i)^{-1}$ belong to $C^{t+m}(\bar U^i)$ [resp. $C^{t+m}(\alpha^i(\bar U^i))$]. Put $\tilde u(\tilde{\mathbf{x}}):=u(\mathbf{x})$ for $\tilde{\mathbf{x}}=\alpha^i(\mathbf{x})$, that is $\tilde u=u\circ(\alpha^i)^{-1}$. In a similar way define $\tilde a_{\alpha\beta}$, $\tilde\chi_i$, $\tilde f$, $\tilde d_i$. In

$$a(\chi_i u,v)=\sum_{\alpha,\beta}\int_{\alpha^i(U^i)} \tilde a_{\alpha\beta}\,D_x^\alpha(\tilde\chi_i\tilde u)\,D_x^\beta\tilde v\,\big|\det(\alpha^i)'\big|^{-1}\,\mathrm{d}\tilde{\mathbf{x}}$$

one can replace the derivatives D_x^α, D_x^β with derivatives with respect to the new coordinates, since $t \geq 0$ and thus obtain the bilinear form

$$a_i(\tilde{\chi}_i\tilde{u}, \tilde{v}) = \sum_{\alpha,\beta} \int_{\alpha^i(U^i)} \widehat{a}_{\alpha\beta} D^\alpha(\tilde{\chi}_i\tilde{u}) D^\beta \tilde{v} \, \mathrm{d}\tilde{\mathbf{x}} \,,$$

which again is $H_0^m(\alpha^i(U^i))$-coercive and whose new coefficients $\widehat{a}_{\alpha\beta}$ satisfy the conditions corresponding to (9.18). As in (ii), $\widehat{a}_{\alpha\beta}$ can be continued to $\mathbb{R}^n_+ \supset \alpha^i(U^i)$ such that the resulting bilinear form $\bar{a}_i(\cdot,\cdot)$ is $H_0^m(\mathbb{R}^n_+)$-coercive. One can apply Theorem 9.14 to

$$\bar{a}_i(\tilde{\chi}_i\tilde{u}, \tilde{v}) = a_i(\tilde{\chi}_i\tilde{u}, \tilde{v}) = a(\chi_i u, v) = (\chi_i f + d_i, v)_0 = \Big(\frac{\tilde{\chi}_i \tilde{f} + \tilde{d}_i}{|\det (\alpha^i)'|} \,, \tilde{v} \Big)_{L^2(\mathbb{R}^n_+)}$$

for all $\tilde{v} \in H_0^m(\mathbb{R}^n_+)$, which yields $\tilde{\chi}_i\tilde{u} \in H^{m+s}(\mathbb{R}^n_+)$ and

$$|\tilde{\chi}_i\tilde{u}|_{m+s} \leq \tilde{C}_s \Big[\, |\tilde{f}|_{s-m} + |\tilde{d}_i|_{s-m} + |\tilde{\chi}_i\tilde{u}|_m \Big] .$$

Transforming back (cf. Theorems 6.35 and 6.43g) yields the estimates $|\chi_i u|_{m+s} \leq C_s \left[\, |f|_{s-m} + |d_i|_{s-m} + |\chi_i u|_m \right] \leq C_s \left[\, |f|_{s-m} + C_d \, |u|_m + C \, |u|_m \right]$, and thus

$$|\chi_i u|_{m+s} \leq C_s \left[\, |f|_{s-m} + |u|_m \, \right] \quad \text{for all } 1 \leq i \leq N,\ 0 \leq s \leq 1. \tag{9.21b}$$

(iv) (9.21a,b) hold for $s \leq 1$. Since $|u|_{m+s} = |\sum_i \chi_i u|_{m+s} \leq \sum_i |\chi_i u|_{m+s}$, estimate (9.19) has been proved for $s \leq 1$. If the conditions of the theorem allow an $s \in (1,2]$, one proves (9.19) as follows. Since $u \in H^{m+1}(\Omega)$ has been proved already, one can estimate the forms $d_i(u,v)$ after further integration by parts via $|d_i(u,v)| \leq C_d \, |u|_{m+1} \, |v|_{m-s}$. Accordingly, $d_i(u,v) = (d_i, v)_0$ with $d_i \in H^{-m+s}(\Omega)$ and $|d_i|_{-m+s} \leq C_d \, |u|_{m+1}$. If one inserts the above estimate (9.19) for $s = 1$, one obtains (9.21a,b) and hence also (9.19) for $1 < s \leq 2$. Further induction yields (9.19) for admissible $s \in (k, k+1]$.

(v) The case of inhomogeneous boundary values is treated as in Theorem 9.17. ■

Analogously one may prove the next theorem.

Theorem 9.20 (natural boundary conditions). *Let $\Omega \in C^{t+1}$ with $t \geq 0$. If $0 \leq s \leq t \in \mathbb{N}$ or $0 \leq s < t \notin \mathbb{N}$, $\mathbb{R}^n_+$ in Theorem 9.18 can be replaced by Ω.*

Corollaries 9.9–9.11 transfer mutatis mutandis to Theorems 9.19 and 9.20.

Corollary 9.21. Let Ω and the coefficients of $a(\cdot,\cdot)$ satisfy the conditions in Theorem 9.19, resp. 9.20. An eigenfunction, i.e., a solution $u \in V$ ($V = H_0^m(\Omega)$ in the case of Theorem 9.19, $V = H^1(\Omega)$ in the case of Theorem 9.20) of

$$a(u,v) = 0 \quad \text{for all } v \in V, \quad u \neq 0$$

belongs to $H^{m+s}(\Omega)$ for all $0 \leq s \leq t \in \mathbb{N}$ or $0 \leq s < t \notin \mathbb{N}$.

Proof. Use Theorem 9.19 (9.20) for $f = 0$ (and $\varphi = 0$). ■

According to Theorem 6.48 (Sobolev's lemma) one obtains sufficient conditions via $C^{k+\lambda}(\overline{\Omega}) \supset H^{k+\lambda+n/2}(\Omega)$ for u to be a classical solution from $C^{k+\lambda}(\overline{\Omega})$.

The minimal conditions for $u \in C^{k+\lambda}(\overline{\Omega})$ result from a different theoretical approach, which goes back to Schauder [258]. The following theorem, for example, can be found in Miranda [205, §V] and Wienholtz–Kalf–Kriecherbauer [306, §8]. It shows the $C^{k+\lambda}$-regularity of the operator L in (5.1).

Theorem 9.22. *Let $k \geq 2$, $0 < \lambda < 1$. Let $\Omega \in C^{k+\lambda}$ be a bounded domain. Let the differential operator $L = \sum a_{ij}\partial^2/\partial x_i \partial x_j + \sum a_i \partial/\partial x_i + a$ be uniformly elliptic in Ω (i.e., (5.4a) holds). Let*

$$a_{ij}, a_i, a \in C^{k-2+\lambda}(\overline{\Omega}) \text{ and } f \in C^{k-2+\lambda}(\overline{\Omega}), \ \varphi \in C^{k+\lambda}(\Gamma).$$

Then the boundary-value problem $Lu = f$ in Ω, $u = \varphi$ on Γ either has a unique (classical) solution $u \in C^{k+\lambda}(\overline{\Omega})$, or there exists a finite-dimensional eigenspace $\{0\} \neq E \subset C^{k+\lambda}(\overline{\Omega})$ such that for all $e \in E$ the following holds: $Le = 0$ in Ω, $e = 0$ on Γ. If $a \geq 0$, the first alternative always holds.

The condition $\Omega \in C^{t+m}$ in Theorem 9.19 is stronger than necessary. For the Dirichlet problem the Lipschitz-continuity of the boundary is already sufficient to obtain the following result.

Theorem 9.23 (Nečas [210]). *Let $\Omega \in C^{0,1}$ be a bounded domain. Let the bilinear form (9.5) be $H_0^m(\Omega)$-coercive. Let the following hold:*

$$0 < s < t \leq 1/2.$$

The coefficients $a_{\alpha\beta} \in L^\infty(\Omega)$ must belong to $C^t(\overline{\Omega})$ if $|\beta| = m$. Then the weak solution $u \in H_0^m(\Omega)$ of the problem

$$a(u,v) := \int_\Omega f(\mathbf{x}) v(\mathbf{x}) \mathrm{d}\mathbf{x} \qquad \text{for all } v \in H_0^m(\Omega)$$

with $f \in H^{-m+s}(\Omega)$ belongs to $f \in H^{-m+s}(\Omega)$ and satisfies the estimate (9.19).

The condition of $H_0^m(\Omega)$-coercivity can be replaced by that of uniform ellipticity (7.4) (cf. Theorems 7.11, 7.13). The statement of Theorem 9.23 cannot be extended to $s \geq 1/2$ since then $u \in H_0^{m+s}(\Omega)$ would contain another boundary condition.

The *proof of Theorem 9.23* uses an isomorphism $R = R^\star$ (related to $R_{h,i}$ in the proof of Theorem 9.12) between $H_0^{m+s}(\Omega)$ and $H_0^m(\Omega)$ and also between $H_0^m(\Omega)$ and $H_0^{m-s}(\Omega)$ such that the form $b(u,v) := a(Ru, Rv)$ is $H_0^{m+s}(\Omega)$-coercive. It is necessary to prove that $\tilde{b}(u,v) := a(u, R^2 v)$ is also $H_0^{m+s}(\Omega)$-coercive. We know $f \in H^{-m+s}(\Omega)$ implies $\tilde{f} := R^2 f \in H^{-m-s}(\Omega)$. Each solution of $a(u,v) = (f,v)_0$ is also a solution of

$$a(u, R^2\tilde{v}) = \tilde{b}(u,\tilde{v}) = (\tilde{f}, \tilde{v})_0 = (f, R^2\tilde{v})_0$$

so that $u \in H_0^{m+s}(\Omega)$ follows.

9.1.5 Regularity for Convex Domains and Domains with Corners

A domain Ω is *convex* if with $\mathbf{x}', \mathbf{x}'' \in \Omega$, $\mathbf{x}' + t(\mathbf{x}'' - \mathbf{x}')$ belongs to Ω for all $0 \le t \le 1$. Convex domains in particular belong to $C^{0,1}$, but permit stronger regularity statements than Theorem 9.23.

Theorem 9.24 (Kadlec [165]). *Let Ω be bounded and convex. Let the bilinear form (9.5) be $H_0^1(\Omega)$-coercive. Let the coefficients of the principal part be Lipschitz-continuous:*

$$a_{\alpha\beta} \in C^{0,1}(\overline{\Omega}) \qquad \text{for all } |\alpha| = |\beta| = 1,$$

for the remaining ones let the following hold:

$$D^\gamma a_{\alpha\beta} \in L^\infty(\Omega) \qquad \text{for all } \alpha, \beta, \gamma \text{ with } \gamma \le |\beta|,\ |\alpha| + |\beta| \le 1.$$

Then every weak solution $u \in H_0^1(\Omega)$ of the problem

$$a(u, v) := \int_\Omega f(\mathbf{x}) v(\mathbf{x}) \mathrm{d}\mathbf{x} \qquad \text{for all } v \in H_0^1(\Omega)$$

with $f \in L^2(\Omega)$ belongs to $H^2(\Omega) \cap H_0^1(\Omega)$ and satisfies the estimate

$$|u|_2 \le C_1 \left[|f|_0 + |u|_1 \right]. \tag{9.22}$$

The constant C_1 depends only on the diameter of Ω.

Concerning the proof we refer to the original paper or to Grivard [123, §3]. In addition, we give an explicit proof for the special case treated in Corollary 9.25.

In Section 8.5.4 H^2-regularity was required. A generalisation of (9.22) in the form of H^{m+1}-regularity for the biharmonic differential equation with $m = 2$ is known for convex polygons (cf. Blum—Rannacher [42]). For the Poisson equation the inequality (9.22) the constants can be quantified explicitly.

Corollary 9.25. For the solution of the Poisson equation $-\Delta u = f \in L^2(\Omega)$ in a convex domain Ω with $u = 0$ on Γ, the following holds:

$$\sqrt{\sum_{|\alpha|=2} |D^\alpha u|_0^2} \le |f|_0 . \tag{9.23}$$

As in Lemma 8.54 one shows that the left-hand side of (9.23) is a norm of $H^2(\Omega) \cap H_0^1(\Omega)$ which is equivalent to $|\cdot|_2$. Thus, (9.22) follows. As a model for the proof of Theorem 9.24 we carry out the proof for Corollary 9.25 for the case $\Omega \subset \mathbb{R}^2$.

Proof. (i) First let us assume that the convex domain is smooth: $\Omega \in C^\infty$. According to Theorem 9.19 $-\Delta u = f \in L^2(\Omega)$ has a solution $u \in H^2(\Omega) \cap H_0^1(\Omega)$. We want to show

$$|||u|||_2^2 := \sum_{|\alpha|=2} |D^\alpha u|_0^2 \le |\Delta u|_0^2 \quad \text{for all } u \in H^2(\Omega)\cap H_0^1(\Omega). \tag{9.24}$$

It suffices to prove (9.24) that for all u in the dense subset $\{u \in C^\infty(\Omega) : u = 0 \text{ on } \Gamma\} = C^\infty(\Omega)\cap H_0^1(\Omega)$. Integration by parts yields

$$\begin{aligned}
\iint_\Omega |\Delta u|^2 \,\mathrm{d}x\mathrm{d}y &= \iint_\Omega \left(u_{xx}^2 + u_{yy}^2 + 2u_{xx}u_{yy}\right) \mathrm{d}x\mathrm{d}y \\
&= \iint_\Omega \left(u_{xx}^2 + u_{yy}^2 - 2u_{xxy}u_y\right) \mathrm{d}x\mathrm{d}y + 2\int_\Gamma u_{xx}u_y n_y \mathrm{d}\Gamma \\
&= \iint_\Omega \left(u_{xx}^2 + u_{yy}^2 + 2u_{xy}^2\right) \mathrm{d}x\mathrm{d}y + 2\int_\Gamma u_{xx}u_y n_y \mathrm{d}\Gamma - 2\int_\Gamma u_{xy}u_y n_x \mathrm{d}\Gamma \\
&= |||u|||_2^2 + 2\int_\Gamma \left(u_{xx}n_y - u_{xy}n_x\right) u_y \mathrm{d}\Gamma,
\end{aligned}$$

where $\mathbf{n} = \binom{n_x}{n_y}$ is the normal vector. The tangent direction is given by $\mathbf{t} = \binom{-n_y}{n_x}$. Then $u_{xx}n_y - u_{xy}n_x = -(u_x)_t$ is the negative tangential derivative. Now u_x and u_y can be expressed in terms of u_t and u_n. Since both u and u_t vanish on Γ, $u_x = n_x u_n$ and $u_y = n_y u_n$. Hence the boundary integral becomes

$$\begin{aligned}
2\int_\Gamma \left(u_{xx}n_y - u_{xy}n_x\right) u_y \mathrm{d}\Gamma &= -2\int_\Gamma \left(n_x u_n\right)_t n_y u_n \mathrm{d}\Gamma \\
&= -\int_\Gamma \left[2u_n^2 n_y \left(n_x\right)_t + n_x n_y \left(u_n^2\right)_t\right] \mathrm{d}\Gamma.
\end{aligned}$$

Integration by parts of the second term yields

$$-\int_\Gamma \left[2u_n^2 n_y \left(n_x\right)_t + n_x n_y \left(u_n^2\right)_t\right] \mathrm{d}\Gamma = \int_\Gamma u_n^2 \left[\left(n_y\right)_t n_x - \left(n_x\right)_t n_y\right] \mathrm{d}\Gamma.$$

The bracketed expression in the last display is the curvature in $\mathbf{x} \in \Gamma$, which for a convex domain is always nonnegative. (9.23) has thus been proved.

(ii) Every convex domain Ω can be approximated monotonically by convex $\Omega_\nu \in C^\infty$:

$$\Omega_1 \subset \Omega_2 \subset \ldots \subset \Omega, \qquad \bigcup\nolimits_\nu \Omega_\nu = \Omega.$$

We interpret $V_\nu := \{u \in H_0^1(\Omega) : \mathrm{supp}(u) \subset \overline{\Omega}_\nu\}$ as Ritz–Galerkin space $V_\nu \subset H_0^1(\Omega)$ (cf. Footnote 8 on page 183). Each $u \in C_0^\infty(\Omega)$ lies in V_μ for sufficiently large μ. With $C_0^\infty(\Omega)$, $\bigcup_\nu V_\nu$ is therefore also a dense subset of $H_0^1(\Omega)$. For every ν, the Ritz–Galerkin problem provides the solution $u_\nu \in H_0^1(\Omega_\nu)$ of $-\Delta u_\nu = f$ in Ω_ν, $u_\nu = 0$ on $\partial\Omega_\nu$. In $\Omega\backslash\Omega_\nu$, u_ν may be continued by $u_\nu = 0$. Theorem 8.24 proves $|u_\nu - u|_1 \to 0$, where $u \in H_0^1(\Omega)$ is the solution of $-\Delta u = f$ in Ω. Theorem 9.30 will show that for every μ the restriction of u on Ω_μ belongs to $H^2(\Omega_\mu)$. For each $v \in V_\mu \subset V_\nu$, $\nu \ge \mu$, we have

$$\left|\int_{\Omega_\mu} u_{xx} v \mathrm{d}x\mathrm{d}y\right| = \left|\int_{\Omega_\mu} u_x\, v_x \,\mathrm{d}x\,\mathrm{d}y\right| = \left|\lim_{\substack{\nu\to\infty\\ \nu\geq\mu}} \int_{\Omega_\mu} (u_\nu)_x\, v_x \,\mathrm{d}x\,\mathrm{d}y\right|$$
$$= \left|\lim_{\nu\to\infty} \int_{\Omega_\mu} (u_\nu)_x\, v_x \,\mathrm{d}x\,\mathrm{d}y\right| = \left|\lim_{\nu\to\infty} \int_{\Omega_\mu} (u_\nu)_{xx}\, v \,\mathrm{d}x\,\mathrm{d}y\right|$$
$$\leq \sup_{\nu\geq\mu} \|(u_\nu)_{xx}\|_{L^2(\Omega_\mu)}\, \|v\|_{L^2(\Omega_\mu)}\,.$$

From this one infers

$$\sum_{|\alpha|=2} \|D^\alpha u\|^2_{L^2(\Omega_\mu)} \leq \sup_{\nu\geq\mu} \sum_{|\alpha|=2} \|D^\alpha u_\nu\|^2_{L^2(\Omega_\mu)} \leq \|f\|^2_{L^2(\Omega_\mu)} \leq \|f\|^2_{L^2(\Omega)}$$

(cf. (9.24)), and obtains (9.23). ∎

Which role the H^2-regularity is playing for the $H^1_0(\Omega)$-projection on a subspace $V_h \subset H^1_0(\Omega)$ is shown in the next exercise.

Exercise 9.26. Let the subspace $V_h \subset H^1_0(\Omega)$ satisfy

$$\inf\{|u-v|_1 : v \in V_h\} \leq C_0 h\, |u|_2 \qquad \text{for all } u \in H^2(\Omega)\cap H^1_0(\Omega)$$

(cf. (8.63)). Let the Poisson problem be H^2-regular (according to Theorem 9.23 convexity of Ω is sufficient). Let $Q_V : H^1_0(\Omega) \to V_h \subset H^1_0(\Omega)$ be the orthogonal projection on V_h with respect to $|\cdot|_1$. Show that there exists a C_1 such that

$$|u - Q_V u|_0 \leq C_1 h\, |u|_1 \qquad \text{for all } u \in H^1_0(\Omega).$$

Hint: (i) With the Poisson problem the boundary-value problem $-\Delta u + u = f$ in Ω and $u = 0$ on Γ are also H^2-regular (cf. Remark 9.2). The corresponding bilinear form $a(\cdot,\cdot)$ is the scalar product in $H^1_0(\Omega)$.

(ii) Q_V agrees with the Ritz projection S_h for $a(\cdot,\cdot)$.

(iii) Use Corollary 8.66.

In a certain sense, the $L^2(\Omega)$-orthogonal projection Q_0 plays a dual role. By definition the approximation is the best possible: $|u - Q_0 u|_0 \leq |u - Q_V u|_0$. The estimate of the approximation error $|u - Q_0 u|_0$ does not require any regularity assumption.

Remark 9.27. Let $Q_0 : L^2(\Omega) \to V_h$ be the $L^2(\Omega)$-orthogonal projection onto the space V_h containing at least linear functions on a triangulation. Then there holds

$$|u - Q_0 u|_0 \leq Ch\, |u|_1 \qquad \text{for all } u \in H^1(\Omega), \tag{9.25}$$

where h is the maximal grid size.

Proof. Interpolation of the inequalities $\|I - Q_0\|_{0\leftarrow 0} = 1$ and $\|I - Q_0\|_{0\leftarrow 2} \leq Ch^2$ (cf. (8.54) for $k = 0$) yields $\|I - Q_0\|_{0\leftarrow 1} \leq C'h$. ∎

While $|Q_V|_{1\leftarrow 1} = 1$ is trivial, the correspondent inequality $|Q_0|_{1\leftarrow 1} \leq C$ is not at all obvious. Bank–Yserentant [27] prove the following result.[3]

Theorem 9.28. *Under natural conditions on the finite-element grid (details in [27]) the stability* $|Q_0|_{1\leftarrow 1} \leq C$ *holds.*

In connection with finite elements one often considers polygonal domains Ω. Since polygons belong to $C^{0,1}$, the Dirichlet problem, according to Theorem 9.23, is H^{1+s}-regular with $0 \leq s < 1/2$. If the polygon is convex (i.e., if the inner angles are $\leq \pi$) then as in Theorem 9.24 one has H^2-regularity ($m = 1$). One obtains results between $H^{3/2}$ and H^2 if the maximal inner angle of the polygon lies between π and 2π (cf. Schatz–Wahlbin [255]). If Ω has a reentrant corner the boundary-value problem can no longer be H^2-regular (cf. Example 2.4; with an inner angle α the solution belongs to $H^{1+s}(\Omega)$ for $s < \frac{\pi}{\alpha}$). Stronger regularity properties may be obtained, however, if special compatibility conditions are satisfied in the corners (cf. Kondrat'ev [173]).

Example 9.29. Let u be the solution of the Poisson equation $-\Delta u = f$ in the square $\Omega = (0,1)\times(0,1)$ with $u = 0$ on Γ. Only for $s < 3$ does $f \in H^{s-2}(\Omega)$ lead to $u \in H^s(\Omega)$. Under the additional compatibility condition that f vanish at all corners,

$$f(0,0) = f(0,1) = f(1,0) = f(1,1) = 0,$$

however, one can also conclude, for $f \in H^{s-2}(\Omega)$ with $3 < s < 4$, that $u \in H^s(\Omega)$.

9.2 Regularity in the Interior

Up to now all regularity statements have referred to the entire domain Ω. Moreover, the kind of regularity is uniform in all parts of Ω. This approach is not realistic. The weaker regularity shown in Examples 2.3 and 2.4 is caused by the corner singularity. Obviously, the point singularity $r^{2/3}\sin((2\varphi - \pi)/3)$ described on page 253 is smooth distant from the origin.

The Laplace equation in $\Omega \subset \mathbb{R}^2$ serves as simplest example for illustrating the interior regularity. Identifying Ω with a subset of $\mathbb{C}$, we can interpret the solutions of $-\Delta u = 0$ as holomorphic functions (cf. Example 1.3). Let $\rho = \rho(z_0) > 0$ be the distance of $z_0 \in \Omega$ from the boundary Γ. The derivatives of u are characterised by the Cauchy formula

$$\frac{\mathrm{d}^k u(z_0)}{\mathrm{d}z^k} = \frac{k!}{2\pi\mathrm{i}} \int_{\partial K_\rho(z_0)} u(\zeta)\, \frac{\mathrm{d}\zeta}{(\zeta - z_0)^{k+1}} \qquad (K_\rho(z_0) \subset \Omega)$$

and can be bounded by $\frac{k!}{\rho^k}\,\|u\|_\infty$. The larger the distance of z_0 from the boundary, the smaller is the bound of the derivative. Note that the shape of the boundary does not matter.

[3] Under stronger conditions Crouzeix–Thomée [80] prove the stability of Q_0 with respect to the L^p and $W^{p,1}$ norms.

9.2.1 Estimates

In the following we study the regularity of the solution u in $\Omega_0 \subset\subset \Omega$. The next theorem shows that the estimate does not depend on the smoothness of the boundary nor on the kind of boundary condition.

Theorem 9.30. *Let $\Omega_0 \subset\subset \Omega_1 \subset \Omega$ and $s \geq 0$. Let the bilinear form (9.5) be $H_0^m(\Omega_1)$-coercive. For the coefficients assume condition (9.18) with Ω be replaced by Ω_1 and with $t \geq s \in \mathbb{N}$ or $t > s$. Let $u \in V \subset H^m(\Omega)$ be a weak solution of the problem $a(u,v) = \int_\Omega f v \mathrm{d}\mathbf{x}$ $(v \in V)$, where the restriction $f|_{\Omega_1}$ belongs to $H^{-m+s}(\Omega_1)$. Then the restriction of u to Ω_0 belongs to $H^{m+s}(\Omega_0)$ and satisfies*

$$\|u\|_{H^{m+s}(\Omega_0)} \leq C(s, \Omega_0, \Omega_1, \Omega) \left[\|f\|_{H^{-m+s}(\Omega_1)} + \|u\|_{H^m(\Omega)} \right] .$$

Proof. A special covering of Ω is given by $U^0 = \Omega_1$, $U^1 = \Omega \backslash \Omega_0$. Thus we obtain the assertion from part (ii) of the proof of Theorem 9.19. ∎

Remark 9.31. In the case of differential equations with constant coefficients and right-hand side[4] Theorem 9.30 holds for all $m + s$, i.e., $u \in C^\infty(\Omega_0)$.

The (multiple) derivatives are bounded by negative powers of the distance $\rho(\mathbf{x}) = \operatorname{dist}(\mathbf{x}, \Gamma)$. This behaviour can be compensated by a weighted norm $\|\rho^{\omega(\alpha)} D^\alpha u\|_{L^2(\Omega)}$. Estimates of this kind are used, e.g., by Melenk [204, §1.4.1].

9.2.2 Behaviour of the Singularity and Green's Function

Singularity functions are examples of solutions of $Lu = 0$ with a singularity only at $\mathbf{x} = \xi$. In the case of constant coefficients they are explicitly known and are used, for example, in the integral equation method. Their behaviour is called *asymptotically smooth*. More precisely, a function $s(\cdot,\cdot)$ which is infinitely differentiable in $\hat{\Omega} := (\Omega \times \Omega) \setminus \{(\mathbf{x}, \mathbf{x}) : \mathbf{x} \in \Omega\}$ is called asymptotically smooth if the following inequality holds:[5]

$$|D_x^\alpha D_y^\beta s(\mathbf{x}, \mathbf{y})| \leq c_{\mathrm{as}}(\alpha + \beta) \, |\mathbf{x} - \mathbf{y}|^{-|\alpha| - |\beta| - \sigma} \quad \text{for} \quad \begin{cases} (\mathbf{x}, \mathbf{y}) \in \hat{\Omega}, \\ \alpha, \beta \in \mathbb{N}_0^d, \\ \alpha + \beta \neq 0, \end{cases} \tag{9.26a}$$

with some $\sigma \in \mathbb{R}$ and

$$c_{\mathrm{as}}(\nu) = C \, \nu! \, |\nu|^r \, \gamma^{|\nu|} \qquad (\nu \in \mathbb{N}_0^d), \tag{9.26b}$$

where C, r, γ are suitable constants. The value of σ depends on the order of the singularity of s.

[4] It is sufficient that these quantities are constant in Ω_1 with $\Omega_0 \subset\subset \Omega_1 \subset \Omega$.

[5] In (9.26a) the case $\alpha + \beta = 0$ is excluded since for $r > 0$ the factor $c_{\mathrm{as}}(0)$ vanishes. A logarithmic singularity s only satisfies (9.26a) for $\alpha + \beta \neq 0$.

Translation invariant singularity functions as those in (2.4a) only depend on $\mathbf{x}-\mathbf{y}$. Since $D_y^\beta s = (-1)^{|\beta|} D_x^\beta s$ inequality (9.26a) reduces to

$$|D_x^\alpha s(\mathbf{x},\mathbf{y})| \le c_{\mathrm{as}}(\alpha)\,|\mathbf{x}-\mathbf{y}|^{-|\alpha|-\sigma} \quad \text{for } (\mathbf{x},\mathbf{y}) \in \hat{\Omega},\ 0 \neq \alpha \in \mathbb{N}_0^d\,. \tag{9.26c}$$

Interpolation error estimates often require directional derivatives $D_t = \sum_{\nu=1}^n t_i \frac{\partial}{\partial x_i}$ with a unit vector $\mathbf{t} \in \mathbb{R}^n$, $|\mathbf{t}| = 1$. Correspondingly, the inequality becomes

$$|D_{t,x}^k s(\mathbf{x},\mathbf{y})| \le C\,p!\,p^r\,\gamma^k\,|\mathbf{x}-\mathbf{y}|^{-k-\sigma} \quad ((\mathbf{x},\mathbf{y}) \in \hat{\Omega},\ k \in \mathbb{N},\ |\mathbf{t}| = 1). \tag{9.26d}$$

The constants in (9.26d) are explicitly known for the singularity function in (2.4a). More generally, the following statement holds for $s(\mathbf{x},\mathbf{y}) = |\mathbf{x}-\mathbf{y}|^{-a}$.

Theorem 9.32. *Let $\mathbf{t} \in \mathbb{R}^n$ with $|\mathbf{t}| = 1$. Then the directional derivatives satisfy*

$$\left| D_{t,x}^k\,|\mathbf{x}-\mathbf{y}|^{-a} \right| \le k!\,\frac{k^{a-1}+\mathcal{O}(k^{a-2})}{\Gamma(a)}\,|\mathbf{x}-\mathbf{y}|^{-k-a} \tag{9.26e}$$

for all $\mathbf{x}-\mathbf{y} \in \mathbb{R}^n$, $\mathbf{x} \neq \mathbf{y}$ and all $k \in \mathbb{N}$. $\Gamma(\cdot)$ is the Gamma function.

Proof. We refer to Hackbusch [140, §E.1.1]. ∎

Consider a function $f(\cdot,\cdot)$ defined on $\Omega \times \Omega$. In numerical applications it might be helpful if its restriction to the Cartesian product $X \times Y \subset \Omega \times \Omega$ with $\mathrm{dist}(X,Y) > 0$ can be approximated by the separable expression

$$\sum_{\nu=1}^r f_{1\nu}(\mathbf{x})\,f_{2\nu}(\mathbf{y}) \tag{9.27}$$

with a small number of terms r.

For fixed $\mathbf{y} \in X$ one can interpolate $f(\cdot,\mathbf{y})$ with respect to the first argument. Choosing the product interpolation by polynomials of degree k, the interpolation error is bounded by

$$\frac{C}{k!}\left[c\,\mathrm{diam}(X)\right]^k \max_{\mathbf{x}\in X, 1\le i\le n} |\partial_{x_i}^k f(\mathbf{x},\mathbf{y})| \qquad \text{(cf. [140, §B.3.2])}.$$

Inserting the estimate (9.26e) together with $|\mathbf{x}-\mathbf{y}| \ge \mathrm{dist}(X,Y)$, we obtain the error bound η^k with $\eta := c\frac{\mathrm{diam}(X)}{\mathrm{dist}}$. If X and Y are chosen so that $\eta < 1$, the interpolation error decays exponentially in k. The interpolating polynomial has the form $\sum_{\|\alpha\|_\infty \le k} c_\alpha \mathbf{x}^\alpha$, where the coefficients c_α depend on $\mathbf{y}$: $\sum_\alpha c_\alpha(\mathbf{y})\mathbf{x}^\alpha$. After a suitable renumbering, this sum corresponds to (9.27) with $r = k^n$.

The (sufficiently large) distance $\mathrm{dist}(X,Y)$ corresponds to the *interior* regularity. In the case of the Green function $G(\mathbf{x},\mathbf{y})$ one cannot expect the same smoothness since $Y \subset \overline{\Omega}$ may touch the boundary. Even if $Y \subset\subset \Omega$, the convergence rate

is determined by $\min\{\operatorname{dist}(Y,\Omega),\operatorname{dist}(X,Y)\}$ instead of $\operatorname{dist}(X,Y)$. The smoothness of the Green function is minimal if the coefficients of the differential operator are not smooth. Nevertheless, one can prove a separable approximation where the exponential convergence rate of (9.27) does not depend on the smoothness of the coefficients. The proof is mainly based on Lemma 9.33 and the Poincaré inequality.

Let the boundary-value problem be given by the bilinear form

$$a(u,v)=\int_\Omega \sum_{i,j=1}^{n} a_{ij}(\mathbf{x})\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i}\mathrm{d}\mathbf{x} \quad \begin{cases} \text{for } u,v\in H_0^1(\Omega) \text{ with} \\ \Omega\in C^{0,1},\, a_{ij}\in L^\infty(\Omega). \end{cases} \tag{9.28a}$$

The coefficients a_{ij} form the symmetric matrix $A(\mathbf{x})$. The uniform ellipticity is quantified by

$$0<\lambda_{\min}\le\lambda\le\lambda_{\max} \quad \begin{cases} \text{for all eigenvalues } \lambda \text{ of } A(\mathbf{x}) \\ \text{in almost all } \mathbf{x}\in\Omega. \end{cases} \tag{9.28b}$$

$$\text{with} \qquad \kappa := \lambda_{\max}/\lambda_{\min}. \tag{9.28c}$$

Because of the lacking smoothness ($a_{ij}\in L^\infty(\Omega)$!) the interior regularity is reduced to the following *Caccioppoli inequality* for the gradients in the interior.

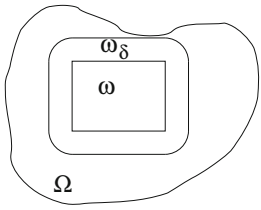


Fig. 9.1 Subdomains $\omega\subset\omega_\delta$ in Ω.

Lemma 9.33. *Let (9.28a–c) hold and $\omega\subset\subset\Omega$ with $\delta:=\operatorname{dist}(\omega,\partial\Omega)$. Let the support*[6] *of $f\in H^{-1}(\Omega)$ be contained in $\overline{\Omega}\backslash\omega_\delta$, where*

$$\omega_\delta := \{\mathbf{x}\in\Omega : \operatorname{dist}(\mathbf{x},\omega)<\delta\}\subset\Omega$$

(cf. Figure 9.1). Let u satisfy[7] *$a(u,v)=f(v)$ for all $v\in H_0^1(\Omega)$. Then we have*

$$\|\nabla u\|_{L^2(\omega)} \le \frac{2\sqrt{\kappa}}{\delta}\|u\|_{L^2(\omega_\delta)}. \tag{9.29}$$

Proof. Choose a cut-off function $\eta\in C^1(\overline{\Omega})$ with $0\le\eta\le 1$ in Ω, $\eta=1$ in ω, and $\operatorname{supp}(\eta)\subset\subset\omega_\delta$. For any $\varepsilon>0$, η can be chosen to that $|\nabla\eta|\le(1+\varepsilon)/\delta$ in $\omega_\delta\backslash\omega$. $\operatorname{supp}(\eta)\subset\subset\omega_\delta$ implies $\eta^2 u=0$ in $\overline{\Omega}\backslash\omega_\delta\supset\partial\omega_\delta$ and

$$\begin{aligned} 0=f(v)=a(u,v)&=\int_{\omega_\delta}(\nabla u)^\top A(\mathbf{x})\,\nabla(\eta^2 u)\mathrm{d}\mathbf{x}\\ &=2\int_{\omega_\delta}\eta\,u(\nabla u)^\top A(\mathbf{x})\,(\nabla\eta)\,\mathrm{d}\mathbf{x}+\int_{\omega_\delta}\eta^2(\nabla u)^\top A(\mathbf{x})\,(\nabla u)\,\mathrm{d}\mathbf{x}\,. \end{aligned}$$

In following chain of inequalities we use the above identity, $\|A\|\le\lambda_{\max}$, $|\nabla\eta|\le(1+\varepsilon)/\delta$, and Schwarz' inequality:

[6] A functional $f\in X'$ has the support K, if K is the smallest set with the property $f(v)=0$ for all $v\in X$ with $\operatorname{supp}(v)\subset\overline{\Omega}\backslash K$.

[7] This implies the weak formulation of $Lu=0$ in ω_δ.

$$\int_{\omega_\delta} \eta^2 \|A^{1/2}\nabla u\|^2 \mathrm{d}\mathbf{x} = \int_{\omega_\delta} \eta^2 (\nabla u)^\top A(\nabla u)\mathrm{d}\mathbf{x} = 2\left|\int_{\omega_\delta} \eta u(\nabla u)^\top A(\nabla \eta)\mathrm{d}\mathbf{x}\right|$$
$$\leq 2\int_{\omega_\delta} \eta\,|u|\,\|A^{1/2}\nabla\eta\|\,\|A^{1/2}\nabla u\|\,\mathrm{d}\mathbf{x}$$
$$\leq 2(1+\varepsilon)\,\frac{\sqrt{\lambda_{\max}}}{\delta}\int_{\omega_\delta} |u|\,\eta\,\|A^{1/2}\nabla u\|\,\mathrm{d}\mathbf{x}$$
$$\leq 2(1+\varepsilon)\,\frac{\sqrt{\lambda_{\max}}}{\delta}\sqrt{\int_{\omega_\delta}\eta^2\|A^{1/2}\nabla u\|^2\mathrm{d}\mathbf{x}}\;\|u\|_{L^2(\omega_\delta)}.$$

Division by $\sqrt{\int_{\omega_\delta}\eta^2\|A^{1/2}\nabla u\|^2\mathrm{d}\mathbf{x}} = \|\eta\,A^{1/2}\nabla u\|_{L^2(\omega_\delta)}$ yields

$$\|\eta\,A^{1/2}\nabla u\|_{L^2(\omega_\delta)} \leq 2(1+\varepsilon)\frac{\sqrt{\lambda_{\max}}}{\delta}\|u\|_{L^2(\omega_\delta)}.$$

Since $\eta = 1$ in ω, we conclude that

$$\|\nabla u\|_{L^2(\omega)} = \|\eta\nabla u\|_{L^2(\omega)} \leq \|\eta\nabla u\|_{L^2(\omega_\delta)} \leq \lambda_{\min}^{-1/2}\|\eta A^{1/2}\nabla u\|_{L^2(\omega_\delta)}.$$

Altogether the inequality (9.29) follows with an additional factor $1+\varepsilon$ for all $\varepsilon > 0$, hence also for $\varepsilon = 0$. ∎

The existence of Green's function is proved for $n \geq 3$ by Grüter–Widman [125] together with the bound $|G(\mathbf{x},\mathbf{y})| \leq \frac{C_G}{\lambda_{\min}}|\mathbf{x}-\mathbf{y}|^{2-n}$ ($C_G = C_G(\kappa)$ with κ in (9.28c), $\lambda_{\min}$ in (9.28b)). For $n = 2$, Doltzmann–Müller [89] prove the existence of G and $|G(\mathbf{x},\mathbf{y})| \leq \frac{C_G}{\lambda_{\min}}\log|\mathbf{x}-\mathbf{y}|$.

Theorem 9.34. *Let (9.28a–c) hold and $\Omega \in C^{0,1}$. Let $X, Y \subset \overline{\Omega}$ be subdomains with*

$$X \text{ convex}, \quad \operatorname{diam}(X) \leq \eta \operatorname{dist}(X,Y) \text{ for some } \eta > 0.$$

Then the Green function has a separable expansion

$$G(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{\infty} u_i^{(k)}(\mathbf{x})\,v_i^{(k)}(\mathbf{y}) \qquad \text{for } x \in X,\ y \in Y. \tag{9.30}$$

The Green function and the partial sum $G_k(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{k} u_i^{(k)}(\mathbf{x})\,v_i^{(k)}(\mathbf{y})$ define maps $\mathcal{G}, \mathcal{G}_k \in \mathcal{L}\left(L^2(Y), L^2(X)\right)$ via $\mathcal{G}f = \int_X \mathcal{G}(\cdot,\mathbf{y})f(\mathbf{y})\,\mathrm{d}\mathbf{y}$. The error estimate

$$\|\mathcal{G} - \mathcal{G}_k\|_{L^2(X)\leftarrow L^2(Y)} \leq \varepsilon_k \|\mathcal{G}\|_{L^2(X)\leftarrow L^2(\Omega\setminus X)} \quad \text{with } \varepsilon_k \leq c_1\exp(-c_2 k^{c_3})$$

shows exponential convergence of the series (9.30). The constants are $c_1 \approx 1$, $c_2 \approx \left[2\,\mathrm{e}\,c_{\mathrm{appr}}\sqrt{\kappa_C}\,(\eta+2)\right]^{d/(d+1)}$, and $c_3 = \frac{1}{d+1}$.

The proof can be found in [140, §11.3] as well as in Bebendorf–Hackbusch [33] with generalisation by Bebendorf [31, 32]. See also the literature mentioned at the end of Section 9.3.6.

9.3 Regularity Properties of Difference Equations

The convergence estimates for difference equations in §4.4 read, for example, $\|u-u_h\|_\infty \le Ch^2\,\|u\|_{C^4(\overline{\Omega})}$ under the condition that $u \in C^4(\overline{\Omega})$ or $u \in C^{3,1}(\overline{\Omega})$. This regularity assumption is frequently not satisfied (cf. Examples 2.3–2.4). Note that the differentiation order in C^4 and C^0 differs by four, while the error is only proportional to the square of h. In contrast, the finite-element error estimate $|u-u^h|_k \le Ch^{2-k}|u|_2$ $(k=0,1)$ is optimal in the sense that the difference of the differentiation orders involved in $|\cdot|_k$ and $|\cdot|_2$ coincides with the power of h. In this section we aim at similar estimates for difference solutions. To obtain the latter, one needs to replace the stability estimate $\|L_h^{-1}\|_2 \le C$ (or $\|L_h^{-1}\|_\infty \le C$), which corresponds to $L^{-1} \in L(L^2(\Omega), L^2(\Omega))$, by stronger estimates which correspond to $L^{-1} \in L(H^{-1}(\Omega), H_0^1(\Omega))$ or $L^{-1} \in L(L^2(\Omega), H^2(\Omega))$.

Results of this kind are proved by Auzinger [12], Dryja [91], Emmrich–Grigorieff [95], Hackbusch [129, 130], and Lapin [181]. The monograph Jovanović–Süli [163] contains further literature about this subject.

9.3.1 Discrete H^1-Regularity

We define an infinite grid Q_h and a grid Ω_h contained in $\Omega \subset \mathbb{R}^n$:

$$Q_h := \{\mathbf{x} \in \mathbb{R}^n : \; x_i = \nu_i h,\; \nu_i \in \mathbb{Z}\}\,, \qquad \Omega_h := \Omega \cap Q_h.$$

A grid function v_h defined on Ω_h is extended to Q_h by $v_h = 0$:

$$v_h(\mathbf{x}) := 0 \qquad \text{for all } \mathbf{x} \in Q_h \backslash \Omega_h. \tag{9.31}$$

We denote the vector space of all grid functions v_h with (9.31) by $L_h^2 = L_h^2(\Omega_h)$. The Euclidean norm is now called the L_h^2-norm:

$$|v_h|_0 := \|v_h\|_{L_h^2} := \sqrt{h^n \sum_{\mathbf{x}\in Q_h} |v_h(\mathbf{x})|^2}.$$

It comes from the scalar product

$$(v_h, w_h)_0 := (v_h, w_h)_{L_h^2} := h^n \sum_{\mathbf{x}\in Q_h} v_h(\mathbf{x})\, w_h(\mathbf{x})\,.$$

The discrete analogue of $H_0^1(\Omega)$ is H_h^1 with the norm

$$|v_h|_1 := \|v_h\|_{H_h^1} := \sqrt{|v_h|_0^2 + \sum_{i=1}^n \left|\partial_i^+ v_h\right|_0^2} \qquad (\text{note that } v_h = 0 \text{ on } Q_h\backslash\Omega_h)\,,$$

where ∂_i^+ is the forward difference in the x_i direction. The dual norm reads

$$|v_h|_{-1} := \|v_h\|_{H_h^{-1}} := \sup\big\{\, |(v_h, w_h)_0| \,/\, |w_h|_1 : w_h \neq 0 \text{ satisfies (9.31)}\big\}.$$

The associated matrix norms $\|L_h\|_{H_h^1 \leftarrow H_h^{-1}} = |L_h|_{1\leftarrow -1}$, $\|L_h\|_{H_h^1 \leftarrow L_h^2} = |L_h|_{1\leftarrow 0}$, etc., are defined by

$$|L_h|_{i\leftarrow -j} := \sup\left\{ \frac{|L_h v_h|_i}{|v_h|_j} : 0 \neq v_h \text{ satisfies (9.31)} \right\} \quad \text{for } i,j \in \{-1,0,1\}\,.$$

Exercise 9.35. Show that (a) $|L_h|_{0\leftarrow 0}$ is the spectral norm of L_h (cf. §4.3).
(b) The following *inverse estimates* hold:

$$|v_h|_i \le C_{ij} h^{j-i} |v_h|_j \qquad \text{for } 1 \ge i \ge j \ge -1. \tag{9.32}$$

The difference operator L_h yields the bilinear form

$$a_h(v_h, w_h) := (L_h v_h, w_h)_{L_h^2}\,.$$

$a_h(\cdot,\cdot)$ is said to be H_h^1-*elliptic* if a $C_E > 0$ exists such that

$$a_h(u_h, u_h) \ge C_E\, |u_h|_1^2 \qquad \text{for all } u_h \text{ and all } h > 0. \tag{9.33a}$$

Correspondingly, $a_h(\cdot,\cdot)$ is said to be H_h^1-*coercive* if there exist $C_E > 0$ and $C_K \in \mathbb{R}$ with

$$a_h(u_h, u_h) \ge C_E\, |u_h|_1^2 - C_K\, |u_h|_0^2 \qquad \text{for all } u_h \text{ and all } h > 0.$$

As defined in Section 4.5, L_h (resp. $a_h(\cdot,\cdot)$) is said to be L_h^2-*stable* if

$$\big|L_h^{-1}\big|_{0\leftarrow 0} \le C_0 \qquad \text{for all } h > 0. \tag{9.33b}$$

We call L_h H_h^1-*regular* if [8]

$$\big|L_h^{-1}\big|_{1\leftarrow -1} \le C_1 \qquad \text{for all } h > 0.$$

Exercise 9.36. (a) H_h^1-regularity implies L_h^2-stability.
(b) $|L_h|_{i\leftarrow -j} = |L_h^{\mathsf{T}}|_{-j\leftarrow -i}$ for all $i,j \in \{-1,0,1\}$.
(c) If L_h is H_h^1-regular, then so is L_h^{T}.
(d) If L_h and L_h^{T} are stable with respect to $|\cdot|_\infty$, i.e., $\big\|L_h^{-1}\big\|_\infty \le C_\infty$ and $\|L_h^{-\mathsf{T}}\|_\infty \le C_1$, then L_h^2-stability (9.33b) follows with $C_0 := \sqrt{C_1 C_\infty}$.
(e) H_h^1-ellipticity implies H_h^1-regularity.

The following statement resembles the alternative in Theorem 6.107.

[8] This definition is slightly different from the H^1-regularity (9.3) in the continuous case.

Theorem 9.37. *If $a_h(\cdot,)$ is H_h^1-coercive and if L_h is L_h^2-stable, then L_h is also H_h^1-regular.*

Proof. (i) Let u_h be arbitrary and set $f_h := L_h^{\mathsf{T}} u_h$. Then the identity $a_h(u_h,u_h) = (u_h, L_h^{\mathsf{T}} u_h)_0 = (u_h,f_h)_0$ holds. Coercivity provides

$$\begin{aligned}|u_h|_1^2 &\le \left[a_h(u_h,u_h) + C_K\,|u_h|_0^2\right]/C_E = \left[(u_h,f_h)_0 + C_K\,|u_h|_0^2\right]/C_E \\ &\le C'\left[\,|f_h|_0 + C_K\,|u_h|_0\,\right]|u_h|_0\ .\end{aligned}$$

On the basis of the stability estimate $|u_h|_0 \le C_0\,|f_h|_0$ we obtain $|u_h|_1^2 \le C''\,|f_h|_0^2$. From this one infers $\left|L_h^{-\mathsf{T}}\right|_{1\leftarrow 0} \le \sqrt{C''} =: C^\star$ and hence $\left|L_h^{-1}\right|_{0\leftarrow -1} \le C^\star$ (cf. Exercise 9.36b).

(ii) Now let $f_h = L_h u_h$. According to part (i), one has $|u_h|_0 \le C^\star\,|f_h|_{-1}$. By estimating

$$|a_h(u_h,u_h)| = |(f_h,u_h)_0| \le |f_h|_{-1}\,|u_h|_1$$

through $\frac{1}{2}C_E\,|u_h|_1^2 + \frac{1}{2}C_E^{-1}\,|f_h|_{-1}^2$ according to (5.34) one obtains the stability

$$\begin{aligned}|u_h|_1^2 &\le \frac{a_h(u_h,u_h) + C_K\,|u_h|_0^2}{C_E} \le \frac{\frac{1}{2}C_E\,|u_h|_1^2 + \frac{1}{2}C_E^{-1}\,|f_h|_{-1}^2 + C_K\,|u_h|_0^2}{C_E} \\ &\le \frac{1}{2}\,|u_h|_1^2 + \frac{\frac{1}{2}C_E^{-1} + C_K\,(C^*)^2}{C_E}\,|f_h|_{-1}^2\end{aligned}$$

so that $|L_h^{-1}|_{1\leftarrow -1} \le C_1$ with $C_1 = \frac{\sqrt{1+2C_K(C^\star)^2C_E}}{C_E}$. ∎

Instead of the L_h^2-stability in Theorem 9.37 one can also assume the solvability of the continuous problem and a consistency condition (cf. Corollary 11.38).

By analogy with Lemma 7.12 the following may be proved.

Exercise 9.38. If L_h is H_h^1-coercive, and if $\delta L_h := a_0 + \sum_i b_i\,\partial_i^{\pm} + \sum_i \partial_i^{\pm} c_i$, with $|a_0|$, $|b_i|$, $|c_i| \le \text{const}$, contains at most first differences, then $L_h + \delta L_h$ is also H_h^1-coercive.

According to Exercise 9.38 it is sufficient to investigate the principal part of a difference operator as to its coercivity.

Example 9.39. Let $\Omega \subset \mathbb{R}^2$ be bounded. On Ω_h let L_h. be given by the difference method (5.19) (with $a = 0$). For this purpose assume that the coefficients a_{11}, a_{22} are extended to $\mathbb{R}^2$ (only those values are essential which appear in the following equation for $(x,y) \in \Omega_h$):

$$(L_h u_h)(x,y) = \partial_x^-\left(a_{11}\big(x+\tfrac{h}{2},y\big)\,\partial_x^+ u_h(x,y)\right) + \partial_y^-\left(a_{22}\big(x,y+\tfrac{h}{2}\big)\,\partial_y^+ u_h(x,y)\right),$$

where $u_h(x,y) = 0$ for $(x,y) \in Q_h \backslash \Omega_h$ according to (9.31). Here, let $-a_{11} \ge \varepsilon > 0$, $-a_{22} \ge \varepsilon > 0$ in $\overline{\Omega}$. Then $a_h(\cdot,\cdot)$ is H_h^1-elliptic and L_h is H_h^1-regular.

Proof. (i) For arbitrary $v_h,\ w_h$ defined on Q_h the following rules of summation by parts hold:

$$\left(v_h, \partial^+_{x_i} w_h\right)_0 = -\left(\partial^-_{x_i} v_h, w_h\right)_0, \qquad \left(v_h, \partial^-_{x_i} w_h\right)_0 = -\left(\partial^+_{x_i} v_h, w_h\right)_0 .$$

(ii) In the scalar product $(L_h u_h, w_h)_0$ the values $(L_h u_h)(x,y)$ for $(x,y) \in Q_h \backslash \Omega_h$ are multiplied by $w_h(x,y) = 0$ (cf. (9.31)). Therefore at these points the coefficients a_{11}, a_{22} can be defined arbitrarily. Hence it follows that

$$\begin{aligned} a_h(u_h,u_h) &= (L_h u_h, u_h)_0 \\ &= -\left(a_{11}\left(\cdot + \tfrac{h}{2}, \cdot\right) \partial^+_x u_h, \partial^+_x u_h\right)_0 - \left(a_{22}\left(\cdot, \cdot + \tfrac{h}{2}\right) \partial^+_y u_h, \partial^+_y u_h\right)_0 \\ &\geq \varepsilon \left[\left|\partial^+_x u_h\right|^2_0 + \left|\partial^+_y u_h\right|^2_0 \right]. \end{aligned}$$

As in Lemma 6.29 the latter expression is bounded from below by $\varepsilon \varepsilon^2_\Omega |u_h|^2_1$ so that (9.33a) follows. The H^1_h-regularity results from Exercise 9.36e. ∎

Exercise 9.40. Let $\Omega \subset \mathbb{R}^2$ be bounded. Let the equation

$$(a_{11}u_x)_x + (a_{12}u_y)_x + (a_{12}u_x)_y + (a_{22}u_y)_y = f$$

be given on Ω_h by the difference stars (5.19), (5.20), where $u_h = 0$ in $Q_h \backslash \Omega_h$. Let the differential equation be uniformly elliptic in $\overline{\Omega}$: $a_{ii} < 0,\ a_{11}a_{22} - a^2_{12} > 0$. Further, let $a_{ij} \in C^0(\overline{\Omega})$ hold. Show that for sufficiently small h the associated matrix L_h is H^1_h-regular; for all $h > 0$, L_h is H^1_h-coercive. *Hint:* Uniform ellipticity also implies $-\sum a_{ij}\xi_i\xi_j \geq \varepsilon |\xi|^2$ for some $\varepsilon > 0$.

Lemma 6.29 and its proof transfer without difficulty to the discrete case.

Lemma 9.41. *For bounded domains the norms*

$$|v_h|_{1,0} := \sqrt{\sum_{i=1}^{n} \left|\partial^+_i v_h\right|^2}$$

and $|v_h|_1$ *are equivalent:* $|v_h|_1 \geq |v_h|_{1,0} \geq \varepsilon_\Omega |v_h|_1$.

The H^1_h-coercive difference methods constructed so far remain H^1_h-coercive if differences of lower order are added (cf. Exercise 9.38) or if the principal term $(a_{11}u_x)_x + \ldots$ is replaced by $a_{11}u_{xx} + \ldots$ with $a_{11} \in C^1(\overline{\Omega})$. The above difference methods are described by the same difference operator regardless whether the grid points are close to or far from the boundary. The homogeneous Dirichlet boundary condition is discretised by (9.31): '$u_h = 0$ on $Q_h \backslash \Omega_h$'. If one wants to approximate the boundary condition more accurately, one needs to select special discretisations in the points near the boundary of Ω (cf. Sections 4.8.1–2). One thus obtains an irregularity which complicates the proof of H^1_h-regularity as, e.g., in Example 9.39.

We begin with the one-dimensional case.

Lemma 9.42. *Let L_h be the matrix of the one-dimensional Shortley–Weller discretisations of $-u'' = f$ in Ω, $u = 0$ on $\partial\Omega$:*

$$h^{-2}\left\{\frac{2}{s_\ell s_r}u_h(x) - \frac{2}{s_\ell\,(s_\ell+s_r)}u_h(x-s_\ell h) - \frac{2}{s_r\,(s_\ell+_r)}u_h(x+s_r h)\right\} = f(x) \tag{9.34}$$

for $x \in \Omega_h$. Here $0 < s_\ell,\ s_r \le 1$ (cf. (4.89)) have the value 1 except for the first [last] grid point $x \in \Omega_h$, where they are defined by $x - s_\ell h \in \partial\Omega$ $[x + s_r h \in \partial\Omega]$. At the boundary points $\xi \in \partial\Omega$ we set $u_h(\xi) = 0$. For arbitrary $\Omega \subset \mathbb{R}$ there holds

$$(v_h, L_h v_h)_0 \ge c\,\left|\partial^+ v_h\right|_0^2 \qquad \text{with } c := \sqrt{3} - \tfrac{3}{4} \ge 0.982$$

for all v_h with $v_h = 0$ on $Q_h\backslash\Omega_h$. Thus for bounded Ω, L_h is H_h^1-regular.

Proof. (i) First, let Ω be assumed to be connected. Let the grid points of Ω_h be

$$x_j := x_0 + jh \in \Omega \qquad \text{for} \quad j = 0, \dots, k > 0.$$

Let the boundary points of Ω_h be $x_0 - s_{\ell,0}h$ and $x_k + s_{r,k}h$ with $s_{\ell,0}, s_{r,k} \in (0,1]$. The other factors of equation (9.34) in $x = x_i$ are $s_{\ell,i} = s_{r,i} = 1$. Taking into account $v_h = 0$ on $\mathbb{R}\backslash\Omega$, we obtain the following identity:

$$\begin{aligned}(v_h, L_h v_h)_0 &= h\sum_{j=0}^{k} v_h(x_j)\,(L_h v_h)(x_j)\\ &= \left|\partial^+ v_h\right|_0^2 + \frac{1}{h}\left\{\begin{array}{l} v_h(x_0)\left[\left(\frac{2}{s_\ell}-2\right)v_h(x_0) + \left(1-\frac{2}{1+s_\ell}\right)v_h(x_1)\right]\\ +\,v_h(x_k)\left[\left(\frac{2}{s_r}-2\right)v_h(x_k) + \left(1-\frac{2}{1+s_r}\right)v_h(x_{k-1})\right]\end{array}\right\},\end{aligned}$$

where $s_\ell := s_{\ell,0}$ and $s_r := s_{r,k}$. For the proof we write L_h as $-\partial^-\partial^+ + \delta$. The identity $(v_h, -\partial^-\partial^+ v_h)_0 = |\partial^+ v_h|_0^2$ is already shown in the proof of Example 9.39. The perturbation δ is a difference operator with $\delta v_h = 0$ for all points except x_0 and x_k. Therefore $(v_h, \delta v_h)_0$ produces the expression $\frac{1}{h}\{\dots\}$.

Because $v_h(x_{-1}) = 0$ we may write $v_h(x_0) = h\partial^+ v_h(x_{-1})$ and $v_h(x_1) = h[\partial^+ v_h(x_{-1}) + \partial^+ v_h(x_0)]$. Therefore the first part of $\frac{1}{h}\{\dots\}$ becomes

$$\begin{aligned}&\frac{1}{h}v_h(x_0)\left[\left(\frac{2}{s_\ell}-2\right)v_h(x_0) + \left(1-\frac{2}{1+s_\ell}\right)v_h(x_1)\right] \qquad (9.35a)\\ &\quad = h\frac{1-s_\ell}{1+s_\ell}\left[\frac{2+s_\ell}{s_\ell}\left(\partial^+ v_h(x_{-1})\right)^2 - \partial^+ v_h(x_{-1})\partial^+ v_h(x_0)\right].\end{aligned}$$

For $\alpha := \partial^+ v_h(x_{-1})$ and $\beta := \partial^+ v_h(x_0)$ apply $-\alpha\beta \ge -\frac{\lambda}{2}\alpha^2 - \frac{1}{2\lambda}\beta^2$ with $\frac{\lambda}{2} = (2+s_\ell)/s_\ell$ (cf. (5.34)): $\frac{2+s_\ell}{s_\ell}\alpha^2 - \alpha\beta \ge -\frac{1/4}{\lambda/2}\beta^2 = -\frac{1}{4}\frac{s_\ell}{s_\ell+2}\beta^2$. Since

$$\frac{s_\ell(1-s_\ell)}{4(1+s_\ell)(2+s_\ell)} \le \frac{1}{16\sqrt{3}+28} < 0.018 \quad (\text{maximum at } s_\ell = \tfrac{\sqrt{3}-1}{2}),$$

the expression (9.35a) is bounded by $\ge -\frac{1}{16\sqrt{3}+28}h\left|\partial^+ v_h(x_0)\right|^2$ from below. Treating the second term similarly, one obtains

$$(v_h, L_h v_h)_0 \geq \left|\partial^+ v_h\right|_0^2 - \frac{1}{16\sqrt{3}+28}\left[h\left(\partial^+ v_h(x_0)\right)^2 + h\left(\partial^+ v_h(x_{k-1})\right)^2\right]$$
$$\geq c\left|\partial^+ v_h\right|_0^2 \tag{9.35b}$$

with $c = 1 - \frac{1}{16\sqrt{3}+28} = \sqrt{3} - \frac{3}{4} > 0.982$.

(ii) In part (i) $k > 0$ is assumed. For $k = 0$ we have

$$(v_h, L_h v_h)_0 = \frac{1}{h}\frac{2}{s_\ell s_r}\left(v_h(x_0)\right)^2 = \frac{h}{s_\ell s_r}\left[\left(\partial^+ v_h(x_0)\right)^2 + \left(\partial^+ v_h(x_{-1})\right)^2\right]$$
$$\geq \left|\partial^+ v_h\right|_0^2,$$

so that again (9.35b) is valid.

(iii) For an arbitrary Ω consisting of intervals $I_i = (a_i, b_i)$ $(i \in \mathbb{Z})$ with $b_i \leq a_{i+1}$ summation of the single terms yields the desired estimate. ∎

Theorem 9.43. *Let L_h be the matrix associated with the Shortley–Weller discretisation of the Poisson equation in $\Omega \subset \mathbb{R}^n$ (cf. Section 4.8.1). Then L_h is H_h^1-coercive. If Ω is bounded, L_h is also H_h^1 -elliptic and H_h^1-regular.*

Proof. Split $L_h = L_h^x + L_h^y$ into x- and y-differences (analogously for $n > 2$). Lemma 9.42 proves $(v_h, L_h^x v_h)_0 \geq c\left|\partial_x^+ v_h\right|_0^2$ and $(v_h, L_h^y v_h)_0 \geq c|\partial_y^+ v_h|_0^2$. Therefore $(v_h, L_h v_h)_0 \geq c[\,|\partial_x^+ v_h|_0^2 + |\partial_y^+ v_h|_0^2\,] \geq c\left[\,|v_h|_1^2 - |v_h|_0^2\,\right]$ shows that L_h is H_h^1-coercive. By Lemma 9.41 H_h^1 -ellipticity holds for bounded domains. ∎

Theorem 9.44. *Let Ω be bounded. Let the Poisson equation be discretised by the five-point formula (4.95c) in the interior points and by the interpolation (4.96) in near-boundary points. The corresponding matrix is H_h^1-regular.*

Proof. The corresponding, one-dimensional formulae, except for the scaling factor $s_r + s_\ell \leq 2$, agree with (9.34) so that the proof of Theorem 9.43 can easily be carried over. ∎

Theorem 9.43 and 9.44 can be strengthened in the following way.

Corollary 9.45. Let L_h be as in Theorem 9.43 or 9.44. Define the diagonal matrix $D_h = \operatorname{diag}\{d(\mathbf{x}) : \mathbf{x} \in \Omega_h\}$ by

$$d(\mathbf{x}) = \min\{2s_\ell s_r, 2s_o s_u, 1\} \qquad (\mathbf{x} \in \Omega_h),$$

where s_ℓ, s_r, s_o, s_u come from (4.89), respectively (4.96), respectively. Evidently, $d(\mathbf{x}) = 1$ holds for far-boundary points $\mathbf{x} \in \Omega_h$. The matrix

$$L_h' := D_h L_h$$

belonging to the re-scaled system $L_h' u_h = f_h' := D_h f_h$ is also H_h^1-regular: $|{L_h'}^{-1}|_{1\leftarrow -1} \leq C$.

Proof. In the Shortley–Weller case (4.89) one has to add the correction terms (9.35a) for the x and y directions. One verifies that they are bounded from below by $-\frac{1}{2}[\,|\partial_x^+ v_h|_0^2 + |\partial_y^+ v_h|_0^2\,]$ so that

$$(v_h, L_h' v_h)_0 \geq \frac{1}{2}\left[\,|\partial_x^+ v_h|_0^2 + |\partial_y^+ v_h|_0^2\,\right] \geq \varepsilon\,|v_h|_1^2 \qquad \text{with } \varepsilon > 0. \quad \blacksquare$$

An analogous estimate holds for the difference schemes in §4.8.2.

9.3.2 Consistency

In the following we carry over the estimate $|u^h - u|_1 \leq Ch\,|u|_2 \leq C'h\,|f|_0$ which holds for finite-element solutions, to difference methods. To this end one needs to prove the *consistency condition*

$$|L_h R_h - \check{R}_h L|_{-1\leftarrow 2} := \|L_h R_h - \check{R}_h L\|_{H_h^{-1}\leftarrow H^2(\Omega)} \leq C_K h \tag{9.36}$$

for suitable *restrictions*

$$R_h : H^2(\Omega) \cap H_0^1(\Omega) \to H_h^2, \qquad \check{R}_h : L^2(\Omega) \to L_h^2.$$

To construct the restrictions we first continue $u \in H^2(\Omega)$ in $\bar{u} := E_2 u \in H^2(\mathbb{R}^2)$. According to Theorem 6.58c one assumes

$$\bar{u} := E_2 u = u \text{ in } \Omega, \quad \|E_2 u\|_{H^2(\mathbb{R}^2)} \leq C\,\|u\|_{H^2(\Omega)} \text{ for all } u \in H^2(\Omega). \tag{9.37}$$

An analogous continuation $E_0 : L^2(\Omega) \to L^2(\mathbb{R}^2)$ with

$$\bar{f} := E_0 f = f \text{ in } \Omega, \quad \|E_0 f\|_{L^2(\mathbb{R}^2)} \leq C\,\|f\|_{L^2(\Omega)} \quad \text{for all } f \in L^2(\Omega)$$

is given, for example, by $\bar{f} = 0$ on $\mathbb{R}^2 \backslash \Omega$. Its dual map

$$E_0^* : L^2(\mathbb{R}^2) \to L^2(\Omega)$$

is the restriction to Ω. Let the averaging operators $\sigma_h^x, \sigma_h^y : C_0^\infty(\mathbb{R}^2) \to C_0(\mathbb{R}^2)$ be defined by

$$(\sigma_h^x u)(x,y) := \frac{1}{h}\int_{-h/2}^{h/2} u(x+\xi, y)\mathrm{d}\xi, \quad (\sigma_h^y u)(x,y) := \frac{1}{h}\int_{-h/2}^{h/2} u(x, y+\eta)\mathrm{d}\eta. \tag{9.38}$$

The restrictions $R_h,\ \check{R}_h$ are chosen as follows:[9]

$$R_h := \sigma_h^x \sigma_h^y E_2, \text{ i.e., } (R_h u) = \frac{1}{h^2} \int_{-h/2}^{h/2} \int_{-h/2}^{h/2} \bar{u}(x+\xi, y+\eta) \mathrm{d}\xi \mathrm{d}\eta,\ \bar{u} := E_2 u, \tag{9.39a}$$

$$\check{R}_h := \left(\sigma_h^x \sigma_h^y\right)^2 E_0, \tag{9.39b}$$

$$\text{i.e., } (\check{R}_h u)(x,y) = \frac{1}{h^4} \int_{-h/2}^{h/2} \int_{-h/2}^{h/2} \int_{-h/2}^{h/2} \int_{-h/2}^{h/2} \bar{f}(x+\xi+\xi', y+\eta+\eta')\, \mathrm{d}\xi\, \mathrm{d}\xi'\, \mathrm{d}\eta\, \mathrm{d}\eta'$$

with $\bar{f} := E_0 f$. The characteristic properties of the convolutions σ_h^x and σ_h^y are the subject of the next exercise. Part (a) describes the correspondence to divided differences and derivatives. Part (d) connects the norms of L_h^2 and $L^2(\mathbb{R}^2)$.

Exercise 9.46. Let $\hat{\partial}_x$ be the symmetric difference operator $(\hat{\partial}_x u)(x,y) := \frac{1}{h}[u(x+\frac{h}{2}, y) - u(x-\frac{h}{2}, y)]$. $\hat{\partial}_y$ is defined analogously. Show that

(a) $\sigma_h^x \sigma_h^y = \sigma_h^y \sigma_h^x, \quad \hat{\partial}_x = \frac{\partial}{\partial x} \sigma_h^x = \sigma_h^x \frac{\partial}{\partial x}, \quad \hat{\partial}_y = \frac{\partial}{\partial y} \sigma_h^y = \sigma_h^y \frac{\partial}{\partial y}$.

(b) The averaging operators are selfadjoint: $(\sigma_h^x)^* = \sigma_h^x,\ (\sigma_h^y)^* = \sigma_h^y$.

(c) $\|\sigma_h^x\|_{H^k(\mathbb{R}^2) \to H^k(\mathbb{R}^2)} \le 1$ and $\|\sigma_h^y\|_{H^k(\mathbb{R}^2) \to H^k(\mathbb{R}^2)} \le 1$ hold in particular for $k = 0, \pm 1, 2$.

(d) $\|\sigma_h^x \sigma_h^y v\|_{L_h^2} \le \|v\|_{L^2(\mathbb{R}^2)}$ for all $v \in L^2(\mathbb{R}^2)$.

(e) If $a \in C^{0,1}(\mathbb{R}^2)$ then $\|a \sigma_h^x \sigma_h^y - \sigma_h^x \sigma_h^y a\|_{L_h^2 \leftarrow L^2(\mathbb{R}^2)} \le Ch \|a\|_{C^{0,1}(\mathbb{R}^2)}$. Here we use the notation

$$\left(a \sigma_h^x u\right)(\mathbf{x}) = a(\mathbf{x}) \left(\left(\sigma_h^x u\right)(\mathbf{x})\right) \quad \text{and} \quad \left(\sigma_h^x a u\right)(\mathbf{x}) = \left(\sigma_h^x (a u)\right)(\mathbf{x}).$$

(f) $\|a \left(\sigma_h^x\right)^\nu \left(\sigma_h^y\right)^\mu - \left(\sigma_h^x\right)^\nu \left(\sigma_h^y\right)^\mu a\|_{L_h^2 \leftarrow L^2(\mathbb{R}^2)} \le h\,(\nu+\mu)\, \|a\|_{C^{0,1}(\mathbb{R}^2)}$ for $\nu, \mu \in \mathbb{N}$.

(g) $\|u - \left(\sigma_h^x\right)^\nu \left(\sigma_h^y\right)^\mu u\|_{H^k(\mathbb{R}^2)} \le C_{\nu\mu} h \|u\|_{H^{k+1}(\mathbb{R}^2)}$ for $u \in H^{k+1}(\mathbb{R}^2)$.

Consider the differential operator

$$L = \sum_{i,j=1}^n a_{ij}(\mathbf{x}) \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^n a_i(\mathbf{x}) \frac{\partial^2}{\partial x_i} + a(\mathbf{x}). \tag{9.40a}$$

First we assume that $\Omega = \mathbb{R}^n$ and discretise L with the regular difference operator

$$L_h = \sum_{i,j=1}^n a_{ij}(\mathbf{x}) \partial_{x_i}^{\pm} \partial_{x_j}^{\pm} + \sum_{i=1}^n a_i(\mathbf{x}) \partial_{x_i}^{\pm} + a(\mathbf{x}) \tag{9.40b}$$

with an arbitrary combination of the $\pm$-signs.

[9] The following definitions refer to the two-dimensional case. The generalisation to $\Omega \subset \mathbb{R}^n$ is obvious.

Lemma 9.47. *Let $\Omega = \mathbb{R}^n$. Let $a_{ij}, a_j, a \in C^{0,1}(\mathbb{R}^n)$. Let L and L_h be given by (9.40a,b). Then the consistency estimate (9.36) holds.*

Proof. To simplify the notation let us assume that $n = 2$. Let

$$\check{R}_h a_{11} u_{xx}(\mathbf{x}) = \left(\sigma_h^x \sigma_h^y\right)^2 a_{11} u_{xx}(\mathbf{x}) = a_{11}(\mathbf{x}) \left(\sigma_h^x \sigma_h^y\right)^2 u_{xx}(\mathbf{x}) - \delta_1$$

hold with $\|\delta_1\|_{L_h^2} \leq Ch \|a_{11}\|_{C^{0,1}(\mathbb{R}^2)} |u|_2$ (cf. Exercise 9.46f,c). Let the term corresponding to $a_{11}u_{xx}$ in (9.40b) be, for example, $a_{11}(\mathbf{x})\partial_x^+\partial_x^+$. Exercise 9.46a shows that

$$\begin{aligned} a_{11} \left(\sigma_h^x \sigma_h^y\right)^2 u_{xx}(x,y) &= a_{11}\hat{\partial}_x\hat{\partial}_x\sigma_h^y\sigma_h^y u(x,y) = a_{11}\partial_x^+\partial_x^+\sigma_h^y\sigma_h^y u(x-h,y) \\ &= a_{11}\partial_x^+\partial_x^+\sigma_h^y\sigma_h^y u(x,y) - a_{11}\partial_x^+\delta_2 \end{aligned}$$

with

$$\delta_2(x,y) := -\partial_x^+\sigma_h^y\sigma_h^y[u(x-h,y)-u(x,y)] = -\sigma_h^x\sigma_h^y\sigma_h^y[u_x(x-h,y)-u_x(x,y)]$$

and

$$\|\delta_2\|_2 \leq \|\sigma_h^y[u_x(x-h,y) - u_x(x,y)]\|_{L^2(\mathbb{R}^2)} \leq h\,|u|_2$$

(cf. Exercise 9.46d,c). Since $\hat{\partial}_x\hat{\partial}_x = \partial_x^+\partial_x^-$, the error term δ_2 does not appear if one also approximates $a_{11}u_{xx}$ by $a_{11}(x)\partial_x^+\partial_x^-$. Finally, one obtains

$$\begin{aligned} a_{11}\partial_x^+\partial_x^+\sigma_h^y\sigma_h^y u &= a_{11}\partial_x^+\partial_x^+ R_h u + a_{11}\partial_x^+\partial_x^+ \left[\sigma_h^y\sigma_h^y - \sigma_h^x\sigma_h^y\right] u \\ &= a_{11}\partial_x^+\partial_x^+ R_h u + a_{11}\partial_x^+\sigma_h^x\sigma_h^y \left[\sigma_h^y - \sigma_h^x\right] u_x \\ &= a_{11}\partial_x^+\partial_x^+ R_h u - a_{11}\partial_x^+\delta_3 \end{aligned}$$

with $\|\delta_3\|_{L_h^2} = \|\sigma_h^x\sigma_h^y[\sigma_h^y - \sigma_h^x]u_x\|_{L_h^2} \leq \|[\sigma_h^y - \sigma_h^x]u_x\|_{L^2(\mathbb{R}^2)} \leq Ch\,|u|_2$ (cf. Exercise 9.46d,g). Putting this altogether one obtains

$$\left[a_{11}\partial_x^\pm\partial_x^\pm R_h - \check{R}_h a_{11}\frac{\partial^2}{\partial x^2}\right] u = \delta_1 + a_{11}\,\partial_x^\pm\left(\delta_2 + \delta_3\right).$$

For the first error term the following holds:

$$\|\delta_1\|_{H_h^{-1}} \leq \|\delta_1\|_{L_h^2} \leq Ch\,\|a_{11}\|_{C^{0,1}(\mathbb{R}^2)}\,|u|_2\,. \tag{9.41a}$$

For arbitrary $v_h \in H_h^1$ we have

$$\left(v_h, a_{11}\partial_x^\pm(\delta_2+\delta_3)\right)_{L_h^2} = -\left(\partial_x^\mp[a_{11}v_h], \delta_2+\delta_3\right)_{L_h^2}\,.$$

From $a \in C^{0,1}(\mathbb{R}^2)$ it follows that

$$\left\|\partial_x^\pm[av_h]\right\|_{L_h^2} \leq \|a\|_{C^{0,1}(\mathbb{R}^2)}\,\|v_h\|_{L_h^2} + \|a\|_{C^0(\mathbb{R}^2)}\,\|v_h\|_{H_h^1} \leq C\,\|v_h\|_{H_h^1}\,,$$

so that

$$\left\|a_{11}\,\partial_x^{\pm}(\delta_2+\delta_3)\right\|_{H_h^{-1}} = \sup_{0\neq v_h\in H_h^1}\left|\left(\partial_x^{\mp}[a_{11}v_h]\,,\delta_2+\delta_3\right)_{L_h^2}\right| / \left|v_h\right|_1$$
$$\leq C\left\|\delta_2+\delta_3\right\|_{L_h^2} \leq C'h\left|u\right|_2. \tag{9.41b}$$

From (9.41a,b) we have $\|a_{11}\partial_x^{\pm}\partial_x^{\pm}R_h - \check{R}_h a_{11}\frac{\partial^2}{\partial x^2}\|_{H_h^{-1}\leftarrow H^2(\mathbb{R}^2)} \leq Ch$. Analogously, one shows

$$\left\|a_{ij}\partial_{x_i}^{\pm}\partial_{x_j}^{\pm}R_h - \check{R}_h a_{ij}\frac{\partial^2}{\partial x_i\partial x_j}\right\|_{H_h^{-1}\leftarrow H^2(\mathbb{R}^2)} \leq Ch. \tag{9.41c}$$

For $a_i\partial_{x_i}^{\pm}R_h - \check{R}_h a_i\frac{\partial}{\partial x_i}$ similar reasoning results in an $\mathcal{O}(h)$-estimate for the norms $\|\cdot\|_{H_h^{-1}\leftarrow H^1(\mathbb{R}^2)}$ and $\|\cdot\|_{L_h^2\leftarrow H^2(\mathbb{R}^2)}$. Both are upper bounds for $\|\cdot\|_{H_h^{-1}\leftarrow H^2(\mathbb{R}^2)}$ such that

$$\|a_i\partial_{x_i}^{\pm}R_h - \check{R}_h a_i\partial/\partial x_i\|_{H_h^{-1}\leftarrow H^2(\mathbb{R}^2)} \leq Ch. \tag{9.41d}$$

Likewise

$$\|aR_h - \check{R}_h a\|_{H_h^{-1}\leftarrow H^2(\mathbb{R}^2)} \leq Ch. \tag{9.41e}$$

Statement (9.36) follows from (9.41c–e). ∎

When generalising the consistency estimate to more general domains $\Omega\subset\mathbb{R}^2$, the following difficulty arises. The entries of the matrix L_h according to (4.89) or (4.96) are not bounded by Ch^{-2}. Rather, at near-boundary points the inverse of the distance to the boundary point enters, and this distance may be arbitrarily small. One way around this, would be to formulate the discretisation so that the distances between boundary points and near-boundary points remain, for example, $\geq h/2$. A second possibility would be a suitable definition of R_h so that the product L_hR_h appearing in (9.36) can be estimated (cf. Hackbusch [130]). Here we choose a third option: L_h is replaced by the re-scaled matrix $L_h' = D_hL_h$ from Corollary 9.45.

Theorem 9.48. *Let $\Omega\in C^2$ (or Ω convex) be bounded. For the discretisation of $Lu=f$ for $L=-\Delta$ on Ω with $u=0$ on Γ use the discretisation $L_hu_h=f_h$ according to (4.89) or (4.96). Let $L_h' = D_hL_h$ be defined as in Corollary 9.45. Then the consistency estimate*

$$|L_h'R_h - D_h\check{R}_hL|_{-1\leftarrow 2} \leq Ch \tag{9.42}$$

holds.

Here, the matrix L_h from (4.89) or (4.96) is only taken as an example. The proof will show that the estimate (9.36), respectively (9.42), also holds for other L_h if $(L_hu_h)(\mathbf{x})$, $\mathbf{x}$ near the boundary, represents a second difference. First, two lemmata are needed.

Let $\gamma_h\subset\Omega_h$ be the set of near-boundary points. If v_h is a grid function defined on Ω_h, then we denote by $v_h|_{\gamma_h}$ the restriction to γ_h :

$$\left(v_h|_{\gamma_h}\right)(\mathbf{x}) = v_h(\mathbf{x}) \quad \text{for } \mathbf{x}\in\gamma_h, \qquad \left(v_h|_{\gamma_h}\right)(\mathbf{x}) = 0 \quad \text{for } \mathbf{x}\in\Omega_h\backslash\gamma_h.$$

Lemma 9.49. *Let $\Omega \in C^{0,1}$ be bounded. Then there exists a $C = C(\Omega)$ independent of h, such that*

$$\left|v_h|_{\gamma_h}\right|_0 \le Ch\left|v_h|_{\gamma_h}\right|_1. \tag{9.43}$$

Proof. From $\Omega \in C^{0,1}$ follows: there exist numbers $K \in \mathbb{N}$ and $h_0 > 0$ such that for all $\mathbf{x} \in \gamma_h$, with $h \le h_0$, not all grid points $\{\mathbf{x} + (\nu h, \mu h) : -K \le \nu, \mu \le K\}$ lie in Ω. Let for example $\mathbf{x} + (\nu_0 h, \mu_0 h) \notin \Omega$. Then we define the functions $w_h^{\nu\mu}$ $(-K \le \nu, \mu \le K)$ in $\mathbf{x}$ by $w_h^{\nu_0\mu_0}(\mathbf{x}) := v_h(\mathbf{x})$, $w_h^{\nu\mu}(\mathbf{x}) = 0$ for $(\nu, \mu) \ne (\nu_0, \mu_0)$. Note that $\sum_{\nu,\mu=-K}^{K} w_h^{\nu\mu} = v_h|_{\gamma_h}$ is a decomposition of $v_h|_{\gamma_h}$ with $w_h^{\nu\mu}(\mathbf{x}) = v_h(\mathbf{x})$ or $w_h^{\nu\mu}(\mathbf{x}) = 0$.

Without loss of generality we assume $\nu > 0$ and $\mu > 0$. Starting at $\mathbf{x} \in \Omega_h$ we define a chain of points $\{\mathbf{x}^0, \ldots, \mathbf{x}^{\nu+\mu}\}$ which first proceeds horizontally,

$$\mathbf{x}^0 = \mathbf{x},\ \mathbf{x}^1 = \mathbf{x} + (h, 0)\,,\ \ldots,\ \mathbf{x}^\nu = \mathbf{x} + (\nu h, 0)\,,$$

and then vertically,

$$\mathbf{x}^\nu = \mathbf{x} + (\nu h, 0)\,,\ \mathbf{x}^{\nu+1} = \mathbf{x} + (\nu h, h)\,,\ \ldots,\ \mathbf{x}^{\nu+\mu} = \mathbf{x} + (\nu h, \mu h)\,.$$

By definition we have either $w_h^{\nu\mu}(\mathbf{x}) = 0$ or $\mathbf{x}^{\nu+\mu} \notin \Omega$, i.e., $w_h^{\nu\mu}(\mathbf{x}^{\nu+\mu}) = 0$ (cf. (9.31)). In both case the estimate

$$\begin{aligned}
|w_h^{\nu\mu}(\mathbf{x})| &\le \left|w_h^{\nu\mu}(\mathbf{x}^0) - w_h^{\nu\mu}(\mathbf{x}^{\nu+\mu})\right| \\
&= h\left|\frac{1}{h}\left[w_h^{\nu\mu}(\mathbf{x}^0) - w_h^{\nu\mu}(\mathbf{x}^1)\right] + \frac{1}{h}\left[w_h^{\nu\mu}(\mathbf{x}^1) - w_h^{\nu\mu}(\mathbf{x}^2)\right] + \ldots\right| \\
&\le h\left[\begin{array}{l}\left|\partial_x^- w_h^{\nu\mu}(\mathbf{x}^1)\right| + \left|\partial_x^- w_h^{\nu\mu}(\mathbf{x}^2)\right| + \ldots + \left|\partial_x^- w_h^{\nu\mu}(\mathbf{x}^\nu)\right| \\ + \left|\partial_y^- w_h^{\nu\mu}(\mathbf{x}^{\nu+1})\right| + \ldots + \left|\partial_y^- w_h^{\nu\mu}(\mathbf{x}^{\nu+\mu})\right|\end{array}\right]
\end{aligned}$$

holds and thus $|w_h^{\nu\mu}|_0 \le h[\sqrt{\nu}\,|\partial_x^+ w_h^{\nu\mu}|_0 + \sqrt{\mu}\,\left|\partial_y^+ w_h^{\nu\mu}\right|_0] \le \sqrt{2K}h\,|w_h^{\nu\mu}|_1$. Summation over ν, μ yields estimate (9.43) for $v_h|_{\gamma_h} = \sum w_h^{\nu\mu}$. ∎

In most cases $\mathbf{x} \in \gamma_h$ already has a direct neighbour in $\mathbb{R}^2 \backslash \Omega$ such that (9.43) follows with $C = \sqrt{2}$. This holds in particular for convex domains.

Lemma 9.50. *Let R_h be defined by (9.39a). Let E_2 satisfy (9.37). Then we have*

$$(R_h u)(\boldsymbol{\xi}) \le Ch\,\|E_2 u\|_{H^2(K_{h/2}(\boldsymbol{\xi}))} \quad \text{for all } \boldsymbol{\xi} \in \Gamma,\ u \in H_0^1(\Omega) \cap H^2(\Omega), \tag{9.44}$$

where $K_{h/2}(\boldsymbol{\xi}) := \left\{\boldsymbol{\xi} + \mathbf{x} \in \mathbb{R}^2 : \mathbf{x} \in (-h/2, h/2) \times (-h/2, h/2)\right\}$.

Proof. (i) Let $Q = (-\frac{1}{2}, \frac{1}{2}) \times (-\frac{1}{2}, \frac{1}{2})$. For $v \in H^2(Q)$ one shows

$$\left|v(\mathbf{0}) - \int_Q v(\mathbf{x})\mathrm{d}\mathbf{x}\right| \le C\sqrt{\int_Q \left(v_{xx}^2 + 2v_{xy}^2 + v_{yy}^2\right)\mathrm{d}x\mathrm{d}y}\,,$$

since the left-hand side vanishes for linear functions $v(x,y) = \alpha + \beta x + \gamma y$. The proof is similar to the one for (8.50). Transforming from the unit square Q to $Q_h := (-\frac{h}{2}, \frac{h}{2}) \times (-\frac{h}{2}, \frac{h}{2})$ gives

$$\left| h^2 v(\mathbf{0}) - \int_{Q_h} v(\mathbf{x}) \mathrm{d}\mathbf{x} \right| \le C h^3 \sqrt{\int_Q \left(v_{xx}^2 + 2 v_{xy}^2 + v_{yy}^2 \right) \mathrm{d}x \mathrm{d}y} \,. \tag{9.45}$$

(ii) Let $\xi \in \Gamma$. Statement (9.44) follows from (9.45) with $v(\mathbf{x}) := \bar{u}(\boldsymbol{\xi} + h\mathbf{x}) = (E_2 u)(\boldsymbol{\xi} + h\mathbf{x})$, since $u(\boldsymbol{\xi}) = 0$ for $\boldsymbol{\xi} \in \Gamma$. ∎

Proof of Theorem 9.48. (i) Inequality (9.42) is proved if

$$\left| \left(v_h, \left[L_h' R_h - D_h \check{R}_h L \right] u \right)_0 \right| \le Ch \tag{9.46a}$$

for all $v_h \in H_h^1$ and $u \in H^2(\Omega) \cap H_0^1(\Omega)$ with $|v_h|_1 = |u|_2 = 1$. To this end v_h is split into

$$v_h = v_h' + v_h'' \qquad \text{with} \quad v_h' := v_h|_{\gamma_h} .$$

In part (ii) we show

$$\left| \left(v_h'', \left[L_h' R_h - D_h \check{R}_h L \right] u \right)_0 \right| \le C_1 h. \tag{9.46b}$$

The other steps of the proof, (iii) and (iv), yield

$$\left| \left(v_h', \left[L_h' R_h - D_h \check{R}_h L \right] u \right)_0 \right| \le C_2 h, \tag{9.46c}$$

such that (9.46a) with $C = C_1 + C_2$ follows.

(ii) Lemma 9.49 shows

$$|v_h'|_0 \le C_3 h \, |v_h|_1 = C_3 h \,. \tag{9.46d}$$

For v_h'' one obtains $|v_h''|_1 \le |v_h|_1 + |v_h'|_1 = 1 + |v_h'|_1$. The inverse estimate (9.32) yields $|v_h'|_1 \le C h^{-1} |v_h'|_0 \le C C_3$, thus

$$|v_h''|_1 \le C_4 := 1 + C C_3 \,. \tag{9.46e}$$

Let $\hat{L}_h$ be the (regular) difference operator on the infinite grid $Q_h = \{(\nu h, \mu h) : \nu, \mu \in \mathbb{Z}\}$. Since the support of v_h'' is $Q_h \backslash \gamma_h$, we have

$$(v_h'', L_h' w_h)_0 = (v_h'', \hat{L}_h w_h)_0 \qquad \text{for all } w_h .$$

Furthermore, $(v_h'', D_h w_h)_0 = (v_h'', w_h)_0$. This proves the first equality in

$$\begin{aligned} \left| \left(v_h'', \left[L_h' R_h - D_h \check{R}_h L \right] u \right)_0 \right| &= \left| \left(v_h'', \left[\hat{L}_h R_h - D_h \check{R}_h L \right] \bar{u} \right)_0 \right| \\ &\le C_5 h \, |v_h''|_1 \, |\bar{u}|_2 \le C_5 h C_4 C_6 =: C_1 h \,, \end{aligned}$$

where $\bar{u} := E_2 u$ is the continuation of u to $\mathbb{R}^2$. Further inequalities result from Lemma 9.47, (9.46e), and

$$|\bar{u}|_2 = \|\bar{u}\|_{H^2(\mathbb{R}^2)} \le C_6 \|u\|_{H^2(\Omega)} = C_6 \tag{9.46f}$$

(cf. (9.37)). Theorem 6.58c guarantees the existence of an extension $\bar{u}$ with (9.46f) if $\Omega \in C^2$. Another sufficient condition for (9.46f) is the convexity of Ω.

(iii) The left-hand side of (9.46c) splits into $(v_h', L_h' R_h u)_0$ and $(v_h', D_h R_h L u)_0$. The first part is estimated in part (iv). Exercise 9.46d and (9.46d) yield the second term

$$\left|(v_h', D_h \check{R}_h L u)_0\right| \le |v_h'|_0 \left|D_h \check{R}_h L u\right|_0 \le C_3 h C' |Lu|_0 \le C_7 h |u|_2 = C_7 h. \tag{9.46g}$$

(iv) We set $w_h := (L_h' R_h u)|_{\gamma_h}$. Since the support of v_h' is contained in γ_h, we have $(v_h', L_h' R_h u)_0 = (v_h', w_h)_0$. L_h' contains differences with respect to the x and y directions. Accordingly we write $w_h = w_h^x + w_h^y$. In the following we limit ourselves (a) to the Shortley–Weller discretisation, (b) to the term w_h^x, and (c) to the case that

$$\mathbf{x} \in \gamma_h, \quad \mathbf{x}^r = \mathbf{x} + (s_r h, 0) \in \Omega_h \quad \text{(i.e., } s_r = 1\text{)}, \quad \mathbf{x}^\ell = \mathbf{x} - (s_\ell h, 0) \in \Gamma.$$

The other cases should be treated analogously. We set

$$\hat{u} := R_h u = \sigma_h^x \sigma_h^y E_2 u \in H^2(\mathbb{R}^2).$$

The Shortley–Weller difference in the x direction reads

$$\hat{w}_h^x(\mathbf{x}) = d_{\mathbf{x}} \left[\frac{\hat{u}(\mathbf{x}) - \hat{u}(\mathbf{x}^r)}{h s_r} - \frac{\hat{u}(\mathbf{x}^\ell) - \hat{u}(\mathbf{x})}{h s_\ell} \right] / \frac{(s_\ell + s_r)\, h}{2},$$

where $d_{\mathbf{x}}$ is the diagonal element of D_h. Since in the equation $L_h' u_h = f_h$ the variables $u_h(\boldsymbol{\xi})$, $\boldsymbol{\xi} \in \Gamma$ (for example, $\boldsymbol{\xi} = \mathbf{x}^\ell$), have already been eliminated, $w_h^x(\mathbf{x})$ has the form

$$w_h^x(\mathbf{x}) = \hat{w}_h^x(\mathbf{x}) + \frac{2 d_{\mathbf{x}}}{h^2 s_\ell\, (s_\ell + s_r)} \hat{u}(\mathbf{x}^\ell)$$

The factor $2d_{\mathbf{x}}/[h^2 s_\ell (s_\ell + s_r)]$, due to the definition of D_h, remains bounded by $4h^{-2}$. From Lemma 9.50 and $K_{h/2}(\mathbf{x}^\ell) \subset K_{3h/2}(\mathbf{x})$ one infers

$$|w_h^x(\mathbf{x}) - \hat{w}_h^x(\mathbf{x})| \le 4h^{-2} \left|\hat{u}(\mathbf{x}^\ell)\right| \le C h^{-1} \|\bar{u}\|_{H^2(K_{3h/2}(\mathbf{x}))}. \tag{9.46h}$$

Since $s_r = 1$, the second divided difference $\hat{w}_h^x$ in $\mathbf{x} = (x, y)$ can be represented by

$$\hat{w}_h^x(x, y) = d_{\mathbf{x}} \int_{-h s_\ell}^{h} g(t) \hat{u}_{xx}(x + t, y) \mathrm{d}t \quad \text{with}$$

$$g(t) = \begin{cases} 2\,(t-h) \,/\, \big(h^2 + s_\ell h^2\big) & \text{for } 0 \le t \le h, \\ -2\,(s_\ell h + t) \,/\, \big(s_\ell h^2\,(1+s_\ell)\big) & \text{for } -s_\ell h \le t \le 0, \\ 0 & \text{otherwise.} \end{cases}$$

From this one infers

$$\begin{aligned} |\hat{w}_h^x(x,y)| &\le d_{\mathbf{x}} \sqrt{\int_{-hs_\ell}^{h} g^2(t)\mathrm{d}t} \sqrt{\int_{-hs_\ell}^{h} \hat{u}_{xx}^2(x+t,y)\mathrm{d}t} \\ &\le \frac{2}{h} \sqrt{\int_{-h}^{h} \hat{u}_{xx}^2(x+t,y)\mathrm{d}t} \le \frac{C}{h} \left\|\bar{u}\right\|_{H^2(K_{3h/2}(\mathbf{x}))} \end{aligned} \tag{9.46i}$$

(cf. (6.11a)). From (9.46h,i) and the corresponding estimate for w_h^y we obtain the bound $|w_h(\mathbf{x})| \le C_8 h^{-1} \left\|\bar{u}\right\|_{H^2(K_{3h/2}(x))}$, such that

$$\begin{aligned} |w_h|_0^2 &= h^2 \sum_{\mathbf{x}\in\gamma_h} |w_h(\mathbf{x})|^2 \le C_8^2 \sum_{\mathbf{x}\in\gamma_h} \left\|\bar{u}\right\|_{H^2(K_{3h/2}(x))} \\ &\le 9C_8^2 \left\|\bar{u}\right\|_{H^2(\mathbb{R}^2)} \le \left(3C_8C_6\right)^2 =: C_9^2 \,. \end{aligned}$$

From this follows

$$|(v_h', L_h' R_h u)_0| \le |(v_h', w_h)_0| \le |v_h'|_0 \, |w_h|_0 \le C_3 h C_9 =: C_{10} h \,. \tag{9.46j}$$

(9.46g) and (9.46j) yield the required inequality (9.46c). ∎

Remark 9.51. The proof steps for Theorem 9.48 can be carried out in the same manner for more general difference equations (for example, with variable coefficients, as in Lemma 9.47).

9.3.3 Optimal Error Estimates

In the following we compare the discrete solution $u_h = L_h^{-1} f_h$ with the restriction $u_h^\star := R_h u$ of the exact solution $u = L^{-1} f$. From the representation

$$\begin{aligned} u_h - u_h^\star &= L_h^{-1} f_h - R_h u = L_h^{-1}\big(f_h - \check{R}_h f\big) + L_h^{-1} \check{R}_h f - R_h u \\ &= L_h^{-1}\big(f_h - \check{R}_h f\big) - L_h^{-1}\big(L_h R_h - \check{R}_h L\big) u \end{aligned} \tag{9.47}$$

one immediately obtains the following result.

Theorem 9.52. *Let $u \in H^2(\Omega)$ hold for the solution of $Lu = f$. Let the right-hand side f_h of the discrete equation $L_h u_h = f_h$ be chosen so that*

$$\big|f_h - \check{R}_h f\big|_{-1} \le C_f h \,. \tag{9.48}$$

If, furthermore, L_h is H^1_h-regular, and if the consistency condition (9.36) holds, then u_h satisfies the error estimate

$$|u_h - u_h^\star|_1 \le C_1 \left(C_f + C_K\,|u|_2\right) h\,. \tag{9.49}$$

Proof. $|u_h - u_h^\star|_1 \le |L_h^{-1}|_{1\leftarrow -1}(|f_h - \check{R}_h f|_{-1} + |L_h R_h - \check{R}_h L|_{-1\leftarrow 2}\,|u|_2)$. ∎

Corollary 9.53. (a) Inequality (9.48) holds in particular if one chooses $f_h := \check{R}_h f$. (b) The choice $f_h(\mathbf{x}) := f(\mathbf{x})$ for $\mathbf{x} \in \Omega_h$ (cf. (4.13b)) leads to (9.48) if $f \in C^{0,1}(\overline{\Omega})$ or $f \in H^2(\Omega)$. In these cases the following even holds:

$$|f_h - \check{R}_h f|_0 \le C\|f_h - \check{R}_h f\|_\infty \le C' h \|f\|_{C^{0,1}(\overline{\Omega})} \text{ resp. } |f_h - \check{R}_h f|_0 \le Ch^2 |f|_2. \tag{9.50}$$

(c) In Theorem 9.52 one can replace the H^1_h-regularity of L_h by that of $L'_h = D_h L_h$ (cf. Corollary 9.45), (9.36) by (9.42), and (9.48) by

$$|D_h f_h - D_h \check{R}_h f|_{-1} \le C_f h\,.$$

Proof. The proof of (9.50) is based on the inequality (9.45). ∎

Error estimates of order $\mathcal{O}(h^2)$ can be derived in the same way if one has consistency conditions of second order. These are, for example, (9.51a) or (9.51b):

$$|L_h R_h - \check{R}_h L|_{-2\leftarrow 2} \le Ch^2, \tag{9.51a}$$

$$|L_h R_h - \check{R}_h L|_{-1\leftarrow 3} \le Ch^2. \tag{9.51b}$$

Remark 9.54. If $\Omega_h = Q_h$ (i.e., $\Omega = \mathbb{R}^2$) or $\Omega = (x', x'') \times (y', y'')$, the inequalities (9.51a,b) can be shown in a way similar to Lemma 9.47.

Example 9.55. The difference method in Example 4.53 shows quadratic convergence. This case can be analysed as follows. Using Remark 9.54 one shows (9.51a). In the following section we prove the H^2_h-regularity $|L_h^{-1}|_{2\leftarrow 0} \le C$ which for the symmetric matrix under discussion, L_h, is equivalent with $\left|L_h^{-1}\right|_{0\leftarrow -2} \le C$ (cf. (6.33)). Expression (9.47) leads to

$$|u_h - u_h^\star|_0 \le Ch^2\,|u|_2\ .$$

The corresponding estimate

$$|u_h - u_h^\star|_1 \le Ch^2\,|u|_3\ ,$$

which is based on (9.51b), fails due to the fact that the solution in Example 4.53 does not belong to $H^3(\Omega)$ (cf. Example 9.29).

The verification of (9.51b) in the presence of irregular discretisations at the boundary becomes more complicated.

9.3.4 $H_{0,h}^{m+\theta}$-Regularity for $-1/2 < \theta < 1/2$

Higher regularity cannot be expected for domains with reentrant corners like the L-shaped domain in Example 2.4. However, in the continuous case Theorem 9.23 ensures $H_0^{m+\theta}(\Omega)$-regularity for $|\theta| < 1/2$. In an analogous way, here we define and prove $H_{0,h}^{m+\theta}$-regularity of L_h (cf. Hackbusch [129]).

First we have to define the norm of $H_{0,h}^s$ for non-integer s as a discrete analogue of $H_0^s(\Omega)$. As in §6.2.4 we use the Fourier transform for the definition.

The grid Q_h consists of the grid points $\boldsymbol{\nu}h$ with multi-indices $\boldsymbol{\nu} \in \mathbb{Z}^n$. A grid function $u_h : \mathbb{Z}^n \to \mathbb{C}$ has the nodal values $u_{\boldsymbol{\nu}}$. These values define the 2π-periodic Fourier series

$$\hat{u}_h(\boldsymbol{\xi}) := \sum_{\boldsymbol{\nu}\in\mathbb{Z}^n} u_{\boldsymbol{\nu}}\, \mathrm{e}^{\mathrm{i}\langle\boldsymbol{\nu},\boldsymbol{\xi}\rangle} \qquad (\boldsymbol{\xi} \in \Pi := [-\pi,\pi]^n),$$

where $\langle\boldsymbol{\nu},\boldsymbol{\xi}\rangle = \sum_{i=1}^n \nu_i\xi_i$. The Euclidean norm $|u_h|_0 = \sqrt{h^n \sum_{\mathbf{x}\in Q_h} |v_h(\mathbf{x})|^2}$ (cf. §9.3.1) can be expressed by $\hat{u}_h$ because of Parseval's equality

$$|u_h|_0 = [h/(2\pi)]^{n/2}\, \|\hat{u}_h\|_{L^2(\Pi)}\,.$$

For general $s \in \mathbb{R}$ we set

$$|u_h|_s = [h/(2\pi)]^{n/2} \left\| \left[1 + h^{-2} \sum_{j=1}^n \sin^2(\xi_j/2) \right]^{s/2} \hat{u}_h \right\|_{L^2(\Pi)}.$$

$H_{0,h}^s = H_{0,h}^s(Q_h)$ is the space of all u_h with finite norm $|u_h|_s$. For $\Omega_h \subset Q_h$ we define

$$H_{0,h}^s(\Omega_h) := \left\{ u_h \in H_{0,h}^s : u_h(\mathbf{x}) = 0 \text{ for all } \mathbf{x} \in Q_h\backslash\Omega_h \right\}.$$

In the following we use $H_{0,h}^{m+\theta}(\Omega_h)$ for $|\theta| < 1/2$. The property $u_h = 0$ on $Q_h\backslash\Omega_h$ corresponds to the fact that in the continuous case functions in $H_0^s(\Omega)$ for $s + \frac{1}{2} \notin \mathbb{N}$ can be continuously extended to by zero on $\mathbb{R}^n\backslash\Omega$ (cf. Lions–Magenes [194, page 60]).

In the continuous case, the domain and its boundary can be (piecewise) described by functions (cf. Definition 6.51). However, this is more involved for the discrete point set Ω_h. Instead we may formulate a cone condition[10] whose details are in [129, page 76]. A sufficient property for the cone condition are $\Omega \in C^{0,1}$ and $\Omega_h = Q_h \cap \Omega$.

[10] Concerning the use of cone conditions for Sobolev spaces we refer to Adams [1, pages 79ff] or Remark 6.87.

We represent the difference scheme is the form

$$L_h = \sum_{|\boldsymbol{\alpha}|,|\boldsymbol{\beta}|\le m} \sum_{\boldsymbol{\gamma},\boldsymbol{\delta}\in\mathbb{Z}^d} \partial^{\boldsymbol{\alpha}}\, T^{\boldsymbol{\gamma}}\, c_{\boldsymbol{\alpha}\boldsymbol{\beta}\boldsymbol{\gamma}\boldsymbol{\delta}}(\cdot,h)\, T^{\boldsymbol{\delta}}\, \partial^{\boldsymbol{\beta}},$$

where $T^{\boldsymbol{\gamma}}$ is the translation $(T^{\boldsymbol{\gamma}}u)(\mathbf{x}) = u(\mathbf{x}+\boldsymbol{\gamma}h)$ or, equivalently, $(T^{\boldsymbol{\gamma}}u)_{\boldsymbol{\nu}} = u_{\boldsymbol{\nu}+\boldsymbol{\gamma}}$, and $\partial^{\alpha} = \prod_{j=1}^{n}(\partial_h^-)^{\alpha_j}$ are the backward differences. The *characteristic function* belonging to the principal part of L_h is

$$p(\mathbf{x},\boldsymbol{\xi}) = \sum_{|\boldsymbol{\alpha}|,|\boldsymbol{\beta}|=m} \sum_{\boldsymbol{\gamma},\boldsymbol{\delta}\in\mathbb{Z}^d} c_{\boldsymbol{\alpha}\boldsymbol{\beta}\boldsymbol{\gamma}\boldsymbol{\delta}}(\mathbf{x},0)\, \mathrm{e}^{-\mathrm{i}\langle\boldsymbol{\xi},\boldsymbol{\gamma}+\boldsymbol{\delta}\rangle} \prod_{j=1}^{n} \left[1-\mathrm{e}^{\mathrm{i}\xi_j}\right]^{\alpha_j+\beta_j}.$$

L_h is said to be *elliptic* if

$$\Re\mathfrak{e}\, p(\mathbf{x},\boldsymbol{\xi}) \ge \varepsilon \left[\sum\nolimits_{j=1}^{n} \sin^2(\xi_j/2)\right]^m \quad \text{for all } \mathbf{x}\in\mathbb{R}^n,\ \boldsymbol{\xi}\in\Pi$$

(cf. Thomée–Westergren [289, Lemma 2.3]).

Theorem 9.56. *Let Ω_h satisfy the cone condition mentioned above. Let $|\theta| < 1/2$. The difference operator is assumed to be elliptic, its coefficients must be bounded and satisfy*

$$c_{\boldsymbol{\alpha}\boldsymbol{\beta}\boldsymbol{\gamma}\boldsymbol{\delta}} \in C^{\kappa}(\Omega) \quad \textit{if} \left\{\begin{array}{l} |\boldsymbol{\alpha}| = m \textit{ and } \theta > 0, \\ |\boldsymbol{\alpha}| = |\boldsymbol{\beta}| = m, \\ |\boldsymbol{\beta}| = m \textit{ and } \theta < 0, \end{array}\right\} \quad \textit{with} \left\{\begin{array}{l} \kappa > |\theta| > 0 \\ \textit{or} \\ \kappa \ge \theta = 0. \end{array}\right.$$

Finally, let $\left|L_h^{-1}\right|_{0\leftarrow 0} \le \mathrm{const}$, let $p(\mathbf{x},\boldsymbol{\xi})$ be real-valued.[11] *Then L_h is $H_{0,h}^{m+\theta}(\Omega_h)$-regular, i.e., $\|L_h^{-1}\|_{H_{0,h}^{m+\theta}(\Omega_h)\leftarrow H_{0,h}^{-m+\theta}(\Omega_h)} \le \mathrm{const}$.*

The proof can be found in [129]. In [129, §2.5] we discuss modifications of the difference scheme L_h at near-boundary points. This also includes the Shortley–Weller discretisation.

9.3.5 H_h^2-Regularity

Under suitable conditions on Ω, $Lu = f \in L^2(\Omega)$ has a solution $u \in H^2(\Omega)\cap H_0^1(\Omega)$: $|u|_2 \le C\,|f|_0$ (cf. for example, Theorem 9.24). For the discrete solution of $L_h u_h = f_h$ the corresponding question arises: does $|u_h|_2 \le C\,|f_h|_0$ hold? First one has to define the norm $|\cdot|_2$ of H_h^2.

If $\Omega = \mathbb{R}^2$ or if the boundary Γ coincides with grid lines, one can define[12]

[11] If p is complex-valued, the statement still holds for $|\theta| \le \theta_0$ with sufficiently small $\theta_0 < \frac{1}{2}$.

[12] The exact proof requires that the line between two neighboured grid points do not intersect the boundary. This holds for convex domains and is very likely for smooth domains. Otherwise in the

$$|u_h|_2 := \sqrt{|u_h|_1^2 + h^2 \sum_{\mathbf{x}\in\Omega_h} \left[\left|\partial_x^+ \partial_x^- u_h(\mathbf{x})\right|^2 + \left|\partial_y^+ \partial_y^- u_h(\mathbf{x})\right|^2 \right]}.$$

However, in general, one has to use irregular differences at the boundary. For far-boundary points $\mathbf{x} \in \Omega_h \backslash \gamma_h$ set $D_{xx} u_h(\mathbf{x}) = \partial_x^+ \partial_x^- u_h(\mathbf{x})$. Let $\mathbf{x} = (x, y) \in \gamma_h$ be a near-boundary point with a left neighbour $\mathbf{x}^\ell = (x - s_\ell h, y) \in \Gamma$ and a regular right one: $\mathbf{x}^r = (x + h, y) \in \Omega_h$. Then we use the divided difference (4.87), but scaled with a factor corresponding to $d(\mathbf{x})$ in Corollary 9.45:

$$D_{xx} u_h(\mathbf{x}) := 2 s_\ell \frac{\dfrac{u_h(x+h,y) - u_h(x,y)}{h} - \dfrac{u_h(x,y) - u_h(x - s_\ell h, y)}{s_\ell h}}{(1 + s_\ell)\, h}.$$

Together with analogously defined difference quotients D_{yy} and D_{xy} we set

$$|u_h|_2 := \sqrt{|u_h|_1^2 + h^2 \sum_{\mathbf{x}\in\Omega_h} \left[|D_{xx} u_h(\mathbf{x})|^2 + |D_{yy} u_h(\mathbf{x})|^2 + |D_{xy} u_h(\mathbf{x})|^2 \right]}.$$

The scaling by the factor corresponding to $d(\mathbf{x})$ guarantees the inverse estimate

$$|\cdot|_2 \le C h^{-1} |\cdot|_1 . \tag{9.52}$$

In the following we analyse the diagonally scaled matrix $L_h' := D_h L_h$ corresponding to the matrix in Corollary 9.45.

The matrix L_h' is said to be H_h^2-*regular* if $\left|L_h'^{-1}\right|_{2\leftarrow 0} \le C$. An equivalent formulation is $|u_h|_2 \le C\, |f_h|_0$ for $u_h = L_h'^{-1} f_h,\ f_h \in L_h^2$.

Exercise 9.57. Let L_h be H_h^2-regular, let $L_h + \delta L_h$ be H_h^1-regular with a perturbation bounded by $|\delta L_h|_{0\leftarrow 1} \le C$ for all h. Show that $L_h + \delta L_h$ is also H_h^2-regular.

For the proof of H_h^2-regularity one can—in analogy to the proof of Corollary 9.25—partially sum the scalar product $(L_h u_h, L_h u_h)_0$ in order to show the equivalence of $|f_h|_0^2 = |L_h u_h|_0^2$ and $|u_h|_2^2$. This technique, however, can only be applied to a rectangle Ω. Here we use a simpler proof which is also applicable to general domains.

Let $P_h : L_h^2 \to L^2(\Omega)$ be the following piecewise constant interpolation:

$$\big(P_h u_h\big)(\mathbf{x}) := \begin{cases} \big(\hat{P}_h u_h\big)(\mathbf{x}) & \text{if } \mathbf{x} \in \Omega, \\ 0 & \text{if } \mathbf{x} \notin \Omega, \end{cases} \tag{9.53a}$$

$$\hat{P}_h u_h(x,y) := u_h(x',y') \quad \text{if} \quad \begin{cases} (x',y') \in \Omega_h \text{ with} \\ x' - h/2 < x \le x' + h/2, \\ y' - h/2 < y \le y' + h/2, \end{cases} \tag{9.53b}$$

where $u_h(\mathbf{x}) = 0$ for $\mathbf{x} \in Q_h \backslash \Omega_h$ (cf. (9.31)).

---

definition. e.g., of the H_h^1-norm one has to replace the term $\left|\partial_x^+ u_h(x,y)\right|$ by $|u_h(x,y)/h|$ if the line beween (x,y) und $(x+h,y)$ cuts the boundary.

Lemma 9.58. *Let $\Omega \in C^{0,1}$. The following estimates hold:*

$$|\check{R}_h P_h - I|_{-1\leftarrow 0} \le Ch, \tag{9.54a}$$
$$|P_h|_{0\leftarrow 0} \le C, \tag{9.54b}$$
$$|R_h|_{2\leftarrow 2} \le C. \tag{9.54c}$$

Proof. (i) The estimate (9.54a) is equivalent to $|P_h^\star \check{R}_h^\star - I|_{0\leftarrow 1} \le Ch$ (cf. (6.32)). Thus we must show

$$|w_h|_0 \le Ch\,|u_h|_1 \qquad \text{for} \quad w_h := (P_h^\star \check{R}_h^\star - I)u_h.$$

From the following Exercise 9.59 we obtain the expression

$$\hat{w}_h(\mathbf{x}) = \left(\sigma_h^x \sigma_h^y \left[\sigma_h^x \sigma_h^y - I\right] \hat{P}_h u_h\right)(\mathbf{x}) \quad \text{for } \hat{w}_h := \left(\hat{P}_h^\star R_h^\star - I\right) u_h, \ \ \mathbf{x} \in \Omega_h.$$

Set $w := (\sigma_h^x \sigma_h^y - I)\hat{P}_h u_h$. Exercise 9.46d shows $|\hat{w}_h|_0 \le |w|_{L^2(\mathbb{R}^2)}$. For every $\boldsymbol{\xi} \in (0,1) \times (0,1)$ define the grid function

$$w_{h,\boldsymbol{\xi}} \in L_h^2 \qquad \text{with} \quad w_{h,\xi}(\mathbf{x}) := w(\mathbf{x} + \boldsymbol{\xi}) \quad \text{for } \mathbf{x} \in Q_h.$$

One may check that $w_{h,\boldsymbol{\xi}}(x,y)$ is a weighted sum of first differences $u_h(x,y) - u_h(x',y')$ where $(x',y') \in \{(x \pm h, y), (x, y \pm h), (x \pm h, y \pm h)\}$. Thus we have $|w_{h,\boldsymbol{\xi}}|_0 \le h\,|u_h|_1$ for all $\boldsymbol{\xi}$ from which one infers

$$\begin{aligned} |w|_0^2 &= \int_{\mathbb{R}^2} |w(\mathbf{x})|^2 \,\mathrm{d}\mathbf{x} = h^2 \int_{(0,1)\times(0,1)} \sum_{\mathbf{x}\in\Omega_h} |w(\mathbf{x} + \boldsymbol{\xi} h)|^2 \,\mathrm{d}\boldsymbol{\xi} \\ &= \int_{(0,1)\times(0,1)} |w_{h,\xi}|_0^2 \,\mathrm{d}\boldsymbol{\xi} \le h^2\,|u_h|_1\,, \end{aligned}$$

thus $|\hat{w}_h|_0 \le |w|_0 \le h\,|u_h|_1$. From

$$w_h - \hat{w}_h = (P_h^\star - \hat{P}_h^\star)\check{R}_h^\star u_h = (P_h^\star - \hat{P}_h^\star)\sigma_h^x \sigma_h^y \hat{P}_h u_h = (P_h^\star - \hat{P}_h^\star)\sigma_h^x \sigma_h^y \hat{P}_h (u_h|_{\hat{\gamma}_h})$$

with $\hat{\gamma}_h = \{\mathbf{x} \in \Omega_h : K_{3h/2}(\mathbf{x}) \cap \Gamma \ne \emptyset\}$ we infer

$$\left|w_h - \hat{w}_h\right|_0 \le \left|(P_h^\star - \hat{P}_h^\star)\sigma_h^x \sigma_h^y\right|_{0\leftarrow 0} \left|\hat{P}_h\right|_{0\leftarrow 0} |u_h|_{\hat{\gamma}_h}|_0 \le C\,|u_h|_{\hat{\gamma}_h}|_0 \le C' h\,|u_h|_1$$

(see (ii) and Lemma 9.49, in which γ_h may be replaced by $\hat{\gamma}_h$). Together with $|\hat{w}_h|_0 \le h\,|u_h|_1$ we obtain $|w_h|_0 \le (C'+1)h\,|u_h|_1$.

(ii) (9.54b) with $C = 1$ follows from $|P_h u_h|_0 = |u_h|_0$.

(iii) Second differences of $\hat{u} = R_h u$ have already been estimated in (9.46i). (9.46i) implies $|u_h|_2 \le C\,|u|_2$, i.e., (9.54c). ∎

Exercise 9.59. Show that the adjoint operators for $\hat{P}_h$ and $\check{R}_h$ are

$$\hat{P}_h^\star : L^2(\mathbb{R}^2) \to L_h^2 \quad \text{and} \quad \check{R}_h^\star : L_h^2 \to L^2(\mathbb{R}^2)$$

with

$$\left(\hat{P}_h^\star u\right)(\mathbf{x}) = \left(\sigma_h^x \sigma_h^y u\right)(\mathbf{x}) \text{ for } \mathbf{x} \in Q_h, \quad \check{R}_h^\star = E_0^* \sigma_h^x \sigma_h^y \hat{P}_h$$

Furthermore, $(\hat{P}_h^\star \hat{P}_h u_h)(\mathbf{x}) = (\sigma_h^x \sigma_h^y \hat{P}_h u_h)(\mathbf{x}) = u_h(\mathbf{x})$ holds for all $\mathbf{x} \in Q_h$.

The identity

$$L_h'^{-1} = R_h L^{-1} P_h - L_h'^{-1} \left[\left(L_h' R_h - \check{R}_h L \right) L^{-1} P_h + \left(\check{R}_h P_h - I \right) \right]$$

yields the estimate

$$\begin{aligned} \left| L_h'^{-1} \right|_{2\leftarrow 0} \le |R_h|_{2\leftarrow 2} \left| L^{-1} \right|_{2\leftarrow 0} |P_h|_{0\leftarrow 0} + |I|_{2\leftarrow 1} |L_h'^{-1}|_{1\leftarrow -1} \times \\ \times \left[|L_h' R_h - \check{R}_h L|_{-1\leftarrow 2} |L^{-1}|_{2\leftarrow 0} |P_h|_{0\leftarrow 0} + |\check{R}_h P_h - I|_{-1\leftarrow 0} \right]. \end{aligned}$$

The inverse estimate (9.52) yields $|I|_{2\leftarrow 1} \le Ch^{-1}$ for the identity $I : H_h^1 \to H_h^2$. Together with the inequalities (9.54a–c) the next statement follows.

Theorem 9.60. *Let L_h' be H_h^1-regular and assume the consistency (9.36). Suppose that Ω be convex or from C^2, and let L be H^2-regular (i.e., $|L^{-1}|_{2\leftarrow 0} \le C$). Then L_h' is also H_h^2-regular.*

For a convex domain Auzinger [12] proves a quantitative estimate, a discrete analogue of inequality (9.23), for the Shortley–Weller discretisation of the Poisson problem.

Exercise 9.61. If $\Omega \subset \mathbb{R}^n$ ($n \le 3$) is bounded and L_h is H_h^2-regular, then L_h is stable with respect to the row-sum norm: $\|L_h^{-1}\|_\infty \le C$. *Hint:* Use $C^{-1} |u_h|_0 \le \|u_h\|_\infty \le C|u_h|_2$.

9.3.6 Interior Regularity

Also the interior regularity can be transferred to difference equations. Details are for example in Thomée [285] and Thomée–Westergren [289]. Elliptic systems are discussed by Bube–Strikwerda [58]. Let Ω be the domain of the boundary-value problem and Ω_h the grid for which at least the far-boundary points are regular. Let

$$\Omega_h' \subset \Omega_h$$

be an *interior* subset, i.e., the distance of Ω_h' and $\Gamma = \partial\Omega$ is bounded from below by a positive constant. By assumption Ω_h' is a regular grid so that divided differences approximating higher derivatives can be constructed. For instance, the fourth

x-derivative $\partial_x^+ \partial_x^- \partial_x^+ \partial_x^- u_h$ is defined at all $\mathbf{x} \in \Omega'_h$. The interior regularity results will imply that indeed $\partial_x^+ \partial_x^- \partial_x^+ \partial_x^- u_h$ approximates $\partial^4 u / \partial x^4$, although u_h is only accurate of first or second order.

Corresponding results are also true for finite-element discretisations in a subdomain $\Omega' \subset\subset \Omega$, provided that the triangulation in Ω'' with

$$\Omega' \subset\subset \Omega'' \subset \Omega$$

is regular (e.g., a square-grid triangulation). However, since one usually uses finite-element discretisations with irregular triangulations, it is not possible to computed higher derivatives of the solution in a direct way.

The approximation of higher derivatives described above, of course requires these derivatives to exist, i.e., the coefficients of the differential operator must be sufficiently smooth. Even if this is not the case, the statements of Theorem 9.34 can still be transferred to (irregular) finite-element discretisations as proved by Faustmann [98] and Faustmann–Melenk–Praetorius [99].

Chapter 10
Special Differential Equations

Abstract If the boundary-value problems have special properties, one often uses special discretisations for them. We give two examples. In **Section 10.1** the principal part has jumping coefficients. Starting from the variational formulation, one obtains a strong formulation for each subdomain in which the coefficients are smooth. In addition, one gets transition equations at the inner boundary γ. Finite-element methods should use a triangulation which follows γ. Finally, in §10.1.4, we discuss the case that coefficients of terms different from the principal part are discontinuous. Typically the differential operators in fluid dynamics are nonsymmetric because of a derivative of first order. If this convections term becomes dominant, we obtain a singularly perturbed problem which is discussed in **Section 10.2**. In this case other discretisation variants are appropriate. In the case of difference method there is a conflict between stability and consistency conditions. Usual finite-element discretisation have similar difficulties. A remedy is the streamline-diffusion method explained in §10.2.3.2.

10.1 Differential Equations with Discontinuous Coefficients

10.1.1 Formulation

The selfadjoint differential equation

$$-\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}(\mathbf{x})\frac{\partial}{\partial x_j}u\right) + a(\mathbf{x})u(\mathbf{x}) = f \qquad \text{in } \Omega \tag{10.1a}$$

(cf. (5.18)) often occurs in physics. It can also be written as

$$-\operatorname{div}(\mathbf{A}(\mathbf{x})\,\nabla u) + au = f \quad \text{with} \quad \mathbf{A}(\mathbf{x}) := (a_{ij}(\mathbf{x}))_{i,j=1,\ldots,n}\,,$$

so that for $a = 0$ and $f = 0$ we have the *conservation law* $\operatorname{div}\phi = 0$ for $\phi := \mathbf{A}\,\nabla u$.

W. Hackbusch, *Elliptic Differential Equations*, Springer Series
in Computational Mathematics 18, DOI 10.1007/978-3-662-54961-2_10

In the applications in physics the coefficients a_{ij} are, in general, constants of the material. The functions a_{ij} can be varying, if the constitution of the material depends on position. As soon as we have several materials in contact with each other, the coefficients $a_{ij}(\mathbf{x})$ may be discontinuous on the boundary of contact γ (see Figure 10.1).

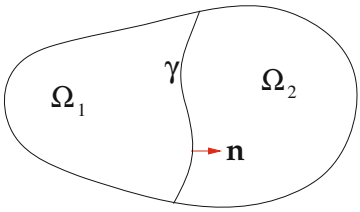


Fig. 10.1 Interior boundary $\gamma = \Gamma_1 \cap \Gamma_2$.

The equation (10.1a) can be understood in a classical sense only if $a_{ij} \in C^1(\Omega)$. For discontinuous a_{ij} one has to use the variational formulation. If one supplements equation (10.1a) with the Dirichlet condition

$$u = 0 \qquad \text{on } \Gamma\,, \tag{10.1b}$$

then the weak formulation is written

$$\text{find } u \in H_0^1(\Omega) \quad \text{with} \quad a(u,v) = (f,v)_{L^2(\Omega)} \quad \text{for all } v \in H_0^1(\Omega), \tag{10.2a}$$

where

$$a(u,v) := \int_\Omega \left[a(\mathbf{x})u(\mathbf{x})v(\mathbf{x}) + \sum_{i,j=1}^{n} a_{ij}(\mathbf{x}) \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} \right] \mathrm{d}\mathbf{x}\,. \tag{10.2b}$$

Note that equations (10.2a,b) are defined for arbitrary $a_{ij} \in L^\infty(\Omega)$ with positive-definite matrix $\mathbf{A} = (a_{ij})$. For $a(\cdot) \geq 0$ and bounded Ω the solvability of (10.2a) is guaranteed.

In the sequel we shall assume the situation to be as in Figure 10.1: Ω is divided by the boundary line γ into two subregions Ω_1 and Ω_2. Let the coefficients a_{ij} be piecewise smooth: $a_{ij} \in C^1(\Omega_k)$ for $k = 1, 2$. Along γ the coefficients may be discontinuous so that the one-sided boundary values

$$a_{ij}^{(1)}(\mathbf{x}) := \lim_{\Omega_1 \ni \mathbf{y} \to \mathbf{x}} a_{ij}(\mathbf{x}), \qquad a_{ij}^{(2)}(\mathbf{x}) := \lim_{\Omega_2 \ni \mathbf{y} \to \mathbf{x}} a_{ij}(\mathbf{x}) \qquad (\mathbf{x} \in \gamma)$$

may be different. In addition let us assume that the solution u is continuous, $u \in C^0(\overline{\Omega})$, but only that it is piecewise smooth, $u \in C^1(\Omega_1)$ and $u \in C^1(\Omega_2)$. The one-sided boundary values of the derivatives are denoted $u_{x_j}^{(1)}(\mathbf{x})$ and $u_{x_j}^{(2)}(\mathbf{x})$ ($\mathbf{x} \in \gamma$). With these assumptions integration by parts of $\int_\Omega \sum_{i,j=1}^n a_{ij}(\mathbf{x}) \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} \mathrm{d}\mathbf{x}$ gives the result

$$\int_{\Omega_1} a_{ij}\, v_{x_i} u_{x_j} \mathrm{d}\mathbf{x} = -\int_{\Omega_1} v\,(a u_{x_j})_{x_i} \mathrm{d}\mathbf{x} + \int_\gamma v\, a_{ij}^{(1)} u_{x_j}^{(1)}\, n_i \,\mathrm{d}\Gamma\,.$$

Since $\mathbf{n} = (n_1, n_2, \ldots)$ in Figure 10.1 is the outgoing normal to Ω_1, but the ingoing normal with respect to Ω_2, in Ω_2 there results

$$\int_{\Omega_2} a_{ij}\, v_{x_i} u_{x_j} \mathrm{d}\mathbf{x} = -\int_{\Omega_2} v\,(a_{ij} u_{x_j})_{x_i} \mathrm{d}\mathbf{x} - \int_{\gamma} v\, a_{ij}^{(2)} u_{x_j}^{(2)}\, n_i \,\mathrm{d}\Gamma\,.$$

Putting this together one has

$$a(u,v) = \int_{\Omega} \left[a(\mathbf{x})u(\mathbf{x}) - \sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}(\mathbf{x}) \frac{\partial u}{\partial x_j} \right) \right] v(\mathbf{x}) \mathrm{d}\mathbf{x}$$
$$- \int_{\gamma} \sum_{i,j=1}^{n} \left[a_{ij}^{(2)} u_{x_j}^{(2)} - a_{ij}^{(1)} u_{x_j}^{(1)} \right] n_i\, v \,\mathrm{d}\Gamma\,.$$

Thus from the variational equation (10.2a) there follows in addition to the differential equation (10.1a) also the *transition equation*

$$\sum_{i,j=1}^{n} a_{ij}^{(1)} u_{x_j}^{(1)} n_i = \sum_{i,j=1}^{n} a_{ij}^{(2)} u_{x_j}^{(2)} n_i \qquad \text{on } \gamma\,. \tag{10.3}$$

$B_k := \sum_{i,j=1}^{n} a_{ij}^{(k)} n_i \frac{\partial}{\partial x_j}$ for $k = 1, 2$ are the corresponding conormal derivatives (cf. §5.2.1).

If the coefficients are discontinuous, then in general the solution u of (10.2a) does not belong to $C^2(\Omega)$, but has discontinuous derivatives in γ. Only the tangential derivative along γ may be continuous. Even if neither the matrix function $\mathbf{A}$ nor ∇u are continuous, this holds for the product $\mathbf{A}\nabla u$ as seen in (10.3). This proves the following lemma.

Lemma 10.1. *Assume $a_{ij} \in C^1(\Omega_k)$, $k = 1, 2$. If the weak solution u of (2a) in Ω is continuous and piecewise differentiable in Ω_1 and Ω_2, then it is the classical solution of the differential equation (10.1a) in $\Omega_1 \cup \Omega_2 = \Omega \backslash \gamma$. In addition to fulfilling the boundary condition (10.1b) on Γ the solution also satisfies the transition condition (10.3) on the interior boundary γ.*

Example 10.2. Let $\xi \in (0,1)$. Suppose the coefficient of $a(u,v) := \int_0^1 a(x) u' v' \mathrm{d}x$ are given by $a(x) = 1$ on the interval $(0, \xi)$ and $a(x) = 2$ on $(\xi, 1)$. For this one-dimensional example the point ξ plays the role of the curve γ. The solution of the equation (10.2a) with $f(x) = 1$ is

$$u(x) = \frac{1}{2}\left[\frac{1+\xi^2}{1+\xi} x - x^2 \right] \qquad \text{on } \Omega_1 = (0, \xi),$$
$$u(x) = \frac{1}{4}\left[\frac{1 + 2\xi - \xi^2}{1+\xi} (1-x) - (1-x)^2 \right] \qquad \text{on } \Omega_2 = (\xi, 1).$$

It satisfies $-au'' = 1$ on $(0,\xi) \cup (\xi, 1)$, $u(0) = u(1) = 0$ and the transition equation $1 \cdot u'(\xi - 0) = 2 \cdot u'(\xi + 0)$.

As mentioned at the beginning, the differential equation can be written in the form $\operatorname{div} \phi = f$ on $\Omega_1 \cup \Omega_2$, where $\phi := \mathbf{A}(\mathbf{x}) \nabla u$ on Ω and $u = 0$ on Γ. The

transition condition (10.3) means that $\langle \phi, \mathbf{n} \rangle$ is continuous on γ. Since $\langle \phi, \mathbf{t} \rangle$ is also continuous for any direction $\mathbf{t}$ tangential to γ, ϕ is continuous on γ and thus throughout Ω.

The regularity proofs of Section 9.1 can be carried over to the present situation. The smoothness assumptions upon the coefficients of $a(\cdot, \cdot)$ are in each case to be required piecewise on Ω_1 and Ω_2; in addition the dividing line (or hypersurface) γ must be sufficiently smooth. Then there results the piecewise regularity $u \in H^{m+s}(\Omega_1)$ and $u \in H^{m+s}(\Omega_2)$, instead of $u \in H^{m+s}(\Omega)$. Bringing in the transition condition (10.3), one obtains $\phi = \mathbf{A}(\mathbf{x}) \nabla u \in H^{m+s-1}(\Omega)$ on the whole region.

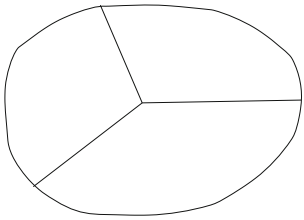


Fig. 10.2 Domain with crossing interior boundaries.

If the interior boundary γ has corners or if as in Figure 10.2 different interior boundaries intersect, additional corner singularities occur (cf. page 253).

10.1.2 Finite-Element Discretisation

When one discretises using finite elements there are the following difficulties.

Remark 10.3. Let $\mathcal{T}$ be a general triangulation. Linear or bilinear elements give a finite-element solution with the error estimate $|u - u^h|_1 = \mathcal{O}(h^{1/2})$. The $L^2(\Omega)$ error bound is in general no better than $|u - u^h|_0 = \mathcal{O}(h)$.

Proof. $\Omega_\gamma := \cup \{T \in \mathcal{T} : T \cap \gamma \neq \emptyset\}$ consists of all finite elements that are in contact with γ. Since u has discontinuous derivatives on γ, one can have no better estimate on Ω_γ than $\nabla u - \nabla v = \mathcal{O}(1)$ $(v \in V_h)$. The surface area of Ω_γ is of the order $\mathcal{O}(h)$, so that there results $|u - v|_1 = \mathcal{O}(h^{1/2})$. For the bound on $|u - u^h|_0$ see the following example. ∎

Example 10.4. In Example 10.2 choose $\xi = \frac{1+h}{2}$ and discretise using piecewise linear elements of length h. Then there results an error of u at $x = 1/2$ of $u(x) - u^h(x) = \alpha h + \mathcal{O}(h^2)$ with $\alpha \approx 0.00463$. Since $u - u^h$ behaves linear in $(0, 1/2)$, one obtains the error $\mathcal{O}(h)$ not only for the L^2 norm $|u - u^h|_0$ but also for the maximum norm $|u - u^h|_\infty$ and for the L^1 norm $\|u - u^h\|_{L^1(0,1)}$.

The usual bounds on the errors $|u - u^h|_1 = \mathcal{O}(h)$ and $|u - u^h|_0 = \mathcal{O}(h^2)$ can however be attained. To do this one must adjust the geometry of the triangulation to the curve γ. If the dividing line γ is piecewise linear, one must choose the triangulation such that γ coincides with sides of triangles. Note that the error estimates in Section 8.5.3 require only the smoothness of the restriction of u to the finite elements $T \in \mathcal{T}$. If γ is curvilinear then it may be approximated by isoparametric elements (cf. §8.6.3).

Another possibility is the discretisation by *mixed* finite elements (cf. §8.9.1). The two components of the solution are u and $v := \mathbf{A} \nabla u$.

10.1.3 Discretisation by Difference Schemes

In the case of Example 10.4 a similar assertion holds for the difference scheme (5.19).

Example 10.5. Let the discontinuity position ξ in Example 10.2 be a grid point, i.e., $\frac{\xi}{h} \in \mathbb{N}$. Then the difference scheme (5.19), which here takes the form

$$\frac{1}{h^2}\left[a\left((\nu+\tfrac{1}{2})h\right)(u_{\nu+1}-u_\nu) - a\left((\nu-\tfrac{1}{2})h\right)(u_\nu - u_{\nu-1})\right] = 1$$

$(1 \le \nu \le \frac{1}{h} - 1)$, is suited to the equation from Example 10.2. In general the error is of the order $\mathcal{O}(h^2)$. Since the solution of the differential equation is piecewise quadratic here, it is in fact exactly given by the difference solution. On the other hand, if ξ is not a grid point then the error is of the order $\mathcal{O}(h)$.

In the two-dimensional case one obtains $\mathcal{O}(h^2)$ difference solutions if γ coincides with lines of the grid. Otherwise the error worsens to $\mathcal{O}(h)$. Another possible form of difference approximation consists in approximating the differential equations separately in the regions Ω_1 and Ω_2 and then, to handle the unknown values on γ to discretise the transition condition (10.3).

10.1.4 Discontinuous Coefficients of the First and Zeroth Derivatives

In the following the principal part is harmless (we choose the Laplace operator), whereas the other coefficient $\mathbf{b}^I$, $\mathbf{b}^{II}$, and a may be discontinuous on the interior boundary γ:

$$-\sum_{i,j=1}^{n} \Delta u + \left\langle \mathbf{b}^I(\mathbf{x}), \nabla u\right\rangle - \sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(b_i^{II}(\mathbf{x})u\right) + a(\mathbf{x})u(\mathbf{x}) = f \qquad \text{in } \Omega.$$

The variational formulation for the solution $u \in H_0^1(\Omega)$ is

$$a(u,v) := \int_\Omega \left[\langle \nabla u, \nabla v\rangle + \left\langle \mathbf{b}^I(\mathbf{x}), \nabla u\right\rangle v + u\left\langle \mathbf{b}^{II}(\mathbf{x}), \nabla v\right\rangle + auv\right] \mathrm{d}\mathbf{x} = f(v)$$

for all $v \in H_0^1(\Omega)$. Also in this case the second derivatives of u are discontinuous across γ. Assuming $\mathbf{b}^{II} \in C^1(\Omega_1) \cap C^1(\Omega_2)$ and $\mathbf{b}^I \in C^0(\Omega_1) \cap C^0(\Omega_2)$ in the situation of Figure 10.1, integration by parts shows that u satisfies the equations

$$-\Delta u + \left\langle \mathbf{b}^I, \nabla u\right\rangle - \sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(b_i^{II} u\right) + au = f \qquad \text{in } \Omega_1 \text{ and } \Omega_2$$

and the transition equation

$$u^{(1)} = u^{(2)}, \quad \frac{\partial u^{(1)}}{\partial n} + \left\langle \mathbf{b}^{II\,(1)}, \mathbf{n} \right\rangle u^{(1)} = \frac{\partial u^{(2)}}{\partial n} + \left\langle \mathbf{b}^{II\,(2)}, \mathbf{n} \right\rangle u^{(2)} \quad \text{on } \gamma.$$

Here $\cdot^{(1)}$ and $\cdot^{(2)}$ denote the limits on γ from the side of Ω_1 and respectively Ω_2. $\mathbf{n}$ is the normal directions of one of the subdomains.

Concerning the discretisation the statement in §10.1.2 can be repeated.

10.2 A Singular Perturbation Problem

10.2.1 The Convection-Diffusion Equation

In the following we shall consider the boundary-value problem

$$-\varepsilon \Delta u + \sum_{i=1}^{n} c_i \frac{\partial u}{\partial x_i} = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \Gamma \tag{10.4}$$

for $\varepsilon > 0$. One calls $-\varepsilon \Delta u$ the *diffusion term* and $\sum c_i \frac{\partial u}{\partial x_i} = \langle \mathbf{c}, \nabla u \rangle$ the *convection term*. For small ε the convection term dominates. The corresponding variational formulation is

$$\int_\Omega \left[\varepsilon \langle \nabla u, \nabla v \rangle + \sum_{i=1}^{n} c_i \frac{\partial u}{\partial x_i} v \right] \mathrm{d}\mathbf{x} = \int_\Omega f\, v \, \mathrm{d}\mathbf{x}. \tag{10.5}$$

Equation (10.4) is elliptic for all $\varepsilon > 0$. The unique solvability is a consequence of Theorem 5.11 (cf. also Exercise 10.6). We denote the solution by u_ε.

In the following we analyse what happens when ε becomes small, i.e., if the convection term is dominating.

Exercise 10.6. Let the coefficients c_i be constants. Equation (10.4) can be transformed by

$$v(\mathbf{x}) := u(\mathbf{x}) \exp\Big(-\sum\nolimits_i c_i x_i/(2\varepsilon) \Big) = u(\mathbf{x}) \exp\big(-\langle \mathbf{c}, \mathbf{x} \rangle /(2\varepsilon) \big)$$

into the symmetric $H_0^1(\Omega)$-elliptic equation

$$-\varepsilon \Delta v + \left(\sum_{i=1}^{n} \frac{c_i^2}{4\varepsilon} \right) v = f(\mathbf{x}) \cdot \mathrm{e}^{-\langle \mathbf{c}, \mathbf{x} \rangle /(2\varepsilon)} \quad \text{in } \Omega, \qquad v = 0 \quad \text{on } \Gamma.$$

Concerning the convergence of u_ε as $\varepsilon \to 0$, we first give a negative result.

Remark 10.7. u_ε cannot converge for $\varepsilon \to 0$ with respect to the norm $\|\cdot\|_{H^1(\Omega)}$.

Proof. Otherwise we can perform the limit $\varepsilon \to 0$ in (10.5). Then $u_0 := \lim u_\varepsilon$ satisfies the variational problem $\int_\Omega \sum_{i=1}^n c_i \frac{\partial u_0}{\partial x_i} v \mathrm{d}\mathbf{x} = \int_\Omega f v \mathrm{d}\mathbf{x}$ and the equation

$$c_i \frac{\partial u_0}{\partial x_i} = f \qquad \text{in } \Omega, \tag{10.6}$$

which is a differential equation of first order. This equation is of hyperbolic type, but not compatible with the boundary condition $u = 0$ on Γ (cf. §1.4), i.e., equation (10.6) has, in general, no solution with $u_0 = 0$ on Γ. ∎

Equation (10.6) is called the '*reduced equation*'. Equations (10.4) and (10.6) differ in the 'perturbation term' $-\varepsilon \Delta u$. Since the equations, (10.4) and (10.6), are of *different types* one speaks of a *singular* perturbation.

Another example of a singularly perturbed equation is the *reaction-diffusion equation* $-\varepsilon \Delta u + au = f$ (cf. Schatz–Wahlbin [257]). Here the reduced equation is the algebraic equation $au = f$. A transition from the elliptic to the parabolic type occurs for $-\varepsilon u_{xx} - u_{yy} + u_x = f$. Also the fourth order equation $\varepsilon \Delta^2 u - \Delta u = f$ is singularly perturbed. Although also the reduced equation $-\Delta u = f$ is elliptic, different boundary condition are required for $\varepsilon > 0$ and $\varepsilon = 0$.

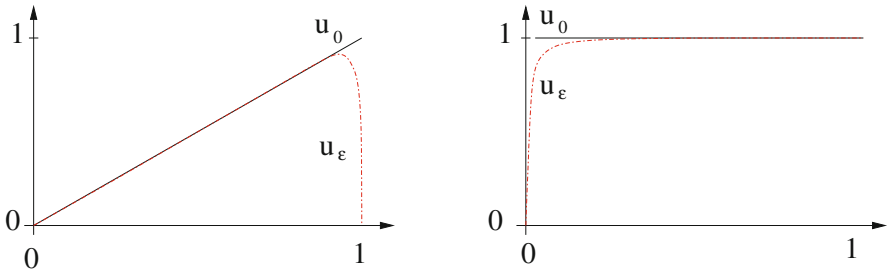


Fig. 10.3 (a) Solution from Example 10.8a and (b) solution from Example 10.8b.

The following example will show that $u_0 = \lim u_\varepsilon$ exists in Ω and satisfies equation (10.6), but that the boundary condition $u_0 = 0$ is only satisfied on a part Γ_0 of the boundary Γ.

Example 10.8. (a) The solution of the ordinary boundary-value problem

$$-\varepsilon u'' + u' = 1 \quad \text{in } \Omega := (0,1), \qquad u(0) = u(1) = 0, \tag{10.7a}$$

is $u_\varepsilon(x) = x - (\mathrm{e}^{x/\varepsilon} - 1)/(\mathrm{e}^{1/\varepsilon} - 1)$. On $[0,1)$, $u_\varepsilon(x)$ converges pointwise to $u_0(x) = x$. This function satisfies the reduced equation (10.6), $u' = 1$, and the left boundary condition $u_0(0) = 0$, but not $u_0(1) = 0$.

(b) The solution of

$$-\varepsilon u'' - u' = 0 \quad \text{in } \Omega := (0,1), \qquad u(0) = 0,\ u(1) = 1, \tag{10.7b}$$

is $u_\varepsilon(x) = (1-\mathrm{e}^{-x/\varepsilon})/(1-\mathrm{e}^{-1/\varepsilon})$. In this case, $u_0(x) := \lim_{\varepsilon\to 0} u_\varepsilon(x) = 1$ satisfies equation (10.6): $-u' = 0$ and $u_0(1) = 1$, but not the left boundary condition at $x = 0$.

Which boundary condition is fulfilled depends for equations (10.7a,b) on the sign of the convection term $\pm u'$. In the many-dimensional case the decisive factor is the direction of the vector $\mathbf{c} = (c_1, \ldots, c_n)$, which is also called the *characteristic direction*.

In Figure 10.3a,b the solutions of Example 10.8 are sketched. In the interior, u_ε is close to u_0; only in the neighbourhood of $x = 1$ (Figure 10.3a) [resp. $x = 0$, Figure 10.3b] does u_ε differ from u_0, in order to satisfy the second boundary condition. These neighbourhoods, in which the derivatives of u_ε attain order $\mathcal{O}(\frac{1}{\varepsilon})$, are called *boundary layers*. For Example 10.8 the thickness of the boundary layer is of order $\mathcal{O}(\varepsilon)$.

Exercise 10.9. The interval $[1-\xi, 1]$, in which the function $(\mathrm{e}^{x/\varepsilon}-1)/(\mathrm{e}^{1/\varepsilon}-1)$ exceeds the value $\eta \in (0,1)$, has the thickness $\xi = \mathcal{O}(\varepsilon\,|\log\eta|)$.

A detailed analysis of singularly perturbed problems can be found, e.g., in Roos–Stynes–Tobiska [246] and Großmann–Roos–Stynes [124, §6]).

10.2.2 Stable Difference Schemes

The special case of equation (10.4) for $\mathbf{c} = (1, 0)$ is

$$-\varepsilon\Delta u + u_x = f \quad \text{in } \Omega, \qquad u = \varphi \quad \text{on } \Gamma. \tag{10.8}$$

The symmetric difference formula (5.10) for equation (10.8) is

$$L_h = h^{-2}\begin{bmatrix} & -\varepsilon & \\ -\varepsilon - h/2 & 4\,\varepsilon & -\varepsilon + h/2 \\ & -\varepsilon & \end{bmatrix}. \tag{10.9}$$

For a fixed ε, L_h is consistent of order 2. From Corollary 5.19 there follows the next remark.

Remark 10.10. As soon as $h < 2\varepsilon$, the difference scheme (10.9) leads to an M-matrix.

The property of being an M-matrix was used in Section 5.1.4 to demonstrate the solvability of the discrete equation. It turns out that the difference equations are also solvable for $h \geq 2\varepsilon$. However, with increasing h/ε there develops an instability which is made clear in Table 10.1.

ε x	0	1/32	2/32	3/32	4/32	…	30/32	31/32	1
0.5	1	0.93	0.87	0.81	0.75	…	0.04	0.02	0
0.1	1	0.97	0.95	0.92	0.89	…	0.22	0.13	0
0.05	1	0.99	0.97	0.96	0.94	…	0.47	0.31	0
0.02	1	0.99	0.99	0.98	0.98	…	0.82	0.73	0
0.01	1	1.00	0.99	0.99	0.99	…	0.87	1.11	0
1_{10}-3	1	1.03	0.99	1.04	0.99	…	0.23	1.89	0
1_{10}-4	1	4.95	0.95	5.00	0.90	…	0.08	5.88	0
1_{10}-5	1	48.6	0.94	48.7	0.88	…	0.06	49.6	0
1_{10}-7	1	4859.5	0.94	4859.6	0.88	…	0.06	4860.4	0

Table 10.1 Values $u_{\varepsilon,h}(x, 1/2)$ of the solution of equation (10.8).

Table 10.1 shows the values $u_{\varepsilon,h}(x, 1/2)$ of the difference solution $u_{\varepsilon,h}$ of equation (10.8) in the unit square $\Omega = (0,1)\times(0,1)$ with $f = 0$ and $\varphi(x,y) = (1-x)\sin(\pi y)$. The grid step is $h = 1/32$. For $\varepsilon > h/2 = 1/64 = 0.015625$, L_h is an (irreducibly diagonally dominant) M-matrix. Because of the maximum principle (cf. Remark 4.37) the values lie between $\min\varphi = 0$ and $\max\varphi = 1$. The solution of the reduced equation (10.6) is

$$u_0(x,y) = \lim_{\varepsilon\to 0} u_\varepsilon(x,y) = \sin\pi y.$$

Accordingly the values $u_{\varepsilon,h}(x, 1/2)$ from the first part of Table 10.1 approach the limiting value $\sin\pi/2 = 1$.

In the second part of the table we have $\varepsilon < \frac{h}{2}$. Thus L_h is not an M-matrix. Notice that $u_{\varepsilon,h}(31/32, 1/2) > 1$, i.e., even the maximum principle does not hold. In addition one can see that $u_{\varepsilon,h}(x, 1/2)$ converges to $1-x$ for the even multiples $x = \nu h$, and thus not to $u_0(x,y) = \sin\pi y$. For the intervening odd multiples $x = \nu h$ an oscillation of amplitude $\mathcal{O}(h^2/2\varepsilon)$ develops.

The resulting difficulty is similar to that for initial value problems for stiff (ordinary) differential equations: If one keeps ε constant and lets h go to zero, the convergence assertions of §5.1.4 hold. However, if ε is very small, the condition $h < 2\varepsilon$, without which one does not obtain a reasonable solution, cannot in practice be satisfied.

One way out of this difficulty has already been described in §5.1.4. One must approximate the convection term $\sum_{i=1}^{n} c_i \frac{\partial u}{\partial x_i}$ (here for $n = 2$) by suitable *one-sided* differences (5.15):

$$L_h = h^{-2}\begin{bmatrix} & -\varepsilon + hc_2^- & \\ -\varepsilon - hc_1^+ & 4\varepsilon + h|c_1| + h|c_2| & -\varepsilon + hc_1^- \\ & -\varepsilon - hc_2^+ & \end{bmatrix} \text{ with } \begin{cases} c_i^+ := \max\{0, c_i\}, \\ c_i^- := \min\{0, c_i\}. \end{cases} \tag{10.10}$$

In the case of $c_1, c_2 \geq 0$, i.e., $L_h = h^{-2}\begin{bmatrix} & -\varepsilon & \\ -\varepsilon - hc_1 & 4\varepsilon + hc_1 + hc_2 & -\varepsilon \\ & -\varepsilon - hc_2 & \end{bmatrix}$, only backward differences occur.

Remark 10.11. If one discretises equation (10.4) using (10.10) then one obtains, for all $\varepsilon > 0$ and $h > 0$, an M-matrix L_h. For fixed ε, the scheme has consistency order 1.

Exercise 10.12. The discretisation of equation (10.7b) corresponding to (10.10) is $L_h = h^{-2}[-\varepsilon \;\; 2\varepsilon+h \;\; -\varepsilon-h]$ and this gives the discrete solution

$$u_h(x) = \frac{1-(1+h/\varepsilon)^{-x/h}}{1-(1+h/\varepsilon)^{-1/h}}.$$

If one applies the difference operator (10.10) to a smooth function one obtains the Taylor series

$$L_h u = Lu - \frac{h}{2}\left[\, |c_1|\, u_{xx} + |c_2|\, u_{yy} \,\right] + \mathcal{O}(h^2). \tag{10.11}$$

The $\mathcal{O}(h)$ term $\frac{h}{2}\left(|c_1|\, u_{xx} + |c_2|\, u_{yy}\right)$ is called the *numerical viscosity* (or the *numerical diffusion*). In the sense of a backward error analysis one can interpret the discrete solution as the approximation of a differential equation with the additional term $\frac{h}{2}\left(|c_1|\, u_{xx} + |c_2|\, u_{yy}\right)$. This term depends on the *streamline* $\mathbf{c} = (c_1, c_2)$. If, e.g., $c_2 = 0$ (convection in x-direction), the numerical viscosity is a second derivative in x-direction only.

A second remedy consists in replacing the parameter ε by a discretisation using $\varepsilon_h \geq \varepsilon$. If one chooses

$$\varepsilon_h := \max\left\{\varepsilon, \frac{h}{2}\,|c_1|\,, \frac{h}{2}\,|c_2|\right\} \qquad \text{or} \qquad \varepsilon_h := \varepsilon + \frac{h}{2}\,\|\mathbf{c}\|_\infty\,, \tag{10.12}$$

then the symmetric difference method

$$L_h = \varepsilon_h h^{-2}\begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix} + \frac{1}{2}h^{-1}\begin{bmatrix} & c_2 & \\ -c_1 & 0 & c_1 \\ & -c_2 & \end{bmatrix} \tag{10.13}$$

leads to an M-matrix. It is true that the convection term has been discretised to a second order of consistency, but the error of the diffusion term is $\mathcal{O}(\varepsilon_h - \varepsilon)$, which for the practically relevant case $h > 2\varepsilon/\,\|\mathbf{c}\|_\infty$ amounts to $\mathcal{O}(h)$. Instead of (10.11) one has

$$L_h u = Lu - (\varepsilon_h - \varepsilon)\Delta u + \mathcal{O}(h^2).$$

The difference $-(\varepsilon_h - \varepsilon)\Delta u$ is called the *artificial viscosity*. Different from (10.11) the consistency is deteriorated with respect to all directions.

In the one-dimensional case the methods using numerical and artificial viscosities do not differ.

Remark 10.13. In the one-dimensional case the difference formulae (10.10) and (10.13) coincide, if one chooses ε_h according to the second alternative in (10.12).

10.2.3 Finite Elements

The difficulties described in the previous section are not restricted to difference methods. First we discuss the standard Galerkin methods and its shortcomings. In §10.2.3.2 the streamline-diffusion method is presented which is a variant of the Galerkin method adapted to the problem.

10.2.3.1 Standard Galerkin Method

The weak formulation of the convection-diffusion equation is

$$\text{find } u \in H_0^1(\Omega) \text{ with } a(u,v) = f(v) \text{ for all } v \in H_0^1(\Omega), \text{ where} \tag{10.14a}$$

$$a(u,v) := \int_\Omega \left[\varepsilon \langle \nabla u(\mathbf{x}), \nabla v(\mathbf{x})\rangle + \langle \mathbf{c}, \nabla u(\mathbf{x})\rangle\, v(\mathbf{x}) + auv\right] \mathrm{d}\mathbf{x}\,, \tag{10.14b}$$

$f(v) = \int_\Omega f(\mathbf{x})v(\mathbf{x})\mathrm{d}\mathbf{x}$, and $\mathbf{c} = \mathbf{c}(\mathbf{x}) = (c_i)_{i=1,\ldots,n}$. In the first part we limit ourselves to $a = 0$ in (10.14b).

Exercise 10.14. Show that: (a) Linear finite elements on a square-grid triangulation applied in the case of equation (10.8) give a discretisation that is identical with the difference method

$$L_h = \varepsilon \begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix} + h \begin{bmatrix} 0 & -1/6 & 1/6 \\ -1/3 & 0 & 1/3 \\ -1/6 & 1/6 & 0 \end{bmatrix} \tag{10.15}$$

(cf. Exercise 8.42).

(b) For bilinear elements (cf. Exercise 8.46) one obtains

$$L_h = \frac{\varepsilon}{3} \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix} + \frac{h}{12} \begin{bmatrix} -1 & 0 & +1 \\ -4 & 0 & +4 \\ -1 & 0 & +1 \end{bmatrix}.$$

(c) One-dimensional linear elements for $-\varepsilon u'' + u' = f$ lead to the central difference formula

$$L_h = \varepsilon h^{-1} [-1 \quad 2 \quad -1] + [-1 \quad 0 \quad 1]\,.$$

The Exercises 10.14b,c show that finite-element methods correspond to central difference formulae, and thus can equally well lead to instability.

The method of artificial viscosity corresponds to the finite-element solution of the equation $-\varepsilon_h \Delta u + \langle \mathbf{c}, \nabla u\rangle = f$ for appropriate ε_h. As in Exercise 10.14b one can show the following statement.

Remark 10.15. If one sets $\varepsilon_h := \max\{\varepsilon, |c_1|h, |c_2|h\}$ and uses bilinear elements as in Exercise 10.14b then the discretisation of equation (10.4) leads to an M-matrix.

On the other hand the matrix (10.15) has different signs in the sub-diagonal and in the super-diagonal, so that it is not possible to have an M-matrix for any value of ε_h and $\varepsilon > 0$.

The analogues of one-sided differences are more difficult to construct. One approach is to combine a finite-element method for the diffusion term with a (one-sided) difference method for the convection term (cf. Thomasset [284, §2.4]).

A second possibility is the generalisation of the Galerkin method to the Petrov–Galerkin method, in which the discrete solution of the general equation (8.1) is defined by

$$\text{find } u \in V_h, \text{ so that } \quad a(u,v) = f(v) \quad \text{for all } v \in W_h$$

(cf. Fletcher [102, §7.2], Thomasset [284, §2.2], and §8.9.10.3).

A third possibility is the streamline-diffusion method explained below.

10.2.3.2 Streamline-Diffusion Method

We follow the presentation in Roos–Stynes–Tobiska [246, §III.3.2.1] and start from the weak formulation $a(u,v) = f(v)$ in (10.14a,b). The (weak) solution u satisfies the equation $-\varepsilon\Delta u + \langle \mathbf{c}, \nabla u\rangle + au = f$ in $H^{-1}(\Omega)$. Assuming $f \in L^2(\Omega)$, we can multiply both sides by the expression $\omega(\mathbf{x}) \cdot \langle \mathbf{c}, \nabla v\rangle \in L^2(\Omega)$, where $v \in H_0^1(\Omega)$ is a test function and ω a weight function:

$$\int_\Omega \omega(\mathbf{x}) \, (-\varepsilon\Delta u + \langle \mathbf{c}, \nabla u\rangle + au) \, \langle \mathbf{c}, \nabla v\rangle \, \mathrm{d}\mathbf{x} = \int_\Omega f(\mathbf{x}) \, \langle \mathbf{c}, \nabla v\rangle \, \mathrm{d}\mathbf{x}\,.$$

Adding this equation to the original variational problem (10.14a), we obtain

$$\begin{aligned} &a^{\mathrm{SD}}(u,v) = f^{\mathrm{SD}}(v) \quad \text{for all } v \in H_0^1(\Omega) \qquad \text{with} \qquad (10.16)\\ &a^{\mathrm{SD}}(u,v) := \int_\Omega \Big\{ \varepsilon \, \langle \nabla u, \nabla v\rangle + \langle \mathbf{c}, \nabla u\rangle \, v + auv \\ &\qquad\qquad + \omega(\mathbf{x}) \, \big[-\varepsilon\Delta u + \langle \mathbf{c}, \nabla u\rangle + au \, \big] \, \langle \mathbf{c}, \nabla v\rangle \Big\} \, \mathrm{d}\mathbf{x}, \\ &f^{\mathrm{SD}}(v) := \int_\Omega f(\mathbf{x}) \, (v + \langle \mathbf{c}, \nabla v\rangle) \, \mathrm{d}\mathbf{x}\,. \end{aligned}$$

Note that in spite of this modification the solution of (10.14a) is also a solution of the new equation (10.16).

The discretisation by finite elements leads to a problem. In general, $V_h \subset V = H_0^1(\Omega)$ holds, but not $V_h \subset H_0^1(\Omega) \cap H^2(\Omega)$. Therefore Δu^h does not belong to $L^2(\Omega)$. In the interior of each triangle T of the finite-element triangulation $\mathcal{T}$ the restriction $u^h|_T$ is a polynomial, so that $\Delta u^h|_T$ is again a polynomial and therefore smooth. Across the edges of the triangles the derivative is discontinuous, so that the second derivative Δu^h is a distribution with support in $\bigcup_{T\in\mathcal{T}} \partial T$. The discretisation

presented now is ignoring the distributional parts in Δu^h; it only involves integrals of $\Delta u^h|_T$. The weight function is chosen as a constant in each triangle. Hence each triangle T is associated with a factor ω_T. We are looking for a solution $u^h \in V_h$ of

$$a_h^{\mathrm{SD}}(u^h, v) = f_h^{\mathrm{SD}}(v) \qquad \text{for all } v \in V_h \qquad \text{with} \tag{10.17}$$

$$a_h^{\mathrm{SD}}(u^h, v) := \int_\Omega \Big[\varepsilon \left\langle \nabla u^h, \nabla v \right\rangle + \left\langle \mathbf{c}, \nabla u^h \right\rangle v + a u^h v + \sum_{T \in \mathcal{T}} \omega_T \int_T \left(\varepsilon \Delta u^h + \left\langle \mathbf{c}, \nabla u^h \right\rangle + a u^h \right) |_T \, \langle \mathbf{c}, \nabla v \rangle \Big] \mathrm{d}\mathbf{x}$$

$$f_h^{\mathrm{SD}}(v) := \int_\Omega f(\mathbf{x}) \left(v + \langle \mathbf{c}, \nabla v \rangle \right) \mathrm{d}\mathbf{x}.$$

Because the distributional part in Δu^h is omitted, the finite-element solution of (10.14a)—different from the continuous case—does *not* coincide with the solution of (10.17). Quite the contrary, the discretisation (10.17) will possess better stability properties as we shall see.

Remark 10.16. In the standard case of piecewise linear finite elements, $\Delta u^h = 0$ holds on each triangle.

Discretisation (10.17) is called the *streamline-diffusion method*[1] because of the term $\langle \mathbf{c}, \nabla v \rangle$. It corresponds to the stabilisation by one-sided differences in the case of difference methods. To see this connection we consider the regular square-grid triangulation from Figure 8.3 and piecewise linear finite elements. The streamline in x direction is characterised by $\mathbf{c} = (c_1, 0)$, i.e., $\left\langle \mathbf{c}, \nabla u^h \right\rangle = c_1 \partial u^h / \partial x$. We assume c_1 to be constant. By Remark 10.16 the Δu^h-term vanishes. Let T be a triangle with the vertices (x, y), $(x + h, y)$, $(x, y + h)$. Since u^h and v are linear on T, the derivatives are constant and coincide with the difference quotients:

$$\int_T \left\langle \mathbf{c}, \nabla u^h \right\rangle \langle \mathbf{c}, \nabla v \rangle \, \mathrm{d}\mathbf{x} = \int_T c_1^2 \frac{\partial u^h}{\partial x} \frac{\partial v}{\partial x} \mathrm{d}x\mathrm{d}y = \frac{c_1^2}{2} h^2 \, \partial_{x,h}^+ u^h(x, y) \, \partial_{x,h}^+ (x, y).$$

For the adjacent triangle T' with the corners (x, y), $(x, y + h)$, $(x + h, y + h)$ we have

$$\int_{T'} \left\langle \mathbf{c}, \nabla u^h \right\rangle \langle \mathbf{c}, \nabla v \rangle \, \mathrm{d}\mathbf{x} = \frac{c_1^2}{2} h^2 \, \partial_{x,h}^+ u^h(x, y + h) \, \partial_{x,h}^+ v(x, y + h).$$

Therefore the finite-element matrix $\mathbf{L}$ contains an additional part corresponding to the difference star

$$\frac{c_1^2}{2} \begin{bmatrix} 0 & 0 & 0 \\ -1 & 2 & -1 \\ 0 & 0 & 0 \end{bmatrix}.$$

[1] Another name is streamline upwind Petrov–Galerkin, abbreviated by SUPG.

This shows that the numerical diffusion is the second difference in the streamline direction which is the direction $\mathbf{c} = (c_1, 0)$. The one-sided difference (10.10) has the same effect if $c_2 = 0$.

The stability of the streamline-diffusion method will be proved by an inequality resembling the V-ellipticity. For this purpose we need some assumptions. Let the coefficients $\mathbf{c}$ and a in (10.14b) satisfy

$$a - \frac{1}{2} \operatorname{div} \mathbf{c} \geq c_0 > 0 \qquad \text{in } \Omega. \tag{10.18a}$$

In the case of constant coefficients this condition is equivalent to $a > 0$. The significance of inequality (10.18a) can be seen from the next statement.

Lemma 10.17. *Assume (10.18a). Then for all $u \in H_0^1(\Omega)$ the following estimate holds:*

$$\int_\Omega \big[au + \langle \mathbf{c}, \nabla u \rangle \big] u \, \mathrm{d}\mathbf{x} \geq c_0 \, \|u\|^2_{L^2(\Omega)} .$$

Proof. Since $u^2 = 0$ on $\partial\Omega$ the assertion follows from

$$\int_\Omega c_i \frac{\partial u}{\partial x_i} u \mathrm{d}\mathbf{x} = \frac{1}{2} \int_\Omega c_i \frac{\partial u^2}{\partial x_i} \mathrm{d}\mathbf{x} = -\frac{1}{2} \int_\Omega u^2 \frac{\partial c_i}{\partial x_i} \mathrm{d}\mathbf{x}. \qquad \blacksquare$$

The later estimates require a bound of a on each triangle $T \in \mathcal{T}$:

$$c_T := \max_{\mathbf{x} \in T} |a(\mathbf{x})| \ . \tag{10.18b}$$

Set

$$h_T := \text{diameter of the triangle } T. \tag{10.18c}$$

The test function $v \in V_h$ restricted to T is a polynomial of a fixed degree. The inverse inequality states that there is some μ with

$$\|\Delta v\|_{L^2(T)} \leq \frac{\mu}{h_T} \|\nabla v\|_{L^2(T)} \qquad \text{for all } T \in \mathcal{T}_h \text{ and all } v \in V_h, \tag{10.18d}$$

where $\|\nabla v\|^2_{L^2(T)} = \sum_{i=1}^n \|\partial v / \partial x_i\|^2_{L^2(T)}$. According to Remark 10.16, piecewise linear elements satisfy inequality (10.18d) with $\mu = 0$.

In the case of the bilinear form $a_h^{\mathrm{SD}}(\cdot, \cdot)$,

$$|||v|||_{\mathrm{SD}} := \sqrt{\varepsilon \, \|\nabla v\|^2_{L^2(\Omega)} + c_0 \, \|v\|^2_{L^2(\Omega)} + \sum_{T \in \mathcal{T}} \omega_T \, \|\langle \mathbf{c}, \nabla v \rangle\|^2_{L^2(T)}}$$

plays the role of the energy norm.

Theorem 10.18 (stability). *Assume (10.18a–d) with the constants defined therein. Choose weights* ω_T *satisfying*

$$0 < \omega_T \le \min\left\{\frac{c_0}{c_T^2}, \frac{h_T^2}{\varepsilon\mu}\right\} \qquad \textit{for all } T \in \mathcal{T}. \tag{10.19}$$

Then the bilinear form $a_h^{\mathrm{SD}}(\cdot,\cdot) : V_h \times V_h \to \mathbb{R}$ *satisfies the inequality*

$$a_h^{\mathrm{SD}}(v,v) \ge \frac{1}{2}|||v|||_{\mathrm{SD}}^2 \qquad \textit{for all } v \in V_h\,. \tag{10.20}$$

Proof. Using Lemma 10.17, one obtains

$$\begin{aligned} a_h^{\mathrm{SD}}(v,v) \ge{}& \varepsilon\,\|\nabla v\|_{L^2(\Omega)}^2 + c_0\,\|v\|_{L^2(\Omega)}^2 + \sum_{T\in\mathcal{T}} \omega_T\,\|\langle \mathbf{c}, \nabla v\rangle\|_{L^2(T)}^2 \\ &+ \sum_{T\in\mathcal{T}} \omega_T \int_T (\varepsilon\Delta v + av)\,|_T\ \langle \mathbf{c}, \nabla v\rangle\,\mathrm{d}\mathbf{x}\,. \end{aligned}$$

We apply the inequality $\alpha\beta \le \frac{1}{2}\alpha^2 + \frac{1}{2}\beta^2$ to the terms of the last sum:

$$\begin{aligned} &\left|\sum_{T\in\mathcal{T}} \omega_T \int_T (\varepsilon\Delta v + av)\,|_T\ \langle \mathbf{c}, \nabla v\rangle\,\mathrm{d}\mathbf{x}\right| \\ &\quad\le \sum_{T\in\mathcal{T}} \omega_T \left[\frac{1}{2}\,\|\varepsilon\Delta v + av\|_{L^2(T)}^2 + \frac{1}{2}\,\|\langle \mathbf{c}, \nabla v\rangle\|_{L^2(T)}^2\right] \\ &\quad\le \sum_{T\in\mathcal{T}} \omega_T \left[\|\varepsilon\Delta v\|_{L^2(T)}^2 + \|av\|_{L^2(T)}^2 + \frac{1}{2}\,\|\langle \mathbf{c}, \nabla v\rangle\|_{L^2(T)}^2\right] \\ &\quad\le \sum_{T\in\mathcal{T}} \omega_T \left[\varepsilon^2\,\|\Delta v\|_{L^2(T)}^2 + c_T^2\,\|v\|_{L^2(T)}^2 + \frac{1}{2}\,\|\langle \mathbf{c}, \nabla v\rangle\|_{L^2(T)}^2\right]. \end{aligned}$$

Combining both inequalities, we prove the desired estimate. ∎

For $\mu = 0$ we have $\frac{h_T^2}{\varepsilon\mu} = \infty$ and thus $\min\left\{\frac{c_0}{c_T^2}, \frac{h_T^2}{\varepsilon\mu}\right\} = \frac{c_0}{c_T^2}$ in (10.19). The following lemma describes the consistency of the streamline-diffusion method.

Lemma 10.19. *Let the assumptions of Theorem 10.18 and (10.19) be valid. In the case of* $\mu = 0$ *we explicitly require* $\varepsilon\omega_T \le Ch_T^2$.[2] V_h *is assumed to contain polynomials of degree* $\le k$ *for some* $k \ge 1$. *Let the solution* u *of (10.5) belong to* $H^{k+1}(\Omega)$. *The map* $R_h : H^{k+1}(\Omega) \cap H_0^1(\Omega) \subset V \to V_h$ *is the interpolation at the nodes. Then the consistency error is*

$$|||R_h u - u^h|||_{\mathrm{SD}} \le Ch^k \sqrt{\sum_{T\in\mathcal{T}} \left[\varepsilon + \omega_T + h_T^2/\omega_T + h_T^2\right] \|u\|_{H^{k+1}(T)}^2}. \tag{10.21}$$

[2] Otherwise this inequality follows from (10.19).

Proof. (i) Inserting $R_h u - u^h$ into (10.20) yields the first inequality in

$$\frac{1}{2}|||R_h u - u^h|||_{\mathrm{SD}}^2 \le a_h^{\mathrm{SD}}(R_h u - u^h, R_h u - u^h) = a_h^{\mathrm{SD}}(R_h u - u, R_h u - u^h). \tag{10.22a}$$

The second equality is based on the projection property $a_h^{\mathrm{SD}}(u^h - u, v) = 0$ for all $v \in V_h$, which results from combining (10.16) and (10.17).

(ii) On each triangle $T \in \mathcal{T}$ we have $\|R_h u - u\|_{H^1(T)} \le C h^k \|u\|_{H^{k+1}(T)}$. Piecewise estimation of the terms in $a_h^{\mathrm{SD}}(R_h u - u, R_h u - u^h)$ together with Schwarz' inequality gives

$$\int_\Omega \varepsilon \left\langle \nabla\left(R_h u - u\right), \nabla\left(R_h u - u^h\right)\right\rangle \mathrm{d}\mathbf{x} \le \sqrt{\varepsilon}\, h^k\, |u|_{k+1}\, |||R_h u - u^h|||_{\mathrm{SD}}\,, \tag{10.22b}$$

$$\begin{aligned}
&\int_\Omega \left[\langle \mathbf{c}, \nabla\left(R_h u - u\right)\rangle \left(R_h u - u^h\right) + a\left(R_h u - u\right)\left(R_h u - u^h\right)\right] \mathrm{d}\mathbf{x} \\
&\quad = \int_\Omega \left[\left(a - \operatorname{div}\mathbf{c}\right)\left(R_h u - u\right)\left(R_h u - u^h\right) - \left(R_h u - u\right)\left\langle \mathbf{c}, \nabla\left(R_h u - u^h\right)\right\rangle\right] \mathrm{d}\mathbf{x} \\
&\quad \le \left\{ C \sqrt{\sum_{T\in\mathcal{T}} \|u\|_{L^2(T)}^2} + \sqrt{\sum_{T\in\mathcal{T}} \frac{1}{\omega_T} \|u\|_{L^2(T)}^2} \right\} |||R_h u - u^h|||_{\mathrm{SD}} \\
&\quad \le C' h^k \sqrt{\sum_{T\in\mathcal{T}} h_T^2 \left(1 + 1/\omega_T\right) \|u\|_{H^{k+1}(T)}^2}\; |||R_h u - u^h|||_{\mathrm{SD}}\,,
\end{aligned} \tag{10.22c}$$

$$\begin{aligned}
&\left| \sum_{T\in\mathcal{T}} \omega_T \int_T \begin{bmatrix} -\varepsilon \Delta\left(R_h u - u\right) \\ + \langle \mathbf{c}, \nabla\left(R_h u - u\right)\rangle + a\left(R_h u - u\right) \end{bmatrix} \left\langle \mathbf{c}, \nabla\left(R_h u - u^h\right)\right\rangle \mathrm{d}\mathbf{x} \right| \\
&\quad \le C \sum_{T\in\mathcal{T}} \sqrt{\omega_T} \left[\varepsilon h_T^{k-1} + h_T^k + h_T^{k+1}\right] \|u\|_{H^{k+1}(T)} \sqrt{\omega_T} \left\|\left\langle \mathbf{c}, \nabla\left(R_h u - u^h\right)\right\rangle\right\|_{L^2(T)} \\
&\quad \underset{\varepsilon\omega_T \le C h_T^2}{\le} C' \sqrt{\sum_{T\in\mathcal{T}} \left(\varepsilon + \omega_T\right) h_T^{2k} \|u\|_{H^{k+1}(T)}^2}\; |||R_h u - u^h|||_{\mathrm{SD}}\,.
\end{aligned} \tag{10.22d}$$

Combining the estimates (10.22a–d) after division by $|||R_h u - u^h|||_{\mathrm{SD}}$ proves (10.21). ∎

The proven inequality (10.21) suggests choosing the weights ω_T so that the three quantities ε, ω_T, h_T^2/ω_T are of similar size, where however we have to respect the inequalities (10.19). This leads to

$$\omega_T = \begin{cases} \delta_0 h_T & \text{if } \|\mathbf{c}\|_{L^\infty(T)}\, h_T > 2\varepsilon \;\text{ (dominating convection),} \\ \delta_0 h_T^2/\varepsilon & \text{if } \|\mathbf{c}\|_{L^\infty(T)}\, h_T \le 2\varepsilon \;\text{ (dominating diffusion).} \end{cases} \tag{10.23}$$

Stability together with the consistency of the method implies the convergence result of the next theorem (cf. Roos–Stynes–Tobiska [246, Theorem 3.30]).

Theorem 10.20 (convergence). *Let the assumptions of Lemma 10.19 as well as (10.23) be satisfied. Then the solution u^h of the streamline-diffusion method (10.17) satisfies the error estimate*

$$|||u - u^h|||_{\mathrm{SD}} \leq C \left(\sqrt{\varepsilon} + \sqrt{h} \right) h^k \, |u|_{k+1} \, .$$

Proof. Combine

$$|||R_h u - u^h|||_{\mathrm{SD}} \leq C \left(\sqrt{\varepsilon} + \sqrt{h} \right) h^k \, |u|_{k+1}$$

from (10.21) and the inequality

$$|||R_h u - u|||_{\mathrm{SD}} \leq C \left(\sqrt{\varepsilon} + \sqrt{h} \right) h^k \, |u|_{k+1}$$

following from $\|R_h u - u\|_{H^1(T)} \leq C h^k \, \|u\|_{H^{k+1}(T)}$. ∎

See also Knabner–Angermann [172, §9.2] and Brezzi–Marini–Süli [56]. Superconvergence properties of the streamline-diffusion method are discussed by Zhou–Rannacher [318].

Chapter 11
Elliptic Eigenvalue Problems

Abstract If $(L - \lambda I)\, u = f$ is not solvable, the Riesz–Schauder theory states that there are eigenvalues and eigenvectors. The weak formulation of the eigenvalue problem and some basic terms are discussed in **Section 11.1**. Since we do not require the system to be symmetric, also the adjoint problem must be treated. **Section 11.2** is devoted to the finite-element discretisation by a family $\{V_h : h \in H\}$ of subspaces. Theorems 11.13 and 11.15 state an important result: Each eigenvalue λ_0 of L is associated to a sequence of discrete eigenvalues converging to λ_0, and vice versa. The corresponding error estimates are given in §11.2.3 for the case of simple eigenvalues. A related estimate for the eigenfunctions is provided by Lemma 11.23. Finally, Theorem 11.24 presents an improved error estimate of the eigenvalues by means of the eigenfunctions. The Riesz–Schauder theory also states that the equation $(L - \lambda I)u = f$ can even be solved for eigenvalues λ if f satisfies suitable side conditions. These equations are treated in §11.2.4. **Section 11.3** discusses the discretisation by difference schemes. Also in this case similar results can be obtained.

11.1 Formulation of Eigenvalue Problems

The classical formulation of an eigenvalue problem reads

$$Le = \lambda e \quad \text{in } \Omega, \qquad B_j\, e = 0 \quad \text{on } \Gamma \qquad (j = 1, \ldots, m)\,. \tag{11.1}$$

Here L is an elliptic differential operator of order $2m$, and B_j are boundary operators. A solution e of (11.1) is called an *eigenfunction* if $e \neq 0$. In this case, λ is the *eigenvalue* associated with e. Since in general eigenvalues are complex, the underlying field is $\mathbb{C}$ throughout this chapter.

As in Chapter 7, one can replace the classical representation (11.1) by a variational formulation, with a suitable sesquilinear form $a(\cdot, \cdot) : V \times V \to \mathbb{C}$ taking the place of $\{L, B_j\}$:

W. Hackbusch, *Elliptic Differential Equations*, Springer Series
in Computational Mathematics 18, DOI 10.1007/978-3-662-54961-2_11

$$\text{find } e \in V \quad \text{with} \quad a(e,v) = \lambda\,(e,v)_0 \quad \text{for all } v \in V. \tag{11.2a}$$

$(u,v)_0 = \int_\Omega u\,\overline{v}\,\mathrm{d}\mathbf{x}$ is the $L^2(\Omega)$-scalar product. Strictly speaking one ought to replace $(\cdot,\cdot)_0$ by $(\cdot,\cdot)_U$ where U is the Hilbert space of the Gelfand triple $V \subset U \subset V'$ (cf. §6.3.3). But here we limit ourselves to the standard case $U = L^2(\Omega)$.

The adjoint eigenvalue problem is formulated as

$$\text{find } e^* \in V \quad \text{with} \quad a(v,e^*) = \lambda\,(v,e^*)_0 \quad \text{for all } v \in V. \tag{11.2b}$$

Definition 11.1. Let $\lambda \in \mathbb{C}$. By $E(\lambda)$ one denotes the subspace of all $e \in V$ which satisfy equation (11.1) [resp. (11.2a)]. $E(\lambda)$ is called the *eigenspace* for λ. With $E^*(\lambda)$ one denotes the corresponding eigenspace of equation (11.2b). λ is called an *eigenvalue* if $\dim E(\lambda) \geq 1$.

Theorems 6.107 and 7.14 already contain the following statements.

Theorem 11.2. *Let $V \subset L^2(\Omega)$ be continuously, densely and compactly embedded (for example, $V = H_0^m(\Omega)$ with bounded Ω). Let $a(\cdot,\cdot)$ be V-coercive. Then the problems (11.2a,b) have countably many eigenvalues $\lambda \in \mathbb{C}$ which may only have an accumulation point at ∞. For all $\lambda \in \mathbb{C}$ we have*

$$\dim E(\lambda) = \dim E^*(\lambda) < \infty.$$

Exercise 11.3. In addition to the assumptions of Theorem 11.2 let $a(\cdot,\cdot)$ be symmetric. Show that all eigenvalues are real and problems (11.2a) and (11.2b) are identical so that $E(\lambda) = E^*(\lambda)$.

For ordinary differential equations of second order (i.e., $\Omega \subset \mathbb{R}^1$, $m = 1$) it is known that all eigenvalues are simple: $\dim E(\lambda) = 1$. This statement is incorrect for partial differential equations as the following example shows.

Example 11.4 (multiplicity of eigenvalues). The Poisson equation $-\Delta e = \lambda e$ in the rectangle $(0,a) \times (0,b)$ with Dirichlet boundary values $e = 0$ on Γ, has the eigenvalues

$$\lambda = (\nu\pi/a)^2 + (\mu\pi/b)^2 \qquad \text{for all } \nu,\mu \in \mathbb{N}.$$

The associated eigenfunction reads

$$e(x,y) = e^{\nu,\mu}(x,y) := \sin(\nu\pi\frac{x}{a})\sin(\mu\pi\frac{y}{b}).$$

In the case of the square $a = b$ one obtains for $\nu \neq \mu$ eigenvalues $\lambda = \lambda^{\nu,\mu} = \lambda^{\mu,\nu}$, which have multiplicity at least 2 since $e^{\nu,\mu}$ and $e^{\mu,\nu}$ are linearly independent eigenfunctions for the same eigenvalue. A triple eigenvalue, for example, exists for $a = b$, $\lambda = 50\pi^2/a^2$: $E(\lambda) = \operatorname{span}\{e^{1,7}, e^{7,1}, e^{5,5}\}$.

The eigenfunctions $e \in E(\lambda)$, by definition, belong to V. The regularity investigations of Section 9.1 immediately result in a stronger regularity.

Theorem 11.5. *Let* $V = H_0^m(\Omega)$ *with* $m \geq 1$, *or* $V = H^m(\Omega)$ *with* $m = 1$. *Under the assumptions of Theorems 9.19 [resp. 9.20] we have* $E(\lambda) \subset H^{m+s}(\Omega)$.

Proof. Along with $a(\cdot,\cdot)$, the form

$$a_\lambda(u,v) := a(u,v) - \lambda(u,v)_0$$

also satisfies the assumptions. Since $a_\lambda(e,v) = 0$ for $e \in E(\lambda)$ and $v \in V$, the statement follows from Corollary 9.21. ∎

Besides the standard form (11.2a) there are *generalised eigenvalue problems*. An example is the *Steklov problem* (cf. Steklov [273])

$$-\Delta e = 0 \quad \text{in } \Omega, \qquad \partial e/\partial n = \lambda\, e \quad \text{on } \Gamma,$$

whose variational formulation reads

$$e \in H^1(\Omega) \text{ satisfies } \int_\Omega \langle \nabla e, \nabla v\rangle \mathrm{d}\mathbf{x} = \lambda \int_\Gamma ev \mathrm{d}\Gamma \quad (v \in H^1(\Omega)).$$

One can show that all eigenvalues are real and that the statements of Theorem 11.2 hold.

11.2 Finite-Element Discretisation

11.2.1 Discretisation

Let $V_h \subset V$ be a (finite-element) subspace. The Ritz–Galerkin (resp. finite-element) discretisations of the eigenvalue problems (11.2a,b) read

$$\text{find} \quad e^h \in V_h \quad \text{with} \quad a(e^h, v) = \lambda_h (e^h, v)_0 \quad \text{for all } v \in V_h, \tag{11.3a}$$

$$\text{find} \quad e^{*h} \in V_h \quad \text{with} \quad a(v, e^{*h}) = \lambda_h (v, e^{*h})_0 \quad \text{for all } v \in V_h. \tag{11.3b}$$

The discrete eigenspaces $E_h(\lambda_h)$, $E_h^\star(\lambda_h)$ are spanned by the solutions of the problems (11.3a) [resp. (11.3b)]. As in Theorem 11.2, $\dim E_h(\lambda_h) = \dim E_h^\star(\lambda_h)$ holds. If $a(\cdot,\cdot)$ is symmetric, then $E_h(\lambda_h) = E_h^\star(\lambda_h)$.

As in §8.2, the formulation (11.3a,b) can be transcribed into matrix notation.

Remark 11.6. Let $\mathbf{e}$ and $\mathbf{e}^\star$ be the coefficient vectors for $e^h = P\mathbf{e}$ and $e^{\star h} = P\mathbf{e}^\star$ (cf. (8.6)). The eigenvalue problems (11.3a,b) are equivalent to

$$\mathbf{L}\mathbf{e} = \lambda_h \mathbf{M}\mathbf{e}, \qquad \mathbf{L}^{\mathsf{H}}\mathbf{e}^\star = \overline{\lambda_h}\, \mathbf{M}\mathbf{e}^\star, \tag{11.3′}$$

where the system matrix $\mathbf{L}$ is defined as in Theorem 8.5 and the mass matrix $\mathbf{M}$ by (8.91). Since in general $\mathbf{M} \neq \mathbf{I}$, (11.3′) represents a *generalised eigenvalue problem*.

Exercise 11.7. Show that (a) $\mathbf{M}$ is positive definite and possesses a decomposition $\mathbf{M} = \mathbf{A}^{\mathsf{H}}\mathbf{A}$ (for example, $\mathbf{A}$ is the square root $\mathbf{M}^{1/2}$ or the Cholesky factor).

(b) The first problem in (11.3′) is equivalent to the ordinary eigenvalue problem $\tilde{\mathbf{L}}\,\tilde{\mathbf{e}} = \lambda_h \tilde{\mathbf{e}}$ with $\tilde{\mathbf{L}} := (\mathbf{A}^{\mathsf{H}})^{-1}\mathbf{L}\mathbf{A}^{-1}$, $\tilde{\mathbf{e}} = \mathbf{A}\,\mathbf{e}$. The second problem in (11.3′) corresponds to $\tilde{\mathbf{L}}^{\mathsf{H}}\,\tilde{\mathbf{e}}^{\star} = \overline{\lambda_h}\,\tilde{\mathbf{e}}^{\star}$ with $\tilde{\mathbf{e}}^{\star} = \mathbf{A}\mathbf{e}^{\star}$.

When investigating convergence, one must watch out for the following difficulties:

(i) A uniform approximation of all the eigenvalues and eigenfunctions by discrete eigenvalues and eigenfunctions is impossible since the infinitely many eigenvalues of (11.2a) are set against the only finitely many of (11.3a). It is only possible to characterise a fixed eigenvalue λ of (11.2a) as an accumulation point of discrete eigenvalues $\{\lambda_h\}$, and to set up estimates for $\{\lambda_h\}$. For this purpose one needs a family of subspaces $\{V_h : h \in H\}$ with $H \ni h \to 0$ (cf. Remark 8.1(ii)).

(ii) If λ and λ_h are the continuous, resp. discrete, eigenvalues then $\dim E(\lambda) = \dim E_h(\lambda_h)$ need not hold. It is preferable to limit oneself to the case of *simple eigenvalues* where $\dim E(\lambda) = \dim E_h(\lambda_h) = 1$. If $\dim E(\lambda) = k > 1$, it may well be that the multiple eigenvalue λ is approximated by several, pairwise different discrete eigenvalues

$$\lambda_h^{(i)} \quad (i = 1, \ldots, k) \quad \text{with} \quad \dim E(\lambda) = \sum_{i=1}^{k} \dim E_h(\lambda_h^{(i)}).$$

The error estimates of $|\lambda - \lambda_h^{(i)}|$ are then generally worse than for simple eigenvalues. Only for the mean value $\hat{\lambda}_h := \frac{1}{k}\sum_{i=1}^{k} \lambda_h^{(i)}$ does one obtain the usual estimates (cf. Babuška–Aziz [16, page 338]).

Different from the original form (11.1) (and the discretisation by difference schemes, cf. §11.3) the discrete problem is a generalised eigenvalue problem involving the mass matrix $\mathbf{M}$. The replacement of the matrix $\mathbf{M}$ by the *diagonal* matrix $\hat{\mathbf{M}}$ with

$$\hat{\mathbf{M}}_{ii} := \sum\nolimits_k \mathbf{M}_{ik}$$

is called *mass lumping*. A comparison of the exact discrete eigenvalue problem (11.3′) with the simplified lumped problem $\mathbf{L}\mathbf{e} = \lambda_h \hat{\mathbf{M}}\mathbf{e}$ is given by Armentano–Durán [7]. Mass lumping may also be used in connection with Galerkin finite-element methods for parabolic problems (cf. Thomée [287, §15]).

Exercise 11.8. Let the eigenvalue problem be as in Example 11.4 with $a = b$ and $\lambda = 50\pi^2/a^2$. Let V_h consist of bilinear elements over a square grid. Show that to the given triple eigenvalue λ corresponds a double eigenvalue $\lambda_h^{(1)}$ and a simple eigenvalue distinct from it, $\lambda_h^{(2)}$, with $\lim_{h\to 0} \lambda_h^{(i)} = \lambda$ $(i = 1, 2)$. The same holds for linear elements of the square-grid triangulation if mass lumping is used. *Hint:* The nodal values of the discrete eigenfunctions agree with the continuous eigenfunctions $e^{1,7}$, $e^{7,1}$, and $e^{5,5}$.

11.2.2 Qualitative Convergence Results

This section concerns the question as to whether $\lambda_h \to \lambda$ and $e^h \to e$ for $h \to 0$. The rate of convergence will be discussed in §11.2.3. The basic assumptions are the following:

$$a(\cdot,\cdot) : V \times V \to \mathbb{C} \qquad \text{be } V\text{-coercive,} \tag{11.4a}$$

$$\text{let } V \subset L^2(\Omega) \quad \text{be continuously, densely, and compactly embedded,} \tag{11.4b}$$

so that the Riesz–Schauder theory is applicable (Theorem 11.2). Furthermore, let a sequence of subspaces V_{h_i} ($h_i \to 0$) be given which increasingly approximate V (cf. (8.24a)):

$$\lim_{h\to 0} \inf\{\|u - v\|_V : v \in V_h\} = 0 \qquad \text{for all } u \in V. \tag{11.4c}$$

We define

$$a_\lambda(u,v) := a(u,v) - \lambda\,(u,v)_0 : V \times V \to \mathbb{C}, \tag{11.5a}$$

$$\omega(\lambda) := \inf_{u\in V,\ \|u\|_V=1} \sup_{v\in V,\ \|v\|_V=1} |a_\lambda(u,v)|\,, \tag{11.5b}$$

$$\omega_h(\lambda) := \inf_{u\in V_h,\ \|u\|_V=1} \sup_{v\in V_h,\ \|v\|_V=1} |a_\lambda(u,v)|\,. \tag{11.5c}$$

The interpretation of the quantities $\omega(\lambda)$ and $\omega_h(\lambda)$ is given in the next exercise.

Exercise 11.9. Let L and L_h be the operators associated with $a(\cdot,\cdot) : V \times V \to \mathbb{C}$ and $a_h(\cdot,\cdot) : V_h \times V_h \to \mathbb{C}$ (cf. (8.15)). Show that

(a) If λ is not an eigenvalue we have

$$\omega(\lambda) = 1/\|\,(L - \lambda I)^{-1}\,\|_{V\leftarrow V'}, \quad \omega_h(\lambda) = 1/\|\,(L_h - \lambda I)^{-1}\,\|_{V_h\leftarrow V_h'} \tag{11.6}$$

(cf. §8.2.3.1).

(b) $\omega(\lambda)$ and $\omega_h(\lambda)$ are continuous in $\lambda \in \mathbb{C}$.

(c) If in (11.5b,c) one replaces $a_\lambda(u,v)$ by $a_\lambda(v,u)$ one obtains the variables $\omega^\star(\lambda)$ and $\omega_h^\star(\lambda)$ which correspond to the adjoint problem. The following holds: $\omega^\star(\lambda) = \omega(\lambda)$ and $\omega_h^\star(\lambda) = \omega_h(\lambda)$. *Hint*: Use Lemma 6.109.

With the aid of (11.6) and Theorem 6.107 one proves the following connection between $\omega(\lambda)$, $\omega_h(\lambda)$, and the eigenvalue problems.

Remark 11.10. Let (11.4a,b) hold. λ is an eigenvalue of (11.2a) if and only if $\omega(\lambda) = 0$, and λ_h is eigenvalue of (11.3a) if and only if $\omega_h(\lambda_h) = 0$.

Lemma 11.11. *Let $a(\cdot,\cdot)$ be V-coercive (cf. (11.4a)). Then there exists a $\mu \in \mathbb{R}$ such that $a_\mu(\cdot,\cdot)$ is V-elliptic. In addition, then $\omega(\mu) \geq C_E$ and $\omega_h(\mu) \geq C_E$ with $C_E > 0$ from Definition 6.105.*

Proof. Coercivity implies $a(x,x) \geq C_E \|x\|_V^2 - C_K(x,x)_0$ for all $x \in V$. The inequality $\omega(\mu) \geq \inf a_\mu(x,x)/\|x\|_V^2 \geq C_E$ holds for $\mu := -C_K$. Restriction to V_h shows that $\omega_h(\mu) \geq C_E$. ∎

Lemma 11.12. *Let $\Lambda \subset \mathbb{C}$ be compact. Let (11.4a–c) hold. Then there exist numbers $C > 0$ and $\eta(h) > 0$, independent of $\lambda \in \Lambda$ with $\lim \eta(h) = 0$ such that*

$$\omega_h(\lambda) \geq C\omega(\lambda) - \eta(h), \qquad \omega(\lambda) \geq C\omega_h(\lambda) - \eta(h). \tag{11.7}$$

Proof. Let the operators $Z = Z(\lambda) : V \to V$ and $Z_h = Z_h(\lambda) : V \to V_h$ be defined as follows:

$$z := Z(\lambda)u \in V \text{ is the solution of } a_\mu(z,v) = (\lambda - \mu)\,(u,v)_0 \quad \text{for } v \in V, \tag{11.8a}$$

$$z^h := Z_h(\lambda)u \in V \text{ is the solution of } a_\mu(z^h,v) = (\lambda - \mu)\,(u,v)_0 \quad \text{for } v \in V_h, \tag{11.8b}$$

where μ is chosen according to Lemma 11.11. It shows that there is a C_Z with

$$\|Z(\lambda)\|_{V \leftarrow V'} \leq C_Z, \qquad \|Z_h(\lambda)\|_{V \leftarrow V'} \leq C_Z \qquad \text{for all } \lambda \in \Lambda. \tag{11.8c}$$

From

$$\begin{aligned} a_\lambda(u,v) = a(u,v) \,-\, \lambda\,(u,v)_0 &= a(u,v) - \mu\,(u,v)_0 - (\lambda - \mu)\,(u,v)_0 \\ &= a_\mu(u,v) - a_\mu(z,v) \,=\, a_\mu(u-z,v) \end{aligned}$$

with $z := Z(\lambda)u$ and from the definition of $\omega(\lambda)$ one concludes that

$$\omega(\lambda)\,\|u\|_V \leq \sup_{v \in V,\ \|v\|_V = 1} |a_\lambda(u,v)| = \sup_{v \in V,\ \|v\|_V = 1} |a_\mu(u-z,v)| \leq C_S\,\|u-z\|_V \tag{11.8d}$$

with $C_S := \|L - \mu I\|_{V' \leftarrow V}$. For arbitrary $u, v \in V_h$ one infers from the analogous equation $a_\lambda(u,v) = a_\mu(u - z^h, v)$ for $z^h = Z_h(\lambda)u$, Lemma 11.11, and (11.8d) that

$$\begin{aligned} \sup_{v \in V_h,\ \|v\|_V = 1} |a_\lambda(u,v)| = \sup_{v \in V_h,\ \|v\|_V = 1} \left|a_\lambda(u - z^h, v)\right| &\geq C_E\left\|u - z^h\right\|_V \\ &\geq C_E\left[\,\|u-z\|_V - \|z - z^h\|_V\right] \\ &\geq C_E\left[C_S^{-1}\omega(\lambda) - \|Z - Z_h\|_{V \leftarrow V}\right]\|u\|_V \qquad \text{for all } u \in V_h\,. \end{aligned}$$

This yields the first part of (11.7) with the constant $C = C_E/C_S > 0$ and $\eta(h) = C_E\|Z - Z_h\|_{V \leftarrow V}$, provided that

$$\lim_{h \to 0} \sup_{\lambda \in \Lambda} \|Z - Z_h\|_{V \leftarrow V} = 0. \tag{11.8e}$$

The proof of (11.8e) is carried out indirectly. Its negation reads: there exist $\varepsilon > 0$, $\lambda_i \in \Lambda$, $h_i \to 0$ with $\|Z(\lambda_i) - Z_{h_i}(\lambda_i)\|_{V \leftarrow V} \geq \varepsilon > 0$. Then there exist $u_i \in V$ with

$$\|u_i\|_V = 1, \qquad \|[Z(\lambda_i) - Z_h(\lambda_i)]\, u_i\|_V \geq \varepsilon/2 > 0. \tag{11.8f}$$

Due to (11.4b) and the compactness of Λ there exists a subsequence $\lambda_j \in \Lambda$, $u_j \in V$ with $\lim \lambda_j = \lambda^\star$, $\lim u_j = u^\star \in V'$ (convergence in V'). (11.4c) and Theorem 8.24 show $\big\|[Z(\lambda^\star) - Z_{h_j}(\lambda^\star)]u^\star\big\|_V \to 0$. Together with (11.8c) we obtain

$$\begin{aligned}
&\big\|\big[Z(\lambda_j) - Z_{h_j}(\lambda_j)\big]\, u_j\big\|_V \\
&\leq \|[Z(\lambda_j) - Z(\lambda^*)]\, u_j\|_V + \big\|\big[Z_{h_j}(\lambda^*) - Z_{h_j}(\lambda_j)\big]\, u_j\big\|_V \\
&\quad + \|Z(\lambda^*)\,[u_j - u^*]\|_V + \big\|Z_{h_j}(\lambda^*)\,[u_j - u^*]\big\|_V + \big\|\big[Z(\lambda^*) - Z_{h_j}(\lambda^*)\big]\, u^*\big\|_V \\
&\leq 2C\,|\lambda_j - \lambda^*| + 2C_Z\,\|u_j - u^*\|_{V'} + \big\|\big[Z(\lambda^*) - Z_{h_j}(\lambda^*)\big]\, u^*\big\|_V \to 0
\end{aligned}$$

in contradiction to (11.8f).

For the proof of the second part of (11.7) one replaces (11.8d) and the following estimate by

$$\omega_h(\lambda)\|u^h\|_V \leq \sup_{v \in V_h,\ \|v\|_V = 1} \big|a_\mu(u^h - z^h, v)\big| \leq C_S \|u^h - z^h\|_V \text{ for all } u^h \in V_h$$

and

$$\begin{aligned}
\sup_{v \in V_h,\ \|v\|_V = 1} \big|a_\lambda(u^h, v)\big| &= \sup_{v \in V_h,\ \|v\|_V = 1} \big|a_\mu(u^h - z, v)\big| \geq C_E \big\|u^h - z\big\|_V \\
&\geq C_E\big[C_S^{-1}\omega_h(\lambda) - \|Z(\lambda) - Z_h(\lambda)\|_{V \leftarrow V}\big]\, \big\|u^h\big\|_V \qquad \text{for all } u^h \in V_h.
\end{aligned}$$

Let $u \in V$ with $\|u\|_V = 1$ be selected such that

$$\sup_{v \in V,\ \|v\|_V = 1} |a_\lambda(u, v)| = \inf_{u \in V,\ \|u\|_V = 1} \sup_{v \in V,\ \|v\|_V = 1} |a_\lambda(u, v)| = \omega(\lambda).$$

Since $\sup |a_\lambda(u - u^h, v)| \leq C_S \|u - u^h\|_V$, it follows for arbitrary $u^h \in V_h$ that

$$\omega(\lambda) \geq C\omega_h(\lambda) - \|Z(\lambda) - Z_h(\lambda)\|_{V \leftarrow V} - C_S\, \big\|u - u^h\big\|_V\,.$$

From (11.4c) and (11.8e) follows the second part of (11.7). ∎

A corollary to Lemma 11.12 is Theorem 8.29. If problem (8.1) for all $f \in V'$ is solvable, then $\lambda = 0$ cannot be an eigenvalue, i.e., $\omega(0) > 0$. Thus it follows for ε_N in (8.23) that

$$\varepsilon_N \geq \omega_N(0) \geq \varepsilon := \frac{1}{2} C\omega(\lambda) > 0$$

for sufficiently large $N \in \mathbb{N}'$.

A second corollary concerns the convergence of the discrete eigenvalues.

Theorem 11.13. *Let (11.4a–c) hold. If λ_h $(h \to 0)$ are discrete eigenvalues of (11.3a) with $\lambda_h \to \lambda_0$ then λ_0 is an eigenvalue of (11.2a).*

Proof. If λ_0 were not an eigenvalue then $\omega(\lambda) \geq \eta_0 > 0$ would be in the ε-neighbourhood of $K_\varepsilon(\lambda_0)$, since $\omega(\lambda)$ is continuous (cf. Exercise 11.9b). There would exist $h_0 > 0$ such that $\eta(h) \leq C\eta_0/2$ for all $h \leq h_0$ (C and $\eta(h)$ from (11.7)). For all $\lambda_h \in K_\varepsilon(\lambda_0)$ with $h \leq h_0$ the contradiction follows from (11.7):

$$0 = \omega_h(\lambda_{h_i}) \geq C\omega(\lambda_h) - \eta(h) \geq C\eta_0 - \tfrac{1}{2}C\eta_0 = \tfrac{1}{2}C\eta_0 > 0. \qquad \blacksquare$$

Lemma 11.14 (minimum principle). *Let (11.4a,b) hold. The functions $\omega(\lambda)$ and $\omega_h(\lambda)$ in the interior of $\Lambda \subset \mathbb{C}$ have no proper positive minimum.*

Proof. Let L be the operator associated with $a(\cdot,\cdot)$. Let $\lambda^\star$, with $\omega(\lambda^\star) > 0$, be an arbitrary point in the interior of Λ. Since $\omega(\lambda)$ is continuous (cf. Exercise 11.9b), for sufficiently small $\varepsilon > 0$ we have $K_\varepsilon(\lambda^\star) \subset \Lambda$ and $\omega(\lambda) > 0$ in $K_\varepsilon(\lambda^\star)$. Thus $(L - \lambda I)^{-1}$ is defined in $K_\varepsilon(\lambda^\star)$ and holomorphic. For arbitrary but fixed $u, v \in V'$ the function $\Omega(\lambda) := \left\langle (L - \lambda I)^{-1} u, v \right\rangle_{V \times V'}$ is also holomorphic in Λ. Since absolute values of holomorphic functions attain their maximum on the boundary, we have $|\Omega(\lambda)| \leq \max_{\zeta \in \partial K_\varepsilon(\lambda^\star)} |\Omega(\zeta)|$. For $\|u\|_{V'} = \|v\|_{V'} = 1$ the inequality $|\Omega(\zeta)| \leq \| (L - \zeta I)^{-1} \|_{V \leftarrow V'}$ holds and implies

$$|\Omega(\lambda)| \leq \max_{\zeta \in \partial K_\varepsilon(\lambda^\star)} \| (L - \zeta I)^{-1} \|_{V \leftarrow V'} = \max_{\zeta \in \partial K_\varepsilon(\lambda^\star)} \frac{1}{\omega(\zeta)}.$$

Since $\frac{1}{\omega(\lambda)} = \| (L - \lambda I)^{-1} \|_{V \leftarrow V'}$ is the infimum of $|\Omega(\lambda)|$ over all $u, v \in V'$ with $\|u\|_{V'} = \|v\|_{V'} = 1$ we obtain

$$\frac{1}{\omega(\lambda)} = \|(L - \lambda I)^{-1}\|_{V \leftarrow V'} \leq \max_{\zeta \in \partial K_\varepsilon(\lambda^\star)} \| (L - \zeta I)^{-1} \|_{V \leftarrow V'} = \max_{\zeta \in \partial K_\varepsilon(\lambda^\star)} \frac{1}{\omega(\zeta)},$$

i.e., $\omega(\lambda) \geq \min\{\omega(\zeta) : \zeta \in \partial K_\varepsilon(\lambda)\}$ (cf. Exercise 11.9a). Thus, $\omega(\lambda)$ cannot assume a proper minimum in $K_\varepsilon(\lambda^\star)$. On the other hand, $\omega(\lambda)$ cannot be constant, so that a minimum in $K_\varepsilon(\lambda^\star)$ is excluded. For $\omega_h(\lambda)$ the conclusion is the same. $\blacksquare$

The converse of Theorem 11.13 is also valid.

Theorem 11.15. *Let (11.4a–c) hold. Let λ_0 be an eigenvalue of (11.2a). Then there exist discrete eigenvalues λ_h of (11.3a) (for all h) such that $\lim_{h \to 0} \lambda_h = \lambda_0$.*

Proof. Let $\varepsilon > 0$ be arbitrary, but sufficiently small. According to Theorem 11.2, λ_0 is an isolated eigenvalue: $\omega(\lambda) > 0$ for $0 < |\lambda - \lambda_0| \leq \varepsilon$ (ε sufficiently small). Since $\omega(\lambda)$ is continuous and $\partial K_\varepsilon(\lambda_0)$ is compact, we have that $\rho_\varepsilon := \min\{\omega(\lambda) : |\lambda - \lambda_0| = \varepsilon\}$ is positive. Because of (11.7) and $\omega(\lambda_0) = 0$ one obtains for sufficiently small h that $\eta(h) < \rho_\varepsilon C^2/(1 + C)$ and thus

$$\omega_h(\lambda) \geq C\omega(\lambda) - \eta(h) \geq C\rho_\varepsilon - \eta(h) > \frac{\eta(h)}{C} \geq \omega_h(\lambda_0) \quad \text{for all } \lambda \in \partial K_\varepsilon(\lambda_0).$$

Thus $\omega_h(\cdot)$ must have a proper minimum in $K_\varepsilon(\lambda_0)$. By Lemma 11.14 the minimal value is zero. Thus there exists $\lambda_h \in K_\varepsilon(\lambda_0)$ which is a discrete eigenvalue, $\omega_h(\lambda_h) = 0$. ∎

The next theorem describes the convergence of the eigenfunctions.

Theorem 11.16. *Let (11.4a–c) hold. Let $e^h \in E_h(\lambda_h)$ be discrete eigenfunctions with $\|e^h\|_V = 1$ and $\lim \lambda_h = \lambda_0$. Then there exists a subsequence e^{h_i} which converges in V to an eigenfunction $e \in E(\lambda_0)$:*

$$e \in E(\lambda_0), \qquad \|e^{h_i} - e\|_V \to 0 \quad (i \to \infty), \qquad \|e\|_V = 1.$$

Proof. The functions e^h are uniformly bounded in V. Since the subspace $V \subset L^2(\Omega)$ is compactly embedded (cf. (11.4b)), there exists a subsequence e^{h_i} which converges in $L^2(\Omega)$ to an $e \in L^2(\Omega)$:

$$\|e - e^{h_i}\|_{L^2(\Omega)} \to 0 \qquad (i \to \infty). \tag{11.9a}$$

We define $z = Z(\lambda_0)e$, $z^{h_i} = Z_{h_i}(\lambda_0)e$ according to (11.8a,b). According to Theorem 8.24 there exists an $h_1(\varepsilon) > 0$ such that

$$\|z - z^{h_i}\|_V \leq \varepsilon/2 \qquad \text{for } h_i \leq h_1(\varepsilon). \tag{11.9b}$$

The function e^{h_i} is a solution of

$$a_\mu(e^{h_i}, v) = (\lambda_{h_i} - \mu)\left(e^{h_i}, v\right)_0 \qquad \text{for all } v \in V_{h_i}. \tag{11.9c}$$

A combination of $z^{h_i} = Z_{h_j}(\lambda_0)e$ [i.e., (11.8b) for $\lambda = \lambda_0$] and (11.9c) yields

$$a_\mu(z^{h_i} - e^{h_i}, v) = F_i(v) := (\lambda_0 - \mu)\left(e - e^{h_i}, v\right)_0 - (\lambda_{h_i} - \lambda_0)\left(e^{h_i}, v\right)_0$$

for all $v \in V_{h_i}$. Since $\|F_i\|_{V'} \to 0$ because $\lambda_{h_i} \to \lambda_0$ and (11.9a), there exists an $h_2(\varepsilon) > 0$ such that $\|F_i\|_{V'} \leq C_E\,\varepsilon/2$ (C_E from Lemma 11.11) and

$$\|z^{h_i} - e^{h_i}\|_V \leq \varepsilon/2 \qquad \text{for } h_i \leq h_2(\varepsilon). \tag{11.9d}$$

The inequalities (11.9b,e) show that $\|z - e^{h_i}\|_V \leq \varepsilon$ for $h_i \leq \min\{h_1(\varepsilon), h_2(\varepsilon)\}$; thus $\lim_{i\to\infty} e^{h_i} = z$ in V. Therefore $\lim e^{h_i} = z$ in $L^2(\Omega) \subset V$ in $L^2(\Omega) \subset V$ also holds. (11.9a) proves $z = e \in V$ such that $e = z = Z(\lambda_0)e$ becomes $a(e, v) = \lambda_0(e, v)_0$. Therefore $e = \lim e^{h_i}$ is an eigenfunction of (11.2a). In particular, $\|e\|_V = \lim \|e^{h_i}\|_V = 1$. ∎

The selection of a *subsequence* e^{h_i} may have two reasons. If $\dim E(\lambda_0) > 1$, some elements e^{h_i} of the sequence may convergence to one eigenfunction in $E(\lambda_0)$

while other elements converges to another eigenfunction. Even if $\dim E(\lambda_0) = 1$, the eigenvector e^{h_i} may be scaled differently (the normalisation by $\|e^{h_i}\|_V = 1$ does not help since there is more than one, possibly complex, factor of modulus 1). However, after a suitable scaling the complete sequences is converging (cf. Exercise 11.17b).

Exercise 11.17. Let (11.4a–c) hold. Let $\lambda_h, e^h, \lambda_0$, and e be as in Theorem 11.16. Show that

(a) If $\dim E(\lambda_0) = 1$ then also $\lim_{h\to 0} \dim E_h(\lambda_h) = 1$.

(b) Let $\dim E(\lambda_0) = 1$. Then we have $\lim \hat{e}^h = e$ in V for $\hat{e}^h := \frac{1}{(e^h,e)_V} e^h$ if $|(e^h, e)_V| \geq 1/2$ and $\hat{e}^h := e^h$ otherwise.

11.2.3 Quantitative Convergence Results

The geometric and algebraic multiplicities of an eigenvalue λ_0 of (11.2a) agree if[1]

$$\dim \ker(L - \lambda_0 I) = \dim \ker\left((L - \lambda_0 I)^2\right). \tag{11.10}$$

Lemma 11.18. *Let (11.4a,b) and* $\dim E(\lambda_0) = 1$ *hold. Then (11.10) is equivalent to* $(e, e^\star)_0 \neq 0$ *for* $e \in E(\lambda_0)\backslash\{0\},\ 0 \neq e^\star \in E^\star(\lambda_0)\backslash\{0\}$.

Proof. We have $\ker(L - \lambda_0 I) = E(\lambda_0) = \operatorname{span}\{e\}$. $\ker(L - \lambda_0 I)^2 > 1$ holds if and only if there exists a solution $v \in V$ for $(L - \lambda_0 I)v = e$. According to Theorem 6.107c this equation has a solution if and only if $(e, e^\star)_0 = 0$. Hence (11.10) is equivalent to $(e, e^\star)_0 \neq 0$. ∎

Let $E(\lambda_0) = \operatorname{span}\{e\}$, $E^\star(\lambda_0) = \operatorname{span}\{e^\star\}$. Under the assumption (11.10) e and $e^\star$ can be normalised by

$$(e, e^\star)_0 = 1.$$

We define $\hat{V} := \{v \in V : (v, e^\star)_0 = 0\}$, $\hat{V}' = \{v' \in V' : (v', e^\star)_0 = 0\}$. Let $\|\cdot\|_{\hat{V}'}$ be the dual norm for $\|\cdot\|_{\hat{V}} = \|\cdot\|_V$. For problem (11.11):

$$\text{for } f \in \hat{V}' \quad \text{find } u \in \hat{V} \quad \text{with} \quad a_\lambda(u, v) = (f, v)_0 \quad \text{for all } v \in \hat{V}, \tag{11.11}$$

[1] The multiplicities do not coincide if the Jordan normal form of $L - \lambda_0 I$ contains a proper $k \times k$-Jordan block $J = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ & & & 0 \end{bmatrix}$, i.e., if $k \geq 2$. Since $\dim \ker(J) = 1 < \dim \ker(J^2) = 2$, the statement follows.

one defines the variable corresponding to (11.5b):

$$\hat{\omega}(\lambda) := \inf_{u\in\hat{V},\ \|u\|_V=1} \sup_{v\in\hat{V},\ \|v\|_V=1} |a_\lambda(u,v)|\,. \tag{11.12}$$

Lemma 11.19. *Let (11.4a,b), (11.10), and* $\dim E(\lambda_0) = 1$ *hold. Then there exists an* $\varepsilon > 0$ *such that*

$$\hat{\omega}(\lambda) \geq C > 0 \text{ for all } |\lambda - \lambda_0| \leq \varepsilon.$$

Problem (11.11) has exactly one solution $u \in \hat{V}$ *with*

$$\|u\|_V \leq \|f\|_{V'} / \hat{\omega}(\lambda).$$

Proof. Let $\hat{L} : \hat{V} \to \hat{V}'$ be the operator associated with $a(\cdot,\cdot) : \hat{V} \times \hat{V} \to \mathbb{C}$. For $0 < |\lambda - \lambda_0| < \varepsilon$ (ε sufficiently small) $(L - \lambda I)u = f$ has a unique solution $u \in V$. From $f \in \hat{V}'$ follows

$$0 = (f, e^\star)_0 = ([L - \lambda I]\, u, e^\star)_0 = (u, [L - \lambda I]^\star\, e^\star)_0 = (\lambda_0 - \lambda)\,(u, e^\star)_0,$$

i.e., $u \in \hat{V}$. Thus there exists $(\hat{L} - \lambda I)^{-1} : \hat{V}' \to \hat{V}$ as the restriction of $(L - \lambda I)^{-1}$ to $\hat{V}' \subset V'$. For $\lambda = \lambda_0$ problem (11.11) has a unique solution according to Theorem 6.107c. Then there exists $(\hat{L} - \lambda I)^{-1}$ for all $\lambda \in \overline{K_\varepsilon(\lambda_0)}$. According to Remark 11.10 (with $\hat{V}$ instead of V), $\hat{\omega}(\lambda)$ must be positive in $\overline{K_\varepsilon(\lambda_0)}$. The continuity of $\hat{\omega}(\lambda)$ proves $\hat{\omega}(\lambda) \geq C > 0$. In analogy to (11.6) one has

$$\|u\|_V = \|u\|_{\hat{V}} \leq \|f\|_{\hat{V}'} / \hat{\omega}(\lambda).$$

The bound by $\|f\|_{V'} / \hat{\omega}(\lambda)$ results from the next exercise. ∎

Exercise 11.20. Show that $\|f\|_{\hat{V}'} \leq \|f\|_{V'}$ for all $f \in \hat{V}'$.

Lemma 11.21. *Let (11.4a–c),* $\dim E(\lambda_0) = 1$*, and condition (11.10) hold. Let* λ_h *be discrete eigenvalues with* $\lim \lambda_h = \lambda_0$*. According to Exercise 11.17b there exists an* $e^h \in E_h(\lambda_h)$ *and* $e^{\star h} \in E_h^\star(\lambda_h)$ *with*

$$e^h \to e \in E(\lambda_0), \quad e^{\star h} \to e^\star \in E^\star(\lambda_0), \qquad (e, e^\star)_0 = 1.$$

This enables one to construct the space

$$\hat{V}_h := \{v^h \in V_h : (v^h, e^{*h})_0 = 0\}$$

and the variable

$$\hat{\omega}_h(\lambda) := \inf_{u\in\hat{V}_h,\ \|u\|_V=1} \sup_{v\in\hat{V}_h,\ \|v\|_V=1} |a_\lambda(u,v)|\,.$$

Then there exists a constant $C > 0$ *independent of* h *and* $\lambda \in \mathbb{C}$ *such that* $\hat{\omega}_h(\lambda) \geq C\omega_h(\lambda)$*. For sufficiently small* $\varepsilon > 0$ *and* h*,* $\hat{\omega}_h(\lambda) \geq \eta > 0$ *for all* λ *with* $|\lambda - \lambda_0| \leq \varepsilon$*.*

Proof. (i) First statement: there exists $h_0 > 0$ and $\hat{C} > 0$ such that

$$\min\left\{\|v + \alpha e^{\star h}\|_V : \alpha \in \mathbb{C}\right\} \geq \hat{C}\,\|v\|_V \quad \text{for all } v \in \hat{V}_h \text{ and all } H \ni h \leq h_0. \quad (*)$$

The proof is carried out indirectly. The negation reads: there is a sequence $\alpha_i \in \mathbb{C}$, $h_i \to 0$, $v_i \in \hat{V}_{h_i}$ with $\|v_i\|_V = 1$ and $\|v_i + \alpha_i e^{\star h_i}\|_V \to 0$. Thus there exist subsequences with $\alpha_i \to \alpha^\star$, $v_i \to v^\star$ in $L^2(\Omega)$. Evidently, $w_i := v_i + \alpha_i e^{\star h_i}$ must have the limit $w^\star = \lim w_i = v^\star + \alpha^\star e^\star$ in $L^2(\Omega)$. Since

$$\|w^\star\|_{L^2(\Omega)} = \lim \|w_i\|_{L^2(\Omega)} \leq C \lim \|w_i\|_V = 0,$$

it follows that $w^\star = 0$, and thus $v^\star = -\alpha^\star e^\star$. From $0 = \lim(v_i, e^{\star h_i})_0 = (v^\star, e^\star)_0 = -\alpha^\star \|e^\star\|_0^2$ one infers $\alpha^\star = 0$. Thus the contradiction follows from

$$1 = \lim \|v_i\|_V \leq \limsup \|w_i\|_V + \limsup \|\alpha_i e^{\star h_i}\|_V = \lim \|w_i\|_V = 0.$$

(ii) Second statement: $\hat{\omega}_h(\lambda) \geq \hat{C}\omega_h(\lambda)$ with $\hat{C}$ from (i). This follows from

$$\begin{aligned}
\hat{\omega}_h(\lambda) &\underset{(11.12)}{=} \inf_{u \in \hat{V}_h \setminus \{0\}} \sup_{v \in \hat{V}_h \setminus \{0\}} \frac{|a_\lambda(u,v)|}{\|u\|_V \, \|v\|_V} \\
&\underset{(*) \text{ and } a_\lambda(u,v) = a_\lambda(u, v+\alpha e^{\star h}) \text{ for } u \in \hat{V}_h}{\geq} \hat{C} \inf_{u \in \hat{V}_h \setminus \{0\}} \sup_{v \in \hat{V}_h \setminus \{0\}} \max_{\alpha \in \mathbb{C}} \frac{\left|a_\lambda(u, v + \alpha e^{\star h})\right|}{\|u\|_V \, \|v + \alpha e^{\star h}\|_V} \\
&\underset{V_h = \{v + \alpha e^{\star h} : v \in \hat{V}_h, \alpha \in \mathbb{C}\}}{=} \hat{C} \inf_{u \in \hat{V}_h \setminus \{0\}} \sup_{w \in V_h \setminus \{0\}} \frac{|a_\lambda(u,w)|}{\|u\|_V \, \|w\|_V} \\
&\underset{\hat{V}_h \subset V_h}{\geq} \hat{C} \inf_{u \in V_h \setminus \{0\}} \sup_{w \in V_h \setminus \{0\}} \frac{|a_\lambda(u,w)|}{\|u\|_V \, \|w\|_V} = \hat{C}\omega_h(\lambda).
\end{aligned}$$

(iii) Let $\varepsilon > 0$ be chosen such that λ_0 is the only eigenvalue in $\overline{K_\varepsilon(\lambda_0)}$. For sufficiently small h, λ_h is the only discrete eigenvalue in $\overline{K_\varepsilon(\lambda_0)}$. In the proof of Theorem 11.15 we have already used $\omega_h(\lambda) \geq \eta' > 0$ for $\lambda \in \partial K(\lambda_0)$, $h \leq h_0(\varepsilon)$. From part (ii) follows that $\hat{\omega}_h(\lambda) \geq \eta := \eta' C > 0$ for $\lambda \in \partial K_\varepsilon(\lambda_0)$. According to Exercise 11.17a, $a_\lambda(u,v) = (f,v)_0$ $(v \in V_h)$ is solvable for each $f \in V_h$ and all $\lambda \in K_\varepsilon(\lambda_0)$ such that $\hat{\omega}_h(\lambda) = 0$ is excluded. Lemma 11.14 shows that $\hat{\omega}_h(\lambda) \geq \eta > 0$ in $K_\varepsilon(\lambda_0)$. ∎

Exercise 11.22. Let (11.4a–c) hold. Let $d(\cdot, V_h)$ be defined as in (8.21). Show that if h is sufficiently small then there exists a $u^h \in V_h$ with

$$\|u^h - u\|_V \leq 2\, d(u, V_h) \quad \text{with the side condition} \quad \left(u^h - u, v\right)_0 = 0.$$

Lemma 11.23. *Let (11.4a–c) hold. Let λ_0 be an eigenvalue with (11.10) and* $\dim E(\lambda_0) = 1$. *For sufficiently small h there exists $e^h \in E_h(\lambda_h)$ with*

$$\|e^h - e\|_V \leq C\big[\,|\lambda_0 - \lambda_h| + d(e, V_h)\,\big]\,.$$

Proof. Let $z^h := Z_h(\lambda_0)e$ be the solution of (11.8b). Since $e = Z(\lambda_0)e$, one has

$$\|e - z^h\|_V \le C_1 d(e, V_h) \tag{11.13a}$$

(cf. Theorem 8.21). Let μ be the value in Lemma 11.11. For all $v \in V_h$ we have

$$\begin{aligned} a_\mu(z^h - e^h, v) &= a_\mu(z^h, v) - a_\mu(e^h, v) \underset{(11.8b)}{=} (\lambda_0 - \mu)\,(e, v)_0 - a_\mu(e^h, v) \underset{(11.3a)}{=} \\ &= (\lambda_0 \;-\; \mu)\,(e, v)_0 - (\lambda_h - \mu)\,\big(e^h, v\big)_0 \\ &= (\lambda_0 - \lambda_h)\,(e, v)_0 + (\lambda_h - \mu)\,\big(e - z^h, v\big)_0 + (\lambda_h - \mu)\,\big(z^h - e^h, v\big)_0\,, \end{aligned}$$

so that

$$\begin{aligned} a_{\lambda_h}(z^h - e^h, v) &= a_\mu(z^h - e^h, v) - (\lambda_h - \mu)\,\big(z^h - e^h, v\big)_0 \\ &= (\lambda_0 - \lambda_h)\,(e, v)_0 + (\lambda_h - \mu)\,\big(e - z^h, v\big)_0\,. \end{aligned} \tag{11.13b}$$

From Lemma 11.18 and $e^h \to e$, $e^{\star h} \to e^\star$ (cf. Lemma 11.21) we infer $\big|(e, e^{\star h})_0\big| \ge \eta > 0$ and $\big|(e^h, e^{\star h})_0\big| \ge \eta > 0$ for sufficiently small h. Therefore it is possible to scale e^h so that $(z^h - e^h, e^{\star h})_0 = 0$. Hence (11.13b) corresponds to problem (11.11) with the replacements $\hat V \rightsquigarrow \hat V_h$, $u \rightsquigarrow z^h - e^h$, and $f \rightsquigarrow (\lambda_0 - \lambda_h)\,e + (\lambda_h - \mu)\,\big(e - z^h\big)$. Lemma 11.19 proves

$$\begin{aligned} \|z^h - e^h\|_V &\le \hat\omega_h(\lambda_h)^{-1} C\,\big[|\lambda_0 - \lambda_h| + \|e - z^h\|_V\big] \underset{(11.13a)}{\le} \\ &\le C'\,\big[\,|\lambda_0 - \lambda_h| + d(e, V_h)\big]\,. \end{aligned} \tag{11.13c}$$

$\|e^h - e\|_V \le \|e - z^h\|_V + \|z^h - e^h\|_V$ and (11.13a,c) prove the lemma. ∎

Theorem 11.24. *Let (11.4a–c) hold. Let λ_0 be an eigenvalue with (11.10) and* $\dim E(\lambda_0) = 1$. *Let $e \in E(\lambda_0)$, $e^\star \in E^\star(\lambda_0)$, $\|e\|_V = 1$, $(e, e^\star)_0 = 1$. Then there exist discrete eigenvalues λ_h $(h \in H)$ with*

$$|\lambda_0 - \lambda_h| \le C\, d(e, V_h)\, d(e^\star, V_h). \tag{11.14}$$

Proof. Choose u^h according to Exercise 11.22 such that $\|e^\star - u^h\|_V \le 2d(e^\star, V_h)$, $(e^\star - u^h, e)_0 = 0$. Discrete eigenvalues $\lambda_h \to \lambda_0$ exist by Theorem 11.15. In

$$\begin{aligned} 0 &= a_{\lambda_0}(e^h, e^*) = a_{\lambda_h}(e^h, e^*) - (\lambda_0 - \lambda_h)\,(e^h, e^*)_0 \\ &= a_{\lambda_h}(e^h, e^* - u^h) - (\lambda_0 - \lambda_h)\,(e^h, e^*)_0 \\ &= a_{\lambda_h}(e^h - e, e^* - u^h) - (\lambda_0 - \lambda_h)\,\big[(e^h, e^*)_0 - \big(e, e^* - u^h\big)_0\big] \\ &= a_{\lambda_h}(e^h - e, e^* - u^h) - (\lambda_0 - \lambda_h)\,\Big[\underbrace{(e, e^*)_0}_{=1} + \big(e^h - e, e^*\big)_0 - \underbrace{\big(e, e^* - u^h\big)_0}_{=0 \text{ by assumption}}\Big] \end{aligned}$$

we use $a_{\lambda_h}(e^h, v^h) = 0$ for all $v^h \in V_h$ (conclusion from first to second line) and $0 = a_\lambda(e, v) = a_{\lambda_h}(e, v) + (\lambda_h - \lambda_0)(e, v)_0$ for all $v \in V$ (second to third line). The above equation shows

$$\begin{aligned} |\lambda_0 - \lambda_h| &\le C\|e^h - e\|_V \|e^\star - u^h\|_V + |\lambda_0 - \lambda_h|\ \|e^h - e\|_V \|e^*\|_V \\ &\le C'\|e^h - e\|_V \left[\|e^\star - u^h\|_V + |\lambda_0 - \lambda_h|\right]. \end{aligned}$$

By Lemma 11.23 there exists an $e^h \in E_h(\lambda_h)$ such that

$$|\lambda_0 - \lambda_h| \le C'C''\big[\,|\lambda_0 - \lambda_h| + d(e, V_h)\big]\big[\,|\lambda_0 - \lambda_h| + 2d(e^*, V_h)\big].$$

Since $|\lambda_0 - \lambda_h| \to 0$, this quadratic inequality implies (11.14) with $C > 2C'C''$ for sufficiently small h. ∎

Theorem 11.25. *Under the assumptions of Theorem 11.24 there exist for* $e \in E(\lambda_0)$, $e^\star \in E^\star(\lambda_0)$ *discrete eigenfunctions* $e^h \in E_h(\lambda_h)$, $e^{\star h} \in E^\star(\lambda_h)$ *with*

$$\|e^h - e\|_V \le C\, d(e, V_h), \qquad \|e^{\star h} - e^\star\|_V \le C\, d(e^\star, V_h). \tag{11.15}$$

Proof. Insert (11.14) with $d(e^\star, V_h) \le \text{const}$ into Lemma 11.23. The second estimate in (11.15) follows analogously. ∎

In the following, let $V \subset H^1(\Omega)$. Theorem 11.5 proves

$$E(\lambda_0) \subset H^{1+s}(\Omega), \quad E^*(\lambda_0) \subset H^{1+s}(\Omega). \tag{11.16a}$$

Also, let (11.16b) hold (cf. (8.59)):

$$d(u, V_h) \le Ch^s \|u\|_{H^{1+s}(\Omega)} \quad \text{for all } u \in E(\lambda_0) \cup E^*(\lambda_0). \tag{11.16b}$$

Conclusion 11.26. *Let (11.4a,c) and (11.16a,b) hold. Let* λ_0 *be the eigenvalue with (11.10) and* $\dim E(\lambda_0) = 1$. *Then there exists* λ_h, $e^h \in E_h(\lambda_h)$, $e^{\star h} \in E^\star(\lambda_h)$ *such that*

$$|\lambda_0 - \lambda_h| \le Ch^{2s}, \quad \|e^h - e\|_V \le Ch^s, \quad \|e^{\star h} - e^\star\|_V \le Ch^s. \tag{11.17}$$

Occasionally eigenfunctions may have better regularity than is proven for ordinary boundary-value problems. For example, let $-\Delta e = \lambda e$ be in the square $\Omega = (0,1) \times (0,1)$ with $e = 0$ on Γ. First, Theorem 11.5 implies $e \in H^2(\Omega) \cap H_0^1(\Omega)$, thus $e \in C^0(\overline{\Omega})$ (cf. Theorem 6.48). Thus $e = 0$ holds in the corners of Ω. According to Example 9.29 it follows that $e \in H^s(\Omega)$ for all $s < 4$.

As in §8.5.4 one obtains better error estimates for $e - e^h$ in the L^2-norm. The proof is postponed until after Corollary 11.33.

Theorem 11.27. *Let (11.4a–c), (11.10),* $\dim E(\lambda_0) = 1$*, and (11.16a,b) with* $s = 1$ *hold. Let* $a(\cdot,\cdot)$ *and* $a^\star(\cdot,\cdot)$ *be* H^2*-regular, i.e., for*

$$f \in L^2(\Omega), \;\; a_\mu(u,v) = (f,v)_0 \;\; \textit{and} \;\; a_\mu(v,u^\star) = (v,f)_0$$

($v \in V_h$*,* μ *from Lemma 11.11) have solutions* $u,\ u^\star \in H^2(\Omega)$*. Let* $e \in E(\lambda_0)$ *and* $e^\star \in E^\star(\lambda_0)$*. Then there exist* $\lambda_h,\ e^h \in E_h(\lambda_h)$*,* $e^{\star h} \in E^\star(\lambda_h)$ *with*

$$\|e^h - e\|_{L^2(\Omega)} \le Ch^2, \qquad \|e^{\star h} - e^\star\|_{L^2(\Omega)} \le C'h^2.$$

If (11.16a,b) also hold with some $s > 1$*, one must replace* $C'h^2$ *by* $C'h^{1+s}$*.*

11.2.4 Consistent Problems

In Problem (11.11) we have already encountered a singular equation which nevertheless was solvable. Let λ_0 be the only eigenvalue in the disc $\overline{K_r(\lambda_0)}$ (this always holds for sufficiently small r). In the following we require

$$\lambda \in \overline{K_r(\lambda_0)} \quad \text{and} \quad \lambda_0 \text{ be the only eigenvalue in } \overline{K_r(\lambda_0)}. \tag{11.18}$$

The equation

$$a_\lambda(u,v) = (f,v)_0 \qquad \text{for all } v \in V \tag{11.19a}$$

is singular for $\lambda = \lambda_0$. For $\lambda \approx \lambda_0$ equation (11.19a) is ill-conditioned. In the following we are going to show that equation (11.19a) is well-defined and well-conditioned if the right-hand side f lies in the orthogonal complement of $E^\star(\lambda_0)$:

$$f \perp E^\star(\lambda_0) \qquad \text{(i.e., } (f,e^*)_0 = 0 \text{ for all } e^* \in E^\star(\lambda_0)). \tag{11.19b}$$

In the case of $\lambda = \lambda_0$, with u, $u+e$ ($e \in E(\lambda_0)$) is also the solution. The uniqueness of the solution is obtained under the conditions (11.10) and (11.19c):

$$u \perp E^\star(\lambda_0). \tag{11.19c}$$

Remark 11.28. (a) Let (11.18), (11.4a,b), and (11.10) hold. If $\lambda = \lambda_0$, problem (11.19a,b) has the solution set $u + E(\lambda_0)$ so that (11.19c) determines a unique solution. In the case of $\lambda \neq \lambda_0$ the solution is already unique and satisfies (11.19c). There exists a C independent of f and λ such that $\|u\|_V \le C\,\|f\|_V$.

(b) Let $\lambda \neq \lambda_0$ and let the functions u, f be as above. The solution of (11.19a) for $\hat{f} := f + \alpha e$ instead of f is given by $\hat{u} := u + \frac{\alpha}{\lambda_0 - \lambda}\, e$.

Proof. This follows from Lemma 11.19 in which the assumption $\dim E(\lambda_0) = 1$ is not necessary. ∎

The finite-element discretisation of equation (11.19a) reads:

$$\text{find} \quad u^h \in V_h \quad \text{with} \quad a_\lambda(u^h, v) = (f, v)_0 \quad \text{for all } v \in V_h\,. \tag{11.20}$$

In general, equation (11.20) need not be well-defined, even assuming (11.19b). For the sake of simplicity we limit ourselves in the following to simple eigenvalues: $\dim E(\lambda_0) = 1$. Equation (11.20) is replaced by (11.21a):

$$\begin{aligned} \text{find} \quad u^h \in V_h \quad \text{with} \quad & a_\lambda(u^h, v) = (f^{(h)}, v)_0 \quad \text{for all } v \in V_h && \text{(11.21a)}\\ \text{with} \quad & f^{(h)} \perp E_h^\star(\lambda_h), && \text{(11.21b)}\\ & u^h \perp E_h^\star(\lambda_h). && \text{(11.21c)} \end{aligned}$$

Concerning the practical treatment of side conditions we refer to §8.4.6.

Exercise 11.29. Let $\hat{V}_h = V_h \cap E_h^\star(\lambda_h)^\perp$ be as in Lemma 11.21. Show that (11.21a–c) is equivalent to: Find $u^h \in \hat{V}_h$ with $a_\lambda(u^h, v) = (f^{(h)}, v)_0$ for all $v \in \hat{V}_h$ with $f^{(h)} \perp E_h^\star(\lambda_h)$. The latter formulation is well-defined even without the condition $f^{(h)} \perp E_h^\star(\lambda_h)$ and yields the same solution.

Lemma 11.21 proves the next remark.

Remark 11.30. Let (11.4a–c), (11.10), $\dim E(\lambda_0) = 1$, and (11.18) hold. Then there exists an $h_0 > 0$ such that, for all $h \le h_0$ and all $\lambda \in \overline{K_r(\lambda_0)}$, the problem (11.21a,b) has a unique solution $u^h = u^h(\lambda)$ which satisfies the additional conditions (11.21c). Further there exists a C independent of h, λ, and $f^{(h)}$ such that $\|u^h\|_V \le C\|f^{(h)}\|_{V'}$.

As soon as $E_h^\star(\lambda_h) \neq E^\star(\lambda_0)$, f in (11.19b) does not necessarily satisfy (11.21b). If $e^{\star h} \in E_h^\star(\lambda_h)\backslash\{0\}$ is known, one can construct

$$f^{(h)} := Q_{*h} f := f - \frac{(f, e^{\star h})_0}{(e^{\star h}, e^{\star h})_0}\, e^{\star h}. \tag{11.22}$$

$f^{(h)}$ satisfies (11.21b), since Q_{*h} is the orthogonal projection onto $E_h^\star(\lambda_h)^\perp$.

Exercise 11.31. Besides (11.10) assume that

$$u \perp E^*(\lambda_0), \quad \dim E(\lambda_0) = \dim E_h(\lambda_h) = 1, \quad \|e^{\star h}\|_{V'} = 1, \quad (e^h, e^{\star h})_0 = 1.$$

Show that

$$\begin{aligned} d(u, V_h \cap E_h^\star(\lambda_h)^\perp) &= \inf_{V_h \ni v^h \perp E_h^\star(\lambda_h)} \left\|u - v^h\right\|_V \\ &\le C \left[d(u, V_h) + \|u\|_{L^2(\Omega)} \inf_{e^* \in E^*(\lambda_0)} \|e^{*h} - e^*\|_{L^2(\Omega)} \right] \\ &\le C \left[\, d(u, V_h) + \|u\|_V \; d(e^{*h}, E^*(\lambda_0)) \,\right]. \end{aligned}$$

Theorem 11.32. *Let (11.4a–c), (11.10),* $\dim E(\lambda_0) = \dim E_h(\lambda_h) = 1$*, and (11.18) hold. Let* h *be sufficiently small such that (following Remark 11.30) the problem (11.21a–c) is solvable. For the solutions* u *and* u^h *of (11.19a–c) and (11.21a–c) with* $f^{(h)}$ *in (11.22) the error estimate*

$$\left\|u^h - u\right\|_V \le C\left[d(u, V_h) + \|f\|_{V'}\, d(e^{*h}, E^*(\lambda_0))\right] \tag{11.23}$$

holds with C *independent of* f, $f^{(h)}$, *and* h.

Proof. The proof of Theorem 8.21 (Céa lemma) has to be modified. From (11.19a) and (11.21a) follows

$$a_\lambda(u^h - u, v) = (f^{(h)} - f, v)_0 \qquad \text{for all } v \in V_h.$$

Let $v, w \in V_h$ be arbitrary except the side conditions $\|v\|_V = 1$ and $w \perp E_h^*(\lambda_h)$. Then

$$a_\lambda(u^h - w, v) = a_\lambda([u^h - u] + [u - w], v) = (f^{(h)} - f, v)_0 + a_\lambda(u - w, v)$$

can be estimated by

$$|a_\lambda(u^h - w, v)| \le \|f^{(h)} - f\|_{V'} + C\,\|u - w\|_V\,.$$

Definition (11.22) of $f^{(h)}$ yields $\|f^{(h)} - f\|_{V'} \le C'|(f, e^{\star h})_0| = C'|(f, e^{\star h} - e^\star)_0|$ for all $e^\star \in E^*(\lambda_0)$, hence

$$\|f^{(h)} - f\|_{V'} \le C'\,\|f\|_{V'}\, d(e^{*h}, E^*(\lambda_0)).$$

Using $\hat\omega_h(\lambda) \ge \eta > 0$ $(\lambda \in \overline{K_r(\lambda_0)})$ according to Lemma 11.21), we can bound $\|u^h - w\|_V$ by

$$\frac{1}{\eta} \sup_{w \in V_h \cap E_h^*(\lambda_h), \|w\|_V = 1} |a_\lambda(u^h - w, v)| \le \frac{C'}{\eta}\|f\|_{V'}\, d(e^{*h}, E^*(\lambda_0)) + \frac{C}{\eta}\|u - w\|_V\,.$$

$\|u - w\|_V$ is treated by Exercise 11.31 and $\|u\|_V \le C\,\|f\|_{V'}$. The triangle inequality

$$\begin{aligned}\left\|u^h - u\right\|_V &\le \left\|u^h - w\right\|_V + \|u - w\|_V \\ &\le \frac{C'}{\eta}\,\|f\|_{V'}\, d(e^{*h}, E^*(\lambda_0)) + \left(1 + \frac{C}{\eta}\right)\|u - w\|_V\end{aligned}$$

yields the assertion. ∎

Corollary 11.33. If additionally the assumptions $u \in H^{1+s}(\Omega)$, (11.16a), and $d(u, V_h) \le Ch^s\,|u|_{1+s}$ hold then (11.23) yields the estimate

$$\left\|u^h - u\right\|_V \le Ch^s\,|u|_{H^{1+s}(\Omega)}\,.$$

It remains to add the proof of Theorem 11.27.

Proof of Theorem 11.27. For $e \in E(\lambda_0)$ there exists $e^h \in E_h(\lambda_h)$ with $f := e - e^h \perp E^*(\lambda_0)$ and $|f|_1 = \|f\|_V \le Ch^s = Ch$. According to Remark 11.28 the problem $a_{\lambda_0}(v,w) = (v,f)_0$ has a solution $w \perp E^\star(\lambda_0)$ for all $v \in V$. The assumption of regularity yields $w \in H^2(\Omega)$, $|w|_2 \le C\,|f|_0$ such that $w^h \in V_h$ exists with $w^h \perp E_h^\star(\lambda_h)$, $\left|w - w^h\right|_1 \le Ch\,|w|_2 \le C'h\,|f|_0$. The value

$$\begin{aligned} a_{\lambda_0}(f,w^h) &= a_{\lambda_0}(e,w^h) - a_{\lambda_0}(e^h,w^h) = 0 - a_{\lambda_0}(e^h,w^h) \\ &= (\lambda_0-\lambda_h)\,(e^h,w^h)_0 - a_{\lambda_h}(e^h,w^h) = (\lambda_0-\lambda_h)\,(e^h,w^h)_0 \end{aligned}$$

can be bounded by $Ch^2|w^h|_0|e^h|_0 \le C'h^2\,|f|_0$ (cf. (11.17)). From

$$\begin{aligned} |f|_0^2 &= (f,f)_0 = a_{\lambda_0}(f,w) = a_{\lambda_0}(f,w-w^h) + (\lambda_0-\lambda_h)\,(e^h,w^h)_0 \\ &\le C\left[C'h\,|f|_1\,|f|_0 + h^2\,|f|_0^2\right] \end{aligned}$$

and $|f|_1 \le Ch$ one infers $|f|_0 < C'h^2$. The same method is then applied to $|e^\star - e^{\star h}|_0$. ∎

11.3 Discretisation by Difference Methods

In the following we limit ourselves to the case of a difference operator of the order $2m = 2$. The differential equation $Lu = f$ with homogeneous Dirichlet boundary condition is replaced, as in Chapters 4 and 5, by the difference equation $L_h u_h = f_h$. The eigenvalue equations $Le = \lambda e$, $L^\star e^\star = \bar\lambda e^*$ are discretised by

$$L_h e_h = \lambda_h e_h, \qquad L_h^{\mathsf{H}} e_h^* = \bar\lambda_h e_h^*. \tag{11.24}$$

In the following $|\cdot|_i$ for $i = -1,0,1,2$ denotes the continuous or discrete Sobolev norm depending on the context. The general assumptions of the following analysis are:

$$V = H_0^1(\Omega), \qquad \Omega \in C^{0,1} \text{ bounded}, \tag{11.25a}$$

$$a(u,v) = (Lu,v)_0 \quad \text{is } H_0^1(\Omega)\text{-coercive}, \tag{11.25b}$$

$$|L_h R_h - \check R_h L|_{-1\leftarrow 2} \le Ch \qquad \text{(consistency condition)}. \tag{11.25c}$$

Condition (11.25c) has been discussed in §9.3.2. Assume furthermore that L_h is H_h^1-coercive. For suitable $\mu \in \mathbb{R}$

$$L_{\mu,h} := L_h - \mu I \qquad (I\text{: identity matrix})$$

is thus H_h^1-regular:

$$(L_{\mu,h} v_h, v_h)_0 \ge C_E\,|v_h|_1^2 \qquad \text{for all } v_h. \tag{11.25d}$$

Furthermore, let

$$L_\mu := L - \mu I \quad \text{be } H^2(\Omega)\text{-regular,} \tag{11.25e}$$

i.e., $|L_\mu^{-1}|_{2\leftarrow 0} \le C$. The boundedness of L and L_h reads

$$|L|_{-1\leftarrow 1} \le C, \qquad |L_h|_{-1\leftarrow 1} \le C. \tag{11.25f}$$

In (9.53a,b) we have introduced prolongations $\hat{P}_h : L_h^2 \to L^2(\mathbb{R}^2)$ and $P_h : L_h^2 \to L^2(\Omega)$. Now we need a mapping $P_h : H_h^1 \to H_0^1(\Omega)$ (cf. Footnote 9 on page 297):

$$\bar{u}(\mathbf{x}) := \begin{cases} \hat{P}_h u_h(\mathbf{x}) & \text{if } K_{h/2}(\mathbf{x}) \subset \Omega, \\ 0 & \text{otherwise,} \end{cases}$$
$$P_h u_h(\mathbf{x}) := \left(\sigma_h^x \sigma_h^y \bar{u}\right)(\mathbf{x}) \qquad (\mathbf{x} \in \Omega),$$

where $K_{h/2}(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^2 : \|\mathbf{x}-\mathbf{y}\|_\infty < h/2\}$, $\hat{P}_h$ is defined according to (9.53b), and $\sigma_h^x \sigma_h^y$ according to (9.38). Check that $P_h u_h \in H_0^1(\Omega)$ and

$$|P_h|_{1\leftarrow 1} \le C. \tag{11.25g}$$

Let R_h and $\check{R}_h$ be defined as in (9.39a,b). They satisfy

$$|R_h|_{0\leftarrow 0} \le C, \qquad |R_h|_{1\leftarrow 1} \le C, \qquad |\check{R}_h|_{0\leftarrow 0} \le C. \tag{11.25h}$$

Exercise 11.34. Show that:

$$|R_h - \check{R}_h|_{0\leftarrow 1} \le Ch, \ |I - R_h P_h|_{0\leftarrow 1} \le Ch, \ |P_h^* - R_h|_{0\leftarrow 1} \le Ch, \tag{11.25i}$$
$$|I - P_h^* P_h|_{0\leftarrow 1} \le Ch, \ |I - \check{R}_h P_h|_{0\leftarrow 1} \le Ch, \ |I - \check{R}_h^* R_h|_{0\leftarrow 1} \le Ch. \tag{11.25j}$$

The first inequality in (11.25j) is equivalent to

$$|(P_h u_h, P_h v_h)_0 - (u_h, v_h)_0| \le Ch\, |u_h|_0\, |v_h|_1 \ . \tag{11.25j*}$$

Hint: Exercise 9.46, Corollary 6.62, and Lemma 9.49.

Lemma 11.35. *Let (11.25a,c,g–i) hold. (a) It is true that*

$$\lim_{h\to 0} |u - P_h R_h u|_1 = \lim_{h\to 0} |u - \check{R}_h^* R_h u|_1 = 0 \ \textit{for all } u \in H_0^1(\Omega), \tag{11.25k}$$
$$\lim_{h\to 0} |(\check{R}_h - R_h)u|_0 = 0 \qquad \textit{for all } u \in L^2(\Omega), \tag{11.25l}$$
$$\lim_{h\to 0} |[L_{h,\lambda} R_h - \check{R}_h L_\lambda]\, u|_{-1} = 0 \qquad \textit{for all } u \in H_0^1(\Omega),\ \lambda \in \mathbb{C}. \tag{11.25m}$$

(b) If $u \in H_0^1(\Omega)$ *and* $\lim_{h\to 0} |R_h u|_0 = 0$ *then* $u = 0$.

Proof. (a) The proof of (11.25k,m) follows the same pattern, which will be demonstrated for the case of (11.25l). For $\varepsilon > 0$ one has to show $|(\check{R}_h - R_h)u| \le \varepsilon$

for $h \le h(\varepsilon)$. Since $H_0^1(\Omega)$ is dense in $L^2(\Omega)$ there exists a $\tilde{u} \in H_0^1(\Omega)$ with $|u-\tilde{u}|_0 \le \frac{\varepsilon}{2|\check{R}_h - R_h|_{0\leftarrow 0}}$ such that $|(\check{R}_h - R_h)(u-\tilde{u})|_0 \le \frac{\varepsilon}{2}$. By (11.25i) there follows $|(\check{R}_h - R_h)\tilde{u}|_0 \le Ch\,|\tilde{u}|_1 \le \frac{\varepsilon}{2}$ for $h \le h(\varepsilon) := \varepsilon/[2C\,|\tilde{u}|_1]$. Altogether this gives $|(\check{R}_h - R_h)u|_0 \le \varepsilon$.

(b) From (11.25k) one infers $0 = \lim_{h\to 0}(R_h u, P_h^* u)_0 = \lim_{h\to 0}(P_h R_h u, u)_0 = (u,u)_0$ thus $u = 0$. ∎

Exercise 11.36. Let $\Lambda_h := I - \partial_x^+\partial_x^- - \partial_y^+\partial_y^-$ and $\Lambda := I - \Delta$. Show that

$$
\begin{aligned}
&|u|_1^2 = (\Lambda u, u)_0\,, && |u_h|_1^2 = (\Lambda_h u_h, u_h)_0\,,\\
&|w|_{-1}^2 = (v,w)_0 && \text{for } w \in H^{-1}(\Omega) \text{ and } v = \Lambda^{-1}w \in H_0^1(\Omega),\\
&|w_h|_{-1}^2 = (v_h, w_h)_0 && \text{for } w_h \in H_h^{-1} \text{ and } v_h = \Lambda_h^{-1} w_h \in H_h^1,\\
&\lim_{h\to 0} |(\Lambda_h R_h - \check{R}_h \Lambda)u|_{-1} = 0 && \text{for all } u \in H_0^1(\Omega).
\end{aligned}
$$

The following analysis is tailored to the properties of the difference methods under discussion. We have tried to avoid a more abstract theory that is applicable to finite elements as well as difference methods. This kind of approach can be found in Stummel [277] and Chatelin [66].

The variable $\omega_h(\lambda)$ is now defined by

$$\omega_h(\lambda) := \inf_{|u_h|_1=1} \sup_{|v_h|_1=1} |\,(L_{\lambda,h}u_h, v_h)_0\,| \;=\; \inf_{|u_h|_1=1} |L_{\lambda,h}u_h|_{-1}.$$

As in Exercise 11.9 we have

$$\omega_h(\lambda) = \begin{cases} 0 & \text{if } \lambda \text{ is an eigenvalue of } L_h,\\ 1/|L_{\lambda,h}|_{1\leftarrow -1} & \text{otherwise.}\end{cases}$$

The analogue of Lemma 11.12 reads as follows.

Lemma 11.37. *Let $K \subset \mathbb{C}$ be compact. Let (11.25a–d,f) hold. Then there exist variables $C > 0$ and $\eta(h) \to 0$ $(h \to 0)$, independent of $\lambda \in K$, such that*

$$\omega_h(\lambda) \ge C\omega(\lambda) - \eta(h),\;\; \omega(\lambda) \ge C\omega_h(\lambda) - \eta(h) \;\text{ for all } \lambda \in K,\; h > 0. \quad (11.26)$$

Proof. (i) Since K is compact, $\omega(\lambda)$ is continuous, and $\omega_h(\lambda)$ is equicontinuous in λ, it is sufficient to show that $\underline{\lim}_{h\to 0}\,\omega_h(\lambda) \ge C\omega(\lambda)$, $\omega(\lambda) \ge C\,\overline{\lim}_{h\to 0}\,\omega_h(\lambda)$ for all $\lambda \in K$ with $C > 0$.

(ii) Define for $\lambda \in K$ and u_h with $|u_h|_1 = 1$ and $|L_{\lambda,h}u_h|_{-1} = \omega_h(\lambda)$ the variables

$$u := P_h u_h,\;\; z_h := (\lambda - \mu)\,L_{\mu,h}^{-1}u_h,\;\; z := (\lambda-\mu)\,L_\mu^{-1}u_h \;\text{ with } \mu \text{ from (11.25e).}$$

Without loss of generality it can be assumed that $\mu \notin K$. We have

$$|u-z|_1 = |P_h(u_h - z_h) + P_h z_h - z|_1 \le |P_h|_{1\leftarrow 1}\,|u_h - z_h|_1 + |P_h z_h - z|_1\,,$$
$$\begin{aligned}|P_h z_h - z|_1 &= |P_h[z_h - R_h z] - [I - P_h R_h]\,z|_1\\ &\le |P_h|_{1\leftarrow 1}\,|\lambda-\mu|\left|L_{\mu,h}^{-1}[\check{R}_h L_\mu - L_{\mu,h}R_h]L_\mu^{-1}u + L_{\mu,h}^{-1}[I-\check{R}_h P_h]u_h\right|_1\\ &\quad + |[I-P_h R_h]\,z|_1 \to 0 \quad \text{for } h\to 0\end{aligned}$$

(cf. (11.25g,j,m)) so that

$$|u_h - z_h|_1 \ge C_1\,|u-z|_1 - o(1) \quad \text{with } C_1 > 0. \tag{11.27a}$$

As in (11.8d) one obtains

$$|u-z|_1 \ge C_2\omega(\lambda)\,|u|_1 \quad \text{with } C_2 > 0. \tag{11.27b}$$

From

$$\begin{aligned}\omega_h(\lambda) &= |L_{\lambda,h}u_h|_{-1} = |L_{\mu,h}u_h + (\mu-\lambda)\,u_h|_{-1}\\ &\ge (L_{\mu,h}u_h + (\mu-\lambda)\,u_h, u_h)_0 \ge -|\mu-\lambda|\,|u_h|_0^2 + (L_{\mu,h}u_h, u_h)_0\\ &\ge -C_\mu\,|u_h|_0^2 + C_E\,|u_h|_1^2 = -C_\mu\,|u_h|_0^2 + C_E\end{aligned}$$

with $C_\mu := \max\{|\mu-\lambda| : \lambda \in K\} > 0$ follows

$$|u_h|_0^2 \ge C_\mu^{-1}\,[C_E - \omega_h(\lambda)]\,.$$

Either $\omega_h(\lambda) \ge \frac{C_E}{2}$ from which the statement results directly, or $\omega_h(\lambda) \le \frac{C_E}{2}$ which yields

$$|u_h|_0 \ge C_0 = C_0\,|u_h|_1 \qquad \text{with} \quad C_0 := \sqrt{C_E/(2C_\mu)}\,.$$

We want to show that there exists $h_0 > 0$ and $C_P = C_P(C_0)$ such that

$$|u_h|_1 \le C_P\,|P_h u_h|_1 \quad \text{for all } u_h \text{ with } |u_h|_0 \ge C_0\,|u_h|_1 \text{ and } h \le h_0\,. \tag{11.27c}$$

The negation of (11.27c) reads: there exists u_h with $|u_h|_1 = 1$, $|u_h|_0 \ge C_0$, and $|P_h u_h|_1 \to 0$ ($h \to 0$). From $|R_h P_h u_h|_0 \le |R_h P_h u_h|_1 \le C\,|P_h u_h|_1 \to 0$ and

$$|u_h - R_h P_h u_h|_0 \le |I - R_h P_h|_{0\leftarrow 1}\,|u_h|_1 \le Ch \to 0$$

follows $|u_h|_0 \to 0$ in contrast to $|u_h|_0 \ge C_0$. Thus we have (11.27c).

Together with $L_{\mu,h}(u_h - z_h) = L_{\lambda,h}u_h$, (11.25d) and (11.27a–c) yield the first of the inequalities (11.26):

$$\begin{aligned}\omega_h(\lambda) &= |L_{\lambda,h}u_h|_{-1} = |L_{\mu,h}(u_h - z_h)|_{-1} \ge C_E|u_h - z_h|_1\\ &\ge C_E C_1|u-z|_1 - o(1) \ge C_E C_1 C_2\omega(\lambda)|u|_1 - o(1) \ge C\omega(\lambda) - o(1)\end{aligned}$$

with $C := C_E C_1 C_2/C_P > 0$.

(iii) Let $\varepsilon > 0$ be arbitrary; $u \in H_0^1(\Omega)$ with $|u|_1 = 1$ can be selected such that $\omega(\lambda) \geq |L_\lambda u|_{-1} - \varepsilon$. Set $u_h := R_h u$. According to Exercise 11.36 it holds that

$$|L_\lambda u|_{-1}^2 = (v, L_\lambda u)_0 \qquad \text{for } v := \Lambda^{-1} L_\lambda u \in H_0^1(\Omega), \text{ where } \Lambda = I - \Delta,$$
$$|L_{\lambda,h} u_h|_{-1}^2 = (v_h, L_{\lambda,h} u_h)_0 \quad \text{for} \quad v_h := \Lambda_h^{-1} L_{\lambda,h} u_h.$$

From

$$R_h v - v_h = \Lambda_h^{-1} \left[\Lambda_h R_h - \check{R}_h \Lambda\right] \Lambda^{-1} L_\lambda u + \Lambda_h^{-1} \left[\check{R}_h L_\lambda - L_{\lambda,h} R_h\right] u \to 0,$$
$$\check{R}_h L_\lambda u - L_{\lambda,h} u_h = \left[\check{R}_h L_\lambda - L_{\lambda,h} R_h\right] u \to 0,$$
$$(v, L_\lambda u)_0 - \left(R_h v, \check{R}_h L_\lambda u\right)_0 = \left(\left[I - \check{R}_h^* R_h\right] v, L_\lambda u\right)_0 \to 0$$

for $h \to 0$ (cf. (11.25m,k)), one infers $|L_{\lambda,h} u_h|_{-1} \to |L_\lambda u|_{-1}$ and

$$\omega(\lambda) \geq |L_\lambda u|_{-1} - \varepsilon \geq |L_{\lambda,h} u_h|_{-1} - \varepsilon - o(1) \geq \omega_h(\lambda) - \varepsilon - o(1) \quad (h \to 0)$$

for each $\varepsilon > 0$. Thus $\omega(\lambda) \geq \overline{\lim}_{h \to 0} \omega_h(\lambda)$ has been proved. ∎

Conclusion 11.38. *Under the assumptions of Lemma 11.37 the following holds: If there exists L_λ^{-1} for all $\lambda \in K$ then there exists an $h_0 > 0$ such that $L_{\lambda,h}$ for all $\lambda \in K$ and $h \leq h_0$ is H_0^1-regular:*

$$\sup \left\{ \left|L_{\lambda,h}^{-1}\right|_{1 \leftarrow -1} : \lambda \in K,\ h \leq h_0 \right\} \leq C.$$

Proof. We have assumed $\omega(\lambda) > 0$ in K so $\max\{\omega(\lambda) : \lambda \in K\} =: \eta > 0$. Choose h_0 according to Lemma 11.37 such that $\omega_h(\lambda) \geq C\omega(\lambda) - \frac{1}{2} C\eta \geq \frac{1}{2} C\eta$ for $h \leq h_0$. Then $|L_{\lambda,h}^{-1}|_{1 \to -1} \leq 2/(C\eta)$ for all $\lambda \in K,\ h \leq h_0$. ∎

The proof of Theorem 11.13 can be carried over without change.

Theorem 11.39. *Assume (11.25a–d,f). If λ_h $(h \to 0)$ are discrete eigenvalues of problem (11.24) with $\lambda_h \to \lambda_0$ then λ_0 is an eigenvalue of (11.2a).*

Lemma 11.14 and Theorem 11.15 can also be carried over without changes.

Theorem 11.40. *Let (11.25a–d,f) hold. Let λ_0 be an eigenvalue of (11.2a). Then there exist discrete eigenvalues λ_h of (11.24) [for all h] such that $\lim_{h \to 0} \lambda_h = \lambda_0$.*

Theorem 11.41. *Let (11.25a–d,f) hold. Let e_h be discrete eigenfunctions with $|e_h|_1 = 1$ for λ_h, where $\lambda_h \to \lambda_0$ $(h \to 0)$. Then there exists a subsequence e_{h_i} such that $P_{h_i} e_{h_i}$ converges in $H_0^1(\Omega)$ to an eigenfunction $0 \neq e \in E(\lambda_0)$. Further, we have $|e_{h_i} - R_{h_i} e|_1 \to 0$.*

Proof. Because of $|P_h e_h|_1 \leq C$ (cf. (11.25g)) the functions $e^h := P_h e_h$ are uniformly bounded. $H_0^1(\Omega)$ is compactly embedded in $L^2(\Omega)$ (cf. (11.25a) and Theorem 6.86a) such that a subsequence e^{h_i} converges in $L^2(\Omega)$ to an $L^2(\Omega)$: $\left|e^{h_i} - e\right|_0 \to 0$. We have in particular

$$|R_h e - e_h|_0 \leq |R_h (e - e_h) - (R_h P_h - I) e_h|_0 \to 0 \qquad (h = h_i \to 0).$$

Estimate (11.25c) yields

$$|R_h z - z_h|_0 \le |R_h z - z_h|_1 \le Ch\,|e|_0 \to 0$$
$$\text{for} \quad z := (\lambda_0 - \mu)\, L_\lambda^{-1} e, \;\; z_h := (\lambda_0 - \mu)\, L_{\lambda,h}^{-1} \check{R}_h e.$$

From $L_{\mu,h}(z_h - e_h) = (\lambda_0 - \mu)(\check{R}_h e - e_h) + (\lambda_h - \lambda_0)e_h \to 0$ in H_h^{-1} follows $|z_h - e_h|_1 \to 0$ such that $|R_h(e - z)|_0 \to 0$ ($h = h_i \to 0$) results. Lemma 11.35b shows that $e = z \in H_0^1(\Omega)$, and $e = 0$ is excluded because of $|\lim R_{h_i} e|_1 = |\lim e_h|_1 = 1$. ∎

Theorem 11.42. *Let (11.25a–d,f) hold. Let $e_h^\star$ be the solution of the discrete eigenvalue problem $L_h^\star e_h^\star = \bar{\lambda}_h e_h^\star$ with $|e_h^\star|_1 = 1$, $\lim_{h\to 0} \lambda_h = \lambda_0$. Then there exists a subsequence $e_{h_i}^\star$ such that $P_{h_i} e_{h_i}^\star$ in $H_0^1(\Omega)$ converges to an eigenfunction $0 \neq e^\star \in E^\star(\lambda_0)$. Furthermore, $|e_{h_i}^\star - R_{h_i} e^\star|_1 \to 0$.*

Proof Sketch. The proof is not analogous to that of Theorem 11.41 since the consistency condition (11.25a,m) does not necessarily imply the corresponding statements for the adjoint operators. One has to carry out the following steps:

(i) $e^{\star h} := P_h e_h^\star \to e^\star$ converges in $L^2(\Omega)$ for a subsequence $h = h_i \to 0$.

(ii) For $z = \overline{(\lambda_0 - \mu)}\, L_\mu^{\star -1} e^\star$ and $z_h := \overline{(\lambda_0 - \mu)}\, L_{\mu,h}^{\star -1} e_h^\star$ the following holds:

$$z - \check{R}_h^* z_h = \overline{(\lambda_0 - \mu)}\, L_\mu^{\star -1} \left[(e^\star - R_h^* e_h^\star) + \left(R_h^* L_{\mu,h}^* - L_\mu^* \check{R}_h^* \right) L_{\mu,h}^{\star -1} e_h^\star \right].$$

For each $v \in L^2(\Omega)$ we obtain

$$\left(v, z - \check{R}_h^* z_h\right)_0 = (\lambda_0 - \mu) \left\{ \begin{array}{l} \left(L_\mu^{-1} v, e^\star - e_h^\star\right)_0 + \left(\left[P_h^* - R_h\right] L_\mu^{-1} v, e_h^\star \right)_0 \\ + \left(L_{\mu,h}^{-1} \left[L_{\mu,h} R_h - \check{R}_h L_\mu \right] L_\mu^{-1} v, e_h^\star \right)_0 \end{array} \right\} \to 0$$

for $h = h_i \to 0$ (cf. (11.25i,c)).

(iii) $|z_h - e_h^\star|_1 \to 0$ may be inferred from $L_{\mu,h}^\star (z_h - e_h^\star) = \overline{(\lambda_0 - \lambda_h)} e_h^\star \to 0$. In particular, $(v, \check{R}_h^\star z_h - R_h^\star e_h^\star)_0 \to 0$ for each $v \in L^2(\Omega)$.

(iv) $(v, \check{R}_h^\star e_h^\star - R_h^\star e_h^\star)_0 = ([\check{R}_h - R_h]v, e_h^\star)_0 \to 0$ for each $v \in L^2(\Omega)$ (cf. (11.25l)).

(v) $(v, R_h^\star e_h^\star - e^\star)_0 \to 0$ for each $h = h_i \to 0$.

(vi) From (ii) and (v) follows $(v, z - e^\star)_0 = 0$, thus $z = e^\star \in E^\star(\lambda_0)$, where $e^\star \neq 0$. ∎

Exercise 11.43. Carry over Exercise 11.17 to the present situation.

In Lemma 11.21 $\hat{\omega}_h(\lambda)$ is defined. Now set

$$\hat{V}_h := \{u_h : (u_h, e_h^*)_0 = 0\} = \{e_h^*\}^\perp, \qquad L_h^* e_h^\star = \bar{\lambda}_h e_h^\star, \qquad \lambda_h \to \lambda_0,$$
$$\hat{\omega}_h(\lambda) := \inf_{0 \neq u_h \in \hat{V}_h} \sup_{0 \neq v_h \in \hat{V}_h} \left| (L_{\lambda,h} u_h, v_h)_0 \,/\, (|u_h|_1 \, |v_h|_1) \right|.$$

A basic condition for the following is

$$\dim E_h(\lambda_h) = 1, \quad E_h(\lambda_h) = \operatorname{span}\{e_h\}, \quad E_h^*(\lambda_h) = \operatorname{span}\{e_h^*\}. \tag{11.28}$$

Here

$$E_h(\lambda_h) := \left\{u_h \in H_h^1 : L_h u_h = \lambda_h u_h\right\}, \ E_h^*(\lambda_h) := \left\{u_h \in H_h^1 : L_h^* u_h = \bar{\lambda}_h u_h\right\}.$$

Exercise 11.43 shows that (11.28) is valid for $h \le h_0$ if $\dim E(\lambda_0) = 1$.

Lemma 11.44. *Let (11.25a–d,f),* $\dim E(\lambda_0) = 1$*, and (11.10) hold. Then there exist* $h_0 > 0$ *and a* $C > 0$ *that is independent of* $h \le h_0$ *and* $\lambda \in \mathbb{C}$ *such that* $\hat{\omega}_h(\lambda) \ge C\omega_h(\lambda)$ *for* $h \le h_0$*. If* $\varepsilon > 0$ *and* h *are sufficiently small then* $\hat{\omega}_h(\lambda) \ge \eta > 0$ *for all* $|\lambda - \lambda_0| \le \varepsilon$*.*

Proof. (i) There exists a $C > 0$ such that

$$|v_h + \alpha e_h^*|_1 \ge |v_h|_1 / C \qquad \text{for all } v_h \in \hat{V}_h,\ \alpha \in \mathbb{C},\ h > 0.$$

For fixed h the quotient space norm $|||v_h|||_1 := \inf\{|v_h + \alpha e_h^*|_1 : \alpha \in \mathbb{C}\}$ and $|\cdot|_1$ are two norms on $\hat{V}_h$. Because of the equivalence of norms in finite-dimensional vector spaces, the above inequality holds with $C = C(h)$ possibly depending on h. It remains to investigate the behaviour of C for $h \to 0$.

Indirect proof: Assume that there exists a sequence v_{h_i} with $h_i \to 0$, $|v_{h_i}|_1 = 1$, $\alpha_i \in \mathbb{C}$, $w_{h_i} := v_{h_i} + \alpha_i e_{h_i}^\star$, $|w_{h_i}|_1 \to 0$. For a subsequence of $\{h_i\}$

$$\begin{aligned} &\alpha_i \to \alpha^*, \quad P_{h_i} v_{h_i} \to v^*, \quad \text{and} \quad P_{h_i} e_{h_i}^* \to e^* \ne 0 \quad \text{in } L^2(\Omega), \\ &P_{h_i} w_{h_i} \to w^* := v^* + \alpha^* e^* = 0 \end{aligned}$$

converge. From

$$0 = \left(v_{h_i}, e_{h_i}^*\right)_0 = \left(v_{h_i}, \left[I - P_{h_i}^* P_{h_i}\right] e_{h_i}^*\right)_0 + \left(P_{h_i} v_{h_i}, P_{h_i} e_{h_i}^*\right)_0 \to (v, e^*)_0$$

one infers $(v^\star, e^\star)_0 = 0$, $\alpha^\star = (w^\star, e^\star)_0 / (e^\star, e^\star)_0 = 0$, $v^\star = 0$. The contradiction results from $1 = \lim |v_{h_i}|_1 = \lim |w_{h_i}|_1 = 0$.

(b) The rest of the proof runs as in Lemma 11.21. ∎

Lemma 11.45. *Let (11.25a–m),* $\dim E(\lambda_0) = 1$*, and (11.10) hold. One may choose* $0 \ne e \in E(\lambda_0)$ *and* $e_h \in E_h(\lambda_h)$ *so that*

$$|R_h e - e_h|_1 \le C\left[h + |\lambda_0 - \lambda_h|\right] \qquad \textit{for all } h > 0.$$

Proof. We have that $e = (\lambda_0 - \mu) L_\mu^{-1} e \in H^2(\Omega) \cap H_0^1(\Omega)$. Assume also that $z_h := (\lambda_0 - \mu) L_{\mu,h}^{-1} \check{R}_h e$. The inequality (9.49) implies $|R_h e - z_h|_1 \le Ch\,|e|_2$. For sufficiently small h we have $(e_h, e_h^\star) \ne 0$ so that one can scale e_h such that $(e_h - z_h, e_h^\star)_0 = 0$. From

$$\begin{aligned} L_{\lambda,h}\,(e_h - z_h) &= (\lambda_h - \lambda_0)\,\check{R}_h e + (\lambda_h - \mu)\,(z_h - \check{R}_h e) \\ &= (\lambda_h - \lambda_0)\,\check{R}_h e + (\lambda_h - \mu)\,\big[(z_h - R_h e) + (R_h - \check{R})e\big] \end{aligned}$$

one infers that $|e_h - z_h|_1 \le C\big[\,|\lambda_h - \lambda_0| + h\big]$ (cf. Lemma 11.44) since

$$|(R_h - \check{R}_h)e|_{-1} = \mathcal{O}(h).$$

The statement follows from this. ∎

Lemma 11.46. *Under the assumptions of Lemma 11.45 we have the inequality* $|\lambda_h - \lambda_0| \le Ch$.

Proof. Choose $e_h,\ e_h^\star$ such that $(R_h e - e_h, e_h^\star)_0 = (e_h, R_h e^\star - e_h^\star)_0 = 0$. For the Rayleigh quotient $\tilde{\lambda}_h := (L_h R_h e, R_h e^\star)_0/(R_h e, R_h e^\star)_0$ we then have

$$|\tilde{\lambda}_h - \lambda_h| \le C\,|R_h e - e_h|_1\,|R_h e^\star - e_h^\star|_1 \le \varepsilon_h\,[h + |\lambda_h - \lambda_0|]$$

with $\varepsilon_h := C\,|R_h e^\star - e_h^\star|_1$. From (11.25c,j) one infers that

$$\begin{aligned} &(L_h R_h e, R_h e^\star)_0 - (Le, e^*)_0 \\ &= \big(\big[L_h R_h - \check{R}_h L\big]\,e, R_h e^\star\big)_0 + \big(Le, \big[\check{R}_h^* R_h - I\big]\,e^\star\big)_0 = \mathcal{O}(h) \end{aligned}$$

and $|\tilde{\lambda}_h - \lambda_0| = \mathcal{O}(h)$ such that $|\lambda_h - \lambda_0| \le Ch + \varepsilon_h\,|\lambda_h - \lambda_0|$. For sufficiently small h we have $\varepsilon_h \le \frac{1}{2}$ (cf. Theorem 11.42, Exercise 11.43) thus $|\lambda_h - \lambda_0| \le 2Ch$. ∎

Lemmata 11.45 and 11.46 give the next theorem.

Theorem 11.47. *Let (11.25a–m),* $E(\lambda_0) = \operatorname{span}\{e\}$, *and (11.10) hold. For all* $h > 0$ *there exists* $e_h \in E_h(\lambda_h)$ *with* $|R_h e - e_h|_1 \le Ch$.

Since $|R_h e^\star - e_h^\star|_1 = o(1)$ or even $|R_h e^\star - e_h^\star|_1 = \mathcal{O}(h)$, according to Theorem 11.24 one expects that $|\lambda_0 - \lambda_h| = o(h)$ [resp. $\mathcal{O}(h^2)$]. In general, this estimate is false as the following counterexample shows.

Example 11.48. The eigenvalue problem $-u'' + u' = \lambda u$ in $\Omega = (0,1)$ with $u(0) = u(1) = 0$ has the solution $u(x) = \exp(x/2)\sin(\pi x)$. The eigenvalue is $\lambda_0 = \pi^2 + 1/4$. One calculates that the discretisation $-\partial^-\partial^+ u + \partial^+ u = \lambda u$ has the eigenvalue

$$\begin{aligned} \lambda_h &= 2h^{-2}\,[1 - \cos(\pi h)\cosh(\Lambda h) - \mathrm{i}\sin(\pi h)\sinh(\Lambda h)] + h^{-1}\left[\mathrm{e}^{\mathrm{i}\pi h} - 1\right] \\ &= \pi^2 + \frac{1}{4} + \frac{1 - 3\pi^2}{8}h + \mathcal{O}(h^2) \qquad \text{with } \Lambda := \frac{\log(1-h)}{2h}, \end{aligned}$$

such that $|\lambda_0 - \lambda_h|$ turns out no better than $\mathcal{O}(h)$.

11.4 Further Remarks

In the quantitative error estimates of §11.2.3 we often assumed that the eigenvalue is simple. The discussion of *multiple eigenvalues* can be found in Babuška–Osborn [17, 18, 19].

A-posteriori error estimates and adaptive discretisations of boundary-value problems are mentioned in §8.7. Concerning corresponding investigations for eigenvalue problems we refer to Larson [182], Heuveline–Rannacher [151, 152], and Giani–Grubišić–Mięđlar–Ovall [114].

The constants in the error estimates of §11.2.3 are valid for a fixed eigenpair (λ_h, e_h). They are not uniform for all eigenvalues. The larger the eigenvalue λ_h, the worse is the approximation. Statements about the concrete quantitative behaviour are given by Sauter [249] and Yserentant [314].

Example 11.4 explicitly describes the eigenvalues $\lambda_{\nu\mu}$ of the Laplace operator in the unit square. Given some bound Λ one may ask about the number of eigenvalues with $\lambda_{\nu\mu} \le \Lambda$ (counted with respect of their multiplicity).

Exercise 11.49. $\#\{(\nu,\mu) : (\nu^2+\mu^2)\pi^2 \le \Lambda : \nu,\mu \in \mathbb{N}\} \le \frac{\Lambda}{4\pi}$ is the number of eigenvalues below Λ for the above eigenvalue problem.

This result about the asymptotic *eigenvalue distribution* can be generalised. Let $N(\Lambda)$ be the number of all eigenvalues λ of a selfadjoint elliptic operators in $\Omega \subset \mathbb{R}^n$ with $\lambda \le \Lambda$ Then there holds

$$\lim_{\Lambda\to\infty} \frac{N(\Lambda)}{\Lambda^{n/2}} = C$$

with a constant C depending on the spatial dimension n and the volume of Ω (cf. Weyl [304], Courant–Hilbert [78, page 442], Levendorskii [192, Theorem 13.1]).

Chapter 12
Stokes Equations

Abstract Besides differential equations of second or higher order there are systems of q differential equations for q scalar functions. In **Section 12.1** we present the systems of the Stokes and Lamé equations as particular examples and define the ellipticity of such systems. **Section 12.2** starts with the variational formulation of Stokes' equations. The saddle-point structure is discussed in §12.2.2. Solvability of general saddle-point problems is analysed in §12.2.3. The corresponding conditions are verified for the Stokes equations. A reinterpretation in §12.2.5 leads to a V_0-elliptic problem in a special subspace V_0. In **Section 12.3** the finite-element discretisation is studied. Special inf-sup conditions are to be satisfied since otherwise the problem is not solvable or unstable. Examples of stable finite elements are presented in §12.3.3.

12.1 Elliptic Systems of Differential Equations

In Example 1.11 we have already stated the *Stokes equations* for $\Omega \subset \mathbb{R}^2$:

$$-\Delta u_1 + \partial p/\partial x_1 = f_1, \tag{12.1a$_1$}$$

$$-\Delta u_2 + \partial p/\partial x_2 = f_2, \tag{12.1a$_2$}$$

$$-\partial u_1/\partial x_1 - \partial u_2/\partial x_2 = 0. \tag{12.1b}$$

In the case of $\Omega \subset \mathbb{R}^3$ another equation $-\Delta u_3 + \partial p/\partial x_3 = f_3$ needs to be added, and the left-hand side of (12.1b) must be supplemented by $-\partial u_3/\partial x_3$. A representation independent of the dimension can be obtained if one takes together (u_1, u_2) [resp. (u_1, u_2, u_3)] as a vector-valued function u satisfying the equations

$$-\Delta u + \nabla p = f \qquad \text{in } \Omega, \tag{12.2a}$$

$$-\operatorname{div} u = 0 \qquad \text{in } \Omega, \tag{12.2b}$$

Here, div is the *divergence operator*

W. Hackbusch, *Elliptic Differential Equations*, Springer Series
in Computational Mathematics 18, DOI 10.1007/978-3-662-54961-2_12

$$\operatorname{div} u = \sum_{i=1}^{n} \frac{\partial u_i}{\partial x_i}$$

and n is both the dimension of $\Omega \subset \mathbb{R}^n$ and the number of components of $u(\mathbf{x}) \in \mathbb{R}^n$. Only $n \leq 3$ is of physical interest here. In fluid mechanics the Stokes equation describes the flow of an incompressible medium (neglecting the inertial terms) and describes the velocity field. With $\mathbf{x} \in \mathbb{R}^n$, $u_i(\mathbf{x})$ is the velocity of the medium in the x_i direction; the function p denotes the pressure.

Including the inertial terms, we are led to the *Navier–Stokes equations*. This system combines even three difficulties: the indefinite structure of the Stokes system, the nonlinearity, and the singular perturbation (cf. §10.2) for high Reynolds numbers. Monographs on this subject – including the Stokes system – are John [161], Girault–Raviart [117], Ladyženskaja [179, 178], Marion–Temam [200], Temam [282, 281], and Thomasset [284].

Up to now we have not formulated any boundary conditions. In the following we limit ourselves to Dirichlet boundary values:

$$u = 0 \qquad \text{on } \Gamma\,. \tag{12.3}$$

This implies that the flow vanishes at the boundary. No boundary condition is given for p. Since both the pairs (u, p) and $(u, p + \text{const})$ for any value of const satisfy the Stokes equations (12.2a,b), (12.3), one obtains the following statement.

Remark 12.1. p is determined only up to a constant by the Stokes equation (12.2a,b) and the boundary conditions (12.3).

The Stokes equations have been chosen as an example of a *system* of differential equations. It remains to investigate whether equation (12.2a,b) is elliptic in a sense yet to be defined. Even though the functions u_i, for given p, are solutions of the elliptic Poisson equations $-\Delta u_i = f_i - \partial p/\partial x_i$, in determining p, (12.2a,b) do not provide an elliptic equation, in any current sense of elliptic.

A *general system* of q differential equations for q functions $u_1, \ldots, u_q$ can be written in the form

$$\sum_{j=1}^{q} L_{ij}\, u_j = f_i \qquad \text{in } \Omega \subset \mathbb{R}^n \qquad (1 \leq i \leq q) \tag{12.4a}$$

with the differential operators

$$L_{ij} = \sum_{|\alpha| \leq k_{ij}} c_\alpha D^\alpha \qquad (1 \leq i, j \leq q)\,. \tag{12.4b}$$

The equations (12.4a) are summarised as $Lu = f$ where L is the matrix (L_{ij}) of differential operators and $u = (u_1, \ldots, u_q)^{\mathsf{T}}$, $f = (f_1, \ldots, f_q)^{\mathsf{T}}$. The order of the operator L_{ij} is at most k_{ij}. Let the numbers $m_1, \ldots, m_q, m'_j, \ldots, m'_q$ be chosen so

that[1]

$$k_{ij} = m_i + m'_j \qquad (1 \le i, j \le q)\,. \tag{12.5}$$

As the *principal part* of L_{ij} one defines $L_{ij}^P := \sum_{|\alpha|=m_i+m'_j} c_\alpha D^\alpha$. The characteristic polynomial associated with L_{ij} reads:

$$L_{ij}^P(\boldsymbol{\xi}; \mathbf{x}) := \sum_{|\alpha|=m_i+m'_j} c_\alpha(\mathbf{x})\boldsymbol{\xi}^\alpha \qquad (\boldsymbol{\xi} \in \mathbb{R}^n,\ \mathbf{x} \in \Omega)$$

and forms the matrix-valued function

$$L^P(\boldsymbol{\xi}; \mathbf{x}) = \left(L_{ij}^P(\boldsymbol{\xi}; \mathbf{x})\right)_{i,j=1,\dots,q}.$$

Agmon–Douglis–Nirenberg [4] generalise the notion of ellipticity to a system as follows.

Definition 12.2. Let (12.5) hold for $m_i,\ m'_j$. The differential operator $L = (L_{ij})$ is said to be *elliptic* in $\mathbf{x} \in \Omega$ if

$$\det L^P(\boldsymbol{\xi}; \mathbf{x}) \neq 0 \qquad \text{for all } 0 \neq \boldsymbol{\xi} \in \mathbb{R}^n. \tag{12.6a}$$

L is said to be *uniformly elliptic* in Ω if there exists an $\varepsilon > 0$ such that

$$\left|\det L^P(\boldsymbol{\xi}; \mathbf{x})\right| \ge \varepsilon\,|\boldsymbol{\xi}|^{2m} \quad \begin{cases} \text{for all } \mathbf{x} \in \Omega,\ \boldsymbol{\xi} \in \mathbb{R}^n \\ \text{with } 2m := \sum_{i=1}^{q}(m_i + m'_i). \end{cases} \tag{12.6b}$$

To be more precise one should call L elliptic with indices $m_i,\ m'_j$, since the definition does depend on $m_i,\ m'_j$. In connection with this problem, as well as for an additional condition for $\mathbf{x} \in \Gamma,\ q = 2$, see the original paper of Agmon–Douglis–Nirenberg [4]. Further information on this subject can be found in Cosner [75].

Exercise 12.3. (a) Show that the numbers $m_i,\ m'_j$ are not unique. If m_i and m'_j satisfy the inequality (12.5), then so do $m_i - k$ and $m'_j + k$ for any k. The definition of L_{ij}^P is independent of k.

(b) Show that for $q = 1$, i.e., for the case of a single equation one recovers from (12.6a) the Definitions 1.14a [resp. (5.4a)]; (12.6b) corresponds to (5.4b).

(c) For a first-order system (i.e., $k_{ij} = 1,\ m_i = 1,\ m'_j = 0$), (12.6a) coincides with Definition 1.18.

In order to describe the Stokes equations in the form (12.4a,b) we set

$$q := n+1, \qquad u = (u_1, \dots, u_n, p)^{\mathsf{T}}, \qquad f = (f_1, \dots, f_n, 0)^{\mathsf{T}},$$
$$L_{ii} = -\Delta, \qquad L_{iq} = -L_{qi} = \partial/\partial x_i \text{ for } 1 \le i \le n, \qquad L_{ij} = 0 \quad \text{otherwise.}$$

[1] If $k_{ij} < m_i + m'_j$, replace k_{ij} by $k'_{ij} := m_i + m'_j$ and set $c_\alpha := 0$ for $k_{ij} < |\alpha| \le k'_{ij}$.

The orders are $k_{ii} = 2,\ k_{iq} = k_{qi} = 1\ (i \le n),\ k_{ij} = 0$ otherwise. The numbers

$$m_i = m_i' = \left\{ \begin{array}{ll} 1 & \text{for } 1 \le i \le n = q-1, \\ 0 & \text{for } i = q \end{array} \right\}$$

satisfy (12.5). L_{ij}^P coincides with L_{ij} and is independent of $\mathbf{x}$:

$$L_{ii}^P(\boldsymbol{\xi}) = -\left|\boldsymbol{\xi}\right|^2,\ \ L_{iq}^P(\boldsymbol{\xi}) = -L_{qi}^P(\boldsymbol{\xi}) = \xi_i \quad \text{for } i \le n,\ \ L_{ij}^P(\boldsymbol{\xi}) = 0 \ \text{ otherwise.}$$

From this we see

$$\left|\det L^P(\boldsymbol{\xi};\mathbf{x})\right| = \left| \det \begin{bmatrix} -\left|\boldsymbol{\xi}\right|^2 & & & \xi_1 \\ & \ddots & & \\ & & -\left|\boldsymbol{\xi}\right|^2 & \xi_n \\ -\xi_1 & \dots & -\xi_n & 0 \end{bmatrix} \right| = \left|\boldsymbol{\xi}\right|^{2m} \quad \text{with } \ 2m = 2n,$$

so that (12.6b) is satisfied, with $\varepsilon = 1$.

Exercise 12.4 (Lamé system). In elasticity theory[2] the so-called displacement function $u : \Omega \subset \mathbb{R}^3 \to \mathbb{R}^3$ is described by the Lamé system:

$$\mu \Delta u + (\lambda + \mu)\,\mathrm{grad}\,\mathrm{div}\,u = f \quad \text{in } \Omega, \qquad u = \varphi \quad \text{on } \Gamma$$

($\mu,\ \lambda\ >\ 0$). Show that this system of three equations is uniformly elliptic, in particular: $\left|\det L^P(\boldsymbol{\xi})\right| = \mu^2(2\mu + \lambda)\left|\boldsymbol{\xi}\right|^6$.

For the treatment of Stokes equations we will use a variational formulation in the next section. For reasons of completeness we point out the following transformation.

Remark 12.5. Let $n = 2$ and thus $\mathbf{u} = (u_1,\, u_2)$. Because $\mathrm{div}\,\mathbf{u} = 0$ holds for the Stokes solution there exists a so-called *stream function* Φ with

$$u_1 = \frac{\partial \Phi}{\partial x_2}, \qquad u_2 = -\frac{\partial \Phi}{\partial x_1}.$$

Insertion in equations (12.1a$_1$,a$_2$) results in the biharmonic equation

$$\Delta^2 \Phi = \frac{\partial f_2}{\partial x_1} - \frac{\partial f_1}{\partial x_2}.$$

The boundary condition (12.3) means $\nabla\Phi = 0$ on Γ. This is equivalent to $\partial\Phi/\partial n = 0$ and $\partial\Phi/\partial t = 0$ on Γ where $\partial/\partial t$ is the tangential derivative. $\partial\Phi/\partial t = 0$ implies $\Phi = \mathrm{const}$ on Γ. Since the constant may be chosen arbitrarily, one sets $\Phi = \partial\Phi/\partial n = 0$ on Γ.

[2] A detailed discussion of the various differential equations in elasticity theory can be found in Braess [45, Chapter VI, §3].

12.2 Variational Formulation

12.2.1 Weak Formulation of the Stokes Equations

Since $u = (u_1, \ldots, u_n)$ is a vector-valued function, we introduce

$$\mathbf{H}_0^1(\Omega) := H_0^1(\Omega) \times H_0^1(\Omega) \times \ldots \times H_0^1(\Omega) \qquad (n\text{-fold product}) .$$

A corresponding definition holds for $\mathbf{H}^{-1}(\Omega)$, $\mathbf{H}^2(\Omega)$, etc. The norm associated with $\mathbf{H}_0^1(\Omega)$ will again be denoted by $|\cdot|_1$ in the following.

According to Remark 12.1 the pressure component p of the Stokes problem is not uniquely determined. In order to determine uniquely the constant in $p = \tilde{p} + \text{const}$, we standardise p by the requirement $\int_\Omega p \mathrm{d}\mathbf{x} = 0$. That is the reason why in the following p will always belong to the subspace $L_0^2(\Omega) \subset L^2(\Omega)$:[3]

$$L_0^2(\Omega) := \left\{ p \in L^2(\Omega) : \int_\Omega p(\mathbf{x}) \mathrm{d}\mathbf{x} = 0 \right\} .$$

To derive the weak formulation we proceed as in Section 7.1 and assume that $\mathbf{u}$ and p are classical solutions of the Stokes problem (12.2a,b). Multiplication of the i-th equation in (12.2a) with $v_i \in C_0^\infty(\Omega)$ and subsequent integration implies that

$$\begin{aligned} \int_\Omega f_i(\mathbf{x}) v_i(\mathbf{x}) \mathrm{d}\mathbf{x} &= \int_\Omega \left[-\Delta u_i(\mathbf{x}) + \partial p(\mathbf{x}) / \partial x_i \right] v_i(\mathbf{x}) \mathrm{d}\mathbf{x} \qquad (12.7) \\ &= \int_\Omega \left[\langle \nabla u_i(\mathbf{x}), \nabla v_i(\mathbf{x}) \rangle - p(\mathbf{x}) \partial v_i(\mathbf{x}) / \partial x_i \right] \mathrm{d}\mathbf{x} \qquad \text{for } v_i \in C_0^\infty(\Omega) \end{aligned}$$

with $v = (v_i)_{i=1,\ldots,n}$. Summation over i now gives

$$\int_\Omega \left[\langle \nabla u(\mathbf{x}), \nabla v(\mathbf{x}) \rangle - p(\mathbf{x}) \operatorname{div} v(\mathbf{x}) \right] \mathrm{d}\mathbf{x} = \int_\Omega \big\langle f(\mathbf{x}), v(\mathbf{x}) \big\rangle \mathrm{d}\mathbf{x},$$

where the abbreviation

$$\langle \nabla u, \nabla v \rangle := \sum_{i=1}^n \langle \nabla u_i, \nabla v_i \rangle := \sum_{i,j=1}^n \frac{\partial u_i}{\partial x_j} \frac{\partial v_i}{\partial x_j}$$

is used. Equation (12.2b) is then multiplied with some $q \in L_0^2(\Omega)$ and integrated, giving

$$-\int_\Omega q(\mathbf{x}) \operatorname{div} u(\mathbf{x}) = 0 \qquad \text{for all } q \in L_0^2(\Omega).$$

With the bilinear forms

[3] Often, $L^2(\Omega)/\mathbb{R}$ is written instead of $L_0^2(\Omega)$. This denotes the quotient space of $L^2(\Omega)$ over the constant functions. $L^2(\Omega)/\mathbb{R}$ is isomorphic to $L_0^2(\Omega)$. The quotient space norm of an equivalence class $F \in L^2(\Omega)/\mathbb{R}$ coincides with the L^2 norm of the representative $f \in L_0^2(\Omega)$ of F.

$$a(u,v) := \int_\Omega [\langle \nabla u(\mathbf{x}), \nabla v(\mathbf{x})\rangle - p(\mathbf{x}) \operatorname{div} v(\mathbf{x})]\, \mathrm{d}\mathbf{x} \quad \text{for } u, v \in \mathbf{H}_0^1(\Omega), \tag{12.8a}$$

$$b(p,v) := -\int_\Omega p(\mathbf{x}) \operatorname{div} v(\mathbf{x}) \mathrm{d}\mathbf{x} \qquad \text{for } p \in L_0^2(\Omega),\ v \in \mathbf{H}_0^1(\Omega) \tag{12.8b}$$

we obtain the weak formulation of the Stokes problem as (12.9a–c):

$$\text{find} \quad u \in \mathbf{H}_0^1(\Omega) \quad \text{and} \quad p \in L_0^2(\Omega) \quad \text{such that} \tag{12.9a}$$

$$a(u,v) + b(p,v) = f(v) := \int_\Omega \langle f(\mathbf{x}), v(\mathbf{x})\rangle \, \mathrm{d}\mathbf{x} \qquad \text{for all } v \in \mathbf{H}_0^1(\Omega), \tag{12.9b}$$

$$b(q,u) = 0 \qquad \text{for all } q \in L_0^2(\Omega). \tag{12.9c}$$

In (12.9b) we first replace '$v \in H_0^1(\Omega)$' by '$v \in C_0^\infty(\Omega)$'. Since both sides of (12.9b) depend continuously on $v \in H_0^1(\Omega)$ and since $C_0^\infty(\Omega)$ is dense in $H_0^1(\Omega)$, (12.9b) follows for all $v \in H_0^1(\Omega)$.

Remark 12.6. A classical solution $u \in C^2(\overline{\Omega}) \cap H_0^1(\Omega)$, $p \in C^1(\Omega) \cap L_0^2(\Omega)$ of the Stokes problem (12.2a,b), (12.3) is also a weak solution, i.e., a solution of (12.9a–c). If conversely (12.9a–c) has a solution with $u \in C^2(\overline{\Omega})$, $p \in C^1(\overline{\Omega})$, then it is also the classical solution of the boundary-value problem (12.2a,b), (12.3).

Proof. (a) The above considerations prove the first part.

(b) Equation (12.9c) implies $\operatorname{div} u = 0$. Let $i \in \{1, \ldots, n\}$. In equation (12.9b) one can choose v with $v_i \in C_0^\infty(\Omega)$, $v_j = 0$ for $j \neq i$. Integration by parts recovers (12.7) and hence the i-th equation in (12.2a). ∎

12.2.2 Saddle-Point Problems

The situation in (12.9a–c) is a special case of the following problem. We replace the spaces $\mathbf{H}_0^1(\Omega)$ and $L_0^2(\Omega)$ in (12.9a–c) by two Hilbert spaces V and W. Let

$$a(\cdot,\cdot) : V \times V \to \mathbb{R} \quad \text{be a continuous bilinear form on } V \times V, \tag{12.10a}$$

$$b(\cdot,\cdot) : W \times V \to \mathbb{R} \quad \text{be a continuous bilinear form on } W \times V, \tag{12.10b}$$

$$f_1 \in V', \qquad f_2 \in W'. \tag{12.10c}$$

In generalisation of (6.42) we call $b(\cdot,\cdot) : W \times V \to \mathbb{R}$ continuous (or bounded), if there exists a $C_b \in \mathbb{R}$ such that

$$|b(w,v)| \le C_b \|w\|_W \|v\|_V \qquad \text{for all } w \in W,\ v \in V.$$

The objective of this chapter is to solve the problem (12.11):

$$\text{find } v \in V \text{ and } w \in W \text{ with } \begin{cases} a(v,x) + b(w,x) = f_1(x) & \text{for all } x \in V, \\ b(y,v) \qquad\quad\ = f_2(y) & \text{for all } y \in W. \end{cases} \tag{12.11}$$

Formally, (12.11) can be transformed to the form

$$\text{find } u \in X \text{ with } \quad c(u,z) = f(z) \quad \text{for all } z \in X \tag{12.12a}$$

if one sets: $X := V \times W$ and

$$\left.\begin{array}{ll} c(u,z) & := a(v,x) + b(w,x) + b(y,v) \\ f(z) & := f_1(x) + f_2(y) \end{array}\right\} \text{ for } u = \begin{pmatrix} v \\ w \end{pmatrix},\ z = \begin{pmatrix} x \\ y \end{pmatrix}. \tag{12.12b}$$

Exercise 12.7. Show that (a) $c(\cdot,\cdot) : X \times X \to \mathbb{R}$ is a continuous bilinear form. (b) Problems (12.11) and (12.12a,b) are equivalent.

That the variational problems (12.11) and (12.12a) must be handled differently than in Chapter 7 is made clear by the following statement which follows from $c(u,u) = 0$ for all $u = \binom{0}{w}$.

Remark 12.8. The bilinear form $c(\cdot,\cdot)$ in (12.12b) cannot be X-elliptic.

In analogy to (6.47) we set

$$J(v,w) := a(v,v) + 2b(w,v) - 2f_1(v) - 2f_2(w),$$

and therefore $J(v,w) = c(u,u) - 2f(u)$ for $u = \binom{v}{w}$. For $v \in V$, $w \in W$ we know $J(v,w)$ is neither bounded below nor above. Therefore the solution $v^\star,\ w^\star$ of (12.11) does not give a minimum of J; however, under suitable conditions, $v^\star,\ w^\star$ may be a *saddle point* with the property (12.13), i.e., J is minimal with respect of variations in V and maximal for variations in W.

Theorem 12.9. *Let (12.10a–c) hold. Let $a(\cdot,\cdot)$ be symmetric and V-elliptic. The pair $v^\star \in V$ and $w^\star \in W$ is a solution of the problem (12.11) if and only if*

$$J(v^\star,w) \le J(v^\star,w^\star) \le J(v,w^\star) \qquad \textit{for all } w \in W,\ v \in V. \tag{12.13}$$

Another equivalent characterisation is

$$J(v^\star,w^\star) = \min_{v\in V} J(v,w^\star) = \max_{w\in W} \min_{v\in V} J(v,w). \tag{12.14}$$

Proof. (ia) Let $v^\star,\ w^\star$ solve (12.11). Symmetry of $a(\cdot,\cdot)$ gives

$$\begin{aligned} a(v^\star - v, v^\star - v) - 2a(v^\star, v^\star - v) &= -a(v^\star,v^\star) - a(v,v^\star) + a(v^\star,v) + a(v,v) \\ &= a(v,v) - a(v^\star,v^\star) \quad \text{for all } v \in V. \end{aligned}$$

Let $\delta v := v^\star - v$. The expression in brackets in

$$J(v,w^\star) - J(v^\star,w^\star) = a(\delta v,\delta v) + 2\big[\, a(v^\star,\delta v) + b(w^\star,\delta v) - f_1(\delta v)\,\big]$$

vanishes because of (12.11). Since $a(v^\star - v, v^\star - v) > 0$ for all $v^\star \ne v \in V$, the second inequality in (12.13) follows. One also proves the converse as for Theorem 6.104: If $J(v,w^\star)$ is minimal for $v = v^\star$ then the first equation in (12.11) holds.

(ib) If $v^\star$ is a solution of (12.11), then

$$J(v^\star, w^\star) - J(v^\star, w) = 2\,[\,b(w^\star - w, v^\star) - f_2(w^\star - w)\,]$$

vanishes for all w, which proves the first part of (12.13) in the stronger form $J(v^\star, w) = J(v^\star, w^\star)$. For the reverse direction define $w_\pm = w^\star \pm w$. The first part of (12.13) implies

$$0 \le J(v^\star, w^\star) - J(v^\star, w_\pm) = \pm 2\,[b(w, v^\star) - f_2(w)]$$

for both signs, and so $b(w, v^\star) = f_2(w)$. Since $w \in W$ is arbitrary, one obtains (12.11).

(iia) We set

$$j(w) := \min_{v \in V} J(v, w).$$

According to Theorem 6.104, $j(w) = J(v_w, w)$, where $v_w \in V$ is the solution of the upper equation in (12.11). If v_w and $v_{w'}$ are the solutions for w and w', it follows that

$$a(v_w - v_{w'}, x) = F(x) := b(w - w', x) \qquad \text{for all } x \in X.$$

Since $\|F\|_{V'} \le C_b \|w - w'\|_V$ and $\|v_w - v_{w'}\|_V \le C' \|F\|_{V'}$ one obtains

$$\|v_w - v_{w'}\|_V \le C \|w - w'\|_V \qquad \text{for all } w, w' \in W. \tag{12.15a}$$

(iib) By using the definition of v_w in (12.11) we can write:

$$\begin{aligned}
J(v^\star,w^\star) - J(v_w, w) = \quad & \left(\begin{array}{l} \text{since } a(v^\star, v^\star) - a(v_w, v_w) = \\ = -2a(v_w, v_w - v^\star) + a(v_w - v^\star, v_w - v^\star) \end{array} \right) \\
&= [\,a(v^\star, v^\star) + 2b(w^\star, v^\star) - 2f_1(v^\star) - 2f_2(w^\star)\,] \\
&\quad - [\,a(v_w, v_w) + 2b(w, v_w) - 2f_1(v_w) - 2f_2(w)\,] \\
&= 2\,[\,f_1(v_w - v^\star) - b(w, v_w - v^\star) - a(v_w, v_w - v^\star)\,] \\
&\quad + a(v_w - v^\star, v_w - v^\star) + 2\,[\,b(w^\star - w, v^\star) - f_2(w^\star - w)\,] \\
&= a(v_w - v^\star, v_w - v^\star) + 2\,[\,b(w^\star - w, v^\star) - f_2(w^\star - w)\,]\,.
\end{aligned} \tag{12.15b}$$

(iic) Let $(v^\star, w^\star)$ be a solution of (12.11). Because of the second equation in (12.11) the expression in brackets in (12.15b) vanishes and we have

$$J(v^\star, w^\star) = J(v_w, w) + a(v_w - v^\star, v_w - v^\star) \ge J(v_w, w) = j(w).$$

The upper equation in (12.11) gives $v_{w^\star} = v^\star$, and so

$$J(v^\star, w^\star) = j(w^\star) = \max_{w \in W} j(w),$$

i.e., (12.14) holds.

(iid) Now let $(v^\star, w^\star)$ be a solution of (12.14). If in (12.15b) one sets $w = w^\star$ and $v_w = v_{w^\star}$ one obtains from $J(v^\star, w^\star) = j(w^\star)$ that $v^\star = v_{w^\star}$. Hence $a(v_w - v^\star, v_w - v^\star) = a(v_w - v_{w^\star}, v_w - v_{w^\star})$ depends quadratically on $\|w - w^\star\|_V$ (cf. (12.15a)). The variation over $w := w^\star - \lambda y$ ($\lambda \in \mathbb{R}$, $y \in W$ arbitrary) gives

$$0 = \frac{\mathrm{d}}{\mathrm{d}\lambda}\, j(w^\star - \lambda y)|_{\lambda=0} = 2\,[b(y, v^\star) - f_2(y)]\,,$$

and so the second equation in (12.11) is proved. The upper equation in (12.11) has already been established with $v^\star = v_{w^\star}$. ∎

12.2.3 Existence and Uniqueness of the Solution of a Saddle-Point Problem

To make the saddle-point problem (12.11) somewhat more transparent we introduce the operators associated with the bilinear forms:

$$A \in L(V, V') \quad \text{with } a(v, x) = \langle Av, x\rangle_{V'\times V} \quad \text{for all } v, x \in V, \tag{12.16a}$$

$$B \in L(W, V'),\ B^* \in L(V, W') \quad \text{with } b(w, x) = \langle Bw, x\rangle_{V'\times V} = \langle w, B^* x\rangle_{W\times W'}\,, \tag{12.16b}$$

$$C \in L(X, X') \quad \text{with } c(u, z) = \langle Cu, z\rangle_{X'\times X} \quad \text{for all } u, z \in X, \tag{12.16c}$$

Thus problem (12.12a,b) now has the form $Cu = f$, while (12.11) can be written as

$$Av + Bw = f_1, \tag{12.17a}$$

$$B^* v \qquad = f_2. \tag{12.17b}$$

If one assumes the existence of $A^{-1} \in L(V', V)$, one can solve (12.17a) for v:

$$v = A^{-1}\,(f_1 - Bw) \tag{12.18a}$$

and substitute in (12.17b):

$$B^* A^{-1} B w = B^* A^{-1} f_1 - f_2. \tag{12.18b}$$

The invertibility of A is in no way necessary for the solvability of the saddle-point problem (the exact condition is discussed in Theorem 12.12). However, it does simplify the analysis, and does hold true in the case of the Stokes problem.

Remark 12.10. (a) Under the assumptions

$$A^{-1} \in L(V', V), \qquad B^* A^{-1} B \in L(W', W),$$

the saddle-point problem (12.11) [resp. equations (12.17a,b)] are uniquely solvable.

(b) A necessary condition for the existence of $(B^\star A^{-1}B)^{-1}$ is

$$B \in L(W, V') \quad \text{is injective.} \tag{12.19}$$

Proof. (a) Under the assumption of $(B^\star A^{-1}B)^{-1} \in L(W', W)$, (12.18b) is uniquely solvable for w, and then (12.18a) yields v.

(b) The injectivity of $B^\star A^{-1}B$ implies (12.19). ∎

Attention. In general, $B : W \to V'$ is not bijective so that the representation of $(B^\star A^{-1}B)^{-1}$ as $B^{-1}AB^{\star -1}$ is not possible.

The example of the 3×3 matrix $C = \begin{bmatrix} A & B \\ B^\mathsf{T} & 0 \end{bmatrix}$ with $A = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ and $B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ shows that a system of the form (12.17a,b) can be solvable even with a singular matrix A. Therefore the assumption $A^{-1} \in L(V', V)$ is not necessary. A closer look reveals the subspace

$$V_0 := \ker(B^*) = \{v \in V : b(y, v) = 0 \text{ for all } y \in W\} \subset V, \tag{12.20}$$

which, as we noted before, in general is not trivial. The kernel of a continuous mapping is closed so that V, according to Lemma 6.15, can be represented as the sum of orthogonal spaces:

$$V = V_0 \oplus V_\perp \qquad \text{with} \quad V_\perp := (V_0)^\perp . \tag{12.21a}$$

Exercise 12.11. Let (12.21a) hold. Show that (a) the dual space V' can be represented as

$$V' = V_0' \oplus V_\perp' , \tag{12.21b}$$

where

$$V_0' := \{v' \in V' : v'(v) = 0 \text{ for all } v \in V_\perp\},$$
$$V_\perp' := \{v' \in V' : v'(v) = 0 \text{ for all } v \in V_0\}$$

are orthogonal spaces with respect to the dual norm $\|\cdot\|_{V'}$.

(b) The Riesz isomorphism $J_V : V \to V'$ maps V_0 onto V_0' and $V_\perp$ onto $V_\perp'$.

(c) The following holds:

$$\|v'\|_{V'}^2 = \|v_0'\|_{V'}^2 + \|v_\perp'\|_{V'}^2 \qquad \begin{cases} \text{for all } v' = v_0' + v_\perp' \in V', \\ \text{where } v_0' \in V_0',\ v_\perp' \in V_\perp' . \end{cases}$$

The decompositions (12.21a,b) of V and V' define a block decomposition of the operator A:

$$A = \begin{bmatrix} A_{00} & A_{0\perp} \\ A_{\perp 0} & A_{\perp\perp} \end{bmatrix}, \quad \text{with} \quad \begin{cases} A_{00} \in L(V_0, V_0'), & A_{0\perp} \in L(V_\perp, V_0'), \\ A_{\perp 0} \in L(V_0, V_\perp'), & A_{\perp\perp} \in L(V_\perp, V_\perp'). \end{cases}$$

Here, for example, A_{00} is defined as follows:

$$A_{00}v_0 = v_0' \quad \text{for } v_0 \in V_0, \qquad \text{if } Av_0 = v_0' + v_\perp',\ v_0' \in V_0',\ v_\perp' \in V_\perp'\,.$$

The corresponding decomposition of $B^\star$ into $(B_0^\star, B_\perp^\star)$ is written as $(0, B^\star)$, since $B_0^\star = 0$ according to the definition of V_0. Conversely, we have $\mathrm{range}(B) \subset V_\perp'$, so that $B = \binom{0}{B}$. System (12.17a,b) thus becomes

$$\begin{aligned} A_{00}v_0 \,+\, A_{0\perp}v_\perp \qquad &= f_{10}\,, && \text{(12.22a)}\\ A_{\perp 0}v_0 + A_{\perp\perp}v_\perp + Bw &= f_{1\perp}\,, && \text{(12.22b)}\\ B^*v_\perp \qquad &= f_2\,, && \text{(12.22c)} \end{aligned}$$

where $v = v_0 + v_\perp,\ v_0 \in V_0,\ v_\perp \in V_\perp,\ f_1 = f_{10} + f_{1\perp},\ f_{10} \in V_0',\ f_{1\perp} \in V_\perp'$.

Theorem 12.12. *Let (12.10a–c) hold. Let V_0 be defined by (12.20). A necessary and sufficient condition for the unique solvability of the saddle-point problem (12.11) for all $f_1 \in V'$ is the existence of the inverses*

$$A_{00}^{-1} \in L(V_0', V_0) \qquad \textit{and} \quad B^{-1} \in L(V_\perp', W). \tag{12.23}$$

Proof. (i) (12.22a–c) represents a staggered system of equations. (12.23) implies $B^{\star -1} = (B^{-1})^\star \in L(W', V_\perp)$ so that one can solve (12.22c) for $v_\perp = B^{\star -1} f_2$. $v_0 \in V_0$ is obtained from (12.22a): $v_0 = A_{00}^{-1}(f_{10} - A_{0\perp}v_\perp)$. Finally, w results from (12.22b).

(ii) In order to show that (12.23) is necessary, we take $f_{10} \in V_0'$ arbitrary, $f_{1\perp} = 0$ and $f_2 = 0$. By hypothesis here we have a solution $(v_0, v_\perp, w) \in V_0 \times V_\perp \times W$. $B^\star v_\perp = 0$ implies $v_\perp \in V_0$, so that $v_\perp = 0$ because $V_0 \cap V_\perp = \{0\}$. Thus $A_{00}v_0 = f_{10}$ has a unique solution $v_0 \in V_0$ for each $f_{10} \in V_0'$. Since $A_{00}\colon V_0 \to V_0'$ is bijective and bounded, Theorem 6.12 shows that $A_{00}^{-1} \in L(V_0', V_0)$. If one takes $f_{1\perp} \in V_\perp'$ arbitrary and $f_{10} = 0$, $f_2 = 0$, one infers $v_\perp = 0$ and $v_0 = 0$, so that $Bw = f_{1\perp}$ has a unique solution $w \in W$. As we did for A_{00}, one also infers that $B^{-1} \in L(V_\perp', W)$. ∎

The formulation of conditions (12.23) in terms of the bilinear forms results in the *Babuška–Brezzi conditions* (cf. Footnote 3 on page 153):

$$\inf_{v_0 \in V_0,\, \|v_0\|_V = 1} \sup_{x_0 \in V_0,\, \|x_0\|_V = 1} |a(v_0, x_0)| \geq \alpha > 0, \tag{12.24a}$$

$$\sup_{x_0 \in V_0,\, \|x_0\|_V = 1} |a(x_0, v_0)| > 0 \ \text{ for all } \ 0 \neq v_0 \in V_0, \tag{12.24b}$$

$$\inf_{w \in W,\, \|w\|_W = 1} \sup_{x \in V,\, \|v\|_V = 1} |b(w, x)| \geq \beta > 0. \tag{12.24c}$$

Exercise 12.13. Show that (12.24a) [resp. (12.24c)] are equivalent to

$$\sup_{x_0 \in V_0,\, \|x_0\|_V = 1} |a(v_0, x_0)| \geq \alpha \|v_0\|_V \qquad \text{for all } v_0 \in V_0,$$

$$\sup_{x \in V,\, \|v\|_V = 1} |\, b(w, x)\,| \geq \beta \|w\|_W \qquad \text{for all } w \in W.$$

Lemma 12.14. *Let (12.10a,b) hold. Let V_0 be defined by (12.20). Then conditions (12.23) and (12.24a–c) are equivalent. Here we have*

$$\|A_{00}^{-1}\|_{V_0 \leftarrow V_0'} \le 1/\alpha, \qquad \|B^{-1}\|_{W \leftarrow V_\perp'} \le 1/\beta.$$

Proof. Because of (12.21b) and $b(w,x) = 0$ for $x \in V_0$ one can write (12.24c) in the form

$$\inf_{w \in W,\, \|w\|_W = 1} \sup_{x \in V_\perp,\, \|v\|_V = 1} |b(w,x)| \ge \beta > 0. \tag{12.24d}$$

For $0 \ne x \in V_\perp$ one has that $x \notin V_0$, therefore according to (12.20):

$$\sup_{w \in W,\, \|w\|_W = 1} |b(w,x)| > 0 \qquad \text{for all } 0 \ne x \in V_\perp\,. \tag{12.24e}$$

As in the proof of Lemma 6.94, we obtain the equivalence of (12.24a,b) with $A_{00}^{-1} \in L(V_0', V_0)$ and of (12.24d,e) with $B^{-1} \in L(V_\perp', W)$. ∎

Corollary 12.15. (a) Condition (12.24b) becomes superfluous if $a(\cdot,\cdot)$ is symmetric on $V_0 \times V_0$ or if Lemma 6.109 applies.

(b) Each of the following conditions is sufficient for (12.24a,b) and hence also for $A_{00}^{-1} \in L(V_0', V_0)$:

$$a(\cdot,\cdot) : V_0 \times V_0 \to \mathbb{R} \text{ is } V_0\text{-elliptic: } a(v_0,v_0) \ge \alpha \|v_0\|_V^2 \text{ for all } v_0 \in V_0\,, \tag{12.25a}$$

$$a(\cdot,\cdot) : V \times V \to \mathbb{R} \text{ is } V\text{-elliptic.} \tag{12.25b}$$

Proof. (a) As in Lemma 6.109.
(b) (12.25b) implies (12.25a); (12.25a) yields (12.24a,b). ∎

Exercise 12.16. Show that under the assumptions (12.10a), condition (12.23) is also equivalent to the existence of $C^{-1} \in L(X', X)$ (cf. (12.16c)). Find a bound for $\|C^{-1}\|_{X \leftarrow X'}$ in terms of $\|A_{00}^{-1}\|_{V_0 \leftarrow V_0'}$, $\|A\|_{V' \leftarrow V'}$, and $\|B^{-1}\|_{W \leftarrow V_\perp'}$.

12.2.4 Solvability and Regularity of the Stokes Problem

Conditions (12.24a,b) (i.e., $A_{00}^{-1} \in L(V_0', V_0)$) are easy to satisfy for the Stokes problem.

Lemma 12.17. *Let Ω be bounded. Then the forms (12.8a,b) describing the Stokes problem satisfy the conditions (12.10a,b) and (12.24a,b).*

Proof. (12.10a,b) is self-evident. According to Example 7.10, $\int_\Omega \langle \nabla u, \nabla v \rangle \mathrm{d}\mathbf{x}$ is $H_0^1(\Omega)$-elliptic. From this follows the $\mathbf{H}_0^1(\Omega)$-ellipticity of $a(\cdot,\cdot)$. Corollary 12.15b proves (12.24a,b). ∎

It remains to prove condition (12.24c), which for the Stokes problem assumes the form

$$\sup_{u\in\mathbf{H}^1_0(\Omega),\ |u|_1=1} \left| \int_\Omega w(\mathbf{x}) \operatorname{div} u(\mathbf{x}) \mathrm{d}\mathbf{x} \right| \geq \beta |w|_0 \quad \text{for all } w \in L^2_0(\Omega) \tag{12.26a}$$

or equivalently

$$\|\nabla w\|_{\mathbf{H}^{-1}(\Omega)} \geq \beta \|w\|_{L^2(\Omega)} \qquad \text{for all } w \in L^2_0(\Omega). \tag{12.26b}$$

Lemma 12.18. *Sufficient and necessary for (12.26a) is that for each $w \in L^2_0(\Omega)$ there exists $u \in \mathbf{H}^1_0(\Omega)$ such that*

$$w = \operatorname{div} u, \qquad |u|_1 \leq \tfrac{1}{\beta} |w|_0 \,. \tag{12.26c}$$

Proof. (i) For $w \in L^2_0(\Omega)$ select u such that (12.26c) holds and set $\tilde{u} := u/|u|_1$. The left-hand side in (12.26a) is $\geq \int_\Omega w(\mathbf{x}) \operatorname{div} \tilde{u}(\mathbf{x}) \mathrm{d}\mathbf{x} = |w|_0^2 / |u|_1 \geq \beta |w|_0$.

(ii) If (12.26a) holds one infers as in §12.2.3 the bijectivity of $B^\star : V_\perp \to W$ with $\|B^{\star -1}\|_{V_\perp \leftarrow W} \leq 1/\beta$. Therefore $u := B^{\star -1} w$ satisfies condition (12.26c). ■

Nečas [211] proves the following theorem.

Theorem 12.19. *Condition (12.26a) is satisfied if $\Omega \in C^{0,1}$ is bounded. Under this assumption, the Stokes problem*

$$-\Delta u + \nabla p = f, \quad -\operatorname{div} u = g \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \Gamma \tag{12.27}$$

has a unique solution $(u, p) \in \mathbf{H}^1_0(\Omega) \times L^2_0(\Omega)$ for $f \in \mathbf{H}^{-1}(\Omega)$ and $g \in L^2_0(\Omega)$, with

$$|u|_1 + |p|_0 \leq C_\Omega \left[|f|_{-1} + |g|_0 \right] .$$

Remark 12.20. Under the conditions that $n = 2$ and that $\Omega \in C^2$ is a bounded domain, the existence proof can be carried out as follows.

Proof. We need to prove (12.26c). For $w \in L^2_0(\Omega)$ solve

$$-\Delta \varphi = w \quad \text{in } \Omega, \qquad \varphi = 0 \quad \text{on } \Gamma.$$

Theorem 9.19 shows that $\varphi \in H^2(\Omega)$. Since $\nabla\varphi \in \mathbf{H}^1(\Omega)$ and $\mathbf{n}(x) \in C^1(\Gamma)$ it follows that $g := \partial\varphi/\partial n \in H^{1/2}(\Gamma)$ (cf. Theorem 6.58a). From (3.17) one infers that $\int_\Gamma g \mathrm{d}\Gamma = \int_\Omega w \mathrm{d}\mathbf{x} = 0$ since $w \in L^2_0(\Omega)$. Integration of g over Γ yields $G \in H^{3/2}(\Gamma)$ with $\partial G/\partial t = g$, where $\partial/\partial t$ is the tangential derivative. There exists a function $\psi \in H^2(\Omega)$ with $\psi = G$ and $\partial\psi/\partial n = 0$ on Γ and $|\Psi|_2 \leq C\,|G|_{3/2} \leq C'\,|g|_{1/2} \leq C''\,|\varphi|_2 \leq C'''\,|w|_0$. We set

$$u_1 := -\varphi_x - \psi_y, \qquad u_2 := -\varphi_y + \psi_x \,.$$

Clearly, $u_1, u_2 \in H^1(\Omega)$. Let the normal direction at $u_1, u_2 \in H^1(\Omega)$ be $\mathbf{n}(\mathbf{x}) = \binom{n_1}{n_2}$. The tangent direction is thus $\mathbf{t}(\mathbf{x}) := \binom{n_2(\mathbf{x})}{-n_1(\mathbf{x})}$. For $u = \binom{u_1}{u_2}$ one obtains

$$\begin{aligned}
\langle u, \mathbf{n} \rangle &= -\varphi_x n_1 - \psi_y n_1 - \varphi_y n_2 + \psi_x n_2 = -\frac{\partial\varphi}{\partial n} + \frac{\partial\psi}{\partial t} = -g + \frac{\partial G}{\partial t} = 0, \\
\langle u, \mathbf{t} \rangle &= -\varphi_x n_2 - \psi_y n_2 - \varphi_y n_1 + \psi_x n_1 = -\partial\varphi/\partial t - \partial\psi/\partial n = 0,
\end{aligned}$$

since $\varphi = 0$ on Γ implies $\partial\varphi/\partial t = 0$, while $\partial\psi/\partial n = 0$ holds by definition. $\langle u, \mathbf{n}\rangle = \langle u, \mathbf{t}\rangle = 0$ yields $u = 0$ on Γ such that $u = (u_1, u_2)^\mathsf{T} \in \mathbf{H}_0^1(\Omega)$ is proved. One verifies that

$$\operatorname{div} u = \partial u_1/\partial x + \partial u_2/\partial y = (-\varphi_{xx} - \psi_{yx}) + (-\varphi_{yy} + \psi_{xy}) = -\Delta\varphi = w$$

with $|u|_1 \le |\varphi|_2 + |\psi|_2 \le C\,|w|_0$. ∎

The above proof uses the H^2-regularity of the Poisson problem and requires corresponding assumptions on Ω. Theorem 12.19 also assumes $\Omega \in C^{0,1}$. Since the Poisson equation $-\Delta u = f$ is solvable for any domain Ω which is contained in the ball $K_R(0)$, or at least in a strip $\{x \in \mathbb{R}^n : |x_1| < R\}$, resulting in the inequality $|u|_1 \le C_R\,|f|_{-1}$ with C_R only depending on R, one might conjecture that a similar result holds for the Stokes problem. However, there is the following counterexample.

Example 12.21. For $\varepsilon \in (0,1)$ let

$$\Omega_\varepsilon := \{(x,y) : -1 < x < 1,\ 0 < y < \varepsilon + (1-\varepsilon)\,|x|\}$$

be the underlying domain (cf. Figure 12.1). All domains $\Omega_\varepsilon \in C^{0,1}$ are located in the rectangle $(-1,1)\times(0,1)$. Nevertheless, there exists no $\beta > 0$ such that (12.26c) holds for all $\varepsilon > 0$, $w \in L^2(\Omega_\varepsilon)$.

Proof. We select $w \in L_0^2(\Omega_\varepsilon)$ such that $w(x,y) = 1$ for $x > 0$, $w(x,y) = -1$ for $x \le 0$. Let the inequality (12.26c) hold for Ω_ε with $\beta_\varepsilon > 0$. Let $u \in \mathbf{H}_0^1(\Omega_\varepsilon)$ with $|u|_1 \le |w|_0/\beta_\varepsilon$ be chosen according to Lemma 12.18. We continue u by $u = 0$ onto $\mathbb{R}^2$. For the restriction on $x = 0$ we have

$$\|u_1(0,\cdot)\|_{L^2(\mathbb{R})} \le \|u_1(0,\cdot)\|_{H^{1/2}(\mathbb{R})} \le C\,|u|_1 \le C\,|w|_0/\beta_\varepsilon \le 2C/\beta_\varepsilon$$

according to Theorem 6.46. Let $\chi(y) = 1$ for $0 < y < \varepsilon$ and $\chi(y) = 0$ otherwise. Since $u_1(0,y) = u_1(0,y)\chi(y)$ and $|\chi|_0 = \sqrt{\varepsilon}$, we have

$$\left|\int_0^\varepsilon u_1(0,y)\mathrm{d}y\right| = \left|\int_{\mathbb{R}} u_1(0,y)\chi(y)\mathrm{d}y\right| \le |u_1(0,\cdot)|_0\,|\chi|_0 \le 2C\sqrt{\varepsilon}/\beta_\varepsilon\,.$$

Let $\Omega_\varepsilon^+ = \{(x,y) \in \Omega_\varepsilon : x > 0\}$ and $\gamma := \{(x,y) : x = 0,\ 0 < y < \varepsilon\} = \partial\Omega_\varepsilon^+ \backslash \partial\Omega_\varepsilon$. Because $w = 1$ in Ω_ε^+, and because $\operatorname{div} u = w$ we have

$$\begin{aligned}\frac{1}{2} &\le \int_{\Omega_\varepsilon^+} |w(\mathbf{x})|^2\,\mathrm{d}\mathbf{x} = \int_{\Omega_\varepsilon^+} w\operatorname{div} u\mathrm{d}\mathbf{x} = \int_{\Omega_\varepsilon^+} \operatorname{div} u\mathrm{d}\mathbf{x} = \int_\gamma \langle u, \mathbf{n}\rangle\,\mathrm{d}\Gamma \\ &= -\int_\gamma u_1\mathrm{d}\Gamma = -\int_0^\varepsilon u_1(0,y)\mathrm{d}y.\end{aligned}$$

The last two inequalities result in $1/2 \le 2C\sqrt{\varepsilon}/\beta_\varepsilon$, from which we infer that β_ε cannot be bounded from below by some $\beta_0 > 0$ independent of ε. ∎

Exercise 12.22. Construct a domain Ω located on the strip $\mathbb{R} \times (0,1)$ in which the Stokes equations are not solvable. *Hint:* Join the domains $\Omega_{1/\nu}$ $(\nu \in \mathbb{N})$ from Figure 12.1.

In Braess [45, §III, Bemerkung 6.6 in the fifth German edition] one finds an example of a *non*-Lipschitz domain for which the statement of the Theorem 12.19 is not valid.

As for the case of scalar differential equations one obtains stronger regularity of the Stokes solution u, p if one assumes more than $f \in \mathbf{H}^{-1}(\Omega)$.

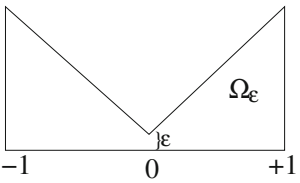


Fig. 12.1 Domain Ω_ε.

Theorem 12.23. *Let Ω be bounded and sufficiently smooth. Let u and p be the (weak) solution of the Stokes problem (12.27) with*

$$f \in H^k(\Omega), \quad g \in H^{k+1}(\Omega) \cap L^2_0(\Omega) \qquad \text{for } k \in \mathbb{N}_0.$$

Then we have $u \in H^{k+2}(\Omega) \cap H^1_0(\Omega)$, $p \in H^{k+1}(\Omega) \cap L^2_0(\Omega)$ and there exists a C depending only on Ω such that

$$|u|_{k+2} + |p|_{k+1} \le C \left[|f|_k + |g|_{k+1} \right].$$

Proof. Cf. Ladyženskaja [179, Chapter III, §5]. ∎

In analogy to Theorem 9.24 it is sufficient to assume the convexity of Ω in order to obtain $u \in H^2(\Omega)$ and $p \in H^1(\Omega)$ from $f \in L^2(\Omega)$ as proved by Kellogg–Osborn [170].

Theorem 12.24. *Let $\Omega \subset \mathbb{R}^2$ be bounded and convex. If $f \in L^2(\Omega)$, then the Stokes equation (12.2a,b) has a unique solution $u \in \mathbf{H}^2(\Omega) \cap \mathbf{H}^1_0(\Omega)$, $p \in H^1(\Omega) \cap L^2_0(\Omega)$, which satisfy the estimate*

$$|u|_2 + |p|_1 \le C\, |f|_0 .$$

For the more general problem (12.27) with $g \neq 0$ in a convex polygonal domain the solution satisfies

$$|u|_2 + |p|_1 \le C \left[\|f\|_{\mathbf{L}^2(\Omega)} + \|g\|_{\mathbf{H}^1_\delta(\Omega)} \right],$$

if $f \in \mathbf{L}^2(\Omega)$ and $g \in L^2_0(\Omega) \cap H^1_\delta(\Omega)$. Here, $H^1_\delta(\Omega)$ is the subspace of $H^1(\Omega)$ with the following (stronger) norm:

$$\|g\|_{H^1_\delta(\Omega)} := \sqrt{\sum_{|\alpha|=1} \|D^\alpha g\|^2_{L^2(\Omega)} + \|\delta^{-1} g\|^2_{L^2(\Omega)}}$$

with $\delta(\mathbf{x}) := \min \{ |\mathbf{x} - \mathbf{e}| : \mathbf{e} \in \Gamma$ corner of the polygon $\Omega \}$.

12.2.5 A V_0-elliptic Variational Formulation of the Stokes Problem

$V_0 \subset \mathbf{H}_0^1(\Omega)$ has been defined in (12.20) by $V_0 := \{u \in \mathbf{H}_0^1(\Omega) : \operatorname{div} u = 0\}$. As kernel for the mapping $B^\star = -\operatorname{div} \in L(\mathbf{H}_0^1(\Omega), L_0^2(\Omega))$, V_0 is a closed subspace of $\mathbf{H}_0^1(\Omega)$, i.e., again a Hilbert space for the same norm $|\cdot|_1$. In the following we investigate the problem

$$\text{find} \quad u \in V_0 \quad \text{with} \quad a(u,v) = f(v) \quad \text{for all } v \in V_0\,, \tag{12.28}$$

where $a(u,v) := \int_\Omega \langle \nabla u(\mathbf{x}), \nabla v(\mathbf{x})\rangle \mathrm{d}\mathbf{x}$. Problem (12.28) has the same form as the weak formulation of the Poisson equations $-\Delta u_i = f_i$ $(1 \le i \le n)$, only here $\mathbf{H}_0^1(\Omega)$ has been replaced by V_0.

Lemma 12.25. *Let Ω be a bounded domain (at least bounded in one direction; cf. Exercise 6.30b). The bilinear form $a(\cdot,\cdot)$ is V_0-elliptic. The constant $C_E > 0$ in $a(u,u) \ge C_E\,|u|_1^2$ depends only on the diameter of Ω. In particular, problem (12.28) has a unique solution $u \in V_0$ with*

$$|u|_1 \le C_E^{-1}\,|f|_{V_0'} \qquad \textit{for } f \in V_0'\,.$$

Proof. The $H_0^1(\Omega)$-ellipticity of $a(\cdot,\cdot)$ (cf. Lemma 12.17) implies the V_0-ellipticity on the subspace $V_0 \subset H_0^1(\Omega)$ (cf. Exercise 6.98a). This implies the other statements (cf. Theorem 6.101). ∎

Theorem 12.26. *Let $\Omega \in C^{0,1}$ be a bounded domain. Assume $f \in \mathbf{H}^{-1}(\Omega)$. Then the solution $u \in V_0$ of problem (12.28) coincides with the solution component u of the mixed formulation (12.3).*

Proof. Let V_0 and $V_\perp = (V_0)^\perp$ be as in (12.21a). According to Exercise 12.11 one can split $f \in \mathbf{H}^{-1}(\Omega)$ in such a way that

$$f = f_0 + f_\perp\,, \quad f_0 \in V_0'\,, \quad f_\perp \in V'\,, \quad \begin{cases} f_0(v_\perp) = 0 & \text{for } v_\perp \in V_\perp\,, \\ f_\perp(v) = 0 & \text{for } v \in V_0\,. \end{cases}$$

In problem (12.28) one can replace $f(v)$ by $f_0(v)$. $u \in V_0 \subset \mathbf{H}_0^1(\Omega)$ results in $-\Delta u \in \mathbf{H}^{-1}(\Omega)$. The part of $-\Delta u$ that belongs to $V_\perp'$ is $g_\perp \in V_\perp'$ with

$$g_\perp(v_\perp) := a(u, v_\perp) \qquad \text{for all } v_\perp \in V_\perp = (V_0)^\perp\,.$$

Theorem 12.19 proves the condition (12.19), which results in the bijectivity of $B : L_0^2(\Omega) \to V_\perp'$ for $B = \nabla$ (cf. Lemma 12.18). $p := B^{-1}(f_\perp - g_\perp)$, by definition, satisfies

$$b(p, v_\perp) = \langle Bp, v_\perp\rangle_{\mathbf{H}^{-1}(\Omega)\times\mathbf{H}_0^{-1}(\Omega)} = f_\perp(v_\perp) - g_\perp(v_\perp) = f_\perp(v_\perp) - a(u, v_\perp)$$

for all $v_\perp \in V_\perp$. Furthermore, since $\operatorname{div} v_0 = 0$ for $v_0 \in V_0$, it follows that

$$b(p, v_0) = 0 \qquad \text{for all } v_0 \in V_0\,.$$

For arbitrary $v \in \mathbf{H}_0^1(\Omega)$, to be split into $v = v_0 + v_\perp$ with $v_0 \in V_0$ and $v_\perp \in V_\perp$, one obtains

$$\begin{aligned} b(p,v) &= b(p,v_0) + b(p,v_\perp) = f_\perp(v_\perp) - a(u,v_\perp) = f_\perp(v_\perp) - a(u,v) + a(u,v_0) \\ &= f(v_\perp) - a(u,v) + f(v_0) = f(v) - a(u,v) \end{aligned}$$

by (12.28). Since also $b(w,u) = 0$ for all $w \in L_0^2(\Omega)$ because $u \in V_0$, the functions u and p satisfy the variational formulation (12.9a–c) of the Stokes problem. ■

Note that problem (12.28) is solvable for all bounded domains although problem (12.9a–c) depends more sensitively on Ω (cf. Example 12.21). Theorem 12.26 shows that only the component p has a domain-dependent bound $|p|_0 \le C_\Omega\, |f|_{-1}$ while $|u|_1 \le C\,|f_0|_{-1} \le C\,|f|_{-1}$ for all $\Omega \subset K_R(0)$.

Strictly speaking, the variational problem (12.28) is not equivalent to the Stokes problem since, for example, for Ω from Exercise 12.22 the Stokes equations are not solvable, whereas problem (12.28) definitely has a solution.

The original formulation (12.9a–c) may be interpreted as equation (12.28), into which one has introduced the *side condition* $\operatorname{div} u = 0$ using the Lagrange function p (cf. §8.4.6).

12.3 Finite-Element Method for the Stokes Problem

12.3.1 Finite-Element Discretisation of a Saddle-Point Problem

One would have an ordinary Ritz–Galerkin discretisation if in the variational formulation (12.28) one were to replace the space V_0 by a finite-dimensional subspace $V_h \subset V_0$. But this is not as easy as it sounds (more details in §12.3.4).

The remaining procedure is thus oriented toward the weak formulation (12.9a–c). The space $X = V \times W$ is replaced with $X_h = V_h \times W_h$. The discrete problem

$$\begin{aligned} &\text{find} \quad v^h \in V_h \quad \text{and} \quad w^h \in W_h \\ &\text{with} \quad \begin{cases} a(v^h, x) + b(w^h, x) = f_1(x) & \text{for all } x \in V_h\,, \\ b(y, v^h) \qquad\qquad\quad = f_2(y) & \text{for all } y \in W_h\,, \end{cases} \end{aligned} \tag{12.29}$$

is called a *mixed Galerkin problem* [resp. a mixed finite-element] since the side condition $B^* v = f_2$ is included by the Lagrange parameter $w^h \in W_h$. To the

formulation (12.12a,b) corresponds the equivalent way of writing (12.29):

$$\text{find} \quad x^h \in X_h \quad \text{with} \quad c(x^h, z) = f(z) \quad \text{for all } z \in X_h. \tag{12.29$'$}$$

In the case of the Stokes equations, (12.9a–c) provides the desired solution which satisfies the side condition $\operatorname{div} u = 0$. However, the finite-element solution of (12.29) generally does not satisfy the condition $\operatorname{div} u^h = 0$. One can view v^h from (12.29) as the solution of a nonconforming finite-element discretisation of (12.28), as is shown in the following exercise.

Exercise 12.27. Let $f_2 = 0$ in the last equation of (12.29); set $V_{0,h} := \{x \in V_h : b(y, x) = 0 \text{ for all } y \in W_h\}$. Show that each solution v^h in (12.29) is also a solution of

$$\text{find} \quad v^h \in V_{0,h} \quad \text{with} \quad a(v^h, x) = f_1(x) \qquad \text{for all } x \in V_{0,h}\,. \tag{12.30}$$

Since in general $V_{0,h} \not\subset V_0$ (cf. (12.20)), (12.30) is a nonconforming discretisation of (12.28).

Let $\{b_1^V, \ldots, b_{N_{V,h}}^V\}$ and $\{b_1^W, \ldots, b_{N_{W,h}}^W\}$ be suitable bases of V_h and W_h, where

$$N_{V,h} := \dim V_h, \qquad N_{W,h} := \dim W_h.$$

Let the coefficients of $v \in V_h$ and $w \in W_h$ be $\mathbf{v}$ and $\mathbf{w}$:

$$v = P_V \mathbf{v} := \sum_{i=1}^{N_{V,h}} v_i\, b_i^V, \qquad w = P_W \mathbf{w} := \sum_{i=1}^{N_{W,h}} w_i\, b_i^W.$$

As in Section 8.2, we prove the following theorem.

Theorem 12.28. *The variational problem (12.29) is equivalent to the system of equations*

$$\begin{bmatrix} \mathbf{A}_h & \mathbf{B}_h \\ \mathbf{B}_h^{\mathsf{T}} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{v} \\ \mathbf{w} \end{bmatrix} = \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \end{bmatrix} \tag{12.31a}$$

where the matrices and vectors are given by

$$\begin{aligned} \mathbf{A}_{h,ij} &:= a(b_j^V, b_i^V), \quad \mathbf{B}_{h,ik} := b(b_k^W, b_i^V) \\ \mathbf{f}_{1,i} \quad &:= f_1(b_i^V), \qquad \mathbf{f}_{2,k} \quad := f_2(b_k^W) \end{aligned} \quad \textit{for} \quad \begin{cases} 1 \le i, j \le N_{V,h}, \\ 1 \le k \le N_{W,h}\,. \end{cases} \tag{12.31b}$$

The connection between (12.29) and (12.31a,b) is given by $v^h = P_V \mathbf{v}$, $w^h = P_W \mathbf{w}$. For $\mathbf{u} := \binom{\mathbf{v}}{\mathbf{w}}$, $\mathbf{f} := \binom{\mathbf{f}_1}{\mathbf{f}_2}$ one obtains the system of equations

$$\mathbf{C}_h \mathbf{u} = \mathbf{f} \qquad \textit{with } \mathbf{C}_h := \begin{bmatrix} \mathbf{A}_h & \mathbf{B}_h \\ \mathbf{B}_h^{\mathsf{T}} & 0 \end{bmatrix}, \tag{12.31a$'$}$$

which corresponds to the formulation (12.29$'$).

12.3.2 Stability Conditions

When selecting the subspaces V_h and W_h, one has to be careful, for even seemingly reasonable spaces lead to singular or unstable systems of equations (12.29). The former occurs in the following example.

Example 12.29. Let the Stokes problem be posed for the L-shaped domain in Figure 8.2. For all three components u_1, u_2, and p we use piecewise linear triangular elements over the triangulation $\mathcal{T}$ which is given by the second or third picture of Figure 8.2. Here, let $u_1 = u_2 = 0$ on Γ and let $p \in W_h$ satisfy the side condition $p \in L_0^2(\Omega)$, i.e., $\int_\Omega p \mathrm{d}\mathbf{x} = 0$. Then the discrete variational problem (12.29) is not solvable.

Proof. The second [third] triangulation in Figure 8.2 has 5 [10] inner nodes $\mathbf{x}^i$, with each of them carrying values $u_1(\mathbf{x}^i)$ and $u_2(\mathbf{x}^i)$. Thus $\dim V_h = 2 \cdot 5 = 10$ [$\dim V_h = 20$]. Because of the side condition $\int_\Omega p \mathrm{d}\mathbf{x} = 0$ the dimension of W_h is smaller by one than the number of inner and boundary nodes: $\dim W_h = 21 - 1 = 20$ [$\dim W_h = 21$]. In both cases the statement results from the following lemma. ■

Lemma 12.30. *It is necessary for the solvability of (12.29) that* $N_{V,h} \geq N_{W,h}$*, i.e.,* $\dim V_h \geq \dim W_h$.

Proof. According to Theorem 12.28 the solvability of (12.29) is equivalent to the nonsingularity of the matrix $\mathbf{C}_h$ in (12.31a′). Elementary considerations show that $\mathrm{rank}(\mathbf{C}_h) \leq N_{V,h} + \mathrm{rank}(\mathbf{B}_h)$. Since $\mathbf{B}_h$ is an $N_{V,h} \times N_{W,h}$-matrix it follows from $N_{V,h} < N_{W,h}$ that $\mathrm{rank}(\mathbf{C}_h) < N_{V,h} + N_{W,h}$, and hence $\mathbf{C}_h$ is singular. ■

Thus an increase in the dimension of W_h does not always lead to an better approximation of $w \in W$. The choices of V_h and W_h must be mutually adjusted. The inequality $\dim V_h \geq \dim W_h$ corresponds to the requirement that B in (12.16b) needs to be injective, but not necessarily to be surjective.

Since $W = L_0^2(\Omega)$, the finite elements in W_h need not be continuous. Consider a tessellation of $\Omega = (0,1) \times (0,1)$ by a regular square grid of step size h. Define W_h by piecewise constant functions, while V_h uses bilinear elements. One checks that $\dim(V_h) = 2(\frac{1}{h} - 1)^2$ and $\dim(W_h) = (\frac{1}{h})^2 - 1$. Hence the condition $\dim V_h \geq \dim W_h$ is satisfied for $h \leq 1/4$. Nevertheless, this ansatz for V_h and W_h leads to an instability (cf. Braess [45, §III.7]).

In order to formulate the necessary stability conditions, we define

$$V_{0,h} := \{v \in V_h : b(y, v) = 0 \text{ for all } y \in W_h\}$$

(cf. Exercise 12.27). $V_{0,h}$ is the discrete analogue of the space V_0 in (12.20). The conditions, which can be traced back to Brezzi [53], read:

$$\inf_{v_0 \in V_{0,h},\ \|v_0\|_V = 1} \quad \sup_{x_0 \in V_{0,h},\ \|x_0\|_V = 1} |a(v_0, x_0)| \geq \alpha_h > 0, \tag{12.32a}$$

$$\inf_{w \in W_h,\ \|w\|_W = 1} \quad \sup_{x \in V_h,\quad \|v\|_V = 1} |\, b(w, x)\,| \geq \beta_h > 0. \tag{12.32b}$$

Theorem 12.31. *Let (12.10a–c) hold and* $\dim V_h < \infty$. *The Brezzi conditions (12.32a,b) are necessary and sufficient for the solvability of the discrete problem (12.29). The solution* $u^h = (v^h, w^h) \in V_h \times W_h$ *satisfies*

$$\|u^h\|_X := \sqrt{\|v^h\|_V^2 + \|w^h\|_W^2} \le C_h \|f\|_{X'} := C_h \sqrt{\|f_1\|_{V'}^2 + \|f_2\|_{W'}^2},$$

where C_h *depends on* α_h, β_h *and on the bounds* C_a *and* C_b *in* $|a(v,x)| \le C_a \|v\|_V \|x\|_V$ *and* $|b(w,x)| \le C_b \|w\|_W \|x\|_V$. *If*

$$\alpha_h \ge \underline{\alpha} > 0, \qquad \beta_h \ge \underline{\beta} > 0$$

holds for all parameters h *of a sequence of discretisations, then the discretisation is said to be stable and* C_h *remains bounded:* $C_h \le \overline{C}$ *for all* h.

Proof. Theorem 12.12 and Lemma 12.14 are applicable with $V_{0,h}$, V_h, W_h instead of V_0, V, W. (12.24a,c) then are the same as (12.32a,b), while (12.24b) follows from (12.24a) because $\dim V_{0,h} < \infty$ (cf. Exercise 6.95). ∎

Condition (12.32a) is trivial for the Stokes problem.

Exercise 12.32. Show that condition (12.32a) is always satisfied with a constant α_h independent of h, if $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ is V-elliptic.

It is important to emphasise the role of the uniform (i.e., h-independent) estimates $\alpha_h \ge \underline{\alpha} > 0$, $\beta_h \ge \underline{\beta} > 0$, and $C_h \le \overline{C}$. Obviously such statements require an infinite family of discretisations. If the bounds α_h or β_h are positive but h-dependent, $C_h \le \text{const} \cdot h^{-\alpha}$ may holds with $\alpha > 0$ (cf. Braess [45, (7.8) in §III]). The later error analysis will lead to a consistency error $\mathcal{O}(h^\kappa)$ with $\kappa > 0$. The final discretisation error is the product $\mathcal{O}(h^{\kappa-\alpha})$. In the case of a weak instability one has $\alpha < \kappa$ and observes a reduced convergence speed. However, if $\alpha \ge \kappa$, the method does not converge.

12.3.3 Stable Finite-Element Spaces for the Stokes Problem

A detailed description of stable finite-element discretisations can be found in the monograph of Brezzi–Fortin [55]. We also refer to Braess [45, §III.7], John [161, §§3.5–3.6], and Girault–Raviart [117, §II].

12.3.3.1 Stability Criterion

For the Stokes problem $V_h \subset H_0^1(\Omega)$ and $W_h \subset L_0^2(\Omega)$ must hold. In a bounded domain $a(\cdot,\cdot)$ is $H_0^1(\Omega)$-elliptic so that (12.32a) is satisfied with $\alpha_h \ge C_E > 0$. It is somewhat more difficult to prove the conditions (12.32b,c), which for the Stokes problem assume the form

$$\sup_{u \in V_h,\ |u|_1 = 1} |b(p,u)| \ge \underline{\beta}\ |p|_0 \qquad \text{for all } p \in W_h \tag{12.33}$$

wherein $\beta > 0$ must be independent of p and h. It is simpler to prove the modified condition

$$\sup_{u\in V_h,\ |u|_0=1} |b(p,u)| \geq \tilde{\beta}\ |p|_1 \qquad \text{for all } p \in W_h \tag{12.34}$$

($\tilde{\beta} > 0$ independent of p and h) in which the roles of the $|\cdot|_0$ and $|\cdot|_1$ norms are interchanged. Since in (12.34) the norm $|p|_1$ occurs, the latter requires $W_h \subset H^1(\Omega)$. This excludes, for example, piecewise constant finite elements.

The following condition (12.35a) is the result of Remark 9.27. The inverse inequality (12.35b) holds for uniform grids, while Theorem 9.28 yields the estimate (12.35c) of the $\mathbf{L}^2(\Omega)$-orthogonal projection Q_0.

Theorem 12.33. *Let $\Omega \in C^{0,1}$ be bounded. Suppose that V_h satisfies*

$$\inf_{u^h\in V_h} |u - u^h|_0 \leq C_1 h\,|u|_1 \qquad \textit{for all } u \in \mathbf{H}_0^1(\Omega), \tag{12.35a}$$

$$|u^h|_1 \leq C_i h^{-1} |u^h|_0 \quad \textit{for all } u^h \in V_h\,, \tag{12.35b}$$

$$|Q_0|_{1\leftarrow 1} \leq C_0 \tag{12.35c}$$

Then condition (12.34) is sufficient for the Brezzi condition (12.32b,c).

Proof. (i) For given $p \in W_h$ there exists $u \in \mathbf{H}_0^1(\Omega)$ with $|u|_1 = 1$ and $b(p,u) \geq \beta\,|p|_0$ (cf. Theorem 12.19 and (12.26a)). According to Exercise 9.26, the orthogonal $\mathbf{L}^2(\Omega)$-projection $u^h := Q_0 u$ on V_h satisfies the conditions

$$u = u^h + e, \qquad |u^h|_1 \leq C_0 |u|_1 = C_0, \qquad |e|_0 \leq C_1 h |u|_1 = C_1 h.$$

Here $|u^h|_1 \leq C_0|u|_1$ follows from (12.35c) and $|e|_0 \leq C_1 h|u|_1$ from (12.35a). From $b(p,u^h) = b(p,u) - b(p,e) \geq \beta\,|p|_0 - b(p,e) \geq \beta\,|p|_0 - |p|_1\,|e|_0 \geq \beta\,|p|_0 - C_1 h\,|p|_1$ and $|u^h|_1 \leq C_0$ one infers

$$\sup_{v^h\in V_h,\ |v^h|_1=1} |b(p,v^h)| \geq \frac{|b(p,u^h)|}{C_0} \geq \frac{\beta\ |p|_0 - C_1 h\,|p|_1}{C_0}. \tag{12.36}$$

(ii) Because of (12.34) and (12.35b) there exists $u^\star \in V_h$ with $|u^\star|_0 = 1$ and

$$|b(p,u^\star)| \geq \beta'\ |p|_1 = \beta'\ |p|_1\,|u^\star|_0 \geq \frac{\beta'}{C_i} h\,|p|_1\,|u^\star|_1\,.$$

From this follows

$$\sup_{v^h\in V_h,\ |v^h|_1=1} \big|b(p,v^h)\big| \geq \hat{\beta} h\,|p|_1 \qquad \text{with} \quad \hat{\beta} := \beta'/C_i\,. \tag{12.37}$$

If one multiplies (12.36) by C_0/C_1 and (12.37) by $1/\hat{\beta}$ the sum reads:

$$\sup_{v^h\in V_h \text{ mit } |v^h|_1=1} \big|b(p,v^h)\big| \geq \underline{\beta}\,|p|_0 \qquad \text{with } \underline{\beta} := \frac{\beta/C_1}{C_0/C_1 + 1/\hat{\beta}} = \frac{\beta\hat{\beta}}{C_0\hat{\beta} + C_1}.$$

Since $\underline{\beta}$ is independent of p and h, (12.33) [i.e., (12.32b,c)] has been proved. ∎

12.3.3.2 Finite-Element Discretisations with the Bubble Function

In the following let Ω be a polygonal domain and $\mathcal{T}_h$ an admissible triangulation. Example 12.29 shows that linear triangular elements make no sense for u and p. We increase the dimension of V_h by adding to it the so-called *bubble functions* or 'bulb functions'.

On the reference triangle $T = \{(\xi,\eta) : \xi,\eta > 0, \xi+\eta < 1\}$ the bubble function is defined by

$$u(\xi,\eta) := \xi\eta(1-\xi-\eta) \quad \text{in } T, \qquad u = 0 \quad \text{otherwise.}$$

The name derives from the fact that u is positive only in T and vanishes on ∂T and outside. The map $\phi : T \to \tilde{T}$ to a general $\tilde{T} \in \mathcal{T}_h$ (cf. Exercise 8.43) results in the expression

$$\tilde{u}_{\tilde{T}}(x,y) := u(\phi^{-1}(x,y)) \tag{12.38}$$

for the bubble function on $\tilde{T}$.

Exercise 12.34. Prove $\int_{\tilde{T}} \tilde{u}_{\tilde{T}}(x,y)\mathrm{d}x\mathrm{d}y = \frac{1}{60}\operatorname{area}(\tilde{T})$ for all $\tilde{T} \in \mathcal{T}_h$.

We set

$$\begin{aligned} V_h^1 :&\ \begin{cases} \text{linear combinations of the linear elements in } H_0^1(\Omega) \\ \text{and the bubble functions for } \tilde{T} \in \mathcal{T}_h, \end{cases} \\ V_h &:= V_h^1 \times V_h^1, \qquad W_h : \text{ linear elements in } L_0^2(\Omega). \end{aligned} \tag{12.39}$$

For the side condition $W_h \subset L_0^2(\Omega)$ see Section 8.4.6. Since $\tilde{u}_{\tilde{T}} \in H_0^1(\Omega)$, we have $V_h \subset \mathbf{H}_0^1(\Omega)$. These finite elements are introduced by Arnold–Brezzi–Fortin [8] under the name *mini elements*.

Theorem 12.35. *Let $\mathcal{T}_h$ be the quasi-uniform triangulation on a bounded polygonal domain Ω. Let V_h and W_h be given by (12.39). Then the stability condition (12.34) is satisfied.*

Proof. Choose an arbitrary $p \in W_h$. On every $\tilde{T} \in \mathcal{T}_h$, ∇p is constant: $\nabla p = (p_x|_{\tilde{T}}, p_y|_{\tilde{T}})$. We set

$$v := \sum_{\tilde{T}\in\mathcal{T}_h} \begin{pmatrix} p_x|_{\tilde{T}}\ \tilde{u}_{\tilde{T}} \\ p_y|_{\tilde{T}}\ \tilde{u}_{\tilde{T}} \end{pmatrix} \in V_h, \quad \tilde{v} := \frac{1}{|v|_0} v \quad (\tilde{u}_{\tilde{T}} \text{ bubble function (12.38) on } \tilde{T}),$$

so that $|\tilde{v}|_0 = 1$. Exercise 12.34 yields

$$\begin{aligned} b(p,v) &= \int_\Omega \langle \nabla p, v\rangle \,\mathrm{d}\mathbf{x} = \sum_{\tilde{T}\in\mathcal{T}_h} \Big(|p_x|_{\tilde{T}}|^2 + |p_y|_{\tilde{T}}|^2\Big) \int_{\tilde{T}} \tilde{u}_{\tilde{T}} \mathrm{d}x\mathrm{d}y \\ &\geq \frac{1}{60} \sum_{\tilde{T}\in\mathcal{T}_h} \int_{\tilde{T}} \Big(|p_x|_{\tilde{T}}|^2 + |p_y|_{\tilde{T}}|^2\Big)\,\mathrm{d}x\mathrm{d}y = \frac{1}{60}\,|\nabla p|_0^2 . \end{aligned}$$

Since $|\nabla p|_0$ and $|p|_1$ are equivalent norms on the subspace $H^1(\Omega) \cap L_0^2(\Omega)$ it follows that $|b(p,v)| \geq C|p|_1 |\nabla p|_0$ and $|b(p,\tilde{v})| \geq C\,|p|_1 |\nabla p|_0 / |v|_0$. In a similar way one shows that $|v|_0 \leq C' |\nabla p|_0$ and obtains $|b(p,\tilde{v})| \geq \tilde{\beta}\,|p|_1$ with $\tilde{\beta} := C/C'$ independent of h. The left-hand side in (12.34) is $\geq |b(p,\tilde{v})|$, so that (12.34) follows. ∎

A general result on the stabilisation by bubble functions can be found in Brezzi–Pitkäranta [57].

The bubble functions are examples of basis functions having only one element as support. Another example are piecewise constant functions. In this case the solution of the system of equations can be simplified. The submatrix restricted to these basis functions is diagonal, so that the corresponding variables can be eliminated without filling the matrix (cf. Braess [45, page 99], Schwarz [262, §3.3.1]). The elimination is also called the *static condensation*.

12.3.3.3 Stable Discretisations with Linear Elements in V_h

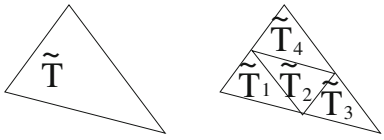


Fig. 12.2 Triangulations T_h and $T_{\frac{h}{2}}$.

If one wants to avoid bubble functions, one must increase the dimension of V_h in some other way. In this section we shall consider for V_h and W_h two different triangulations $\mathcal{T}_{h/2}$ and $\mathcal{T}_h$. By decomposing each $\tilde{T} \in \mathcal{T}_h$ as in Figure 12.2, through halving the sides, into four similar triangles, one obtains $\mathcal{T}_{h/2}$.

We define:

$$\begin{aligned} V_h &\subset \mathbf{H}_0^1(\Omega): \quad \text{linear elements for triangulation } \mathcal{T}_{h/2}, \\ W_h &\subset H^1(\Omega) \cap L_0^2(\Omega): \quad \text{linear elements for triangulation } \mathcal{T}_h, \end{aligned} \tag{12.40}$$

or

$$\begin{aligned} V_h &\subset \mathbf{H}_0^1(\Omega): \quad \text{quadratic elements for triangulation } \mathcal{T}_h, \\ W_h &\subset H^1(\Omega) \cap L_0^2(\Omega): \quad \text{linear elements for triangulation } \mathcal{T}_h. \end{aligned} \tag{12.41}$$

The finite elements in (12.41) are called the *Taylor–Hood elements*.

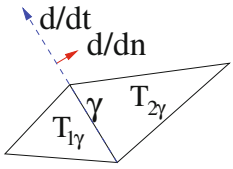


Fig. 12.3 $\gamma, T_{1\gamma}, T_{2\gamma}$.

Theorem 12.36. *Theorem 12.35 holds analogously for $V_h \times W_h$ in (12.40) or (12.41).*

Proof. (i) Let $V_h \times W_h$ be given by (12.40). For each inner triangular side γ of the triangulation $\mathcal{T}_h$ there exist two triangles $T_{1\gamma}, T_{2\gamma} \in \mathcal{T}_h$ with $\gamma = \overline{T_{1\gamma}} \cap \overline{T_{2\gamma}}$ (cf. Figure 12.3). Let $\partial/\partial t$ be the derivative in the direction of γ, let $\partial/\partial n$ be the directional derivative perpendicular to it. There exists a_γ and b_γ with

$$a_\gamma^2 + b_\gamma^2 = 1, \qquad \frac{\partial}{\partial x} = a_\gamma \frac{\partial}{\partial n} + b_\gamma \frac{\partial}{\partial t}, \qquad \frac{\partial}{\partial y} = b_\gamma \frac{\partial}{\partial n} - a_\gamma \frac{\partial}{\partial t}.$$

In contrast to $\partial p/\partial n$, $\partial p/\partial t$ is constant on $T_{1\gamma} \cup T_{2\gamma} \cup \gamma$. We denote its value by $p_t|_\gamma$. The midpoint $\mathbf{x}^\gamma$ of γ is a node of $\mathcal{T}_{h/2}$. We define the piecewise linear function u_γ over $\mathcal{T}_{h/2}$ by its values at the nodes

$$u_\gamma(\mathbf{x}^\gamma) = p_t|_\gamma, \qquad u_\gamma(\mathbf{x}^j) = 0 \quad \text{at the remaining nodes} \tag{12.42a}$$

and set

$$v := \sum_\gamma \begin{pmatrix} b_\gamma \\ -a_\gamma \end{pmatrix} p_t|_\gamma \, u_\gamma \in V_h, \qquad \tilde{v} := \frac{1}{|v|_0} v. \tag{12.42b}$$

The sum $\sum_\gamma$ extends over all interior sides of $\mathcal{T}_h$. In $T_{1\gamma} \cup T_{2\gamma}$ we have $\left\langle \nabla p, \left(\begin{smallmatrix} b_\gamma \\ -a_\gamma \end{smallmatrix}\right) \right\rangle p_t|_\gamma = |p_t|_\gamma|^2$, so that

$$\begin{aligned} b(p,v) &= \sum_\gamma \int_{T_{1\gamma} \cup T_{2\gamma}} \left\langle \nabla p, \left(\begin{smallmatrix} b_\gamma \\ -a_\gamma \end{smallmatrix}\right) p_t|_\gamma \, u_\gamma \right\rangle \mathrm{d}\mathbf{x} = \sum_\gamma |p_t|_\gamma|^2 \int_{T_{1\gamma} \cup T_{2\gamma}} u_\gamma \, \mathrm{d}\mathbf{x} \\ &\geq Ch^2 \sum\nolimits_\gamma |p_t|_\gamma|^2 . \end{aligned}$$

If γ_1, γ_2, γ_3 are the sides of $\tilde{T} \in \mathcal{T}_h$ ($\mathcal{T}_h$ is quasi-uniform!), then $\int_{\tilde{T}} |\nabla p|^2 \, \mathrm{d}\mathbf{x} \leq C' h^2 \sum_{i=1}^3 |p_t|_{\gamma_i}|^2$. From this one infers that $b(p,v) \geq C''' \, |p|_1 \, |\nabla p|_0$, as in the proof of Theorem 12.35, and finishes the proof analogously.

(ii) In the case of quadratic elements given by (12.41) one has the same nodes as in part (i) (cf. Figures 8.8a and 12.2). Use (12.42a,b) to define the quadratic function $v \in V_h$ and carry out the proof as in (i). ∎

12.3.3.4 Error Estimates

In the following, the condition (8.17a) should be replaced by the stability condition (12.32a–c). In the place of the approximation property (8.54′) we now have the inequalities

$$\inf_{v^h \in V_h} \left| v - v^h \right|_1 \leq Ch \, |v|_2 \qquad \text{for all } v \in \mathbf{H}^2(\Omega) \cap \mathbf{H}_0^1(\Omega), \tag{12.43a}$$

$$\inf_{p^h \in W_h} \left| p - p^h \right|_1 \leq Ch \, |p|_1 \qquad \text{for all } p \in H^1(\Omega) \cap L_0^2(\Omega). \tag{12.43b}$$

The following theorem applies to general saddle-point problems, and can be reduced to Theorem 8.21.

Theorem 12.37. *Let $u = \binom{v}{p} \in X = V \times W$ be the solution of (12.11) [resp. (12.32a)]. Let the discrete problem (12.29) with $X_h = V_h \times W_h \subset X$ satisfy the Brezzi condition (12.32a–c) and have the solution $u^h = (v^h, w^h)$. Then there exists a variable C independent of h such that*

$$\|u - u^h\|_X \leq C \inf_{x^h \in X_h} \left| u - x^h \right|_1 . \tag{12.44}$$

Proof. The Brezzi condition (cf. Theorem 12.31) yields $\|u^h\|_X \leq \bar{C}\,\|f\|_{X'}$ for all right-hand sides $f \in X'$ in (12.11), in particular for all $f \in X_h = X_h'$. The above inequality means $\|L_h^{-1}\|_{X_h \leftarrow X_h'} \leq \bar{C}$ for the operator $L_h : X_h \to X_h'$ which belongs to $c(\cdot,\cdot) : X_h \times X_h \to \mathbb{R}$. According to §8.2.3.1, $\|L_h^{-1}\|_{X_h \leftarrow X_h'} \leq \bar{C}$ is equivalent to condition (8.17a) for $c(\cdot,\cdot) : X \times X \to \mathbb{R}$ [instead of $a(\cdot,\cdot) : V \times V \to \mathbb{R}$] with $\varepsilon_N = 1/\bar{C}$. Theorem 8.21 yields the statement (12.44). ∎

For the Stokes problem with $\binom{u}{p}$, $\binom{u^h}{p^h}$ instead of $u = \binom{v}{p}$, $u^h = \binom{v^h}{p^h}$, inequality (12.44) is now rewritten as follows:

$$|u - u^h|_1^2 + |p - p^h|_0^2 \leq C^2 \inf \left\{ |u - v^h|_1^2 + |p - q^h|_0^2 : v^h \in V_h,\, q^h \in W_h \right\} \tag{12.45a}$$

and

$$|u - u^h|_1 + |p - p^h|_0 \leq \sqrt{2} C \inf \left\{ |u - v^h|_1 + |p - q^h|_0 : v^h \in V_h,\, q^h \in W_h \right\}. \tag{12.45b}$$

Theorem 12.38. *Let the Stokes equation (12.2a,b) have a solution with the components $u \in \mathbf{H}^2(\Omega) \cap \mathbf{H}_0^1(\Omega)$, $p \in H^1(\Omega) \cap L_0^2(\Omega)$ (cf. Theorem 12.24). For the subspaces $V_h \subset \mathbf{H}_0^1(\Omega)$ and $W_h \subset L_0^2(\Omega)$ let the Brezzi condition (12.32a–c) and the approximation conditions (12.43a,b) be satisfied. Then the discrete solution u^h, p^h satisfies the estimate*

$$|u - u^h|_1 + |p - p^h|_0 \leq C' h \left[|u|_2 + |p|_1 \right]. \tag{12.46}$$

Proof. Combine inequalities (12.45a,b) and (12.43a,b). ∎

Using the same reasoning as in the second proof for Theorem 8.65 with $c(\cdot,\cdot)$ instead of $a(\cdot,\cdot)$, one proves the following theorem.

Theorem 12.39. *For each $f \in \mathbf{L}^2(\Omega)$, $g \in L_0^2(\Omega) \cap H^1(\Omega)$ let the Stokes problem $-\Delta u + \nabla p = f$, $-\operatorname{div} u = g$ have a solution*

$$u \in \mathbf{H}^2(\Omega) \cap \mathbf{H}_0^1(\Omega), \quad p \in H^1(\Omega) \cap L_0^2(\Omega)$$

with $|u|_2 + |p|_1 \leq C[|f|_0 + |g|_1]$. Under the assumptions of Theorem 12.38 we then have the estimates

$$\begin{aligned} |u - u^h|_0 + |p - p^h|_{-1} &\leq C'\, h\, [\, |u|_1 + |p|_0 \,], \\ |u - u^h|_0 + |p - p^h|_{-1} &\leq C'' h^2 [\, |u|_2 + |p|_1 \,] \end{aligned} \tag{12.47}$$

for the finite-element solutions. Here $|p|_{-1}$ is the dual norm for $H^1(\Omega) \cap L_0^2(\Omega)$.

Corollary 12.40. Combining (12.47) and (12.46) one obtains

$$|u - u^h|_0 + h|p - p^h|_0 \leq C' h^2 \left[|u|_2 + |p|_1 \right].$$

12.3.4 Divergence-Free Elements

We return to the formulation of the Stokes problem in §12.2.5. There the V_0-elliptic problem $a(u,v) = f_1(v)$ has to be solved in the space $V_0 = \{u \in \mathbf{H}^1_0(\Omega) : \operatorname{div} u = 0\} \subset \mathbf{H}^1_0(\Omega)$ of the divergence-free functions (cf. (12.20)). The (conforming) Galerkin method requires a subspace $V_N \subset V_0$.

An example for such a V_N in the case of $\Omega = (0,1)\times(0,1)$ can be as follows. Let P_N be the set of polynomials of degree N with (at least) double zero at $x, y \in \{0,1\}$ and set $V_N := \left\{ \begin{pmatrix} p_y \\ -p_x \end{pmatrix} : p \in P_N \right\}$. Because of the double zero the derivatives p_y and $-p_x$ vanish on $\Gamma = \partial\Omega$. The divergence condition $\operatorname{div} u = 0$ follows from $p_{yx} = p_{xy}$. Instead of these global ansatz functions one would like to use finite elements with small support. The construction of such elements however turns out to be complicated (cf. Fortin [104]).

Exercise 12.41. Let the square $\Omega = (0,1) \times (0,1)$ be triangulated regularly as in Figures 8.2 and 8.5a or decomposed into grid squares. Let $V_h^{(1)} \subset H^1_0(\Omega)$ be the space of the finite triangular elements (cf. (8.36)) [resp. of bilinear elements, cf. (8.39b)]. Define the corresponding subspace for the Stokes problem as $V_h := \{u = (u_1, u_2) : u_1, u_2 \in V_h^{(1)} \text{ and } \operatorname{div} u = 0\} \subset V_0$. Show that V_h contains only the null function.

A remedy is the *nonconforming* discretisation. The Crouzeix–Raviart elements are piecewise linear elements on triangles which are continuous only at the midpoints of the side. If the triangle side lies on the boundary Γ, the homogeneous Dirichlet condition also holds only at the midpoint (cf. Crouzeix–Raviart [79] and Braess [45, pages 170f]).

The V_0-elliptic formulation in §12.2.5 was obtained by incorporating the side condition (12.1b) of the Stokes equations into the subspace V_0. The same idea can be applied to the discrete formulation (12.29) with $f_2 = 0$. The corresponding subspace is

$$V_{0,h} := \left\{ v^h \in V_h : b(y, v^h) = 0 \text{ for all } y \in W_h \right\}.$$

Its elements are called *weakly divergence-free* since $b(y, v^h) = 0$ for all $y \in W_h$ is the weak formulation of $\operatorname{div} v^h = 0$. One difficulty of the practical implementation is the construction of a basis (with possibly small support). Severe problems arise for domains of genus larger than zero. Literature about this subject can be found in Griffiths [122], Gustafson–Hartman [128], Ye–Hall [311], John [161, §4.6], Carrero–Cockburn–Schötzau [64], and Le Borne [186].

Appendix A
Solution of the Exercises

Exercises of Chapter 1

Solution of Exercise 1.5. Introduce the new independent variables

$$\xi = x + y, \qquad \eta = x - y. \tag{A.1}$$

The substitution rule gives

$$\frac{\partial}{\partial x} = \xi_x \frac{\partial}{\partial \xi} + \eta_x \frac{\partial}{\partial \eta}, \qquad \frac{\partial}{\partial y} = \xi_y \frac{\partial}{\partial \xi} + \eta_y \frac{\partial}{\partial \eta}. \tag{A.2}$$

(A.1) implies $\xi_x = \xi_y = \eta_x = 1$ and $\eta_y = -1$. Insertion into (A.2) yields

$$\frac{\partial}{\partial x} = \frac{\partial}{\partial \xi} + \frac{\partial}{\partial \eta}, \qquad \frac{\partial}{\partial y} = \frac{\partial}{\partial \xi} - \frac{\partial}{\partial \eta}.$$

Correspondingly the second derivatives are

$$\frac{\partial^2}{\partial x^2} = \left(\frac{\partial}{\partial \xi} + \frac{\partial}{\partial \eta}\right)^2 = \frac{\partial^2}{\partial \xi^2} + 2\frac{\partial}{\partial \xi}\frac{\partial}{\partial \eta} + \frac{\partial^2}{\partial \eta^2}, \qquad \frac{\partial^2}{\partial y^2} = \frac{\partial^2}{\partial \xi^2} - 2\frac{\partial}{\partial \xi}\frac{\partial}{\partial \eta} + \frac{\partial^2}{\partial \eta^2}.$$

Substituting x, y by $(\xi + \eta)/2, (\xi - \eta)/2$, the function $u(x, y)$ becomes $U(\xi, \eta) := u(\frac{\xi+\eta}{2}, \frac{\xi-\eta}{2})$. The differential equation (1.6) (this is $u_{xx} - u_{yy} = 0$) is now written as

$$0 = \left(\frac{\partial^2}{\partial \xi^2} + 2\frac{\partial}{\partial \xi}\frac{\partial}{\partial \eta} + \frac{\partial^2}{\partial \eta^2}\right)U - \left(\frac{\partial^2}{\partial \xi^2} - 2\frac{\partial}{\partial \xi}\frac{\partial}{\partial \eta} + \frac{\partial^2}{\partial \eta^2}\right)U = 4\frac{\partial}{\partial \xi}\frac{\partial}{\partial \eta}U(\xi, \eta).$$

Hence $\frac{\partial}{\partial \eta}U(\xi, \eta)$ is constant with respect to ξ, i.e.,

$$\frac{\partial}{\partial \eta}U(\xi, \eta) = a(\eta)$$

W. Hackbusch, *Elliptic Differential Equations*, Springer Series
in Computational Mathematics 18, DOI 10.1007/978-3-662-54961-2

only depends on η. Integration over η yields

$$U(\xi,\eta) = U(\xi,0) + \int_0^\eta a(\eta)\mathrm{d}\eta.$$

Therefore the functions

$$\varphi(\xi) := U(\xi,0), \qquad \psi(\eta) := \int_0^\eta a(\eta)\mathrm{d}\eta$$

produce the sum

$$U(\xi,\eta) = \varphi(\xi) + \psi(\eta).$$

Inserting (A.1), we obtain $u(x,y) = U(x+y, x-y) = \varphi(x+y) + \psi(x-y)$, i.e., any solution of (1.6) is of the form (1.7).

Solution of Exercise 1.7. Taking a look at a formulary (e.g., the Oxford Users' Guide to Mathematics [315, page 186]) we find that

$$\frac{1}{\sqrt{4\pi y}} \int_{-\infty}^{\infty} \exp\left(\frac{-t^2}{4y}\right) \mathrm{d}t = 1 \qquad \text{for } y > 0. \tag{A.3}$$

Substitution $t = \xi - x$ yields the reformulation

$$\begin{aligned} u(x,y) &= \frac{1}{\sqrt{4\pi y}} \int_{-\infty}^{\infty} u_0(\xi) \exp\left(\frac{-(x-\xi)^2}{4y}\right) \mathrm{d}\xi \\ &= u_0(x) + \frac{1}{\sqrt{4\pi y}} \int_{-\infty}^{\infty} [u_0(\xi) - u_0(x)] \exp\left(\frac{-(x-\xi)^2}{4y}\right) \mathrm{d}\xi. \end{aligned}$$

For the proof of $\lim_{y \searrow 0} u(x,y) = u_0(x)$ one has to show that the last term tends to zero as $y \searrow 0$.

Let x and $\varepsilon > 0$ be fixed. Because u_0 is continuous there is a $\delta > 0$ so that $|u_0(\xi) - u_0(x)| \le \varepsilon/2$ for all ξ with $|\xi - x| \le \delta$. We split the integral into three terms:

$$\begin{aligned} I_1(x,y) &:= \frac{1}{\sqrt{4\pi y}} \int_{x-\delta}^{x+\delta} [u_0(\xi) - u_0(x)] \exp\left(\frac{-(x-\xi)^2}{4y}\right) \mathrm{d}\xi, \\ I_2(x,y) &:= \frac{1}{\sqrt{4\pi y}} \int_{-\infty}^{x-\delta} [u_0(\xi) - u_0(x)] \exp\left(\frac{-(x-\xi)^2}{4y}\right) \mathrm{d}\xi, \\ I_3(x,y) &:= \frac{1}{\sqrt{4\pi y}} \int_{x+\delta}^{\infty} [u_0(\xi) - u_0(x)] \exp\left(\frac{-(x-\xi)^2}{4y}\right) \mathrm{d}\xi. \end{aligned}$$

The first integral is bounded by

$$
\begin{aligned}
|I_1(x,y)| &\le \frac{1}{\sqrt{4\pi y}} \int_{x-\delta}^{x+\delta} |u_0(\xi) - u_0(x)| \exp\left(\frac{-(x-\xi)^2}{4y}\right) \mathrm{d}\xi \\
&\le \frac{\varepsilon/2}{\sqrt{4\pi y}} \int_{x-\delta}^{x+\delta} \exp\left(\frac{-(x-\xi)^2}{4y}\right) \mathrm{d}\xi \\
&\le \frac{\varepsilon/2}{\sqrt{4\pi y}} \int_{-\infty}^{\infty} \exp\left(\frac{-(x-\xi)^2}{4y}\right) \mathrm{d}\xi \underset{\text{(A.3)}}{=} \frac{\varepsilon}{2}.
\end{aligned}
$$

Set $C := \sup_{x\in\mathbb{R}} |u_0(x)| < \infty$. Then I_2 is bounded by

$$
\begin{aligned}
|I_2(x,y)| &\le \frac{1}{\sqrt{4\pi y}} \int_{-\infty}^{x-\delta} |u_0(\xi) - u_0(x)| \exp\left(\frac{-(x-\xi)^2}{4y}\right) \mathrm{d}\xi \\
&\le \frac{2C}{\sqrt{4\pi y}} \int_{-\infty}^{x-\delta} \exp\left(\frac{-(x-\xi)^2}{4y}\right) \mathrm{d}\xi \underset{\text{substitution by } t=(x-\xi)/\sqrt{4y}}{=} \\
&= \frac{2C}{\sqrt{\pi}} \int_{\delta/\sqrt{4y}}^{\infty} \exp\left(-t^2\right) \mathrm{d}t.
\end{aligned}
$$

Since the improper integral $\int_{-\infty}^{\infty} \exp\left(-t^2\right) \mathrm{d}t$ exists, $\int_R^{\infty} \exp\left(-t^2\right) \mathrm{d}t \to 0$ holds for $R \to \infty$. For a sufficiently small $y > 0$ we have $|I_2(x,y)| \le \frac{\varepsilon}{4}$. Also I_3 has the bound $|I_3(x,y)| \le \frac{\varepsilon}{4}$. Altogether,

$$
\left| \frac{1}{\sqrt{4\pi y}} \int_{-\infty}^{\infty} [u_0(\xi) - u_0(x)] \exp\left(\frac{-(x-\xi)^2}{4y}\right) \mathrm{d}\xi \right| \le \frac{\varepsilon}{2} + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \varepsilon
$$

holds for sufficiently small $y > 0$. Since x and ε are arbitrarily chosen, we have for all x that

$$
\lim_{y \searrow 0} \frac{1}{\sqrt{4\pi y}} \int_{-\infty}^{\infty} [u_0(\xi) - u_0(x)] \exp\left(\frac{-(x-\xi)^2}{4y}\right) \mathrm{d}\xi = 0.
$$

The proof even yields a stronger result.

Remark A.1. If u_0 is uniformly continuous then $\lim_{y\searrow 0} u(\cdot, y) = u_0$ converges uniformly.

Solution of Exercise 1.16. We have to reformulate the differential equation (1.16) by means of derivatives in $\xi_i = \Phi_i(\mathbf{x})$. The chain rule yields

$$
\frac{\partial}{\partial x_i} = \sum_{\alpha=1}^{n} \frac{\partial \xi_\alpha}{\partial x_i} \frac{\partial}{\partial \xi_\alpha} = \sum_{\alpha=1}^{n} \frac{\partial \Phi_\alpha(\mathbf{x})}{\partial x_i} \frac{\partial}{\partial \xi_\alpha}, \tag{A.4a}
$$

while the product rule shows

$$\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j} = \sum_{\alpha,\beta=1}^{n} \frac{\partial \Phi_\alpha(\mathbf{x})}{\partial x_i}\frac{\partial \Phi_\beta(\mathbf{x})}{\partial x_j}\frac{\partial}{\partial \xi_\alpha}\frac{\partial}{\partial \xi_\beta} + \sum_{\alpha=1}^{n} \frac{\partial^2 \Phi_\alpha(\mathbf{x})}{\partial x_i \partial x_j}\frac{\partial}{\partial \xi_\alpha}. \qquad \text{(A.4b)}$$

The principal part $\sum_{i,j=1}^{n} a_{ij}(\mathbf{x})\, u_{x_i x_j}$ becomes

$$\sum_{i,j=1}^{n}\sum_{\alpha,\beta=1}^{n} a_{ij}(\mathbf{x})\,\frac{\partial \Phi_\alpha(\mathbf{x})}{\partial x_i}\frac{\partial \Phi_\beta(\mathbf{x})}{\partial x_j} u_{\xi_\alpha \xi_\beta} + \sum_{i,j=1}^{n}\sum_{\alpha=1}^{n} a_{ij}(\mathbf{x})\frac{\partial^2 \Phi_\alpha(\mathbf{x})}{\partial x_i \partial x_j} u_{\xi_\alpha}$$

with $\mathbf{x} = \Phi^{-1}(\boldsymbol{\xi})$, where only the first double sum belongs to the new principal part. $S := \partial\Phi/\partial\mathbf{x} = (\frac{\partial \Phi_\alpha}{\partial x_i})_{\alpha,i=1,\dots,n}$ is the functional matrix. Therefore the new principal part is $\sum_{\alpha,\beta=1}^{n} b_{\alpha\beta} u_{\xi_\alpha \xi_\beta}$ with

$$B = S A S^{\mathsf{T}}.$$

The signature of a matrix is the triple of the numbers of negative, vanishing, and positive eigenvalues. Sylvester's law of inertia states that A and SAS^{T} have the same signature. Since the definition of types is based on the signature, the type is invariant under the transformation.

Solution of Exercise 1.20. By assumption, $\varphi \in C^0[0,1]$ has a representation as the absolutely convergent Fourier series $\sum_{\nu=1}^{\infty} \alpha_\nu \sin(\nu\pi x)$. Hence $C := \sup|\alpha_\nu|$ is finite. We define the function

$$u(x,y) := \sum_{\nu=1}^{\infty} \frac{\alpha_\nu}{\sinh(\nu\pi)} \sin(\nu\pi x)\sinh(\nu\pi y) \qquad \text{for } 0 \le x \le 1,\ 0 \le y < 1$$

and recall the definition $\sinh(x) = \frac{1}{2}\,(\mathrm{e}^x - \mathrm{e}^{-x})$ of the hyperbolic sine. Since

$$\frac{\sinh(A)}{\sinh(B)} = \exp(A-B)\frac{\exp(2B) - \exp(2(B-A))}{\exp(2B)-1} < \exp(A-B) \quad \text{for } A < B,$$

the coefficients $\beta_\nu(y) := \dfrac{\alpha_\nu}{\sinh(\nu\pi)}\sinh(\nu\pi y)$ decay exponentially for $y < 1$: $|\beta_\nu(y)| \le C\mathrm{e}^{\nu\pi(y-1)}$. Therefore converges $u(x,y) = \sum_{\nu=1}^{\infty} \beta_\nu(y)\sin(\nu\pi x)$ absolutely. Since also $\nu^k\beta_\nu$ decays exponentially, also multiple derivatives converge, i.e., $u(\cdot,\cdot) \in C^\infty([0,1]\times[0,1))$.

Since the Fourier series of φ converges absolutely, $\lim_{y\to 1} u(x,y) = \varphi$ follows. The other boundary values result from $\sin(0) = \sin(\pi) = 0$ and $\sinh(0) = 0$.

It remains to show that the Laplace equation is satisfied for $y < 1$. Because of the exponentially decreasing coefficients differentiation and summation can be interchanged:

$$\Delta u(x,y) = \sum_{\nu=1}^{\infty} \frac{\alpha_\nu}{\sinh(\nu\pi)} \left(\frac{\mathrm{d}^2 \sin(\nu\pi x)}{\mathrm{d}x^2}\sinh(\nu\pi y) + \sin(\nu\pi x)\frac{\mathrm{d}^2 \sinh(\nu\pi y)}{\mathrm{d}y^2} \right)$$

$$= \sum_{\nu=1}^{\infty} \frac{\alpha_\nu (\nu\pi)^2}{\sinh(\nu\pi)} \left(-\sin(\nu\pi x)\sinh(\nu\pi y) + \sin(\nu\pi x)\sinh(\nu\pi y)\right) = 0.$$

Here $(\sinh(t))'' = \sinh(t)$ and $(\sin(t))'' = -\sin(t)$ are used.

Solution of Exercise 1.21. Given sets $X, Y \subset \mathbb{R}$ and a real number $a \in \mathbb{R}$, we introduce the notation

$$X + aY := \{x + ay : x \in X, y \in Y\}, \quad X + a := \{x + a : x \in X\}.$$

First we prove the following statement.

Lemma A.2. *Let $p_1 > p_2 > 0$ be two numbers so that p_1/p_2 is irrational. Assume that the nonempty set A is translation invariant with respect to shifts by p_1 and p_2, i.e., $A + p_1 = A + p_2 = A$. Then the closure is $\bar{A} = \mathbb{R}$.*

Proof. (i) $A + p = A$ implies $A - p = A$ and, more general, $A + mp = A$ for all $m \in \mathbb{Z}$.

(ii) The numbers p_1, p_2 can be extended to a zero sequence $p_k > 0$ with $p_{k+1} \le p_k/2$ for $k \ge 2$ such that $A + p_k = A$ for all k and all ratios p_{k+1}/p_{k+2} are irrational. For this purpose we start with $p_1 > p_2 > 0$. The quotient p_1/p_2 has a unique representation as $n + r$ with $n \in \mathbb{N}_0$ and $r \in (0, 1)$. Set $p_3 := rp_2$. Obviously, $0 < p_3 < p_2$ holds and p_2/p_3 is irrational. From $p_3 = p_1 - np_2$ we infer $A + p_3 = A$. If $p_3 \le p_2/2$, the numbers p_2, p_3 satisfy all required conditions. Otherwise $0 < p_2/2 < p_3 < p_2$ must hold, and $\hat{p}_3 := p_2 - p_3$ satisfies $0 < p_2/2 < \hat{p}_3 < p_2$. Again, one checks that $p_2/\hat{p}_3$ is irrational and $A + \hat{p}_3 = A$.

The construction of p_2, p_3 from p_1, p_2 can be repeated by induction and yields the desired sequence $\{p_k\}$.

(iii) Indirect proof of $\bar{A} = \mathbb{R}$. If $\bar{A} \neq \mathbb{R}$, $\mathbb{R}\backslash\bar{A}$ is open and there is an open interval $I \subset \mathbb{R}\backslash\bar{A}$ of length $L > 0$. Choose some $0 < p_k < L$ from part (ii) and an element $\xi \in A$ (here we need $A \neq \emptyset$). The point set $G := \{\xi + mp_k : m \in \mathbb{Z}\} \subset A$ forms a grid of step size $p_k < L$. Hence at least one point $\xi + mp_k$ lies in the interval I in contradiction to $I \subset \mathbb{R}\backslash\bar{A}$. ∎

The set $A := \mathbb{N} + 2\pi\mathbb{Z}$ is translation invariant with respect 2π, but not with respect to 1, since the inclusion $A + 1 \subset A$ only holds in one direction. By $\mathcal{A}(\cdot)$ we denote the set of accumulation points: $\dot{A} := \mathcal{A}(A)$ and recall their definition: $\alpha \in \dot{A}$ holds if and only if each neighbourhood of α contains infinitely many elements of A.

$A + 2\pi = A$ implies $\dot{A} + 2\pi = \dot{A}$. In addition we have the following statement.

Lemma A.3. $A = \mathbb{N} + 2\pi\mathbb{Z}$ *satisfies* $\dot{A} + 1 = \dot{A}$.

Proof. (i) Obviously, $\mathcal{A}(A + a) = \dot{A} + a$ holds for all $a \in \mathbb{R}$.

(ii) $A + 1 \subset A$ implies $\mathcal{A}(A + 1) \subset \mathcal{A}(A) = \dot{A}$. According to part (i), the reverse inclusion $\mathcal{A}(A + 1) \supset \dot{A}$ is equivalent to $\mathcal{A}(A - 1) \subset \dot{A}$. Let α be an

accumulation point of $A-1$, i.e., each neighbourhood $U \subset (\alpha-\pi, \alpha+\pi)$ of α contains infinitely many elements $a_k - 1$ of $A-1$. The numbers a_k are of the form $n_k + 2\pi m_k$ with $n_k \in \mathbb{N}$ and $m_k \in \mathbb{Z}$. U contains at most one $a_k = n_k + 2\pi m_k$ with $n_k = 1$. The remaining infinitely many $a_k - 1$ satisfy $n_k - 1 \in \mathbb{N}$ and therefore lie in A so that $\alpha \in \dot{A}$. This proves $\mathcal{A}(A-1) \subset \dot{A}$. ∎

One requirement of Lemma A.2 with $A = \mathbb{N}+2\pi\mathbb{Z}$ is still to be proved.

Lemma A.4. *$\dot{A} = \mathcal{A}(\mathbb{N}+2\pi\mathbb{Z})$ is not empty.*

Proof. For each $n \in \mathbb{N}$ there is exactly one $r_n \in [0, 2\pi)$ with $n - r_n \in 2\pi\mathbb{Z}$. Since π is irrational, all r_n are different. By definition of A all r_n belong to A. The infinitely many and bounded r_n must possess an accumulation point so that $\dot{A}$ is not empty. ∎

Now we can apply Lemma A.2 and obtain $\overline{\dot{A}} = \mathbb{R}$. Since $\dot{A}$ is closed (limits of accumulation points are again accumulation points), we obtain $\dot{A} = \mathbb{R}$. Since $\dot{A}$ always satisfies $\dot{A} \subset \bar{A}$ also $\bar{A} = \mathbb{R}$ is proved.

Finally, we state that $\sin\nu = \sin(\nu + 2\pi m)$ holds for all $m \in \mathbb{Z}$. Hence $\{\sin\nu : \nu \in \mathbb{N}\} = \{\sin\alpha : \alpha \in A\} = \sin(A)$. Continuity of $\sin(\cdot)$ shows

$$\overline{\sin(A)} = \sin(\bar{A}) = \sin(\mathbb{R}) = [-1, 1].$$

Exercises of Chapter 2

Solution of Exercise 2.6. Applying the chain rule to $F(\mathbf{y}) := f(U^{\mathsf{T}}(\mathbf{y}-\mathbf{z}))$ yields

$$\frac{\partial^2}{\partial y_i^2} F(\mathbf{y}) = \sum_{\alpha,\beta=1}^{n} \frac{\partial}{\partial x_\alpha} \frac{\partial}{\partial x_\beta} f(\mathbf{x})\, U_{i\alpha}\, U_{i\beta} \qquad \text{with } \mathbf{x} = U^{\mathsf{T}}(\mathbf{y}-\mathbf{z}). \tag{A.5}$$

$H = (\frac{\partial}{\partial x_\alpha}\frac{\partial}{\partial x_\beta} f(\mathbf{x}))_{\alpha,\beta=1}^{n}$ is the Hessian matrix . The right-hand side in (A.5) is a diagonal element of the matrix UHU^{T}. The sum of the diagonal elements defines the trace of the matrix. This proves $\Delta F = \operatorname{trace}(UHU^{\mathsf{T}})$. The rule $\operatorname{trace}(AB) = \operatorname{trace}(BA)$ yields $\operatorname{trace}(UHU^{\mathsf{T}}) = \operatorname{trace}(HU^{\mathsf{T}}U) = \operatorname{trace}(H) = \Delta f$.

Solution of Exercise 2.5. The polar coordinates in (2.2) are $x = r\cos\varphi$, and $y = r\sin\varphi$. The Jacobian matrix is

$$\frac{\partial(x,y)}{\partial(r,\varphi)} = \begin{bmatrix} x_r & x_\varphi \\ x_r & x_\varphi \end{bmatrix} = \begin{bmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{bmatrix}.$$

The inverse function has the inverse Jacobian: $\frac{\partial(r,\varphi)}{\partial(x,y)} = \left(\frac{\partial(x,y)}{\partial(r,\varphi)}\right)^{-1}$, hence

$$\begin{bmatrix} r_x & r_y \\ \varphi_x & \varphi_y \end{bmatrix} = \frac{\partial(r,\varphi)}{\partial(x,y)} = \begin{bmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{bmatrix}^{-1} = \begin{bmatrix} \cos\varphi & \sin\varphi \\ -\frac{1}{r}\sin\varphi & \frac{1}{r}\cos\varphi \end{bmatrix}.$$

This equation defines the factors in the chain rules $\frac{\partial}{\partial x} = r_x \frac{\partial}{\partial r} + \varphi_x \frac{\partial}{\partial \varphi}$ and $\frac{\partial}{\partial y} = r_y \frac{\partial}{\partial y} + \varphi_y \frac{\partial}{\partial \varphi}$:

$$\frac{\partial}{\partial x} = \cos\varphi \frac{\partial}{\partial r} - \frac{1}{r}\sin\varphi \frac{\partial}{\partial \varphi}, \qquad \frac{\partial}{\partial y} = \sin\varphi \frac{\partial}{\partial r} + \frac{1}{r}\cos\varphi \frac{\partial}{\partial \varphi}.$$

We introduce the abbreviations $s := \sin\varphi$, $c := \cos\varphi$. The product rule shows

$$\begin{aligned}
\frac{\partial}{\partial r}\left(c\frac{\partial}{\partial r} - \frac{1}{r}s\frac{\partial}{\partial \varphi}\right) &= c\frac{\partial^2}{\partial r^2} + \frac{1}{r^2}s\frac{\partial}{\partial \varphi} - \frac{1}{r}s\frac{\partial}{\partial r}\frac{\partial}{\partial \varphi}, \\
\frac{\partial}{\partial \varphi}\left(c\frac{\partial}{\partial r} - \frac{1}{r}s\frac{\partial}{\partial \varphi}\right) &= -s\frac{\partial}{\partial r} - \frac{1}{r}c\frac{\partial}{\partial \varphi} + c\frac{\partial}{\partial \varphi}\frac{\partial}{\partial r} - \frac{1}{r}s\frac{\partial^2}{\partial \varphi^2}.
\end{aligned}$$

The double x-derivative is

$$\begin{aligned}
\frac{\partial^2}{\partial x^2} &= \left(c\frac{\partial}{\partial r} - \frac{1}{r}s\frac{\partial}{\partial \varphi}\right)\left(c\frac{\partial}{\partial r} - \frac{1}{r}s\frac{\partial}{\partial \varphi}\right) \\
&= c\left(\frac{s}{r^2}\frac{\partial}{\partial \varphi} + c\frac{\partial^2}{\partial r^2} - \frac{s}{r}\frac{\partial^2}{\partial \varphi \partial r}\right) - \frac{s}{r}\left(-s\frac{\partial}{\partial r} - \frac{c}{r}\frac{\partial}{\partial \varphi} + c\frac{\partial^2}{\partial \varphi \partial r} - \frac{s}{r}\frac{\partial^2}{\partial \varphi^2}\right) \\
&= \frac{1}{r}s^2\frac{\partial}{\partial r} + \frac{2}{r^2}sc\frac{\partial}{\partial \varphi} + c^2\frac{\partial^2}{\partial r^2} - \frac{2}{r}sc\frac{\partial^2}{\partial \varphi \partial r} + \frac{1}{r^2}s^2\frac{\partial^2}{\partial \varphi^2}.
\end{aligned}$$

The analogous result for $\frac{\partial^2}{\partial y^2}$ reads

$$\frac{\partial^2}{\partial y^2} = \frac{1}{r}c^2\frac{\partial}{\partial r} - \frac{2}{r^2}sc\frac{\partial}{\partial \varphi} + s^2\frac{\partial^2}{\partial r^2} + \frac{2}{r}sc\frac{\partial^2}{\partial \varphi \partial r} + \frac{1}{r^2}c^2\frac{\partial^2}{\partial \varphi^2}.$$

Summation yields $\Delta = \frac{1}{r}\frac{\partial}{\partial r} + \frac{\partial^2}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2}{\partial \varphi^2}$.

In principle, the proof of part (b) of the exercise is similar, but more tedious. Instead we recommend a formulary.

The representation $\Delta = \frac{\partial^2}{\partial r^2} + \frac{n-1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}B$ in part (c) can be proved by means of an ansatz $\Delta = \hat{a}\frac{\partial^2}{\partial r^2} + \hat{b}\frac{\partial}{\partial r} + \hat{c}B$ with functions $\hat{a} = \hat{a}(r,\omega)$, $\hat{b} = \hat{b}(r,\omega)$, $\hat{c} = \hat{c}(r,\omega)$ in r and the angle variable ω. The Laplace operator is rotationally symmetric, i.e., $\Delta F(x) = (\Delta f)(Ux)$ holds for $F(x) := f(Ux)$ with a unitary matrix U. This fact implies that the functions $\hat{a}$, $\hat{b}$, $\hat{c}$ cannot depend on the angle variable. Obviously, $F(x) := f(\alpha x)$ for some $\alpha > 0$ leads to $\Delta F(x) = \alpha^2 \Delta f(\alpha x)$. The derivatives with respect to r and to an angle variable ω scale like $F_r = \alpha f_r$, $F_{rr} = \alpha^2 f_{rr}$, and $F_\omega = f_\omega$. This proves that $\hat{a} = a$ is constant, while $\hat{b}(r) = b/r$ and $\hat{c}(r) = c/r^2$. The test with $f(x) = \sqrt{\sum_{i=1}^n x_i^2} = r$ yields $\Delta f = \frac{n-1}{r}$. Since f is rotationally invariant, we obtain $\Delta f = \hat{b}$ so that $b = n-1$. A second test with $f(x) = \sum_{i=1}^n x_i^2 = r^2$ shows $\Delta f = 2n$. From $2n = \Delta f = 2\hat{a} + 2\hat{b}r = 2a + 2b$ one concludes $a = n - b = 1$.

Solution of Exercise 2.10. The underlying domain is the ball $\Omega = K_R(\mathbf{y})$. For any $\boldsymbol{\xi} \in \Omega$ we define

$$\boldsymbol{\xi}' = \mathbf{y} + R^2 |\boldsymbol{\xi} - \mathbf{y}|^{-2} (\boldsymbol{\xi} - \mathbf{y}).$$

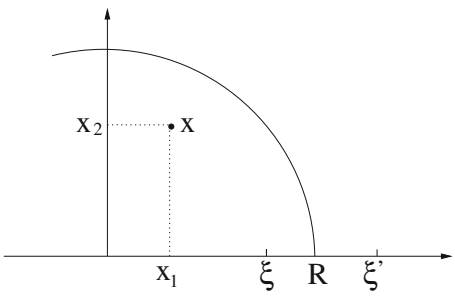


Fig. A.1 Plane containing 0, $\boldsymbol{\xi}$, $\mathbf{x}$.

$\boldsymbol{\xi}'$ is the result of a reflection of $\boldsymbol{\xi}$ at the sphere $\partial\Omega$. Note that $\mathbf{y}$, $\boldsymbol{\xi}$, and $\boldsymbol{\xi}'$ are collinear**.** According to Exercise 2.6, $\mathbf{y}$ may be moved into the origin. Further we rotate the plane spanned by $0, \mathbf{x}$, and $\boldsymbol{\xi}$ so that $\boldsymbol{\xi}$ is the direction of the first axis. We obtain the situation depicted in Figure A.1 with $\mathbf{x} = (x_1, x_2)$, $\boldsymbol{\xi} = (\xi, 0)$, and $\boldsymbol{\xi}' = (\xi', 0)$, where $\xi' = R^2/\xi$. Since we may apply the reflection $\boldsymbol{\xi} \mapsto -\boldsymbol{\xi}$, we assume without loss of generality that $\xi > 0$.

Note that $\boldsymbol{\xi} \in \partial\Omega$ leads to $\boldsymbol{\xi}' = \boldsymbol{\xi}$, i.e., $\xi = \xi' = R$. In this case, $|\mathbf{x} - \boldsymbol{\xi}|$ and $\frac{\xi}{R} \left|\mathbf{x} - \boldsymbol{\xi}'\right|$ coincide and their difference prove part (a): $\gamma(\mathbf{x}, \boldsymbol{\xi}) = 0$.

The first term in $\gamma(\mathbf{x}, \boldsymbol{\xi}) = \dfrac{|\mathbf{x} - \boldsymbol{\xi}|^{2-n} - \left(\frac{|\boldsymbol{\xi} - \mathbf{y}|}{R} \left|\mathbf{x} - \boldsymbol{\xi}'\right|\right)^{2-n}}{(n-2)\,\omega_n}$ $(n > 2)$ is the singularity function $s(\mathbf{x}, \boldsymbol{\xi})$. Up to a constant factor $\frac{|\boldsymbol{\xi} - \mathbf{y}|}{R}$ the second term $\frac{1}{(n-2)\omega_n} \left(\frac{|\boldsymbol{\xi} - \mathbf{y}|}{R} \left|\mathbf{x} - \boldsymbol{\xi}'\right|\right)^{2-n}$ is the singularity function $s(\mathbf{x}, \boldsymbol{\xi}')$. Since the reflected point $\boldsymbol{\xi}'$ lies outside of $\Omega = K_R(\mathbf{y})$, $\Delta s(\mathbf{x}, \boldsymbol{\xi}') = 0$ holds for all $\mathbf{x} \in \Omega$. Thus $\gamma(\mathbf{x}, \boldsymbol{\xi})$ is a fundamental solution in Ω (cf. (2.10)) and part (b) is proved.

The symmetry $\gamma(\mathbf{x}, \boldsymbol{\xi}) = \gamma(\boldsymbol{\xi}, \mathbf{x})$ will be proved in Exercise 3.9 under more general assumptions. For the direct proof we again consider the situation in Figure A.1 where now $\mathbf{x} \in \Omega$ does not lie on the boundary. Both expressions

$$|\mathbf{x} - \boldsymbol{\xi}|^2 = x_1^2 - 2\xi x_1 + \xi^2 + x_2^2 = |\mathbf{x}|^2 - 2\langle \mathbf{x}, \boldsymbol{\xi} \rangle + |\boldsymbol{\xi}|^2 \qquad \text{and}$$

$$\left[\frac{|\boldsymbol{\xi}|}{R} \left|\mathbf{x} - \boldsymbol{\xi}'\right|\right]^2 = \left[\frac{|\boldsymbol{\xi}|}{R}\right]^2 \left(x_1^2 - 2\xi' x_1 + \xi'^2 + x_2^2\right) = \left[\frac{|\boldsymbol{\xi}|}{R}\right]^2 \left(|\mathbf{x}|^2 - 2\langle \mathbf{x}, \boldsymbol{\xi}' \rangle + \left|\boldsymbol{\xi}'\right|^2\right)$$

$$\underset{\xi' = R^2/\xi}{=} \left[\frac{|\boldsymbol{\xi}|\,|\mathbf{x}|}{R}\right]^2 - 2\langle \mathbf{x}, \boldsymbol{\xi} \rangle + R^2$$

are symmetric in $\mathbf{x}$ and $\boldsymbol{\xi}$ so that the assertion of part (c) is proved.

To determine the normal derivative in part (d) again we consider the situation in Figure A.1, i.e., $\frac{\partial}{\partial n_\xi} \gamma(\mathbf{x}, \boldsymbol{\xi}) = -\frac{1}{R\omega_n} \frac{R^2 - |\mathbf{x}|^2}{|\mathbf{x} - \boldsymbol{\xi}|^n}$ is to be shown. At the boundary $\xi = R$ the normal direction coincides with the first axis $x_1 = \xi$, i.e., $\frac{\partial}{\partial n_\xi} = \frac{\partial}{\partial \xi}$. In $\xi = R$ we have

$$\frac{\partial}{\partial \xi} |\mathbf{x} - \boldsymbol{\xi}|^2 = -2x_1 + 2\xi = 2(R - x_1) \qquad (\boldsymbol{\xi} = (\xi, 0))$$

and

$$\frac{\partial}{\partial\xi}\left[\frac{|\boldsymbol{\xi}|}{R}\left|\mathbf{x}-\boldsymbol{\xi}'\right|\right]^2=\frac{\partial}{\partial\xi}\left(\left[\frac{\xi\,|\mathbf{x}|}{R}\right]^2-2x_1\xi+R^2\right)$$
$$=2\xi\left[\frac{|\mathbf{x}|}{R}\right]^2-2x_1=2\left[\frac{|\mathbf{x}|^2}{R}-x_1\right].$$

For $n>2$ we conclude

$$\frac{\partial}{\partial\xi}\left|\mathbf{x}-\boldsymbol{\xi}\right|^{2-n}=\frac{\partial\left[|\mathbf{x}-\boldsymbol{\xi}|^2\right]^{1-\frac{n}{2}}}{\partial\xi}=\frac{1-\frac{n}{2}}{\left[|\mathbf{x}-\boldsymbol{\xi}|^2\right]^{\frac{n}{2}}}\frac{\partial}{\partial\xi}\left|\mathbf{x}-\boldsymbol{\xi}\right|^2=(2-n)\,\frac{R-x_1}{|\mathbf{x}-\boldsymbol{\xi}|^n}$$

and

$$\frac{\partial}{\partial\xi}\left[\frac{|\boldsymbol{\xi}|}{R}\left|\mathbf{x}-\boldsymbol{\xi}'\right|\right]^{2-n}=\frac{\partial}{\partial\xi}\left[\left[\frac{|\boldsymbol{\xi}|}{R}\left|\mathbf{x}-\boldsymbol{\xi}'\right|\right]^2\right]^{1-\frac{n}{2}}\underset{\boldsymbol{\xi}=\boldsymbol{\xi}'}{=}\frac{1-\frac{n}{2}}{\left[|\mathbf{x}-\boldsymbol{\xi}|^2\right]^{\frac{n}{2}}}2\left(\frac{|\mathbf{x}|^2}{R}-x_1\right)$$
$$=(2-n)\,\frac{\frac{|\mathbf{x}|^2}{R}-x_1}{|\mathbf{x}-\boldsymbol{\xi}|^n}.$$

Altogether we obtain

$$\frac{\partial\gamma(\mathbf{x},\boldsymbol{\xi})}{\partial n_\xi}=\frac{1}{(n-2)\,\omega_n}\left[\frac{2-n}{|\mathbf{x}-\boldsymbol{\xi}|^n}\left(R-x_1-\left(\frac{|\mathbf{x}|^2}{R}-x_1\right)\right)\right]=\frac{R^2-|\mathbf{x}|^2}{R\,\omega_n\,|\mathbf{x}-\boldsymbol{\xi}|^n}$$

which proves part (d).

Solution of Exercise 2.12. Assume that $u\in C^0(\overline{\Omega})$ possesses the mean-value property in Ω. For a ball $K_R(\mathbf{x})$ contained in Ω we have

$$u(\mathbf{x})=\frac{1}{\omega_n r^{n-1}}\int_{\partial K_r(\mathbf{x})}u(\boldsymbol{\xi})\,\mathrm{d}\Gamma\qquad\text{for all }0<r\le R.\tag{A.6}$$

The volume and surface integrals are connected by

$$\int_{K_R(\mathbf{x})}u(\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi}=\int_0^R\left(\int_{\partial K_r(\mathbf{x})}u(\boldsymbol{\xi})\mathrm{d}\Gamma\right)\mathrm{d}r.\tag{A.7}$$

The second mean-value property follows from

$$\frac{n}{R^n\omega_n}\int_{K_R(\mathbf{x})}u(\boldsymbol{\xi})\,\mathrm{d}\xi=\frac{n}{R^n\omega_n}\int_0^R\left[\int_{\partial K_r(\mathbf{x})}u(\boldsymbol{\xi})\mathrm{d}\Gamma\right]\mathrm{d}r$$
$$\underset{(A.6)}{=}\frac{n}{R^n\omega_n}\int_0^R\omega_n r^{n-1}u(\mathbf{x})\mathrm{d}r=u(\mathbf{x})\cdot\frac{n}{R^n\omega_n}\omega_n\int_0^R r^{n-1}\mathrm{d}r$$

$$= u(\mathbf{x}) \cdot \frac{n}{R^n \omega_n} \omega_n \frac{R^n}{n} = u(\mathbf{x}).$$

For the reverse direction differentiate (A.7) in R:

$$\frac{d}{dR} \int_{K_R(\mathbf{x})} u(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi} = \int_{\partial K_R(\mathbf{x})} u(\boldsymbol{\xi}) \mathrm{d}\Gamma.$$

Assume the second mean-value property $\int_{K_R(\mathbf{x})} u(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi} = \frac{R^n \omega_n}{n} u(\mathbf{x})$. Hence the derivative is $\int_{\partial K_R(\mathbf{x})} u(\boldsymbol{\xi}) \mathrm{d}\Gamma = R^{n-1} \omega_n u(\mathbf{x})$ and proves the (first) mean-value property.

Solution of Exercise 2.19. The difference $u := u_2 - u_1$ is again a harmonic function in Ω with the boundary value $\varphi := \varphi_2 - \varphi_1 \geq 0$. According to Theorem 2.18 u takes its minimum on the boundary: $u(x) \geq \min_\Omega u = \min_\Gamma u = \min_\Gamma \varphi \geq 0$. This proves part (a).

If $\varphi_1(\mathbf{x}) < \varphi_2(\mathbf{x})$ holds in at least one point $\mathbf{x}_0 \in \Gamma = \partial\Omega$, we have $\varphi \geq 0$ on Γ and $\varphi(\mathbf{x}_0) > 0$. Part (a) shows $u(x) \geq 0$. For an indirect proof assume that

$$\Omega_0 := \{\mathbf{x} \in \Omega : u(\mathbf{x}) = 0\}$$

is not empty. Because $u(\mathbf{x}_0) = \varphi(\mathbf{x}_0) > 0$, Ω_0 cannot coincide with Ω, in particular $\partial\Omega_0 \neq \Gamma = \partial\Omega$. One infers that $\partial\Omega_0 \backslash \Gamma$ is not empty. In $\mathbf{x} \in \partial\Omega_0 \backslash \Gamma \subset \Omega$ we have $u(\mathbf{x}) = 0$. Since $\mathbf{x} \in \Omega$ there is an $R > 0$ with $K_R(\mathbf{x}) \subset \Omega$. The second mean-value property yields $0 = \int_{K_R(\mathbf{x})} u(\boldsymbol{\xi}) \, \mathrm{d}\boldsymbol{\xi}$. Since u is continuous and $u \geq 0$, it follows that $u(\boldsymbol{\xi}) = 0$ for all $\boldsymbol{\xi} \in K_R(\mathbf{x})$. Hence $K_R(\mathbf{x}) \subset \Omega_0$ holds in contradiction to $\mathbf{x} \in \partial\Omega_0$. This proves $\Omega_0 = \emptyset$, thus $u(\mathbf{x}) > 0$ in Ω.

Solution of Exercise 2.24. Let $\Gamma = \partial K_R(\mathbf{y})$. Poisson's integral formula (2.15) defines

$$u(\mathbf{x}) = \frac{R^2 - |\mathbf{x} - \mathbf{y}|^2}{R\omega_n} \int_\Gamma \frac{\varphi(\boldsymbol{\xi})}{|\mathbf{x} - \boldsymbol{\xi}|^n} \mathrm{d}\Gamma_\xi \qquad \text{for } \mathbf{x} \in K_R(\mathbf{y})$$

(note that $K_R(\mathbf{y})$ is an *open* ball). We only assume $\varphi \in L^1(\Gamma)$. The function $|\mathbf{x} - \boldsymbol{\xi}|^{-n}$ is analytic in all components x_i of $\mathbf{x}$. Let $D_\mathbf{x}$ be a k-fold partial derivative. $D_\mathbf{x} |\mathbf{x} - \boldsymbol{\xi}|^{-n}$ has a singularity of order $\mathcal{O}(|\mathbf{x} - \boldsymbol{\xi}|^{-n-k}$ in $\mathbf{x} = \boldsymbol{\xi}$. Since $\mathbf{x} \in K_R(\mathbf{y})$ does not lie on the boundary, $D_\mathbf{x} |\mathbf{x} - \boldsymbol{\xi}|^{-n}$ is bounded on Γ and $\int_\Gamma \varphi(\boldsymbol{\xi}) D_\mathbf{x} |\mathbf{x} - \boldsymbol{\xi}|^{-n} \, \mathrm{d}\Gamma_\xi$ exists. Thus $D_\mathbf{x} u(\mathbf{x})$ exists and has the representation

$$D_\mathbf{x} u(\mathbf{x}) = -\frac{D_\mathbf{x} |\mathbf{x} - \mathbf{y}|^2}{R\omega_n} \int_\Gamma \frac{\varphi(\boldsymbol{\xi}) \mathrm{d}\Gamma_\xi}{|\mathbf{x} - \boldsymbol{\xi}|^n} + \frac{R^2 - |\mathbf{x} - \mathbf{y}|^2}{R\omega_n} \int_\Gamma \varphi(\boldsymbol{\xi}) D_\mathbf{x} \frac{1}{|\mathbf{x} - \boldsymbol{\xi}|^n} \, \mathrm{d}\Gamma_\xi \, .$$

This proves $u \in C^\infty(K_R(\mathbf{y}))$. In particular, the proof on page 22 for $D_\mathbf{x} = \Delta$ shows that $\Delta u(\mathbf{x}) = 0$.

Let φ be continuous at $\mathbf{z} \in \Gamma$ (cf. (2.18)). The integral I_0 can be estimated as in (2.20b). The estimation of I_1 must be modified since φ may be unbounded. We use

$$|I_1| \le \frac{2}{\omega_n} \frac{\delta(\varepsilon)}{(\rho/2)^n} \int_{\Gamma_1} |u(\boldsymbol{\xi}) - \varphi(\mathbf{z})| \, \mathrm{d}\Gamma_\xi \le \frac{2}{\omega_n} \frac{\delta(\varepsilon)}{(\rho/2)^n} \int_{\Gamma} |u(\boldsymbol{\xi}) - \varphi(\mathbf{z})| \, \mathrm{d}\Gamma_\xi$$
$$\le \frac{2}{\omega_n} \frac{\delta(\varepsilon)}{(\rho/2)^n} \left(\int_{\Gamma} |u(\boldsymbol{\xi})| \, \mathrm{d}\Gamma_\xi + |\varphi(\mathbf{z})| \int_{\Gamma} \mathrm{d}\Gamma_\xi \right).$$

The round bracket is bounded since $|u(\boldsymbol{\xi})|$ is integrable on Γ and $|\varphi(\mathbf{z})|$ is finite. Thus (2.20d) follows for a suitably chosen $\delta(\varepsilon)$.

Exercises of Chapter 3

Solution of Exercise 3.4. (a) $f \in C^0(\overline{\Omega} \setminus \{\mathbf{x}_0\})$ has a singularity at $\mathbf{x}_0 \in \Omega \subset \mathbb{R}^n$ bounded by $|f(\mathbf{x})| \le C \, |\mathbf{x} - \mathbf{x}_0|^{-s}$ with $s < n$. Choose R so that $K_R(\mathbf{x}_0) \subset \Omega$. Since $\int_\Omega = \int_{\Omega \setminus K_R(\mathbf{x}_0)} + \int_{K_R(\mathbf{x}_0)}$ it is sufficient to investigate $\int_{K_R(\mathbf{x}_0)} f(\mathbf{x}) \mathrm{d}\mathbf{x}$. According to (2.14) the integral can be estimated by

$$\left| \int_{K_R(\mathbf{x}_0)} f(\mathbf{x}) \mathrm{d}\mathbf{x} \right| \le \int_{K_R(\mathbf{x}_0)} |f(\mathbf{x})| \, \mathrm{d}\mathbf{x} \le C \int_{K_R(\mathbf{x}_0)} |\mathbf{x} - \mathbf{x}_0|^{-s} \, \mathrm{d}\mathbf{x}$$
$$= C \int_0^R \left(\int_{\partial K_r(\mathbf{x})} |\mathbf{x} - \mathbf{x}_0|^{-s} \, \mathrm{d}\Gamma_\xi \right) \mathrm{d}r.$$

Since $|\mathbf{x} - \mathbf{x}_0|^{-s} = r^{-s}$ holds on $\partial K_r(\mathbf{x})$ and $\omega_n r^{n-1}$ is the surface measure of $\partial K_r(\mathbf{x})$ (cf. (2.4b)), it follows that $\int_{\partial K_r(\mathbf{x})} |\mathbf{x} - \mathbf{x}_0|^{-s} \, \mathrm{d}\Gamma_\xi = \omega_n r^{n-1-s}$. The assumption $n - s > 0$ yields

$$\int_{K_R(\mathbf{x}_0)} |f(\mathbf{x})| \, \mathrm{d}\mathbf{x} \le C \int_0^R \omega_n r^{n-1-s} \mathrm{d}r = \frac{C\omega_n}{n-s} R^{n-s},$$

i.e., $f \in L^1(\Omega)$.

(b) $f(\mathbf{x}, \boldsymbol{\xi})$ has a movable singularity at $\mathbf{x}_0(\boldsymbol{\xi})$. Because of

$$|f(\mathbf{x}, \boldsymbol{\xi})| \le C \, |\mathbf{x} - \mathbf{x}_0(\boldsymbol{\xi})|^{-s} \qquad \text{with } s < n$$

part (a) shows that the integral $F(\boldsymbol{\xi}) := \int_\Omega f(\mathbf{x}, \boldsymbol{\xi}) \, \mathrm{d}\mathbf{x}$ exists. It remains to prove $F \in C^0(D)$. Let $\varepsilon > 0$. Continuity of $f(\mathbf{x}, \boldsymbol{\xi})$ in the compact set $G_\varepsilon := \overline{\Omega} \setminus K_\varepsilon(\mathbf{x}_0(\boldsymbol{\xi}))$ shows that also $F_\varepsilon(\boldsymbol{\xi}) := \int_{G_\varepsilon} f(\mathbf{x}, \boldsymbol{\xi}) \, \mathrm{d}\mathbf{x}$ is continuous for $\boldsymbol{\xi} \in D$. The estimate $|F_\varepsilon(\boldsymbol{\xi}) - F(\boldsymbol{\xi})| = \left| \int_{K_\varepsilon(\mathbf{x}_0(\boldsymbol{\xi})) \cap \Omega} f(\mathbf{x}, \boldsymbol{\xi}) \, \mathrm{d}\mathbf{x} \right| \le \frac{C\omega_n}{n-s} \varepsilon^{n-s}$ follows as in part (a) and proves the uniform convergence $F_\varepsilon \to F$ $(\varepsilon \to 0)$ in D. Because of the uniform convergence the limit function is again continuous: $F \in C^0(D)$.

Solution of Exercise 3.9. The functions $u(\mathbf{x}) := g(\mathbf{x}, \mathbf{x}')$ and $v(\mathbf{x}) := g(\mathbf{x}, \mathbf{x}'')$ have their singularities at $\mathbf{x}' \in \Omega$ and $\mathbf{x}'' \in \Omega$. Correspondingly, we remove ε-neighbourhoods from Ω: $\Omega_\varepsilon := \Omega \backslash \overline{K_\varepsilon(\mathbf{x}') \cup K_\varepsilon(\mathbf{x}'')}$. Here $\varepsilon > 0$ must be small enough so that both balls are disjoint and contained in Ω. For smooth functions $u, v \in C^2(\overline{\Omega_\varepsilon})$ the Green formula (2.6b) can be applied:

$$\int_{\Omega_\varepsilon} u(\mathbf{x})\, \Delta v(\mathbf{x})\, \mathrm{d}\mathbf{x} = \int_{\Omega_\varepsilon} v(\mathbf{x})\, \Delta u(\mathbf{x})\, \mathrm{d}\mathbf{x} + \int_{\partial\Omega_\varepsilon} \left[u(\mathbf{x}) \frac{\partial v(\mathbf{x})}{\partial n} - v(\mathbf{x}) \frac{\partial u(\mathbf{x})}{\partial n} \right] \mathrm{d}\Gamma.$$

The first two integrals vanish because $\Delta u = \Delta v = 0$ (cf. (3.5). The boundary of Ω_ε consists of $\partial\Omega$ and of the boundaries $\partial K_\varepsilon(\mathbf{x}')$ and $\partial K_\varepsilon(\mathbf{x}'')$. The boundary values on $\partial\Omega$ are $u = v = 0$. The remaining expressions are

$$\int\limits_{\partial K_\varepsilon(\mathbf{x}')} \left[u(\mathbf{x}) \frac{\partial v(\mathbf{x})}{\partial n} - v(\mathbf{x}) \frac{\partial u(\mathbf{x})}{\partial n} \right] \mathrm{d}\Gamma = \int\limits_{\partial K_\varepsilon(\mathbf{x}'')} \left[v(\mathbf{x}) \frac{\partial u(\mathbf{x})}{\partial n} - u(\mathbf{x}) \frac{\partial v(\mathbf{x})}{\partial n} \right] \mathrm{d}\Gamma. \tag{A.8}$$

On the right-hand side of (A.8) we insert $v(\mathbf{x}) = g(\mathbf{x}, \mathbf{x}'')$ and conclude from (2.11) that

$$\begin{aligned} &\int_{\partial K_\varepsilon(\mathbf{x}'')} \left[v(\mathbf{x}) \frac{\partial u(\mathbf{x})}{\partial n} - u(\mathbf{x}) \frac{\partial v(\mathbf{x})}{\partial n} \right] \mathrm{d}\Gamma \\ &= \int_{\partial K_\varepsilon(\mathbf{x}'')} \left[g(\mathbf{x}, \mathbf{x}'') \frac{\partial u(\mathbf{x})}{\partial n} - u(\mathbf{x}) \frac{\partial g(\mathbf{x}, \mathbf{x}'')}{\partial n_x} \right] \mathrm{d}\Gamma_x = u(\mathbf{x}'') = g(\mathbf{x}'', \mathbf{x}'). \end{aligned}$$

On the left-hand side of (A.8) the roles of u and v are interchanged. Correspondingly, (2.11) yields the value $v(\mathbf{x}') = g(\mathbf{x}', \mathbf{x}'')$, which proves the assertion.

Solution of Exercise 3.12. (a) If $n \geq 3$ we have $s(\mathbf{x}, \mathbf{y}) > 0$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, in particular for $\mathbf{y} \in \Omega$ and $\mathbf{x} \in \partial\Omega$. Since $\Phi(\cdot, \mathbf{y})$ is the solution of $\Delta\Phi(\cdot, \mathbf{y}) = 0$ in Ω and $\Phi(\cdot, \mathbf{y}) = -s(\cdot, \mathbf{y}) < 0$ on $\partial\Omega$, Exercise 2.19b shows $\Phi < 0$ for $\mathbf{x}, \mathbf{y} \in \Omega$. Because $s - g = -\Phi > 0$, the inequality (3.9) is proved.

For $n = 2$ inequality (3.9) is not generally true since $s(\mathbf{x}, \mathbf{y}) = -\frac{1}{\omega_2} \log |\mathbf{x} - \mathbf{y}|$ becomes negative for $|\mathbf{x} - \mathbf{y}| > 1$. However, for domains with a diameter below one, this case does not occur and (3.9) holds true.

The reason of the exceptional behaviour of $n = 2$ is the fact that with $s(\mathbf{x}, \mathbf{y})$ also $s(\mathbf{x}, \mathbf{y}) + C$ for any constant C satisfies the characteristic conditions of a singularity function. As soon as $n > 2$ the constant is uniquely determined by $\lim_{|\mathbf{x}-\mathbf{y}| \to \infty} s(\mathbf{x}, \mathbf{y}) = 0$.

(b) The function $\Psi(\mathbf{x}, \mathbf{y}) := g_2(\mathbf{x}, \mathbf{y}) - g_1(\mathbf{x}, \mathbf{y})$ contains no singularity and is harmonic in Ω_1. Its boundary values are $\Psi(\cdot, \mathbf{y}) = g_2(\cdot, \mathbf{y})$ since $g_1 = 0$ on $\partial\Omega_1$. Inequality (3.8) implies $g_2 \geq 0$ on $\partial\Omega_1$. Since $\Omega_1 \subsetneqq \Omega_2$, there are boundary points of Ω_1 lying in the interior of Ω_2. There $g_2 > 0$ holds (cf. (3.8)). Exercise 2.19b shows that $\Psi > 0$ and thus proves part (b).

Solution of Exercise 3.15. (a) If f is locally Hölder-continuous it is in particular continuous. Since D is compact, f must be bounded, i.e., $C := \|f\|_{C(D)} < \infty$.

Let $\varepsilon(\mathbf{x})$ be the radius of the ball $K_\varepsilon(\mathbf{x})$ appearing in the definition of the local Hölder-continuity. Since $\{K_{\varepsilon(\mathbf{x})}(\mathbf{x}) : \mathbf{x} \in D\}$ is a covering of D, compactness implies that D possesses a finite covering by $K_i := K_{\varepsilon(\mathbf{x}_i)}(\mathbf{x}_i)$, $1 \le i \le m$. For each ball K_i there is a Hölder bound $L_i = \|f\|_{C^\lambda(K_i)}$ of f.

$F(\mathbf{x},\mathbf{y}) := \frac{|f(\mathbf{x})-f(\mathbf{y})|}{|\mathbf{x}-\mathbf{y}|^\lambda}$ is defined on $E := \{(\mathbf{x},\mathbf{y}) \in D\times D : \mathbf{x} \ne \mathbf{y}\}$. We have to show that $L := \sup\{F(\mathbf{x},\mathbf{y}) : (\mathbf{x},\mathbf{y}) \in E\}$ is finite. In the positive case f is globally Hölder-continuous with $\|f\|_{C^\lambda(D)} := \max\{C, L\}$.

There is a sequence $(\mathbf{x}_v, \mathbf{y}_v)$ $(v \in \mathbb{N})$ with $F(\mathbf{x}_v, \mathbf{y}_v) \to L$. By compactness we may choose a subsequence (again denoted by $(\mathbf{x}_v, \mathbf{y}_v)$) converging to $(\mathbf{x}^*, \mathbf{y}^*) \in D \times D$. Two cases are to be distinguished.

Case 1: $\mathbf{x}^* \ne \mathbf{y}^*$. Then $F(\mathbf{x}^*, \mathbf{y}^*) \le 2C/|\mathbf{x}^* - \mathbf{y}^*|^\lambda$ proves that $L < \infty$.

Case 2: $\mathbf{x}^* = \mathbf{y}^*$. There is an index $i \in \{1, \dots, m\}$ with $\mathbf{x}^* \in K_i$. Since K_i is open, $\mathbf{x}^*$ has a positive distance from the boundary ∂K_i. For sufficiently large v all $\mathbf{x}_v$, $\mathbf{y}_v$ lie in K_i so that $F(\mathbf{x}_v, \mathbf{y}_v) \le L_i$ and $L \le \max_{1\le i\le m} L_i < \infty$.

A function is called k-fold *locally* Hölder-continuous of for any $\mathbf{x} \in D$ there is a ball $K_\varepsilon(\mathbf{x})$ with $\varepsilon > 0$ so that the k-fold derivatives belong to $C^\lambda(K_\varepsilon(\mathbf{x}) \cap D)$. In the case of (local) Lipschitz-continuity only the value $\lambda = 1$ is admitted.

(b) For part (b) of the exercise we need the following inequality.

Lemma A.5. $1 - t^s \le (1-t)^s$ *holds for all* $0 \le t \le 1$ *and* $s \ge 0$.

Proof. If $0 < t < 1$ both sides of the inequality are positive. We set $f(t) := (1-t)^s - 1 - t^s$. From $\frac{d}{dt} f'(t) = -s[(1-t)^{s-1} + t^{s-1}] < 0$ we conclude that $f(t) \le f(0) = 0$. The remaining case $t \in \{0, 1\}$ is trivial. ∎

First we consider the case $0 < s < 1$ and investigate $(|\mathbf{x}|^s - |\mathbf{y}|^s) / |\mathbf{x} - \mathbf{y}|^s$. This expression is invariant with respect to a scaling of $\mathbf{x}$ and $\mathbf{y}$. Without loss of generality we may assume that $0 \le t := |\mathbf{y}| \le |\mathbf{x}| = 1$. The reversed triangle inequality (6.1) states $|\mathbf{x} - \mathbf{y}| \ge |\mathbf{x}| - |\mathbf{y}| = 1 - t$. Therefore the previous lemma yields the Hölder estimate $\dfrac{|\mathbf{x}|^s - |\mathbf{y}|^s}{|\mathbf{x} - \mathbf{y}|^s} \le \dfrac{1 - t^s}{(1-t)^s} \le 1$. On $\overline{K_R(0)}$ the function $|\mathbf{x}|^s$ is bounded by R^s. This proves $|\mathbf{x}|^s \in C^s(\overline{K_R(0)})$ with $\big\| |\mathbf{x}|^s \big\|_{C^s(\overline{K_R(0)})} \le \max\{1, R^s\}$.

Since the latter inequality also holds for $s = 1$, $|\mathbf{x}| \in C^{0,1}(\overline{K_R(0)})$ has a norm bounded by $\max\{1, R\}$.

If $1 < s < 2$, $|\mathbf{x}|^s$ is differentiable: $\frac{\partial}{\partial x_i} |\mathbf{x}|^s = s x_i |\mathbf{x}|^{s-2}$. More general, the k-fold derivative of $|\mathbf{x}|^s$ with $s = k + \lambda$, $k \in \mathbb{N}$, $0 < \lambda < 1$ is of the form $F(\mathbf{x}) := p(\mathbf{x}) |\mathbf{x}|^{s-2k}$ with a homogeneous polynomial p of degree k (i.e., $p(t\mathbf{x}) = t^k p(\mathbf{x})$). The function $p(\mathbf{x}) |\mathbf{x}|^{s-2k}$ is homogenous of degree $s - k = \lambda$. Hence the expression $|F(\mathbf{x}) - F(\mathbf{y})| / (|\mathbf{x}| - |\mathbf{y}|)^\lambda$ is scaling invariant. Again, without loss of generality, we may assume $|\mathbf{x}| = 1$ and $t := |\mathbf{y}| \le 1$.

Case 1. Assume $t = 1$ and $\mathbf{x}, \mathbf{y} \in \partial K_1(0)$. On the compact surface $\partial K_1(0)$ the function F is differentiable of any order: $F \in C^\infty(\partial K_1(0))$. This implies $F \in C^\lambda(\partial K_1(0))$ with some Hölder constant L_1.

Case 2. Let $0 \le t < 1$ and $\mathbf{y} = t\mathbf{x}$ with $|\mathbf{x}| = 1$. Homogeneity of F implies $|F(\mathbf{x}) - F(\mathbf{y})| / |\mathbf{x} - \mathbf{y}|^\lambda = |F(\mathbf{x})| \left(1 - t^\lambda\right) / (1-t)^\lambda \le |F(\mathbf{x})| \le L_2$, where $L_2 := \max_{\partial K_1(0)} F < \infty$.

Case 3. Consider the general case $|\mathbf{x}| = 1$ and $|\mathbf{y}| = t \le 1$. Set $\hat{\mathbf{x}} := t\mathbf{x}$. The triangle inequality gives $|\mathbf{x} - \mathbf{y}| \le |\mathbf{x} - \hat{\mathbf{x}}| + |\hat{\mathbf{x}} - \mathbf{y}|$. The terms can be bounded by $|\mathbf{x} - \hat{\mathbf{x}}| \le |\mathbf{x} - \mathbf{y}|$ ($\hat{\mathbf{x}}$ is the projection of $\mathbf{x}$ onto $\partial K_t(0)$) and $|\hat{\mathbf{x}} - \mathbf{y}| \le |\mathbf{x} - \mathbf{y}|$. Therefore we obtain the Hölder estimate

$$\begin{aligned} |F(\mathbf{x}) - F(\mathbf{y})| &\le |F(\mathbf{x}) - F(\hat{\mathbf{x}})| + |F(\hat{\mathbf{x}}) - F(\mathbf{y})| \le L_1 |\mathbf{x} - \hat{\mathbf{x}}|^\lambda + L_2 |\hat{\mathbf{x}} - \mathbf{y}|^\lambda \\ &\le (L_1 + L_2) |\mathbf{x} - \mathbf{y}|^\lambda . \end{aligned}$$

Solution of Exercise 3.17. We start with the first derivative $D_x^{(1,0,\dots,0)} = \partial / \partial x_1$ (derivatives $\partial / \partial x_i$ for $i = 2, \dots, n$ are treated analogously). We define $F(\mathbf{x}) := \int_\Omega f(\boldsymbol{\xi}, \mathbf{x}) \mathrm{d}\boldsymbol{\xi}$ and

$$G(\mathbf{x}) := \int_\Omega \frac{\partial}{\partial x_1} f(\boldsymbol{\xi}, \mathbf{x}) \, \mathrm{d}\boldsymbol{\xi} .$$

Since $f_{x_1} \in C^0(A)$ and $|f_{x_1}(\boldsymbol{\xi}, \mathbf{x})| \le C |\mathbf{x} - \boldsymbol{\xi}|^{-s}$, Exercise 3.4b shows that $G \in C^0(\overline{\Omega})$. Let $\mathbf{e}_1 = (1, 0, \dots, 0)$ be the first unit vector and let $t > 0$. The integrations on the right-hand side of

$$\int_0^t G(\mathbf{x} + \tau \mathbf{e}_1) \mathrm{d}\tau = \int_0^t \int_\Omega \frac{\partial}{\partial x_1} f(\boldsymbol{\xi}, \mathbf{x} + \tau \mathbf{e}_1) \, \mathrm{d}\boldsymbol{\xi} \, \mathrm{d}t$$

can be interchanged since $\left| \frac{\partial}{\partial x_1} f(\boldsymbol{\xi}, \mathbf{x} + \tau \mathbf{e}_1) \right|$ is integrable on $\Omega \times [0, t]$ (theorem of Fubini):

$$\begin{aligned} \int_0^t G(\mathbf{x} + \tau \mathbf{e}_1) \mathrm{d}\tau &= \int_\Omega \int_0^t \frac{\partial}{\partial x_1} f(\boldsymbol{\xi}, \mathbf{x} + \tau \mathbf{e}_1) \mathrm{d}t \mathrm{d}\boldsymbol{\xi} \\ &= \int_\Omega \int_0^t \left[f(\boldsymbol{\xi}, \mathbf{x} + t\mathbf{e}_1) - f(\boldsymbol{\xi}, \mathbf{x}) \right] \mathrm{d}t \mathrm{d}\boldsymbol{\xi} = F(\mathbf{x} + t\mathbf{e}_1) - F(\mathbf{x}). \end{aligned}$$

Since G is continuous, differentiation at $t = 0$ proves $G(\mathbf{x}) = \frac{\partial}{\partial x_1} F(\mathbf{x})$.

This proof can be repeated for the next $k - 1$ derivatives.

Solution of Exercise 3.23. The chain rule gives

$$\begin{aligned} \Delta_\zeta u(\Phi(z)) &= \left(\partial^2 / \partial x^2 + \partial^2 / \partial y^2 \right) u(\xi(x, y), \eta(x, y)) \\ &= u_{\xi\xi} \left(\xi_x^2 + \xi_y^2 \right) + u_{\eta\eta} \left(\eta_y^2 + \eta_x^2 \right) + 2 u_{\xi\eta} \left(\xi_x \eta_x + \xi_y \eta_y \right) \\ &\quad + u_\xi \left(\xi_{xx} + \xi_{yy} \right) + u_\eta \left(\eta_{xx} + \eta_{yy} \right) . \end{aligned}$$

The real and imaginary parts of $\Phi = \xi + \mathrm{i}\eta$ are holomorphic: $\xi_{xx} + \xi_{yy} = \eta_{xx} + \eta_{yy} = 0$. The Cauchy–Riemann differential equations $\xi_x + \eta_y = \xi_y - \eta_x = 0$ imply $\xi_x^2 + \xi_y^2 = \eta_y^2 + \eta_x^2 = |\Phi'|^2$ and $\xi_x \eta_x + \xi_y \eta_y = 0$. Insertion proves the assertion $\Delta_\zeta u(\Phi(z)) = |\Phi'|^2 \left(u_{\xi\xi} + u_{\eta\eta} \right)$.

Exercises of Chapter 4

Solution of Exercise 4.5. $u_h^\mathsf{T} = \left[u_{h,1}^\mathsf{T}, \ldots, u_{h,n-1}^\mathsf{T}\right]$ is the block structure of the vector u_h, where the component $(u_{h,i}^\mathsf{T})_j$ corresponds to the grid point (ih, jh). The block $u_{h,i}$ belongs to the i-th row. For $q_h = L_h u_h$ one chooses the same block decomposition. The definition of L_h shows that $q_{h.i} = h^{-2}\left[Tu_{h.i} - u_{h.i-1} - u_{h.i+1}\right]$ with T in (4.16) (here the terms $u_{h.i\pm1}$ are omitted if $i \pm 1 \in \{0, n\}$). This proves the block structure in (4.16) and part (a).

In the case of (b) the rows contain $n-1$ inner grid points, i.e., T and I are $(n-1)\times(n-1)$ matrices. Since there are $m-1$ row blocks, L_h consists of $(m-1)^2$ blocks.

Solution of Exercise 4.6. Since all neighbours of a 'red' grid point $(x, y) \in \Omega_h^r$ belongs to Ω_h^b and vice versa, the diagonal blocks are diagonal matrices. The remaining coefficients are situated in the off-diagonal blocks A and A^T. Symmetry of L_h shows that one off-diagonal block is the transposed of the other.

Solution of Exercise 4.11. Part (a) is a direct consequence of the definition.

For part (b) assume first that the matrix has the form $A = \begin{bmatrix} A_{11} & A_{12} \\ O & A_{22} \end{bmatrix}$, where $A_{11} \in \mathbb{R}^{I'\times I'}$ and $A_{22} \in \mathbb{R}^{I''\times I''}$ with $n_1 = \#I' \geq 1$ and $n_2 = \#I'' \geq 1$. Choose any indices $\alpha \in I''$ and $\beta \in I'$. For an indirect proof assume that A is irreducible. Then there must be a connection $\alpha = \alpha_0, \alpha_1, \ldots, \alpha_k = \beta$. For at least one pair $(\alpha_{\ell-1}, \alpha_\ell) \in G(A)$ we have $\alpha_{\ell-1} \in I''$ and $\alpha_\ell \in I'$. This cannot be true since the entry $A_{\alpha_{\ell-1},\alpha_\ell}$ belongs to the zero block $O \in \mathbb{R}^{I''\times I'}$. Hence there is no connection and A is not irreducible.

Now assume that A is not irreducible: there is at least one pair $(\alpha, \beta) \notin \overline{G(A)}$. Case 1: $(\gamma, \beta) \notin \overline{G(A)}$ for all $\gamma \in I$. Set $I' := \{\beta\}$, $I'' := I\backslash\{\beta\}$, choose β as first index followed by all indices of I''. Then the first column of A is filled by zeros, in particular $A_{\gamma\beta} = 0$ holds for all $(\gamma, \beta) \in I'' \times I'$.

Case 2: At least one $\gamma \in I$ with $(\gamma, \beta) \in \overline{G(A)}$ exists. Then

$$I' := \left\{\gamma \in I : (\gamma, \beta) \in \overline{G(A)}\right\}, \qquad I'' := I\backslash I'$$

are not empty (note that $\alpha \in I''$). All entries $A_{\delta\gamma}$ with $(\delta, \gamma) \in I'' \times I'$ must vanish, since otherwise $(\delta, \gamma) \in G(A)$ and $(\gamma, \beta) \in \overline{G(A)}$ implies $(\delta, \beta) \in \overline{G(A)}$ in contradiction to $\delta \in I''$. A corresponding ordering of the indices of I' and I'' yields the zero block in $\begin{bmatrix} A_{11} & A_{12} \\ O & A_{22} \end{bmatrix}$.

Solution of Exercise 4.13. Consider the eigenpair (λ, u) of Assertion 2 on page 52. The proof of Criterion 4.12 shows that $|u_\gamma| = 1$ and $|\lambda - a_{\gamma\gamma}| = r_\gamma$ implies $|u_\beta| = 1$ and $|\lambda - a_{\beta\beta}| = r_\beta$ for all $\beta \in I_\gamma$, thus $\lambda \in K_{r_\gamma}(a_{\gamma\gamma}) \cup \bigcap_{\beta\in I_\gamma} \partial K_{r_\beta}(a_{\beta\beta})$. The case $I_\gamma = \emptyset$ is not excluded. Then $\lambda = a_{\gamma\gamma} = r_\gamma = 0$ holds and the latter statement becomes $0 \in \{0\} \cup \emptyset$. The union over all $\gamma \in I$ yields a superset of $\sigma(A)$.

Solution of Exercise 4.16. The assumptions in part (a) imply that for all $\gamma \in I$ there is an $\alpha \in I_\gamma$ (I_γ from Exercise 4.13) with $\partial K_{r_\alpha}(a_{\alpha\alpha}) = \partial K_{r_\alpha}(0) \subset K_1(0)$ since $r_\alpha < 1$. Exercise 4.13 shows that $\sigma(D^{-1}B) \subset K_1(0)$ and hence $\rho(D^{-1}B) < 1$.

In part (b) C is an $n \times n$ matrix with $\rho(C) < 1$. There is a unitary matrix Q so that $C = Q\,R\,Q^{\mathsf{T}}$ holds with an upper triangular matrix R (Schur normal form, cf. Liesen–Mehrmann [193, §14.3], Hackbusch [142, §A.6.1]). We split R into $R = D+T$, where D is the diagonal part and T the strictly upper part. Note that $T^\nu = O$ for $\nu \geq n$. If D and T commuted, the binomial formula $R^\nu = \sum_{\mu=0}^{n-1} \binom{\mu}{n} D^{n-\mu}T^\mu$ would hold and give

$$\|R^\nu\|_\infty \leq \sum_{\mu=0}^{\min\{\nu,n-1\}} \binom{\mu}{n} \|D\|_\infty^{\nu-\mu} \|T\|_\infty^\mu .$$

In the non-commutative case, e.g., the term $nD^{n-1}T^1$ corresponding to $\mu = 1$ becomes $TD^{n-1} + DTD^{n-2} + \ldots + D^{n-1}T$. This sum has the same bound $n\|D\|_\infty^{n-1}\|T\|_\infty$. Therefore the above estimate holds in general. If $\|D\|_\infty = 0$ we conclude $D = O$, $C = T$ and the convergence of $\sum_{\nu=0}^\infty C^\nu = \sum_{\nu=0}^{n-1} T^\nu$ is trivial. Otherwise set $K := \|Q\|_\infty \left\|Q^{\mathsf{T}}\right\|_\infty \sum_{\mu=0}^{n-1} \binom{\mu}{n} \|D\|_\infty^{-\mu} \|T\|_\infty^\mu$ and estimate by $\|C^\nu\|_\infty \leq \|Q\|_\infty \left\|Q^{\mathsf{T}}\right\|_\infty \|R^\nu\|_\infty \leq K\|D\|_\infty^\nu$. Since the diagonal D contains the eigenvalues as diagonal entries, it follows that $\|D\|_\infty = \rho(C) < 1$, and $\|C^\nu\|_\infty \leq K[\rho(C)]^\nu$ proves $C^\nu \to O$. From the representation $\sum_{\nu=0}^{\mu-1} C^\nu = (C-I)^{-1}(C^\mu - I)$ of the finite geometric sum and $C^\mu \to O$ we infer the convergence of the series $\sum_{\nu=0}^\infty C^\nu = (C-I)^{-1}$.

Statement 1) in part (c) is trivial since $(AB)_{ij} = \sum_k A_{ik}B_{kj}$ only contains nonnegative (resp. positive) terms.

A regular diagonal matrix $D \geq O$ has positive diagonal entries and the scaling of $A > O$ into AD does not change the signs (Statement 2).

The first part of Statement 3) follows from $(Av)_i = \sum_k A_{ik}v_k \leq \sum_k A_{ik}w_k = (Aw)_i$. $0 \leq v \leq w$ implies $\|v\|_\infty = \max_i v_i \leq \max_i w_i = \|w\|_\infty$.

$(Au)_i \leq |(Au)_i| = |\sum_k A_{ik}u_k| \leq \sum_k |A_{ik}|\,|u_k| = \sum_k A_{ik}\,|u_k| = (A\,|u|)_i$ is the componentwise Statement 4).

Solution of Exercise 4.22. Definition 4.30 and $|||\,A\,||| = 0$ imply $\|Au\| = 0$, hence $Au = 0$ for all u. Therefore $A = O$ is the zero matrix. Using

$$|||\,\lambda A\,||| = \sup_{u\neq 0} \frac{\|\lambda Au\|}{\|u\|} = \sup_{u\neq 0} \frac{|\lambda|\,\|Au\|}{\|u\|} = |\lambda| \sup_{u\neq 0} \frac{\|Au\|}{\|u\|} = |\lambda|\,|||\,A\,|||$$

and

$$\begin{aligned} |||\,A+B\,||| &= \sup \frac{\|(A+B)\,u\|}{\|u\|} \leq \sup \frac{\|Au\| + \|Bu\|}{\|u\|} \\ &\leq \sup \frac{\|Au\|}{\|u\|} + \sup \frac{\|Bu\|}{\|u\|} \leq |||\,A\,||| + |||\,B\,|||, \end{aligned}$$

we prove the other norm axioms and part (a).

Part (b): $\|Au\| \le |||A||| \, \|u\|$ is trivial for $u = 0$. Let $u \neq 0$. Now the inequality follows from $|||A|||= \sup_{v\neq 0} \frac{\|Av\|}{\|v\|} \ge \frac{\|Au\|}{\|u\|}$.

Let $u \neq 0$ be arbitrary. Then $\|(AB)\,u\| = \|A\,(Bu)\| \le ||| A ||| \, \|Bu\| \le |||A||| \, |||B||| \, \|u\|$ and thus $|||AB|||= \sup \frac{\|(AB)u\|}{\|u\|} \le |||A||| \, |||B|||$ holds.

Furthermore we have $|||I|||= \sup \frac{\|Iu\|}{\|u\|} = \sup 1 = 1$.

Let λ be an eigenvalue and $e \neq 0$ the corresponding eigenvector: $Ae = \lambda e$. Then $\|Ae\| = \|\lambda e\| = |\lambda| \, \|e\|$ implies $|||A|||= \sup_{u\neq 0} \frac{\|Au\|}{\|u\|} \ge \frac{\|Ae\|}{\|e\|} = |\lambda|$.

Solution of Exercise 4.23. Let $||| \cdot |||_\infty$ be the matrix norm associated to the maximum norm. Let $\|\cdot\|_\infty$ be as defined in (4.32). $\|Au\|_\infty = \max_i |\sum_k A_{ik}u_k| \le \max_i \big|\sum_k |A_{ik}|\big| \max_k |u_k| = \|A\|_\infty \|u\|_\infty$ implies $|| A |||_\infty \le \|A\|_\infty$.

Let α be an index with $\sum_k |A_{\alpha k}| = \max_i |\sum_k |A_{ik}|| = \|A\|_\infty$. Define u by $u_k = \operatorname{sign}(A_{\alpha k})$. Since $\|u\|_\infty = 1$ and $A_{\alpha k}u_k = |A_{\alpha k}|$, the reverse inequality also follows: $\|A\|_\infty = \sum_k |A_{\alpha k}| = |\sum_k A_{\alpha k}u_k| = |(Au)_k| \le \|Au\|_\infty = \frac{\|Au\|_\infty}{\|u\|_\infty} \le \sup_{v\neq 0} \frac{\|Av\|_\infty}{\|v\|_\infty} = |||A|||_\infty$. Therefore (4.32) and part (a) are proved.

Part (b) is trivial.

Solution of Exercise 4.26. Part (a) of the proof of Lemma 4.17 uses the splitting $A = D - B$ and shows that $A^{-1} = \left(\sum_{\nu=0}^{\infty} (D^{-1}B)^\nu\right) D^{-1}$ with $D \ge O$ and $B \ge O$. For $A' = D' - B'$ we have $O \le D \le D'$ and $O \le B' \le B$. By $\rho(D'^{-1}B') \le \rho(D^{-1}B) < 1$, the series $A'^{-1} = \left[\sum_{\nu=0}^{\infty} (D'^{-1}B')^\nu\right] D'^{-1}$ converges. $O \le D'^{-1} \le D^{-1}$ and $O \le B' \le B$ imply $A'^{-1} \le A^{-1}$.

Solution of Exercise 4.27. Let A' be the matrix mentioned in the hint. Without loss of generality we may order the index set I so that first the indices of I' appear. The decomposition $I = I' \cup (I \backslash I')$ leads to the block form $A' = \begin{pmatrix} B & O \\ O & D \end{pmatrix}$ with the diagonal matrix $D = \operatorname{diag}\{a_{\alpha\alpha} : \alpha \in I \backslash I'\}$. The M-matrix property $A_{\alpha\beta} \le 0$ for $\alpha \neq \beta$ shows $A \le A'$, and Exercise 4.26 proves $O \le A'^{-1} \le A^{-1}$. The restriction of $A'^{-1} = \begin{pmatrix} B^{-1} & O \\ O & D^{-1} \end{pmatrix} \le A^{-1}$ to the left upper block proves the assertion.

Solution of Exercise 4.28. (i) $\|u\|_2 = \max\{|\langle u, v\rangle| \,/\, \|v\|_2 : v \neq 0\}$ follows from $|\langle u, v\rangle| \le \|u\|_2 \, \|v\|_2$ and $|\langle u, v\rangle| = \|u\|_2^2$ for $v := u$.

(ii) For (b), i.e., $\|A\|_2 = \|A^{\mathsf{T}}\|_2$, we use $\|A\|_2 = \max_u \frac{\|Au\|_2}{\|u\|_2} = \max_{u,v} \frac{|\langle Au, v\rangle|}{\|u\|_2 \|v\|_2}$, insert $\langle Au, v\rangle = \langle u, A^{\mathsf{T}}v\rangle$ and obtain $\max_{u,v} \frac{|\langle u, Av^{\mathsf{T}}\rangle|}{\|u\|_2 \|v\|_2} = \max_v \frac{\|A^{\mathsf{T}}v\|_2}{\|v\|_2} = \|A^{\mathsf{T}}\|_2$.

(iii) Let Q be a unitary matrix. We want to show that

$$\|A\|_2 = \|QA\|_2 = \|AQ\|_2.$$

$\|Qu\|_2 = \|u\|_2$ holds for all u. This shows $\|Q\|_2 = 1$. Inequality (4.31a) implies $\|QA\|_2 \le \|A\|_2$. Analogously $\|AQ\|_2 = \|Q^{\mathsf{T}}A^{\mathsf{T}}\|_2 \le \|A^{\mathsf{T}}\|_2 = \|A\|_2$ holds.

By $Q^{\mathsf{T}}Q = I$ also the reverse inequality is valid: $\|A\|_2 = \|Q^{\mathsf{T}}QA\|_2 \le \|QA\|_2$ respectively $\|A\|_2 = \|AQQ^{\mathsf{T}}\|_2 \le \|AQ\|_2$.

(iv) $\|D\|_2 = \rho(D)$ for diagonal matrices D is trivial.

(v) Symmetric matrices allow a diagonalisation $A = QDQ^{\mathsf{T}}$ with a unitary Q. According to (iii) and (iv), $\|A\|_2 = \|D\|_2 = \rho(A)$. This proves the part (a).

(vi) $\|A\|_2^2 = \left(\max_u \frac{\|Au\|_2}{\|u\|_2}\right)^2 = \max_u \|Au\|_2^2 / \|u\|_2^2$ and $\|Au\|_2^2 = \langle u, A^{\mathsf{T}}Au\rangle$ imply $\|A\|_2^2 \le \|A^{\mathsf{T}}A\|_2$. On the other hand $\|A^{\mathsf{T}}A\|_2 \le \|A^{\mathsf{T}}\|_2 \|A\|_2 = \|A\|_2^2$ follows from (ii). This proves $\|A\|_2 = \sqrt{\|A^{\mathsf{T}}A\|_2}$. The symmetry of $A^{\mathsf{T}}A$ yields part (c): $\|A\|_2 = \sqrt{\rho(A^{\mathsf{T}}A)}$.

(vii) The eigenvalues of $A^{\mathsf{T}}A$ can be estimated by $\|A^{\mathsf{T}}A\|_\infty$. Hence $\|A\|_2^2 = \rho(A^{\mathsf{T}}A) \le \|A\|_\infty \|A^{\mathsf{T}}\|_\infty$ proves part (d).

Solution of Exercise 4.30. (a) A symmetric matrix A has a representation $A = Q^{\mathsf{T}}DQ$ with unitary Q and $D = \operatorname{diag}_{i\in I}\{\lambda_i\}$. In the case of positive eigenvalues we have $\langle Ax, x\rangle = \langle DQx, Qx\rangle$. For $x \neq 0$ and $y := Qx$ we obtain $\langle Ax, x\rangle = \langle Dy, y\rangle \ge (\min_{i\in I}\lambda_i)\cdot\|y\|^2 = (\min_{i\in I}\lambda_i)\cdot\|x\|^2 > 0$. If, however, an eigenvalue is negative, the corresponding eigenvector $x \neq 0$ leads to $\langle Ax, x\rangle = \lambda\|x\|^2 \le 0$.

(b) Let $A' := A|_{I'\times I'}$ with $I' \subset I$ be the principal submatrix corresponding to $I'\times I'$. Each vector $x' \in \mathbb{R}^{I'}$ can be extended to $x \in \mathbb{R}^I$ by $x_\alpha := x'_\alpha$ for $\alpha \in I'$ and $x_\alpha := 0$ for $\alpha \in I\backslash I'$. Obviously, $\langle A'x', x'\rangle = \langle Ax, x\rangle$. Since $x' \neq 0$ implies $x \neq 0$ and $\langle Ax, x\rangle > 0$ it follows that A' is positive definite.

(c) The special case $I' = \{\alpha\}$ in (b) yields part (c).

(d) Let $A = QDQ^{\mathsf{T}}$ with $D = \operatorname{diag}\{\lambda_\alpha : \alpha \in I\}$. Set $D^{\frac12} := \operatorname{diag}\{\sqrt[+]{\lambda_\alpha} : \alpha \in I\}$. $B := QD^{\frac12}Q^{\mathsf{T}}$ is the unique positive-definite root of the matrix equation $B^2 = A$.

(e) The proofs of (b–d) easily carry over to the semidefinite case (alternative: let A be the limit of the positive-definite matrices $A + \varepsilon I$, $\varepsilon \searrow 0$).

Solution of Exercise 4.52. Case $s \in (2,3)$. $u(x\pm h, y) - u(x,y) = \pm\int_0^h u_x(\pm\xi, y)\,\mathrm{d}\xi$ yields

$$\begin{aligned} &u(x+h,y) + u(x-h,y) - 2u(x,y) \\ &\quad= \int_0^h [u_x(x+\xi,y) - u_x(x-\xi,y)]\,\mathrm{d}\xi = \int_0^h \int_{-\xi}^{\xi} u_{xx}(x+t,y)\,\mathrm{d}t\mathrm{d}\xi \\ &\quad= h^2 u_{xx}(x,y) + \int_0^h \int_{-\xi}^{\xi} [u_{xx}(x+t,y) - u_{xx}(x,y)]\,\mathrm{d}t\mathrm{d}\xi. \end{aligned}$$

Using $|u_{xx}(x+t,y) - u_{xx}(x,y)| \le |t|^{s-2}\|u\|_{C^s(\overline{\Omega})}$, we can estimate the remainder by

$$\int_0^h \int_{-\xi}^{\xi} |t|^{s-2}\,\mathrm{d}t\mathrm{d}\xi\,\|u\|_{C^s(\overline{\Omega})} = \frac{2h^s}{(s-1)s}\|u\|_{C^s(\overline{\Omega})}.$$

Together with the analogous inequality for the y direction we obtain

$$\|\Delta_h R_h u - \tilde{R}_h \Delta u\|_\infty \le K_s h^s \|u\|_{C^s(\overline{\Omega})} \quad \text{with } K_s = \frac{4}{(s-1)s}.$$

Case $3 < s < 4$. Use $u_{xx}(x+t,y) - u_{xx}(x,y) = \int_0^t u_{xxx}(x+\tau,y)\mathrm{d}\tau$:

$$\begin{aligned}
&\int_{-\xi}^{\xi} [u_{xx}(x+t,y) - u_{xx}(x,y)]\,\mathrm{d}t \\
&= \int_0^{\xi} [u_{xx}(x+t,y) - u_{xx}(x,y)]\,\mathrm{d}t + \int_0^{\xi} [u_{xx}(x-t,y) - u_{xx}(x,y)]\,\mathrm{d}t \\
&= \int_0^{\xi}\int_0^{t} [u_{xxx}(x+\tau,y) - u_{xxx}(x-\tau,y)]\,\mathrm{d}\tau\mathrm{d}t.
\end{aligned}$$

By $|u_{xxx}(x+\tau,y) - u_{xxx}(x-\tau,y)| \le (2\tau)^{s-3}\|u\|_{C^s(\overline{\Omega})}$ and integration we obtain the estimate $2^{s-3}\frac{1}{s-2}\frac{1}{s-1}\frac{1}{s}h^s$ for each direction, thus $K_s = 2^{s-2}\frac{1}{(s-2)(s-1)s}$.

Case $s = 3$. Use $|u_{xxx}(x+\tau,y) - u_{xxx}(x-\tau,y)| \le 2\|u\|_{C^s(\overline{\Omega})}$ and argue as above. The same estimate holds for $C^{2,1}(\overline{\Omega})$ (cf. Corollary 4.50).

Solution of Exercise 4.54. (a) L_h satisfies the sign conditions (4.21a) and is irreducibly diagonal dominant. Criterion 4.18 proves the M-matrix property.

(b) The same choice $w(x,y) = x(1-x)/2$ as in the proof of (4.36c) yields the estimate $\|L_h^{-1}\|_\infty \le 1/8$. The row-sum norm $\|L_h\|_\infty \le 20h^{-2}/3$ is easy to determine.

Solution of Exercise 4.56. The matrix L_h has the entries $L_{h,\mathbf{x}\boldsymbol{\xi}}$ with $\mathbf{x}, \boldsymbol{\xi} \in \overline{\Omega}_h'$. The entries for $\mathbf{x}, \boldsymbol{\xi} \in \Omega_h$ are the same as in the Dirichlet case, in particular they are symmetric. The five-point formula shows $L_{h,\mathbf{x}\boldsymbol{\xi}} = -h^{-2}$ for $\mathbf{x} \in \Omega_h$ and $\boldsymbol{\xi} \in \Gamma_h'$. The same value holds for $L_{h,\boldsymbol{\xi}\mathbf{x}}$ after the rescaling. The case $\mathbf{x}, \boldsymbol{\xi} \in \Gamma_h'$ is uninteresting since $L_{h,\mathbf{x}\boldsymbol{\xi}} = L_{h,\boldsymbol{\xi}\mathbf{x}} = 0$. The sign conditions (4.21a) follows from (4.62b).

Solution of Exercise 4.63. If $0 < x_1 = \nu h < 1$, the neighbouring grid points $x_1 \pm h$ exist so that the usual difference star $h^{-2}[-1\ 2\ -1]$ appears. For $x_1 = 0$ one starts with the difference $h^{-2}[-1\ 2\ -1]$ and eliminates the value at $-h$ by $h^{-2}[1\ 0\ -1]$, resulting in $h^{-2}[0\ 2\ -2]$.

The block representation of the matrix in Exercise 4.63 defines the standard five-point formula $\begin{bmatrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{bmatrix}$ for interior points. If $x_1 = 0$ the term corresponding to $u(-h, x_2)$ is eliminated by $2h^{-1}\partial_n^0$ and yields $\begin{bmatrix} & -1 & \\ 0 & 4 & -2 \\ & -1 & \end{bmatrix}$. This explains the coefficients $\begin{bmatrix} 4 & -2 \end{bmatrix}$ at the left upper position in the matrix T. For $x_2 = 0$ the analogous elimination leads to the blocks $\begin{bmatrix} T & -2I \end{bmatrix}$ in L_h. Similar for $x_1 = 1$ and $x_2 = 1$. This proves part (a).

The scaling in part (b) transfers T into $\hat{T} = \begin{bmatrix} 2 & -1 & & \\ -1 & 4 & -1 & \\ & \ddots & \ddots & \ddots \\ & & \ddots & \ddots \end{bmatrix}$ and L_h into $L_h = h^{-2} \begin{bmatrix} \frac{1}{2}\hat{T} & -I & & \\ -I & \hat{T} & -I & \\ & \ddots & \ddots & \ddots \\ & & \ddots & \ddots \end{bmatrix}$. Obviously $\hat{T}$ and L_h are symmetric.

Solution of Exercise 4.75. Without loss of generality let $x = 0$. We have

$$\frac{u(x'') - u(x)}{x'' - x} = \frac{u(x'') - u(0)}{x''} = \frac{1}{x''} \int_0^{x''} u'(\xi) \mathrm{d}\xi = \int_0^1 u'(\xi x'') \mathrm{d}\xi$$

and analogously $\frac{u(x)-u(x')}{x-x'} = \int_0^1 u'(\xi x') \mathrm{d}\xi$. The second divided difference becomes

$$\frac{2}{x'' - x'} \int_0^1 \left[u'(\xi x'') - u'(\xi x') \right] \mathrm{d}\xi = \frac{2}{x'' - x'} \int_0^1 \int_{\xi x'}^{\xi x''} u''(t) \mathrm{d}t \mathrm{d}\xi.$$

Using the identity $\frac{2}{x''-x'} \int_0^1 \int_{\xi x'}^{\xi x''} \mathrm{d}t \mathrm{d}\xi = 1$, the right-hand side can be written as $u''(x) + \mathrm{rem}$ with the remainder

$$\mathrm{rem} = \frac{2}{x'' - x'} \int_0^1 \int_{\xi x'}^{\xi x''} \left[u''(t) - u''(0) \right] \mathrm{d}t \mathrm{d}\xi.$$

The estimate $|u''(t) - u''(0)| \le |t| \, \|u\|_{C^3([x',x''])}$ and $\frac{2}{x''-x'} \int_0^1 \int_{\xi x'}^{\xi x''} |t| \, \mathrm{d}t \mathrm{d}\xi = \frac{1}{x''-x'} \int_0^1 \left[(\xi x'')^2 + (\xi x')^2 \right] \mathrm{d}\xi = \frac{x''^2 + x'^2}{x''-x'} \int_0^1 \xi^2 \mathrm{d}\xi = \frac{1}{3} \frac{x''^2 + x'^2}{x''-x'}$ implies the statements of the parts (a,b).

Inserting $x'' = x + h$ and $x' = x - h$, we obtain the statement of part (c).

Solution of Exercise 4.78. Split the system $L_h u_h = q_h$ with $q_h = f_h + \varphi_h$ into $L_h u_h' = f_h$ and $L_h u_h'' = \varphi_h$. In $\|u_h\|_\infty \le \|u_h'\|_\infty + \|u_h''\|_\infty$ we estimate $\|u_h'\|_\infty \le \|L_h^{-1}\|_\infty \|f_h\|_\infty \le \frac{d^2}{8} \max_{\mathbf{x} \in \Omega_h} |f(\mathbf{x})|$ by (4.91). The matrix L_h is weakly diagonal dominant. This allows us to apply Remark 4.37 to L_h: u_h'' attains the maximum on the boundary: $\|u_h''\|_\infty \le \|\varphi_h\|_\infty = \max_{\boldsymbol{\xi} \in \Gamma_h} |\varphi(\boldsymbol{\xi})|$. Together we obtain the desired estimate.

Solution of Exercise 4.80. As in the proof of Theorem 4.79 the consistency error is split into $c_h = c_h^1 + c_h^2$, where c_h^1 arises from the irregular Shortley–Weller approximation at the near-boundary points, while $c_h^2 = \mathcal{O}(h^4)$ for sufficiently smooth u belongs to the mehrstellen method. As in (4.94) c_h^1 leads to $\|w_h^1\|_\infty \le \mathcal{O}(h^3)$, while $\|w_h^2\|_\infty \le \mathcal{O}(h^4)$. Together we obtain $\|u_h - R_h u\|_\infty = \mathcal{O}(h^3)$.

Solution of Exercise 4.82. Let $\mathbf{x} = (x, y) \in \Omega_h$ and $\boldsymbol{\xi} = (x - s_\ell h, y)$. The coefficient of $u_h(\boldsymbol{\xi})$ in (4.96) is $-h^{-2}/s_\ell$. It is an entry of the matrix L_h only if $\boldsymbol{\xi} \in \Omega_h$, which implies $s_\ell = 1$. Therefore the coefficient is $L_{h,\mathbf{x}\boldsymbol{\xi}} = -h^{-2}$ as for the standard five-point formula. The change of the diagonal coefficient $L_{h,\mathbf{xx}}$ does not matter, since this does not disturb the symmetry.

The previous consistency error $|c_h^1(x,y)|_\infty \le \frac{1}{2} s_r s_\ell \, (s_r + s_\ell) \, h^2 \, \|u\|_{C^{1,1}(\overline{\Omega})}$ in the proof of Theorem 4.81 is scaled by $1/(s_r s_\ell)$ and is used for both directions. Together we obtain

$$\tfrac{1}{2} h^2 \, (s_r + s_\ell + s_o + s_u) \, \|u\|_{C^{1,1}(\overline{\Omega})} \le 2h^2 \, \|u\|_{C^{1,1}(\overline{\Omega})} .$$

This bound of $u_h^1 = L_h^{-1} c_h^1$ yields the first term $2h^2 \, \|u\|_{C^{1,1}(\overline{\Omega})}$ in the estimate of Exercise 4.82.

Exercises of Chapter 5

Solution of Exercise 5.2. (a) Since L is elliptic in K, $c(\mathbf{x}) > 0$ holds in K.
(b) By assumption $c(\cdot)$ is continuous and takes its minimum on K. By part (a) the minimum is positive.

Solution of Exercise 5.3. Use (A.4a,b) with $\frac{\partial \Phi_\alpha(\mathbf{x})}{\partial x_i} = S_{\alpha i}$ and $\frac{\partial^2 \Phi_\alpha(\mathbf{x})}{\partial x_i \partial x_j} = 0$. Hence the differential equation in $\hat{\mathbf{x}}$ reads

$$\hat{L} = \sum_{i,j=1}^{n} \hat{a}_{ij}(\hat{\mathbf{x}}) \frac{\partial^2}{\partial \hat{x}_i \partial \hat{x}_j} + \sum_{i=1}^{n} \hat{a}_i(\hat{\mathbf{x}}) \frac{\partial}{\partial \hat{x}_i} + \hat{a}(\hat{\mathbf{x}})$$

with $\hat{A} := (\hat{a}_{ij}) = SAS^\mathsf{T}$, $\hat{\mathbf{a}} := (\hat{a}_i) = S\mathbf{a}$, and $\hat{a} = a$. These quantities are functions of $\hat{\mathbf{x}} = \Phi(\mathbf{x})$. If A is negative definite, then so is $\hat{A}$, i.e., ellipticity in the $\mathbf{x}$-formulation implies ellipticity with respect of $\hat{\mathbf{x}}$.

Uniform ellipticity in Ω requires $-\langle A(\mathbf{x})\boldsymbol{\xi}, \boldsymbol{\xi}\rangle \ge c_0 \, |\boldsymbol{\xi}|^2$ with $c_0 > 0$ for all $\mathbf{x} \in \Omega$. $\hat{A} = SAS^\mathsf{T}$ leads to $-\langle \hat{A}(\hat{\mathbf{x}})\boldsymbol{\xi}, \boldsymbol{\xi}\rangle = -\langle A(\mathbf{x}) S^\mathsf{T}\boldsymbol{\xi}, S^\mathsf{T}\boldsymbol{\xi}\rangle \ge c_0 \left|S^\mathsf{T}\boldsymbol{\xi}\right|^2$. Since S is regular, the smallest eigenvalue λ of SS^T is positive. From $\left|S^\mathsf{T}\boldsymbol{\xi}\right|^2 = \left\langle SS^\mathsf{T}\boldsymbol{\xi}, \boldsymbol{\xi}\right\rangle \ge \lambda \, |\boldsymbol{\xi}|^2$ it follows that $-\langle \hat{A}(\hat{\mathbf{x}})\boldsymbol{\xi}, \boldsymbol{\xi}\rangle \ge \hat{c}_0 \lambda \, |\boldsymbol{\xi}|^2$ with $\hat{c}_0 := c_0 \lambda > 0$, i.e., the differential equation in $\hat{\mathbf{x}}$ is also uniformly elliptic.

Solution of Exercise 5.6. (a) The diagonal entries of AB and BA are $\sum_{j=1}^n a_{ij} b_{ji}$ and $\sum_{j=1}^n b_{ij} a_{ji}$, respectively. Summation over i gives $\sum_{i,j=1}^n a_{ij} b_{ji}$ respectively $\sum_{i,j=1}^n b_{ij} a_{ji}$. Renaming $i \leftrightarrow j$ in the latter case shows the equality.

(b) $a_{ii} \ge 0$ is already shown in Exercise 4.30e. Summation yields $\operatorname{trace}(A) \ge 0$.

(c) According to Exercise 4.30e a positive-semidefinite matrix $B^{1/2}$ exists with the property $(B^{\frac{1}{2}})^2 = B$. If A is positive semidefinite then so is $B^{1/2} A B^{1/2}$, hence $\operatorname{trace}(B^{1/2} A B^{1/2}) \ge 0$ according to (b). Application of (a) shows

$$\operatorname{trace}(B^{1/2} A B^{1/2}) = \operatorname{trace}(A B^{1/2} B^{1/2}) = \operatorname{trace}(AB).$$

Solution of Exercise 5.13. (a) Since $\overline{\Omega}$ is bounded and closed, it is compact. Uniform ellipticity follows from Exercise 5.2.

(b) $v := u_2 - u_1$ satisfies $Lv = 0$ in Ω with boundary values $v = \varphi_1 - \varphi_2$ on Γ. Set $w(\mathbf{x}) := \|\varphi_1 - \varphi_2\|_\infty$. Then $Lw = a(\mathbf{x})\,\|\varphi_1 - \varphi_2\|_\infty \geq Lv$ holds since $a \geq 0$. Lemma 5.10 yields $v \leq w$ in Ω and thus $u_2 - u_1 \leq \|\varphi_1 - \varphi_2\|_\infty$. The analogously provable inequality $-w \leq v$ yields $\|u_2 - u_1\|_\infty \leq \|\varphi_1 - \varphi_2\|_\infty$.

(c) A suitable translation and rotation maps Σ onto the strip

$$\hat{\Sigma} = \{\mathbf{x} \in \mathbb{R}^n : -\delta/2 \leq x_1 \leq \delta/2\},$$

i.e., the affine transformation is $\hat{\mathbf{x}} := \mathbf{x}_0 + \mathbf{S}\mathbf{x}$ with a unitary matrix $\mathbf{S}$. Therefore the constant of the uniform ellipticity is unchanged in the new variables (cf. Exercise 5.3). The constant R in the proof of Theorem 5.12 is $R = \delta/2$ and defines the quantity $M = \mathrm{e}^{2R\alpha} = \mathrm{e}^{\delta\alpha}$.

(d) Apply Theorem 5.12 for $u_1 := u$ and $u_2 := 0$ with $\varphi_2 = 0$ and $f_2 = 0$.

Solution of Exercise 5.14. The transformation does not chance the ellipticity (cf. Exercise 5.3). Since $\hat{a}(\boldsymbol{\xi}) = a(\Phi^{-1}(\mathbf{x}))$, the sign condition $\hat{a} \geq 0$ is unchanged. Hence, also after the transformation, the assumptions of Lemma 5.10 are satisfied.

By assumption Ω is bounded (cf. Lemma 5.10), so that $\overline{\Omega}$ is compact. From this and $\Phi \in C^1(\overline{\Omega}), \Phi^{-1} \in C^1(\overline{\Omega})$ we conclude that the smallest singular value of $S = \partial\Phi/\partial\mathbf{x}$ is positive on $\overline{\Omega}$, i.e., $|S\boldsymbol{\xi}|^2 \geq c_0\,|\boldsymbol{\xi}|^2$ with $c_0 > 0$. As in the solution of Exercise 5.3 this ensures the uniform ellipticity of the transformed equation. Hence Theorem 5.12 is applicable to L'.

Solution of Exercise 5.18. The off-diagonal entries $a_{11} + |a_{12}|$ and $a_{22} + |a_{12}|$ are ≤ 0 because of (5.12). They cannot vanish simultaneously since otherwise $\det(A) = 0$ in contradiction to (5.4a). Hence the matrix graph has a direct connection from (x, y) to $(x \pm h, y)$ or $(x, y \pm h)$. This path can be continued until we reach a near-boundary point (x', y') in which the coefficient $a_{11} + |a_{12}| < 0$ or $a_{22} + |a_{12}| < 0$ belongs to a boundary point. Hence the strict inequality (4.26a) is satisfied in (x', y') as required in Exercise 4.16a.

Solution of Exercise 5.20. The choice

$$a_1(x, y) = \left\{ \begin{array}{l} +6 \;\text{ for } x = 1/3 \\ -6 \;\text{ for } x = 2/3 \end{array} \right\} \quad \text{and } \; a_2(x, y) = a_1(y, x)$$

leads to a difference formula L_h with $L_{h,\mathbf{x}\boldsymbol{\xi}} = 0$ for $\boldsymbol{\xi} \in \Gamma_h$. L_h is singular since $L_h u_h = 0$ holds for the constant grid function $u_h(\mathbf{x}) = 1$ $(\mathbf{x} \in \Omega_h)$.

Solution of Exercise 5.22. (a) The matrix L_h is weakly diagonal dominant. Its star satisfies the sign condition $\begin{bmatrix} \leq 0 & < 0 & \\ < 0 & > 0 & < 0 \\ & < 0 & \leq 0 \end{bmatrix}$ for sufficiently small h. Therefore L_h is irreducible and satisfies the strong diagonal dominance (4.26a) in near-boundary points. Thus L_h is irreducibly diagonal dominant and an M-matrix. Because of the symmetric differences L_h is symmetric and the method has consistency order 2. L_h is positive definite because of Criterion 4.32.

(b) If $a_{12} < 0$, $\partial^2/\partial x \partial y$ is discretised as in (5.13) by the star $\begin{bmatrix} 0 & * & * \\ * & * & * \\ * & * & 0 \end{bmatrix}$.

Solution of Exercise 5.23. One verifies that transposition turns $\partial_x^{\pm}$ into $-\partial_x^{\mp}$ and $\partial_y^{\pm}$ into $-\partial_y^{\mp}$. Furthermore, the position of the coefficient is changed: $a_{11}\partial_x^+\partial_x^-$ becomes $\partial_x^+\partial_x^- a_{11}$ etc. While $a_{11}\partial_x^+\partial_x^- u$ discretises $a_{11}u_{xx}$, the transposed expression $\partial_x^+\partial_x^-\,(a_{11}u)$ is a discretisation of $(a_{11}u)_{xx}$. A comparison with the adjoint differential operator (5.17) shows that L_h^{T} is the discretisation of L'. Since symmetric second differences and unsymmetric first differences are used, L_h^{T} is a discretisation of first consistency order.

Solution of Exercise 5.30. (a) The coefficients in (5.1) must satisfy: $a_{ii}(\mathbf{x}) = a_{ii}(\mathbf{x}^-)$ for $i=1,2$, $a_{12}(\mathbf{x}) = -a_{12}(\mathbf{x}^-)$, $a_1(\mathbf{x}) = a_1(\mathbf{x}^-)$, $a_2(\mathbf{x}) = -a_2(\mathbf{x}^-)$, $a(\mathbf{x}) = a(\mathbf{x}^-)$ with $\mathbf{x}^- = \binom{x_1}{-x_2}$.

(b) In general one can only state that if u is a solution then so is $u^-(x_1,x_2) := u(x_1,-x_2)$. If the solution is unique, one concludes that $u = u^-$, and hence u is symmetric.

Solution of Exercise 5.32. T in (4.16) is to be replaced by $\begin{bmatrix} 4 & -1 & \dots & 0 & -1 \\ -1 & 4 & -1 & & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & & -1 & 4 & -1 \\ -1 & 0 & \dots & -1 & 4 \end{bmatrix}$.

Solution of Exercise 5.36. Clearly $\Delta^2 u = -\Delta v = f$ holds in Ω, and the boundary conditions (5.28) are satisfied. In the case of the boundary conditions (5.27) there is no boundary condition for the solution of $\Delta v = f$, whereas u with $\Delta u = v$ in Ω must fulfil two boundary conditions $u = \varphi_1$ and $\frac{\partial u}{\partial n} = \varphi_2$.

Solution of Exercise 5.40. (a) The coefficients of (5.1) are translated into a_α with

$$m=1, \quad a_{(2,0)} = a_{11}, \quad a_{(0,2)} = a_{22}, \quad a_{(1,1)} = a_{12} + a_{21},$$
$$a_{(1,0)} = a_1, \quad a_{(0,1)} = a_2, \quad a_{(0,0)} = a.$$

$L = \Delta^2 = \partial^4/\partial x^4 + 2\partial^4/\partial x^2\partial y^2 + \partial^4/\partial y^4$ is described by

$$m=2, \quad a_{(4,0)} = 1, \quad a_{(2,2)} = 2, \quad a_{(0,4)} = 1, \quad a_\alpha = 0 \text{ otherwise.}$$

(b) In the case of Δ^2 the left-hand side in (5.30b) is $\xi_1^4 + 2\xi_1^2\xi_2^2 + \xi_2^4 = (\xi_1^2+\xi_2^2)^2 = |\boldsymbol{\xi}|^4$ with $c(\mathbf{x}) = 1$ on the right-hand side.

(c) An elliptic operator of order $2m+1$ ($m \in \mathbb{N}_0$) would be of the form $P(\mathbf{x},\boldsymbol{\xi}) := \sum_{|\alpha|=2m+1} a_\alpha(\mathbf{x})\boldsymbol{\xi}^\alpha \neq 0$ for all $0 \neq \boldsymbol{\xi} \in \mathbb{R}^n$. Since $|\alpha|$ is odd, we have $P(\mathbf{x},\boldsymbol{\xi}) = -P(\mathbf{x},-\boldsymbol{\xi})$. Choose a path in $\mathbb{R}^n\backslash\{0\}$ from $\boldsymbol{\xi}$ to $-\boldsymbol{\xi}$. By continuity there is an intermediate value $0 \neq \boldsymbol{\xi}_0 \in \mathbb{R}^n$ with $P(\mathbf{x},\boldsymbol{\xi}_0) = 0$ in contradiction to ellipticity.

(d) For $|\alpha| = 1$ use

$$a(\mathbf{x})D^\alpha u = D^\alpha\,(a(\mathbf{x})u) - (D^\alpha a(\mathbf{x}))\,u, \quad \text{i.e., } a(\mathbf{x})D^\alpha = D^\alpha a(\mathbf{x}) - (D^\alpha a(\mathbf{x}))$$

for any differentiable function a. Multiple application of this rule shifts the coefficients a_α from (5.29) into the middle position until only terms of the form $D^\beta a_{\alpha\beta}(\mathbf{x})D^\alpha$ with $|\alpha| \le m$ appear. For odd $|\beta|$ one reverses the sign of $a_{\alpha\beta}$ and obtains $(-1)^{|\beta|}D^\beta a_{\alpha\beta}(\mathbf{x})D^\alpha$.

Exercises of Chapter 6

Solution of Exercise 6.3. The triangle inequality states $\|x\| = \|(x-y)+y\| \le \|x-y\| + \|y\|$ which implies $\|x\| - \|y\| \le \|x-y\|$. Interchanging x and y we obtain (6.1). This inequality proves that the norm is continuous, even Lipschitz-continuous.

Solution of Exercise 6.4. Let $\|\cdot\|_1 \le C\,\|\cdot\|_2$. The corresponding balls $K^i_\varepsilon(x) := \{y : \|x-y\|_i \le \varepsilon\}$ satisfy $K^1_{\varepsilon/C} \subset K^2_\varepsilon$, i.e., an open set of the $\|\cdot\|_2$-topology is also open with respect to $\|\cdot\|_1$. If also the reverse inequality $\|\cdot\|_2 \le C\,\|\cdot\|_1$ holds, the open sets coincide.

Solution of Exercise 6.5. (a_1) If T is bounded, $x \to \xi$ implies that

$$\|Tx - T\xi\|_Y \le \|T\,(x-\xi)\| \le \|T\|_{Y\leftarrow X}\,\|x-\xi\|_X \to 0, \quad \text{i.e., } Tx \to T\xi.$$

(a_2) If the operator T is unbounded, there is a sequence x_i with $\|x_i\|_X = 1$ and $c_i := \|Tx_i\|_Y \to \infty$. The quantities $\xi_i := \frac{1}{\sqrt{c_i}} x_i$ lead to $\|T\xi_i\|_Y = \sqrt{c_i} \to \infty$, although $\xi_i \to 0$. Hence, T is not continuous.

(b) For any $0 \ne x \in X$ we have

$$\frac{\|(S+T)x\|_Y}{\|x\|_X} \le \frac{\|Sx\|_Y}{\|x\|_X} + \frac{\|Tx\|_Y}{\|x\|_X} \le \|S\|_{Y\leftarrow X} + \|T\|_{Y\leftarrow X}\,.$$

Also the supremum of the left-hand side satisfies the estimate $\|S+T\|_{Y\leftarrow X} \le \|S\|_{Y\leftarrow X} + \|T\|_{Y\leftarrow X}$. If $\|T\|_{Y\leftarrow X} = 0$, $\|Tx\|_Y = 0$ follows for all x, i.e., T is the zero mapping: $T = 0$.

Solution of Exercise 6.6. (a) If $x = 0$, $\|Tx\|_Y \le \|T\|_{Y\leftarrow X}\,\|x\|_X$ is trivial. Otherwise $\|Tx\|_Y \,/\, \|x\|_X$ can be bounded by the supremum $\|T\|_{Y\leftarrow X}$.

(b) Part (a) shows

$$\|T_1T_2x\|_Z \le \|T_1\|_{Z\leftarrow Y}\,\|T_2x\|_Y \le \|T_1\|_{Z\leftarrow Y}\,\|T_2\|_{Y\leftarrow X}\,\|x\|_X\,,$$

hence $\|T_1\,T_2\,x\|_Z \,/\, \|x\|_X \le \|T_1\|_{Z\leftarrow Y}\,\|T\|_{Y\leftarrow X}$. Since the right-hand side is independent of x, it is also an upper bound of the supremum $\|T_1T_2\|_{Z\leftarrow X}$.

Solution of Exercise 6.8. Let $T_n \in L(X,Y)$ be a Cauchy sequence. For any x, T_nx is also a Cauchy sequence in Y since $\|T_nx - T_mx\|_Y \le \|T_n - T_m\|_{Y\leftarrow X}\,\|x\|_X$. Hence the limit $\lim T_nx$ exists and defines the map $x \mapsto Tx$. It is easy to verify that T is linear. The Cauchy property implies $C := \sup_n \|T_n\|_{Y\leftarrow X} < \infty$ so that $\|Tx\|_Y \,/\, \|x\|_X = \lim_n \|T_nx\|_Y \,/\, \|x\|_X \le C$, i.e., T is bounded: $T \in L(X,Y)$. Therefore $L(X,Y)$ is not only a normed space but also complete, i.e., $L(X,Y)$ is a Banach space.

Solution of Exercise 6.11. The equivalence of the norms implies that the completion yields the same space as a set. Let $x \in X$ be the limit of $x_n \in X_0$ with respect to both norms. Continuity of the norms yields

$$\|x\| = \lim_n \|x_n\| \le C \lim_n |||x_n||| = |||x|||$$

as well as the reverse inequality.

Solution of Exercise 6.13. (a) We denote the embeddings by $I_{Y\leftarrow X}$ and $I_{Z\leftarrow Y}$. Continuity yields $I_{Y\leftarrow X} \in L(X,Y)$ and $I_{Z\leftarrow Y} \in L(Y,Z)$. By Exercise 6.6 the product $I_{Z\leftarrow Y} I_{Y\leftarrow X}$ belongs to $L(X,Z)$ and describes the embedding $I_{Z\leftarrow X}$. The boundedness $\|I_{Z\leftarrow Y}\|_{Z\leftarrow Y} \le C$ is identical with the statement

$$\|y\|_Z = \|I_{Z\leftarrow Y}\, y\|_Z \le C\,\|y\|_Y \qquad \text{for all } y \in Y.$$

(b) In the case of dense embeddings one has to show that X is dense in Z. Let $z \in Z$. For all $\varepsilon > 0$ one has to show that there is an $x \in X$ with $\|z-x\|_Z \le \varepsilon$. There is a $y \in Y$ with $\|z-x\|_Z \le \varepsilon/2$ and an $x \in X$ with $\|y-x\|_Y \le \varepsilon/(2C)$. Together with $\|y-x\|_Z \le C\,\|y-x\|_Y$ from part (a) the assertion follows.

Solution of Exercise 6.14. (b) Inequality (6.6) is invariant with respect to a scaling of x and y, so that without loss of generality $\|x\| = \|y\| = 1$ may be assumed. Since (6.6) is trivial for $(x,y) = 0$, assume $(x,y) \neq 0$. Use

$$0 \le \|x-\lambda y\|^2 = (x-\lambda y, x-\lambda y) = \|x\|^2 - \lambda\cdot(y,x) - \bar\lambda\cdot(x,y) + |\lambda|^2\,\|y\|^2$$

with $\|x\| = \|y\| = 1$ and the choice $\lambda = 1/(y,x)$. Then $0 \le -1 + |\lambda|^2$ and

$$1/|\lambda| = |(x,y)| \le 1 = \|x\|\,\|y\|\,.$$

(a) The triangle inequality follows from

$$\|x+y\|^2 = \|x\|^2 + 2\,\Re\mathfrak{e}\,(x,y) + \|y\|^2 \le \|x\|^2 + 2\,\|x\|\,\|y\| + \|y\|^2 = (\|x\| + \|y\|)^2\,.$$

(c) For sequences $x_n \to x$ and $y_n \to y$ we obtain $|(x,y) - (x_n,y_n)| = |(x-x_n,y) + (x_n,y-y_n)| \le \|x-x_n\|\,\|y_n\| + \|x_n\|\,\|y-y_n\| \to 0$.

Solution of Exercise 6.17. (a) $\Rightarrow$ (b): Let A be dense in X and $0 \neq x \in X$. Then there is an element $a \in A$ with $\|x-a\| \le \|x\|/2$. Hence

$$\begin{aligned}(a,x)_X &= (x,x)_X - (x-a,x)_X \ge \|x\|_X^2 - \|x-a\|_X\,\|x\|_X \\ &\ge \|x\|_X^2 - \frac{1}{2}\,\|x\|_X^2 = \frac{1}{2}\,\|x\|_X^2 > 0.\end{aligned}$$

(b) $\Rightarrow$ (a): Assume that A is not dense. Then $U := \overline{A} \subsetneqq X$. According to Lemma 6.15 there exists a nontrivial subspace $U^\perp \subset X$. For $0 \neq x \in U^\perp$ we have $(a,x)_X = 0$ for all $a \in A \subset U$ in contradiction to statement (b).

Solution of Exercise 6.21. (a) Let $v = D^\alpha u \in L^2(\Omega)$ be the weak derivative and $\varphi \in C^0(\Omega')$ the classical one. For all $w \in C_0^\infty(\Omega') \subset C_0^\infty(\Omega)$ integration by parts yields $(w,\varphi)_0 = (w, D^\alpha u)_0 = (-1)^{|\alpha|}(D^\alpha w, u)_0$. Subtraction of (6.10) shows $(w, \varphi - v)_0 = 0$ for all $w \in C_0^\infty(\Omega')$. Since $A := C_0^\infty(\Omega')$ is dense in $X := L^2(\Omega')$ (cf. Lemma 6.19), Exercise 6.17 is applicable: If $\varphi \neq v$ in Ω', there is a $w \in C_0^\infty(\Omega')$ with $(w,\varphi - v)_0 \neq 0$ in contradiction to the previous characterisation.

(b) By assumption

$$(w, v_\alpha)_0 = (-1)^{|\alpha|} (D^\alpha w, u)_0 \quad \text{and} \quad (\varphi, v_{\alpha+\beta})_0 = (-1)^{|\beta|} \left(D^\beta \varphi, v_\alpha\right)_0$$

holds for all $w, \varphi \in C_0^\infty(\Omega)$. Since $w := D^\beta \varphi$ again belongs to $C_0^\infty(\Omega)$ it follows that $(\varphi, v_{\alpha+\beta})_0 = (-1)^{|\beta|} \left(D^\beta \varphi, v_\alpha\right)_0 = (-1)^{|\beta|} (-1)^{|\alpha|} \left(D^\alpha D^\beta \varphi, u\right)_0 = (-1)^{|\alpha+\beta|} \left(D^{\alpha+\beta} \varphi, u\right)_0$. Hence $v_{\alpha+\beta}$ is the weak $D^{\alpha+\beta}$-derivative of u.

(c) The case $\sigma = 0$ is trivial. The strong derivative $\frac{\partial}{\partial x_i} |x|^\sigma = \sigma x_i |x|^{\sigma-2}$ exists for $x \neq 0$. In $\Omega_\varepsilon := \Omega \backslash K_\varepsilon(0)$ we have

$$\int_{\Omega_\varepsilon} |x|^\sigma D^\alpha w(x) \mathrm{d}\mathbf{x} = -\sigma \int_{\Omega_\varepsilon} x_i |x|^{\sigma-2} w(x) \mathrm{d}\mathbf{x} + \int_{\partial K_\varepsilon(0)} |x|^\sigma n_i w(x) \mathrm{d}\Gamma$$

for $D^\alpha = \frac{\partial}{\partial x_i}$, $w \in C_0^\infty(\Omega)$. Since $\left| x_i |x|^{\sigma-2} \right| \leq |x|^{\sigma-1}$ and $\int_{K_\varepsilon(0)} |x|^{\sigma-1} \mathrm{d}\mathbf{x} = \omega_n \int_0^\varepsilon r^{\sigma+n-2} \mathrm{d}r = \frac{\omega_n}{\sigma+n-1} \varepsilon^{\sigma+n-1} \to 0$ for $\varepsilon \to 0$, $\int_{\Omega_\varepsilon} \to \int_\Omega$ follows. The boundary integral tends to zero, since $\int_{\partial K_\varepsilon(0)} |x|^\sigma \mathrm{d}\Gamma = \omega_n \varepsilon^{\sigma+n-1} \to 0$. It remains to check that the weak derivative $\sigma x_i |x|^{\sigma-2}$ is square-integrable. This requires $2\sigma + n > 2$.

(d) The limit process $\nu \to \infty$ in $(D^\alpha u_\nu, v)_0 = (-1)^{|\alpha|} (D^\alpha w, u_\nu)_0$ for $w \in C_0^\infty(\Omega)$ yields (6.10).

Solution of Exercise 6.24. The classical derivative $u_x = \frac{x}{r^2 \log(r/2)}$ with $r^2 = x^2 + y^2$ exists for $r > 0$ and is discontinuous at $r = 0$. Since $\frac{x}{r^2 \log(r/2)}$ is improperly integrable, it follows as in Exercise 6.21c that $\frac{x}{r^2 \log(r/2)}$ is the weak derivative in $\mathbb{R}^2$. Also $u_x^2 = \mathcal{O}(1/(r^2 \log^2 r))$ is improperly integrable in Ω. Similar for u_y^2. Hence u belongs to $H^1(\Omega)$.

Solution of Exercise 6.30. (a) $u(x) = 1$ belongs to $H^k(\Omega)$ with the norm $|u|_{k,0} = 0$ for $k \geq 1$. If $H^k(\Omega) = H_0^k(\Omega)$, Lemma 6.29 would imply $|u|_k = 0$ in contradiction to $u \neq 0$.

(b) Without loss of generality let $\nu = 1$. The proof of Lemma 6.29 can be applied without modification.

Solution of Exercise 6.40. (a) Integration by parts yields

$$\begin{aligned}(2\pi)^{n/2} \left(\mathcal{F} u_{x_k}\right)(\xi) &= \int_{\mathbb{R}^n} \mathrm{e}^{-\mathrm{i}\langle \xi, x\rangle} \frac{\partial}{\partial x_k} u(x) \mathrm{d}\mathbf{x} = -\int_{\mathbb{R}^n} \left(\frac{\partial}{\partial x_k} \mathrm{e}^{-\mathrm{i}\langle \xi, x\rangle} \right) u(x) \mathrm{d}\mathbf{x} \\ &= \mathrm{i}\xi_k \int_{\mathbb{R}^n} \mathrm{e}^{-\mathrm{i}\langle \xi, x\rangle} u(x) \mathrm{d}\mathbf{x}.\end{aligned}$$

Multiple application proves the assertion.

(b) One verifies that the ratios $\frac{\sum_{|\alpha| \leq k} |\xi^\alpha|^2}{(1+|\xi|^2)^k}$ and $\frac{(1+|\xi|^2)^k}{\sum_{|\alpha| \leq k} |\xi^\alpha|^2}$ are bounded in $\mathbb{R}^n$.

Solution of Exercise 6.44. (a) $\hat{u}(\xi) = (2\pi)^{-1/2} \int_{-1}^{1} \mathrm{e}^{-\mathrm{i}\xi x} \mathrm{d}x = \sqrt{\frac{2}{\pi}} \frac{\sin(\xi)}{\xi}$ is analytic in $\xi = 0$. $(1+\xi^2)^s \, \hat{u}(\xi)^2$ behaves like ξ^{2s-2} as $\xi \to \pm\infty$, hence it is integrable in $\mathbb{R}$ for $s < 1/2$, so that $|u|_s^{\wedge} < \infty$. However, for $s \geq 1/2$ the integral diverges.

(b) For $|x| < 1$ and $y > 1$ we have $u(x) - u(y) = 1$. The integral appearing in the Sobolev–Slobodeckij norm is $\int_1^\infty \frac{\mathrm{d}y}{|x-y|^{1+2s}}$ and has the value $\frac{(1-x)^{-2s}}{2s}$. $\int_{-1}^{1} \frac{\mathrm{d}x}{(1-x)^{2s}}$ exists as improper integral if and only if $s < 1/2$. Note that the integral $\int_1^\infty \int_{-1}^{1} \ldots \mathrm{d}x\mathrm{d}y$ is a part of the integral $\iint\limits_{\Omega\times\Omega} \ldots \mathrm{d}x\mathrm{d}y$ in (6.22a).

Solution of Exercise 6.53. (a) After a rotation the corners of the rectangle and of the L-shaped domain can be described by the Lipschitz-continuous functions $t \mapsto |t|$.

(b) The cut circle cannot be described by a function.

Solution of Exercise 6.66. $\langle TSx, z' \rangle_{Z\times Z'} = \langle Sx, T'z' \rangle_{Y\times Y'} = \langle x, S'T'z' \rangle_{X\times X'}$ proves the assertion.

Solution of Exercise 6.67. Let $T'y' = 0$. If the functional y' does not vanish, there is a $y \in Y$ with $\langle y, y' \rangle_{Y\times Y'} \neq 0$. Since $T \in L(X,Y)$ is surjective, there is an $x \in X$ with $Tx = y$. $0 \neq \langle y, y' \rangle_{Y\times Y'} = \langle Tx, y' \rangle_{Y\times Y'} = \langle x, T'y' \rangle_{X\times X'} = \langle x, 0 \rangle_{X\times X'} = 0$ yields the contradiction. Therefore $y' = 0$ holds and implies that T' is injective.

Solution of Exercise 6.75. $H_0^s(\Omega) \subset L^2(\Omega)$ holds for $s \geq 0$. The embedding is continuous since $\|\cdot\|_{H^s(\Omega)} \leq \|\cdot\|_{L^2(\Omega)}$. Since $C_0^\infty(\Omega)$ is dense in both spaces, the embedding is also dense. Similar for $H^s(\Omega)$ with $C^\infty(\Omega)$ as dense subspace.

Solution of Exercise 6.80. The image of the unit ball is the unit ball itself. According to Remark 6.78 this ball is (relatively) compact if and only if $\dim(X) < \infty$.

Solution of Exercise 6.83. Write T as $T \cdot I$ if $\dim X < \infty$, or as $I \cdot T$ if $\dim Y < \infty$, and apply Lemma 6.82a and Exercise 6.80.

Solution of Exercise 6.92. The definition of the dual operator A' is

$$a(x,y) = \langle Ax, y \rangle_{V'\times V} = \langle Ax, y \rangle_{V'\times V''} = \langle x, A'y \rangle_{V\times V'},$$

where the middle equation uses $V = V''$ (cf. Conclusion 6.69b). Now

$$\langle x, A'y \rangle_{V\times V'} = (A'y)(x) = \langle A'y, x \rangle_{V'\times V} = a^*(y,x)$$

yields assertion (a). For symmetric a we have $\langle Ax, y \rangle_{V'\times V} = \langle A'x, y \rangle_{V'\times V}$ for all $x, y \in V$ and thus $A = A'$.

Solution of Exercise 6.95. If $n := \dim V < \infty$, V is isomorphic to $\mathbb{R}^n$. The map $A : V \to V'$ becomes an $n \times n$ matrix.

Assertion 1: (6.43a) is equivalent to 'A is injective'.

Assertion 2: (6.43b) is equivalent to 'A is surjective'.

For matrices (and general compact operators) the properties *injective*, *surjective,* and *regular* are equivalent. Hence also (6.43a) and (6.43b) are equivalent. Lemma 6.94 implies $\varepsilon' = \varepsilon$.

For an indirect proof of Assertion 1 assume that A is not injective. Then there is some x with $\|x\|_V = 1$ and $Ax = 0$, i.e., $|\langle Ax, y\rangle| = |a(x, y)| = 0$ for all y in contradiction to (6.43a).

For an indirect proof of Assertion 2 assume that $V_A := \operatorname{range}(A) \subsetneqq \mathbb{R}^n$. Then there is a $0 \neq y \perp V_A$ and therefore $|\langle Ax, y\rangle| = 0$ in contradiction to (6.43b).

Solution of Exercise 6.98. (a) Restriction of $|a(v, v)| \geq C_E \|v\|_V^2$ $(v \in V)$ to $W \subset V$ shows $|a(w, w)| \geq C_E \|w\|_V^2 \geq C_E C^2 \|w\|_W^2$ $(x \in W)$, where C is taken from the inequality $\|w\|_W \leq \|w\|_V / C$.

(b) Let $V \backslash \{0\} \ni x = \lim x_i$ with $x_i \in V_0$. The limit in $a(x_i, x_i) / \|x_i\|_V^2 \geq C_E$ proves the inequality for all $x \in V$.

(c) Since $(x, y)_a := a(x, y)$ satisfies all axioms (6.5), the form $(\cdot, \cdot)_a$ is a scalar product which defines the norm $\|x\|_a$ (cf. Exercise 6.14). Continuity of a implies the inequality $\|x\|_a \leq \sqrt{C_S} \|x\|_V$ (cf. (6.42)). The *V-ellipticity* (6.44) yields the reverse inequality $\|x\|_a \geq \sqrt{C_E} \|x\|_V$ ($C_E > 0$). Hence the norms $\|x\|_a$ and $\|x\|_V$ are equivalent.

Solution of Exercise 6.103. Let $y \in V$ be arbitrary with $y = \lim_i y_i$ and $y_i \in V_0$. The limit process in $a(x, y_i) = f(y_i)$ also proves $a(x, y) = f(y)$ for $y \in V$.

Solution of Exercise 6.106. Part (a) is trivial. (b) We have

$$\tilde{a}(x, y) = a(x, y) + C_K \cdot (x, y)_U = \langle Ax, y\rangle_{V' \times V} + \langle x, y\rangle_{V' \times V}$$

for $x, y \in V$ (cf. Remark 6.74). This shows $\tilde{a}(x, y) = \langle \tilde{A}x, y\rangle_{V' \times V}$.

Exercises of Chapter 7

Solution of Exercise 7.1. (a) m-fold integration by parts of $\int_\Omega (Lu) v \mathrm{d}\mathbf{x}$ with $v \in C_0^\infty(\Omega)$ yields the assertion.

(b) The solutions in Example 2.26 are infinitely differentiable, but not bounded. Even if the solution is bounded as, e.g., $u = \text{const} \neq 0$, it does not belong to $L^2(\Omega)$ and hence also not to $H^1(\Omega)$. If $u \in H^1(\Omega)$ holds for unbounded domains, u and ∇u must decay in a suitable way as $|x| \to \infty$. This can be regarded as a boundary condition at ∞.

Solution of Exercise 7.6. If $a_{\alpha 0} \neq 0$ for $\alpha = (1, 0, \ldots)$, $a(u, u)$ contains the additional integral $\int_\Omega a_{\alpha 0} \frac{\partial u}{\partial x_1} u \mathrm{d}\mathbf{x} = \frac{a_{\alpha 0}}{2} \int_\Omega \frac{\partial}{\partial x_1} u^2 \mathrm{d}\mathbf{x}$. Since the antiderivative u^2 vanishes at the boundary, this term does not change $a(u, u)$. More generally, the addition of antisymmetric bilinear forms (i.e., $a(u, v) = -a(v, u)$) does not disturb the V-ellipticity.

Solution of Exercise 7.17. The critical part is $a(u_0, v)$. For fixed $u_0 \in V$ the map $v \mapsto a(u_0, v) \in \mathbb{R}$ is a (continuous) functional with the V'-norm

$$\sup_{v \in V} \frac{|a(u_0, v)|}{\|v\|_V} \underset{(6.42)}{\leq} \sup_{v \in V} \frac{C_S \|u_0\|_V \|v\|_V}{\|v\|_V} = C_S \|u_0\|_V .$$

Together we obtain (7.18).

Solution of Exercise 7.22. Let $u \in H^1(\Omega)$ be the solution. All other H^1-functions satisfying the posed boundary conditions are of the form $u + v$ with $v \in H_0^1(\Omega)$. As in Theorem 6.104 the inequality $J(u+v) = J(u) + a(v,v) \geq J(u) + C_E |v|_1^2$ proves that $J(u)$ is minimal.

Solution of Exercise 7.24. (a) Using the inequalities $\|\cdot\|_{L^2(\Omega)} \leq \|\cdot\|_{H^1(\Omega)}$ and $\|v\|_{L^2(\Gamma)} \leq \|v\|_{H^{1/2}(\Gamma)} \leq C \|v\|_{H^1(\Omega)}$ and

$$\begin{aligned} |f(v)| &\leq \|g\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)} + \|\varphi\|_{L^2(\Gamma)} \|\gamma_\Gamma v\|_{L^2(\Gamma)} \\ &\leq \Big[\|g\|_{H^1(\Omega)} + C \|\varphi\|_{L^2(\Gamma)} \Big] \|v\|_{H^1(\Omega)} , \end{aligned}$$

we obtain the first inequality ($\gamma_\Gamma v$ is the restriction of v on Γ).

(b) For $v \in H^1(\Omega)$ (and therefore $\gamma_\Gamma v \in H^{1/2}(\Gamma)$) we have

$$\int_\Omega g v \mathrm{d}\mathbf{x} = \langle g, v \rangle_{(H^1(\Omega))' \times H^1(\Omega)} \quad \text{and} \quad \int_\Gamma \varphi v \mathrm{d}\mathbf{x} = \langle g, v \rangle_{H^{-1/2}(\Gamma) \times H^{1/2}(\Gamma)} .$$

This proves (7.20b).

Solution of Exercise 7.28. The coefficients are $a_{11} = a_{22} = -1$, $a_{12} = a_{21} = a_{0i} = 0$. According to (7.24), $B = -\sum_{i=1}^n n_i \frac{\partial}{\partial x_i} = -\frac{\partial}{\partial n}$ is the boundary differential operator.

Solution of Exercise 7.37. (a) $a(u,u) \geq c \|u\|_{H^1(\Omega)}^2$ holds for $c := \min\{1, \alpha\} > 0$.

(b) If integration by parts is possible (i.e., if u and Γ are sufficiently smooth), Green's formula (2.6a) yields

$$\int_\Omega [\langle \nabla u, \nabla v \rangle + auv] \mathrm{d}\mathbf{x} = \int_\Omega v[-\Delta u + au] \mathrm{d}\mathbf{x} + \int_\Gamma v \frac{\partial u}{\partial n} \mathrm{d}\Gamma.$$

The above form is equal to $\int_\Omega gv \mathrm{d}\mathbf{x} + \int_\Gamma \varphi v \mathrm{d}\Gamma$. Variation of $v \in H_0^1(\Omega) \subset V$ leads to $-\Delta u + au = g$ in Ω. It remains $\int_\Gamma v \frac{\partial u}{\partial n} \mathrm{d}\Gamma = \int_\Gamma \varphi v \mathrm{d}\Gamma$. Since all $v \in V$ have constant boundary values we obtain the scalar equation $\int_\Gamma \frac{\partial u}{\partial n} \mathrm{d}\Gamma = \int_\Gamma \varphi \mathrm{d}\Gamma$ proving part (b).

Exercises of Chapter 8

Solution of Exercise 8.10. (a) Let $P^\star w = 0$. The definition of $P^\star$ yields $0 = \langle P^\star w, \mathbf{u}\rangle = (w, P\mathbf{u})_U$ for all $\mathbf{u} \in \mathbb{R}^N$. Since $\mathrm{range}(P) = V_N$, the equation $P^\star w = 0$ is equivalent to $w \perp V_N$. Therefore $P^\star$ as a mapping $P^\star|_{V_N} : V_N \to \mathbb{R}^N$ has a trivial kernel, i.e., $P^\star$ is injective. Since $\dim(V_N) = \dim(\mathbb{R}^N)$, $P^\star$ is also surjective, thus bijective.

(b) Let $\mathbf{x} \in \mathbb{R}^N$ and $u \in V_N$. u is the image $u = P\mathbf{y}$ of some $\mathbf{y} \in \mathbb{R}^N$. Hence $((P^{-1})^*\mathbf{x}, u)_U = \langle \mathbf{x}, P^{-1}u\rangle = \langle \mathbf{x}, P^{-1}P\mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{y}\rangle$. According to part (a) there is a $v \in V_N$ with $\mathbf{x} = P^*v$. This shows that

$$\left((P^*)^{-1*}\mathbf{x}, u\right)_U = (v, u)_U = (v, P\mathbf{y})_U = \langle P^*v, \mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{y}\rangle .$$

$((P^{-1})^*\mathbf{x}, u)_U = \left((P^*)^{-1*}\mathbf{x}, u\right)_U$ for all $u \in V_N$ shows $(P^{-1})^*\mathbf{x} = (P^*)^{-1}\mathbf{x}$ and $(P^{-1})^* = (P^*)^{-1}$.

Solution of Exercise 8.17. Part (a) follows from (8.8a). $\mathbf{x} \neq 0$ implies that $u := P\mathbf{x} \neq 0$, so that according to Remark 8.6 $\langle \mathbf{L}\mathbf{x}, \mathbf{x}\rangle = a(u,u) \geq C_E \|u\|_V^2 > 0$ follows. Hence $\mathbf{L}$ is positive definite. Rewrite $J(u) = J(P\mathbf{x})$ as $\langle \mathbf{L}\mathbf{x}, \mathbf{x}\rangle - 2\langle \mathbf{f}, \mathbf{x}\rangle$. This expression is minimal for $\mathbf{x}^* = \mathbf{L}^{-1}\mathbf{f}$, so that $J(u^N) = J(P\mathbf{x}^*)$ gives the minimum.

Solution of Exercise 8.26. u is the solution of $Lu = f$. Remark 8.9 states that $\mathbf{f} = P^*f$ is the right-hand side of equation (8.9) with the solution $\mathbf{u} = \mathbf{L}^{-1}\mathbf{f}$. The Ritz–Galerkin solution is $u^N = P\mathbf{u}$. This proves $S_N = P\,\mathbf{L}^{-1}P^*L$.

Solution of Exercise 8.34. The Gauss quadrature must be exact for polynomials up to degree 1. For $g = 1$ and $g = x - x_i$ one verifies that $f_{i,\mathrm{GQuad}} = \int_{x_{i-1}}^{x_{i+1}} g b_i \mathrm{d}\mathbf{x}$. For equal step sizes h we get $f_{i,\mathrm{GQuad}} := h\, g(x_i)$.

Solution of Exercise 8.35. The coefficients $L_{i,i\pm1} = -1/h$, $L_{ii} = 2/h$ in Remark 8.33 hold for $2 \leq i \leq N-1$, i.e., for the inner grid points. For $i = 1$ the integration over $[0, h]$ yields $L_{00} = 1/h$, $L_{01} = -1/h$. Correspondingly, we have $L_{NN} = 1/h$, $L_{N-1,N} = -1/h$. $\mathbf{L}^\mathsf{T}\mathbf{1} = \mathbf{L}\mathbf{1} = 0$ shows that the constant solution $\mathbf{1}$ belongs to the kernel. Since $\mathrm{rank}(\mathbf{L}) \geq N-1$, $\mathbf{L}\mathbf{u} = \mathbf{f}$ is solvable if $\langle \mathbf{1}, \mathbf{f}\rangle = 0$. For the Neumann problem the functional f is of the form $f(v) = \int_a^b gv\mathrm{d}x + \varphi_a v(a) + \varphi_b v(b)$ (cf. (7.20a)). Inserting the basis functions, we obtain

$$f_1 = \int_a^b g b_1 \mathrm{d}x + \varphi_a, \quad f_i = \int_a^b g b_i \mathrm{d}x$$

for $2 \leq i \leq N-1$ and $f_N = \int_a^b g b_1 \mathrm{d}x + \varphi_b$. Since the sum of all basis functions is equal to the constant 1, the equation is solvable if $\langle \mathbf{1}, \mathbf{f}\rangle = \int_a^b g(x)\mathrm{d}x + \varphi_a + \varphi_b$ vanishes.

Solution of Exercise 8.42. For piecewise linear basis functions the gradient is piecewise constant; more precisely, they take the values 0 and $\pm 1/h$. For diagonally neighboured $\mathbf{x}^i, \mathbf{x}^j$ the products $(b_i)_x(b_j)_x$ and $(b_i)_y(b_j)_y$ contain one vanishing factor so that $L_{ij} = 0$. For horizontally neighboured $\mathbf{x}^i$, $\mathbf{x}^j$ we find $(b_i)_x(b_j)_x = h^{-2}$ and $(b_i)_y(b_j)_y = 0$ in two triangles. Integration yields $L_{ij} = -1$. Analogously $L_{ij} = -1$ results for vertically neighboured $\mathbf{x}^i, \mathbf{x}^j$. Similarly one shows $L_{ii} = 4$.

Solution of Exercise 8.43. (a) The corners of T are $(\xi, \eta) \in \{(0,0), (1,0), (1,0)\}$. Inserting these values into Φ, we obtain the vertices of $\tilde{T}$. Since Φ is linear, the sides [resp. inner area] of T are mapped onto the sides [resp. inner area] of $\tilde{T}$.

(b) $\partial\Phi/\partial\xi = \mathbf{x}^2 - \mathbf{x}^1$ and $\partial\Phi/\partial\eta = \mathbf{x}^3 - \mathbf{x}^1$ define the constant determinant $\det \Phi' = \det \left[\mathbf{x}^2 - \mathbf{x}^1 \;\; \mathbf{x}^3 - \mathbf{x}^1\right]$ as described in (b).

(c) Since $\det \Phi'$ is constant, it can be taken in front of the integral.

Solution of Exercise 8.46. Without loss of generality let $\mathbf{x}^i = (0,0)$. The basis function b_i has the support $[-h,h] \times [-h,h]$ and therein it is defined by $(1 - \frac{|x|}{h})(1 - \frac{|y|}{h})$. The x-derivative is $\frac{\operatorname{sign}(-x)}{h}(1 - \frac{|y|}{h})$. This leads to

$$\int_\Omega \frac{\partial b_i}{\partial x}\frac{\partial b_i}{\partial x}\mathrm{d}x\mathrm{d}y = \int_{[-h,h]\times[-h,h]} \ldots \mathrm{d}x\mathrm{d}y = \int_{-h}^{h} h^{-2}\mathrm{d}x \int_{-h}^{h} (1 - \frac{|y|}{h})^2 \mathrm{d}y = \frac{4}{3}.$$

For the right neighbour $\mathbf{x}^j = (h,0)$ the basis functions in $[0,h]\times[-h,h]$ yields $\partial b_j/\partial x = \frac{1}{h}(1 - \frac{|y|}{h})$; hence

$$\int_\Omega \frac{\partial b_i}{\partial x}\frac{\partial b_j}{\partial x}\mathrm{d}x\mathrm{d}y = \int_{[0,h]\times[-h,h]} \ldots \mathrm{d}x\mathrm{d}y = \int_{0}^{h} -h^{-2}\mathrm{d}x \int_{-h}^{h} (1 - \frac{|y|}{h})^2 \mathrm{d}y = -\frac{2}{3}.$$

In the case of the vertical neighbour $\mathbf{x}^j = (0,h)$ the intersection of the support is $[-h,h] \times [0,h]$ so that

$$\int_\Omega \frac{\partial b_i}{\partial x}\frac{\partial b_j}{\partial x}\mathrm{d}x\mathrm{d}y = \int_{[-h,h]\times[0,h]} \ldots \mathrm{d}x\mathrm{d}y = \int_{-h}^{h} h^{-2}\mathrm{d}x \int_{0}^{h} (1 - \frac{|y|}{h})\frac{y}{h} \mathrm{d}y = \frac{1}{3}.$$

The diagonal neighbour at $\mathbf{x}^j = (h,h)$ yields

$$\int_\Omega \frac{\partial b_i}{\partial x}\frac{\partial b_j}{\partial x}\mathrm{d}x\mathrm{d}y = \int_{[0,h]\times[0,h]} \frac{\partial b_i}{\partial x}\frac{\partial b_j}{\partial x}\mathrm{d}x\mathrm{d}y = \int_{0}^{h} -h^{-2}\mathrm{d}x \int_{0}^{h} (1 - \frac{y}{h})\frac{y}{h} \mathrm{d}y = -\frac{1}{6}.$$

By symmetry reasons one obtains from the part $\int_\Omega \frac{\partial b_i}{\partial x}\frac{\partial b_j}{\partial x}\mathrm{d}x\mathrm{d}y$ the difference star $\frac{1}{6}\begin{bmatrix} -1 & 2 & -1 \\ -4 & 8 & -4 \\ -1 & 2 & -1 \end{bmatrix}$. Analogously $\frac{1}{6}\begin{bmatrix} -1 & -4 & -1 \\ 2 & 8 & 2 \\ -1 & -4 & -1 \end{bmatrix}$ for $\int_\Omega \frac{\partial b_i}{\partial y}\frac{\partial b_j}{\partial y}\mathrm{d}x\mathrm{d}y$. The sum yields the desired result.

Solution of Exercise 8.67. According to Exercise 8.26 $S_N = P\,\mathbf{L}^{-1}P^*L$ is the Ritz projection. The definition of $\mathbf{L}$ in (8.8a) shows that $a^\star$ corresponding to the operator L^* leads to the system matrix $\mathbf{L}^\mathsf{T}$. Replacing L, $\mathbf{L}$ by $L^*, \mathbf{L}^\mathsf{T}$, we obtain the Ritz projection $\hat{S}_h = P\left(\mathbf{L}^\mathsf{T}\right)^{-1} P^*L^*$ for the adjoint problem. (8.67) follows from

$$L^{-1}\hat{S}_h^*L = L^{-1}(P\left(\mathbf{L}^\mathsf{T}\right)^{-1} P^*L^*)^*L = L^{-1}LP\mathbf{L}^{-1}P^*L = S_N.$$

Solution of Exercise 8.71. As a demonstration we determine $a(b_{1i}, b_{2,i-1})$. The second derivatives in $[x_{i-1}, x_i]$ are

$$b''_{1i}(x) = -6h - 12\,(x - x_i) \quad \text{and} \quad b''_{2,i-1} = 2h + 6\,(x - x_i)\,,$$

so that

$$a(b_{1i}, b_{2,i-1}) = \int_{x_{i-1}}^{x_i} b''_{1i}(x)\, b''_{2,i-1}(x)\,\mathrm{d}x = 6h^{-2}.$$

The other coefficients can be computed analogously.

Solution of Exercise 8.78. (a) Let the transformation $\Phi : T \to \tilde{T}$ be linear. Then

$$|\det \Phi'(\xi, \eta)| = \mathrm{area}(\tilde{T})/\mathrm{area}(T).$$

Thus in the case of the square-grid triangulation we have $|\det \Phi'(\xi, \eta)| = h^2$. The constants in the proof of Theorem 8.76 become $C_1 = C_2 = 1$ and $M_{\min} = M_{\max} = 6$. (8.90) implies $\frac{1}{2}\,\|\mathbf{u}\|_h \le \|\mathbf{u}\|_P \le \|\mathbf{u}\|_h$.

(b) Assume $h = 1$ (the later scaling yields the factor h^2). The central coefficient $1/2$ is the integral over six triangles in $[-1, 1] \times [-1, 1]$. By symmetry reasons all four triangles possessing a hypotenuse containing the origin have the value

$$\int_0^1 \int_0^x (1 - x)^2 \mathrm{d}y\mathrm{d}x = \frac{1}{12}\,.$$

The remaining other two triangles lead to the value $\int_0^1 \int_{x-1}^0 (1-x+y)^2 \mathrm{d}y\mathrm{d}x = \frac{1}{12}$. Analogously, one determines the other coefficients.

Solution of Exercise 8.84. By assumption the identity map $I : U_h \to V_h$ has the operator norm $\|I\|_{V_h \leftarrow U_h} \le C_I h^{-m}$ which coincides with $\|I\|_{U_h \leftarrow V_h'}$, i.e.,

$$\|u\|_U \le \|I\|_{U_h \leftarrow V_h'} \|u\|_{V_h'} \le C_I h^{-m} \|u\|_{V'}$$

by Conclusion 8.8. Combining both inequalities proves $\|u\|_V \le C_I^2 h^{-2m} \|u\|_{V'}$.

Solution of Exercise 8.86. (a) In general, functions in V_h are discontinuous across the triangles sides. The weak derivatives are Dirac distributions which do not belong to $L^2(\Omega)$, hence $V_h \not\subset H^1(\Omega)$.

(b) $T \in \mathcal{T}$ satisfies the Poincaré inequality

$$\int_T |u - \bar{u}|^2 \mathrm{d}\mathbf{x} \le \left(\frac{h}{\pi}\right)^2 \int_T |\nabla u|^2 dx \tag{A.9}$$

where $\bar{u} = \frac{1}{\mathrm{area}(T)} \int_T u \mathrm{d}\mathbf{x}$ is the mean value (cf. Payne–Weinberger [217] and [30]). Here we use that T is convex and that h is the upper bound of the diameter of T. Summation over all $T \in \mathcal{T}$ shows $|u-v|_0 \le \frac{h}{\pi} |u|_{1,0} \le \frac{h}{\pi} |u|_1$ and proves part (b).

(c) The basis functions b_i have a support consisting of only one triangle $T_i \in \mathcal{T}$. Therefore $\mathbf{L}$ is a diagonal matrix with $L_{ii} = \mathrm{area}(T_i)$, since $L_{ii} = a_h(b_i, b_i) = \int_{T_i} 1^2 \, \mathrm{d}\mathbf{x}$. Note that this discretisation completely ignores the essential Laplace term because of $\nabla b_i = 0$ on T_i.

Exercises of Chapter 9

Solution of Exercise 9.13. Let T_j^μ ($1 \le j \le n$, $\mu \in \mathbb{Z}$) be the translation operator which is defined by $(T_j^\mu u)(\mathbf{x}) = u(\mathbf{x} + \mu h \mathbf{e}_j)$. The substitution $\mathbf{y} := \mathbf{x} + \mu h \mathbf{e}_j$ applied to

$$\left(T_j^\mu u, v\right)_0 = \int_{\mathbb{R}^n} (T_j^\mu u)(\mathbf{x})\, v(\mathbf{x})\, \mathrm{d}\mathbf{x} = \int_{\mathbb{R}^n} u(\mathbf{x} + \mu h \mathbf{e}_j)\, v(\mathbf{x})\, \mathrm{d}\mathbf{x}$$

yields $\int_{\mathbb{R}^n} u(\mathbf{y}) v(\mathbf{y} - \mu h \mathbf{e}_j) \mathrm{d}\mathbf{y} = (u, T_j^{-\mu} v)_0$. This proves

$$(T_j^\mu)^* = T_j^{-\mu}$$

and the statement in part (a). The power series in part (b) should be well known (cf. [315, page 115]).

(c) Insert $\widehat{T_j^\mu u}(\boldsymbol{\xi}) = \mathrm{e}^{\mathrm{i}\xi_j h} \widehat{u}(\boldsymbol{\xi})$ into $R = h^{-\Theta} \sum_{\mu=0}^\infty \mathrm{e}^{-\mu h} (-1)^\mu \binom{\Theta}{\mu} T_j^\mu$. Then part (b) implies part (c).

(d) We have $\left|\left(1 - \mathrm{e}^{-h+\mathrm{i}th}\right)/h\right|^2 = \left(\frac{1-\mathrm{e}^{-h}}{h}\right)^2 \cos^2 \frac{th}{2} + \left(1+\mathrm{e}^{-h}\right)^2 \left(\frac{\sin(th/2)}{h}\right)^2$. The first term can be bounded — up to positive factors — by 1, the second behaves like t^2, i.e., $\left|\left(1 - \mathrm{e}^{-h+\mathrm{i}th}\right)/h\right|^2 \sim 1 + t^2$. This proves part (d).

(e) We use the equivalent norm in (6.21b):

$$\begin{aligned} |R_{h,j} u|_\tau^\wedge &= |(1 + |\boldsymbol{\xi}|^2)^{\tau/2} \widehat{R_{h,j} u}(\boldsymbol{\xi})|_0 = |(1+|\boldsymbol{\xi}|^2)^{\frac{\tau}{2}} \left[\left(1-\mathrm{e}^{-h+\mathrm{i}\xi_j h}\right)/h\right]^\Theta \widehat{u}(\boldsymbol{\xi})|_0 \\ &\sim \left|(1 + |\boldsymbol{\xi}|^2)^{\tau/2} \left(1 + |\xi_j|^2\right)^{\Theta/2} \widehat{u}(\boldsymbol{\xi})\right|_0 . \end{aligned}$$

Hence

$$|R_{h,j} u|_\tau^\wedge \le C_\Theta |(1 + |\boldsymbol{\xi}|^2)^{(\tau+\Theta)/2} \widehat{u}(\boldsymbol{\xi})|_0 = C_\Theta \, |u|_{\tau+\Theta}^\wedge \, .$$

(f) $\sum_{i=1}^n |(\widehat{R_{h,i} u})(\boldsymbol{\xi})|^2 \gtrsim \sum_{i=1}^n \left(1 + |\xi_j|^2\right)^{\Theta/2} \gtrsim (1 + |\boldsymbol{\xi}|^2)^{\Theta/2}$.

Solution of Exercise 9.16. (a) The right-sided derivatives of ϕu are those of u, i.e., $\frac{\partial^k}{\partial x_n^k}\phi u(\mathbf{x}', 0+) = \frac{\partial^k}{\partial x_n^k} u(\mathbf{x}', 0)$ for $0 \le k \le L-1$. For $x_n < 0$ we have

$$\frac{\partial^k}{\partial x_n^k}\phi u(\mathbf{x}', x_n) = (-1)^k \sum_{\nu=1}^{L} a_\nu \left[\nu^k \frac{\partial^k u}{\partial x_n^k}(\mathbf{x}', -\nu x_n) + \nu^{-k} \frac{\partial^k u}{\partial x_n^k}(\mathbf{x}', -x_n/\nu) \right].$$

For $x_n \nearrow 0$ we obtain

$$\frac{\partial^k}{\partial x_n^k}\phi u(\mathbf{x}', 0-) = (-1)^k \sum_{\nu=1}^{L} a_\nu \left[\nu^k + \nu^{-k} \right] \frac{\partial^k u}{\partial x_n^k}(\mathbf{x}', 0+).$$

Therefore continuity of the derivatives up to order $L-1$ is equivalent to the described system of equations.

(b) If $u \in H^k(\mathbb{R}^n_+)$, also $u(\mathbf{x}', -x_n/\nu)$ and $u(\mathbf{x}', -\nu x_n)$ belong to $H^k(\mathbb{R}^n_-)$ with $\mathbb{R}^n_- = \mathbb{R}^n \backslash \mathbb{R}^n_+$. Since the derivatives are continuous, we see that $\phi u \in H^k(\mathbb{R}^n)$ and $\phi \in L(H^k(\mathbb{R}^n_+), H^k(\mathbb{R}^n))$ for $0 \le k \le L$.

Part (c) follows from the definition $\int_{\mathbb{R}^n} (\phi u)\, v \mathrm{d}\mathbf{x} = \int_{\mathbb{R}^n_+} u\, (\phi^* v)\, \mathrm{d}\mathbf{y}$ of the adjoint operator. The prefactors $\nu^{\pm 1}$ result from the substitutions $\mathbf{y} = (\mathbf{x}', -\nu x_n)$ and $\mathbf{y} = (\mathbf{x}', -x_n/\nu)$, respectively

Part (d) follows from the same arguments as in part (a) for derivatives up to order $L-1$.

(e) According to (d) $u \in H^k(\mathbb{R}^n)$ is mapped into $H_0^k(\mathbb{R}^n_+)$ $(0 \le k \le L-2)$, since the requirements of Corollary 6.61 are satisfied. It is easy to estimate the norm of $\phi^* \in L(H^k(\mathbb{R}^n), H_0^k(\mathbb{R}^n_+))$.

(f) The additional statements $\phi \in L(H^k(\mathbb{R}^n_+), H^k(\mathbb{R}^n))$ for negative k are identical to $\phi^* \in L(H^{-k}(\mathbb{R}^n), H_0^{-k}(\mathbb{R}^n_+))$ from part (e).

Solution of Exercise 9.26. The bilinear form corresponding to the Helmholtz problem $-\Delta u + u = f$ in Ω and $u = 0$ on Γ is $a(u, v) = (u, v)_V$. Let u_h be the Galerkin solution in V_h. The defining equation $a(u_h, v) - f(v) = a(u_h - u, v) = 0$ (for all $v \in V_h$) is equivalent to the V-orthogonality $u - u_h \perp_V V_h$. Hence $u_h = Q_V u$ holds, i.e., Q_V is the Ritz projection S_h of the Helmholtz problem: $Q_V = S_h$ in $H_0^1(\Omega)$. Corollary 8.66 shows $\|I - Q_V\|_{L^2(\Omega) \leftarrow V} \le C_1 h$, i.e., $|u - Q_V u|_0 \le C_1 h |u|_1$ for all $u \in V = H_0^1(\Omega)$.

Solution of Exercise 9.35. (a) The usual Euclidean norm $\|\cdot\|$ and the newly defined $|\cdot|_0$ differ by the scaling factor $h^{n/2}$. This factor cancels in $|L_h|_{0 \leftarrow 0} = \sup \frac{|L_h v_h|_0}{|v_h|_0}$, i.e., $\sup \frac{|L_h v_h|_0}{|v_h|_0} = \sup \frac{\|L_h v_h\|}{\|v_h\|} = \|L_h\|_2$.

Part (b) is an immediate consequence of $|\partial_j^+ u(\mathbf{x})|^2 \le 2 \dfrac{|u(\mathbf{x})|^2 + |u(\mathbf{x} + h\mathbf{e}^j|^2}{h^2}$, where $\mathbf{e}^j$ is the j-th unit vector.

Solution of Exercise 9.36. (a) $|\cdot|_0 \le |\cdot|_1$ implies $\left|L_h^{-1}\right|_{1\leftarrow -1} = \inf_{u_h} \sup_{v_h} \frac{\left|\left\langle L_h^{-1}u_h, v_h\right\rangle\right|}{|v_h|_1 |u_h|_1}$ $\le \inf_{u_h} \sup_{v_h} \left|\left\langle L_h^{-1}u_h, v_h\right\rangle\right| / [\,|v_h|_0\,|u_h|_0\,] = \left|L_h^{-1}\right|_{0\leftarrow 0} = \|L_h^{-1}\|_2$.

Parts (b) and (c) are a consequence of (6.33) since L_h^{T} is dual to L_h.

(d) The spectral norm satisfies $\|A\|_2 = \sqrt{\|A^{\mathsf{T}}A\|_2} = \rho(A^{\mathsf{T}}A)$ (ρ: spectral radius, cf. (4.27)). On the other hand the spectral radius can be bounded by any associated matrix norm: $\rho(A^{\mathsf{T}}A) \le \|A^{\mathsf{T}}A\|_\infty \le \|A^{\mathsf{T}}\|_\infty \|A\|_\infty$.

(e) H_h^1-ellipticity, i.e., $a_h(u_h, u_h) \ge C_E\,|u_h|_1^2$, implies that $\left|L_h^{-1}\right|_{1\leftarrow -1} \le \frac{1}{C_E}$ according to Lemma 6.97.

Solution of Exercise 9.38. $\delta L_h = a_0$ can be estimated by $|\,\langle \delta L_h v_h, v_h\rangle\,| \le \mathrm{const}\,|v_h|_0^2$. In the case of $\delta L_h = \sum_i b_i\,\partial_i^{\pm}$ we have

$$|\,\langle \delta L_h v_h, v_h\rangle\,| \le \mathrm{const}|\partial_i^{\pm} v_h|_0 |v_h|_0 \le \mathrm{const}|v_h|_1 |v_h|_0 \le \varepsilon |v_h|_1^2 + C_\varepsilon |v_h|_0^2$$

(cf. (5.34) with $a = \sqrt{\varepsilon}|v_h|_1$). For $\delta L_h = \partial_i^{\pm} c_i$ use

$$|\,\langle \delta L_h v_h, v_h\rangle = |\,\left\langle c_i v_h, \partial_i^{\mp} v_h\right\rangle \le \mathrm{const}|v_h|_0 |v_h|_1 \le \varepsilon |v_h|_1^2 + C_\varepsilon |v_h|_0^2\,.$$

Altogether we obtain $|\,\langle \delta L_h v_h, v_h\rangle\,| \le 2n\varepsilon |v_h|_1^2 + C'_\varepsilon |v_h|_0^2$. Choose ε with $2n\varepsilon \le C_E/2$, where C_E is the constant in $\langle L_h v_h, v_h\rangle \ge C_E\,|u_h|_1^2 - C_K\,|u_h|_0^2$. Since $\langle (L_h + \delta L_h)v_h, v_h\rangle \ge \frac{C_E}{2}\,|u_h|_1^2 - C'_K\,|u_h|_0^2$, also $L_h + \delta L_h$ is H_h^1-*coercive.*

Solution of Exercise 9.40. For (5.19) the argument is rather simple. The second x-difference $-\partial_x^- a_{11}\partial_x^+$ as a part of L_h leads to $-\langle a_{11}\partial_x^+ v_h, \partial_x^+ v_h\rangle \ge \varepsilon\,|\partial_x^+ v_h|_0^2$ after summation by parts. With the analogous inequality

$$-\left\langle a_{22}\partial_y^+ v_h, \partial_y^+ v_h\right\rangle \ge \varepsilon |\partial_y^+ v_h|_0^2$$

for the y-direction we obtain $\langle L_h v_h, v_h\rangle \ge \varepsilon |\partial_y^+ v_h|_{1,0}^2 \ge \tilde{\varepsilon} |\partial_y^+ v_h|_1^2$ according to Lemma 9.41. In the case of (5.20) summation by part for all terms leads to the sum of $-a_{11}|\partial_x^+ v_h|^2 - a_{22}|\partial_y^+ v_h|^2 - 2a_{12}|\partial_x^+ v_h||\partial_y^+ v_h|$ over all grid points. If the arising coefficients a_{ij} were evaluated at a common argument, the statement would follow from $-\sum a_{ij}\xi_i\xi_j \ge \varepsilon\,|\xi|^2$. In fact, the coefficients are evaluated at arguments differing by h. Now the result follows from the continuity of a_{ij} (the continuity is uniform since $\overline{\Omega}$ is compact).

Solution of Exercise 9.46. (a) $\sigma_h^x \sigma_h^y$ and $\sigma_h^y \sigma_h^x$ produce the same result (9.39a). $(\sigma_h^x u)(x,y)$ can also be written as $\frac{1}{h}\int_{x-h/2}^{x+h/2} u(\xi, y)\mathrm{d}\xi$. The derivative $\frac{\partial}{\partial x}$ with respect to the integral bounds yields the difference quotient $(\hat{\partial}_x u)(x,y)$. On the other hand, we also have

$$(\sigma_h^x \frac{\partial}{\partial x} u)(x,y) = \frac{1}{h}\int_{x-\frac{h}{2}}^{x+\frac{h}{2}} u_x(\xi, y)\mathrm{d}\xi = \frac{1}{h} u(\cdot, y)\Big|_{x-\frac{h}{2}}^{x+\frac{h}{2}} = (\hat{\partial}_x u)(x,y).$$

(b) In the proofs for (b) and (c) we can ignore the y-dependence. Substitution and interchanging integrals give

$$\begin{aligned}(\sigma_h^x u, v)_{L^2(\mathbb{R})} &= \int_{\mathbb{R}} v(x) \frac{1}{h} \int_{-h/2}^{h/2} u(x+\xi) \mathrm{d}\xi \mathrm{d}x = \frac{1}{h} \int_{-h/2}^{h/2} \int_{\mathbb{R}} u(x+\xi) v(x) \mathrm{d}x \mathrm{d}\xi \\ &\underset{x=t-\xi}{=} \frac{1}{h} \int_{-h/2}^{h/2} \int_{\mathbb{R}} u(t) v(t-\xi) \mathrm{d}t \mathrm{d}\xi = \int_{\mathbb{R}} u(t) \frac{1}{h} \int_{-h/2}^{h/2} v(t-\xi) \mathrm{d}\xi \\ &\underset{\xi=-\zeta}{=} (u, \sigma_h^x v)_{L^2(\mathbb{R})} \,.\end{aligned}$$

(c) Assume first $k = 0$. For a fixed x the Schwarz' inequality leads to

$$\begin{aligned}|\sigma_h^x u(x)|^2 &= \frac{1}{h^2} \left| \int_{x-h/2}^{x+h/2} 1 \cdot u(\xi) \mathrm{d}\xi \right|^2 \le \frac{1}{h^2} \int_{x-h/2}^{x+h/2} \mathrm{d}\xi \cdot \int_{x-h/2}^{x+h/2} |u(\xi)|^2 \mathrm{d}\xi \\ &= \frac{1}{h} \int_{x-h/2}^{x+h/2} |u(\xi)|^2 \mathrm{d}\xi.\end{aligned}$$

Integration over x, substitution and interchanging the integrals yield

$$\begin{aligned}\int_{\mathbb{R}} |\sigma_h^x u(x)|^2 \mathrm{d}x &= \frac{1}{h} \int_{\mathbb{R}} \int_{x-h/2}^{x+h/2} |u(\xi)|^2 \mathrm{d}\xi \mathrm{d}x \underset{\xi=x+\zeta}{=} \frac{1}{h} \int_{\mathbb{R}} \int_{-h/2}^{h/2} |u(x+\zeta)|^2 \mathrm{d}\zeta \mathrm{d}x \\ &= \frac{1}{h} \int_{-h/2}^{h/2} \int_{\mathbb{R}} |u(x+\zeta)|^2 \mathrm{d}x \mathrm{d}\zeta \underset{x=t-\zeta}{=} \frac{1}{h} \int_{-h/2}^{h/2} \int_{\mathbb{R}} |u(t)|^2 \mathrm{d}t \mathrm{d}\zeta \\ &= \frac{1}{h} |u|_0^2 \int_{-h/2}^{h/2} \mathrm{d}\zeta = |u|_0^2 \,.\end{aligned}$$

This proves $\|\sigma_h^x\|_{L^2(\mathbb{R}^2) \to L^2(\mathbb{R}^2)} \le 1$. The commutativity with derivatives also shows $\|\sigma_h^x\|_{H^k(\mathbb{R}^2) \to H^k(\mathbb{R}^2)} \le 1$ for all $k \in \mathbb{N}$. Because $(\sigma_h^x)^* = \sigma_h^x$ the same estimate holds for the dual spaces with negative k.

(d) The first inequality in part (c) can be summed over all $x = \nu h$ and yields $\sum_{\nu \in \mathbb{Z}} h |(\sigma_h^x u)(\nu h)|^2 = \int_{\mathbb{R}} |u(x)|^2 \mathrm{d}x$. The analogous treatment of the y-direction gives the result.

(e) Also here it is sufficient to limit ourselves to σ_h^x and functions $a(\cdot)$ in x. We obtain

$$\begin{aligned}\|a \sigma_h^x u - \sigma_h^x (au)\|_{L_h^2} &= \sum_{\nu \in \mathbb{Z}} h \left[\, |(a \sigma_h^x u)(\nu h) - (\sigma_h^x (au))(\nu h)|^2 \, \right] \\ &= \sum_{\nu \in \mathbb{Z}} \left[a(\nu h) \int_{\nu h - h/2}^{\nu h + h/2} |u(\xi)|^2 \mathrm{d}\xi - \int_{\nu h - h/2}^{\nu h + h/2} |\, (au)\,(\xi)|^2 \mathrm{d}\xi \right]\end{aligned}$$

$$= \sum_{\nu\in\mathbb{Z}} \int_{\nu h-h/2}^{\nu h+h/2} |\,[a(\nu h)-a(\xi)]\,u(\xi)|^2 \mathrm{d}\xi.$$

The Lipschitz-continuity ensures $|a(\nu h)-a(\xi)| \le Lh/2$ with $L = \|a\|_{C^{0,1}(\mathbb{R})}$ so that $\|a\sigma_h^x u - \sigma_h^x\,(au)\|_{L_h^2} \le \frac{Lh}{2}\,\|u\|_{L^2(\mathbb{R})}$ follows.

(f) Abbreviate σ_h^x by σ. By $a\sigma^\nu - \sigma^\nu a = \sum_{\mu=0}^{\nu-1} \sigma^{\nu-\mu-1}\,(a\sigma-\sigma a)\,\sigma^\mu$ and the parts (c) and (e) we obtain the assertion.

(g) Integration by parts proves

$$(u-\sigma_h^x u)\,(x) = \frac{1}{h}\int_{-h/2}^{h/2} [u(x)-u(x+\xi)]\,\mathrm{d}\xi = \frac{1}{h}\int_{-h/2}^{h/2}\left[\xi - \mathrm{sign}(\xi)\frac{h}{2}\right] u_x(x+\xi)\mathrm{d}\xi.$$

Schwarz' inequality shows

$$|(u-\sigma_h^x u)\,(x)|^2 \le \frac{h}{12}\int_{-h/2}^{h/2} |u_x(x+\xi)|^2\,\mathrm{d}\xi,$$

since $\int_{-h/2}^{h/2} [\xi - \mathrm{sign}(\xi)h/2]^2\,\mathrm{d}\xi = h^3/12$. Integration over $x\in\mathbb{R}$ leads to

$$\begin{aligned}\int_{\mathbb{R}} |(u-\sigma_h^x u)\,(x)|^2\,\mathrm{d}x &\le \frac{h}{12}\int_{\mathbb{R}}\int_{-h/2}^{h/2} |u_x(x+\xi)|^2\,\mathrm{d}\xi\mathrm{d}x \\ &= \frac{h}{12}\int_{-h/2}^{h/2}\int_{\mathbb{R}} |u_x(x+\xi)|^2\,\mathrm{d}x\mathrm{d}\xi = \frac{h}{12}\int_{-h/2}^{h/2}\mathrm{d}\xi\,\|u_x\|_{L^2(\mathbb{R})}^2 = \frac{h^2}{12}\,\|u_x\|_{L^2(\mathbb{R})}^2,\end{aligned}$$

hence

$$\|u-\sigma_h^x u\|_{L^2(\mathbb{R})} \le C_1 h\,\|u_x\|_{L^2(\mathbb{R})} \le C_1 h\,\|u\|_{H^1(\mathbb{R})} \quad \text{with } C_1 = 1/\sqrt{12}.$$

From $\frac{\partial}{\partial x}\sigma_h^x = \sigma_h^x\frac{\partial}{\partial x}$ in part (a) we infer $\|u-\sigma_h^x u\|_{H^k(\mathbb{R})} \le C_1 h\,\|u\|_{H^{k+1}(\mathbb{R})}$. The same argument for y yields the statement (g) for $\nu=\mu=1$. For more general ν,μ proceed as in part (f).

Solution of Exercise 9.57. In $(L_h+\delta L_h)^{-1} = L_h^{-1} - L_h^{-1}\delta L_h\,(L_h+\delta L_h)^{-1}$ the last term can be bounded by

$$|L_h^{-1}\delta L_h\,(L_h+\delta L_h)^{-1}\,|_{2\leftarrow 0} \le |L_h^{-1}|_{2\leftarrow 0}|\delta L_h|_{0\leftarrow 1}|\,(L_h+\delta L_h)^{-1}\,|_{1\leftarrow 0}\,.$$

Solution of Exercise 9.59. (a) $(P_h u_h, v)_{L^2(\mathbb{R}^2)} = \sum_{\mathbf{x}\in Q_h}\int_{Q_\mathbf{x}} (P_h u_h(\mathbf{x}))\,v(\mathbf{x})\mathrm{d}\mathbf{x}$ holds with $Q_{(x,y)} := [x-h/2, x+h/2]\times[y-h/2, y+h/2]$. By definition, $\int_{Q_\mathbf{x}}\ldots = 0$ if $\mathbf{x}\notin\Omega$ and otherwise $u_h(\mathbf{x})\int_{Q_\mathbf{x}} v(\mathbf{x})\mathrm{d}\mathbf{x} = h^2 u_h(\mathbf{x})$. This shows $(P_h u_h, v)_{L^2(\mathbb{R}^2)} = h^2\sum_{\mathbf{x}\in\Omega_h} u_h(\mathbf{x})\,(\sigma_h^x\sigma_h^y v)\,(\mathbf{x}) = \langle u_h, \sigma_h^x\sigma_h^y v\rangle_{L_h^2}$ and proves $\hat{P}_h^\star u = (\sigma_h^x\sigma_h^y u)|_{\Omega_h}$.

(b) Let $e_{\mathbf{x}} \in L_h^2$ be the unit vector at $\mathbf{x} = (x, y)$ and $\delta_{\mathbf{x}}$ the Delta distribution at $\mathbf{x}$. One verifies that $\hat{P}_h e_{\mathbf{x}} = \sigma_h^x \sigma_h^y \delta_{\mathbf{x}}$ and thus

$$\left(u, E_0^* \sigma_h^x \sigma_h^y \hat{P}_h e_{\mathbf{x}}\right)_{L^2(\mathbb{R}^2)} = \left(u, E_0^* \left(\sigma_h^x \sigma_h^y\right)^2 \delta_{\mathbf{x}}\right)_{L^2(\mathbb{R}^2)} = \left(\left(\sigma_h^x \sigma_h^y\right)^2 E_0 u, \delta_{\mathbf{x}}\right)_{L^2(\mathbb{R}^2)}$$
$$= \left(\left(\sigma_h^x \sigma_h^y\right)^2 E_0 u\right)(\mathbf{x}) = (\tilde{R}_h u)(\mathbf{x}) = \left\langle \tilde{R}_h u, e_{\mathbf{x}} \right\rangle_{L_h^2}.$$

This proves $\tilde{R}_h^\star = E_0^* \sigma_h^x \sigma_h^y \hat{P}_h$.

(c) By part (a) $\hat{P}_h^\star \hat{P}_h u_h = \sigma_h^x \sigma_h^y \hat{P}_h u_h$ holds on Q_h. $\hat{P}_h u_h$ is piecewise constant. The evaluation in $\mathbf{x} \in Q_h$ produces the mean value of a constant, i.e., the value is not changed: $(\sigma_h^x \sigma_h^y \hat{P}_h u_h)(\mathbf{x}) = u_h(\mathbf{x})$.

Solution of Exercise 9.61. The estimate $|u_h|_0 \le C \|u_h\|_\infty$ holds with $C \approx \mathrm{vol}(\Omega)$. The inequality $\|u_h\|_\infty \le C |u_h|_2$ can be proved, e.g., by means of the discrete Green function $g_h(\mathbf{x}, \boldsymbol{\xi})$ with the property $\Delta_h g_h(\cdot, \boldsymbol{\xi}) = e_{\boldsymbol{\xi}}$ ($e_{\boldsymbol{\xi}}$: unit vector at $\boldsymbol{\xi}$). As in the continuous case we have $|g_h(\cdot, \boldsymbol{\xi})|_0 \le C$. From this we conclude

$$u_h(\boldsymbol{\xi}) = \langle u_h, e_{\boldsymbol{\xi}} \rangle_{L_h} = \langle u_h, \Delta_h g_h(\cdot, \boldsymbol{\xi}) \rangle_{L_h} = \langle \Delta_h u_h, g_h(\cdot, \boldsymbol{\xi}) \rangle_{L_h}$$

and $|u_h(\boldsymbol{\xi})| \le |\Delta_h u_h|_0 \, |g_h(\cdot, \boldsymbol{\xi})|_0 \le C_g \, |u_h|_2$. Maximising over all $\boldsymbol{\xi} \in \Omega_h$, we obtain $\|u_h\|_\infty \le C |u_h|_2$.

Hence $\|L_h^{-1}\|_\infty = \sup_{u_h} \frac{\|L_h^{-1} u_h\|_\infty}{\|u_h\|_\infty} \le C^2 \sup_{u_h} \frac{|L_h^{-1} u_h|_2}{|u_h|_0} = C^2 |L_h^{-1}|_{2 \leftarrow 0} \le \mathrm{const}$.

Exercises of Chapter 10

Solution of Exercise 10.6. Inserting v, we prove the assertion.

Solution of Exercise 10.9. $\delta(x) := (\mathrm{e}^{x/\varepsilon} - 1)/(\mathrm{e}^{1/\varepsilon} - 1)$ is the deviation from the reduced solution x. The function δ increases monotonously from $\delta(0) = 0$ to $\delta(1) = 1$. Hence ξ is the solution of the equation $\delta(1 - \xi) = \eta$. The ansatz $\xi = C\varepsilon \, |\log \eta|$ yields $\delta(1 - \xi) = (\eta^C - \mathrm{e}^{-1/\varepsilon})/(1 - \mathrm{e}^{-1/\varepsilon}) = \eta$. Since $\mathrm{e}^{-1/\varepsilon} \ll \varepsilon$ this equation has a solution $C \approx 1$.

Solution of Exercise 10.12. One verifies that $u_h(0) = u_h(1) = 0$ is satisfied and that u_h fulfils the homogeneous difference equation because of $-\varepsilon(1 + h/\varepsilon)^2 + (2\varepsilon + h)\,(1 + h/\varepsilon) - (\varepsilon + h) = 0$.

Solution of Exercise 10.14. The discretisations of the principal part $-\varepsilon\Delta$ are described in Exercises 8.42 and 8.46.

(a) The x-derivative of the basis functions are the piecewise constants $\pm h^{-1}$ or 0. The integral of a test basis function over one triangle has the value $h^2/6$, $h^2/3$, or 0. The combination yields (10.15).

Part (b) follows by an elementary calculation.

The one-dimensional scheme from part (c) can be derived from (b) if one applies L_h from (b) to grid functions which are constant in y-direction.

Exercises of Chapter 11

Solution of Exercise 11.3. By symmetry (11.2a) is identical to $\overline{a(v,e)} = \lambda\overline{(v,e)_0}$ and $a(v,e) = \bar\lambda(v,e)_0$. Hence $e^* = e$ is also an eigenfunction of the adjoint problem to eigenvalue $\bar\lambda$. The test with $v = e$ yields $\lambda(e,e)_0 = a(e,e) = \bar\lambda(e,e)_0$. $(e,e)_0 > 0$ proves $\lambda = \bar\lambda$, i.e., the eigenvalue is real. Since $e \in E(\lambda)$ is arbitrary, we have proved $E(\lambda) \subset E^\star(\lambda)$. Similarly, one shows $E^\star(\lambda) \subset E(\lambda)$.

Solution of Exercise 11.7. (a) The definition (8.91) states that $\mathbf{M} := P^*P$, i.e., $\langle \mathbf{Mx}, \mathbf{x}\rangle = \langle \mathbf{P}^*\mathbf{Px}, \mathbf{x}\rangle = (\mathbf{Px}, \mathbf{Px})_0$. $\mathbf{x} \neq 0$ implies $\mathbf{Px} \neq 0$ and thus $\langle \mathbf{Mx}, \mathbf{x}\rangle > 0$. Hence the symmetric matrix $\mathbf{M}$ is positive definite. The square root $\mathbf{A} := \mathbf{M}^{1/2}$ exists as a positive-definite matrix and satisfies $\mathbf{A}^\mathsf{H}\mathbf{A} = \mathbf{A}^2 = \mathbf{M}$ (cf. Exercise 4.30e). On the other hand the Cholesky decomposition $\mathbf{M} = \mathbf{A}^\mathsf{H}\mathbf{A}$ exists with a lower triangular matrix $\mathbf{A}$ (cf. Quarteroni–Sacco–Saleri [230, §3.4.2]).

(b) Let $\mathbf{M} = \mathbf{A}^\mathsf{H}\mathbf{A}$. Set $\tilde{\mathbf{e}} = \mathbf{Ae}$ and $\tilde{\mathbf{L}} := (\mathbf{A}^\mathsf{H})^{-1}\mathbf{L}\mathbf{A}^{-1}$. Then we have $\tilde{\mathbf{L}}\tilde{\mathbf{e}} = (\mathbf{A}^\mathsf{H})^{-1}\mathbf{Le} = \lambda_h(\mathbf{A}^\mathsf{H})^{-1}\mathbf{Me} = \lambda_h\mathbf{Ae} = \lambda_h\tilde{\mathbf{e}}$. Analogously, $\tilde{\mathbf{e}}^\star = \mathbf{Ae}^\star$ leads to the eigenvalue equation

$$\tilde{\mathbf{L}}^\mathsf{H}\tilde{\mathbf{e}}^\star = (\mathbf{A}^\mathsf{H})^{-1}\mathbf{L}^\mathsf{H}\mathbf{A}^{-1}\mathbf{Ae}^\star = \overline{\lambda_h}(\mathbf{A}^\mathsf{H})^{-1}\mathbf{Me}^\star = \overline{\lambda_h}\mathbf{Ae}^\star = \overline{\lambda_h}\tilde{\mathbf{e}}^\star.$$

Solution of Exercise 11.8. First we consider the bilinear ansatz. According to Exercise 8.46 the finite-element matrix $\mathbf{L}$ is characterised by the stencil $\frac{1}{3}\begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$, while the mass matrix $\mathbf{M}$ is $\frac{h^2}{36}\begin{bmatrix} 1 & 4 & 1 \\ 4 & 16 & 4 \\ 1 & 4 & 1 \end{bmatrix}$ with $h = \frac{a}{n}$. The ansatz $\mathbf{e}^{\nu,\mu}$ as restriction of $\sin(\nu\pi x/a)\sin(\mu\pi y/a)$ to the nodal values turns out to be an eigenfunction of $\mathbf{L}$ and $\mathbf{M}$. The product $3\mathbf{Le}^{\nu,\mu}$ at (x,y) consists of the four second differences:

$$\begin{aligned} -\sin\tfrac{\mu\pi y}{a}\partial_x^+\partial_x^-\sin\tfrac{\nu\pi x}{a} &= \sin\tfrac{\mu\pi y}{a}\left[2\sin\tfrac{\nu\pi x}{a} - \sin\tfrac{\nu\pi(x-h)}{a} - \sin\tfrac{\nu\pi(x+h)}{a}\right] \\ &= 2\left[1-\cos\tfrac{\nu\pi h}{a}\right]\sin\tfrac{\nu\pi x}{a}\sin\tfrac{\mu\pi y}{a} = 4\sin^2\tfrac{\nu\pi h}{2a}\mathbf{e}^{\nu,\mu}(x,y), \end{aligned}$$

the corresponding result $-\sin\frac{\nu\pi x}{a}\partial_y^+\partial_y^-\sin\frac{\mu\pi y}{a} = 4\sin^2\frac{\mu\pi h}{2a}\mathbf{e}^{\nu,\mu}(x,y)$, and the two diagonal ones

$$\begin{aligned} &4\sin\tfrac{\nu\pi x}{a}\sin\tfrac{\mu\pi y}{a} - \sin\tfrac{\nu\pi(x-h)}{a}\sin\tfrac{\mu\pi(y-h)}{a} - \sin\tfrac{\nu\pi(x+h)}{a}\sin\tfrac{\mu\pi(y+h)}{a} \\ &\quad - \sin\tfrac{\nu\pi(x-h)}{a}\sin\tfrac{\mu\pi(y+h)}{a} - \sin\tfrac{\nu\pi(x+h)}{a}\sin\tfrac{\mu\pi(y-h)}{a} \\ &\quad = 4\left(1-\cos\frac{\nu\pi h}{a}\cos\frac{\mu\pi h}{a}\right)\mathbf{e}^{\nu,\mu}(x,y). \end{aligned}$$

Together with $h = a/n$ this demonstrates

$$\mathbf{L}\,\mathbf{e}^{\nu,\mu} = \frac{4}{3}\left(1+\sin^2\frac{\nu\pi}{2n} + \sin^2\frac{\mu\pi}{2n} - \cos\frac{\nu\pi}{n}\cos\frac{\mu\pi}{n}\right)\mathbf{e}^{\nu,\mu}.$$

Similarly one proves

$$\mathbf{Me}^{\nu,\mu} = \frac{h^2}{9}\left(\cos\frac{\nu\pi h}{a} + 2\right)\left(\cos\frac{\mu\pi h}{a} + 2\right)\mathbf{e}^{\nu,\mu}.$$

Hence we have $\mathbf{Le}^{\nu,\mu} = \lambda_h^{\nu,\mu}\mathbf{Me}^{\nu,\mu}$ with

$$\lambda_h^{\nu,\mu} = 12h^{-2}\frac{1 + \sin^2\frac{\nu\pi}{2n} + \sin^2\frac{\mu\pi}{2n} - \cos\frac{\nu\pi}{n}\cos\frac{\mu\pi}{n}}{\left(\cos\frac{\nu\pi h}{a} + 2\right)\left(\cos\frac{\mu\pi h}{a} + 2\right)} \qquad \left(h = \frac{a}{n}\right).$$

Clearly $\lambda_h^{\nu,\mu} = \lambda_h^{\mu,\nu}$ holds, i.e., for $\nu \neq \mu$ there are (at least) double eigenvalues. The asymptotic expansion is

$$\lambda_h^{\nu,\mu} = \frac{\pi^2}{a^2}\left(\nu^2 + \mu^2\right) + \frac{h^2}{12}\frac{\pi^4}{a^4}\left(\nu^4 + \mu^4\right) + \mathcal{O}\big(h^4\big).$$

For $(\nu,\mu) = (1,7)$ the h^2-term is $\frac{1201}{6}\frac{\pi^4}{a^4}h^2$, and $\frac{625}{6}\frac{\pi^4}{a^4}h^2$ for $\nu = \mu = 5$. Hence, $\lambda_h^{5,5} \neq \lambda_h^{1,7} = \lambda_h^{7,1}$.

Now we consider the linear triangular element. According to Exercise 8.42 $\mathbf{L}$ is identical with the five-point discretisation of step size $h = \frac{a}{n}$. As above we obtain

$$\mathbf{L}\,\mathbf{e}^{\nu,\mu} = 4\left(\sin^2\frac{\nu\pi}{2n} + \sin^2\frac{\mu\pi}{2n}\right)\mathbf{e}^{\nu,\mu}.$$

The lumped mass matrix is $\hat{\mathbf{M}} = h^2 I$ so that $\hat{\mathbf{M}}\mathbf{e}^{\nu,\mu} = h^2\mathbf{e}^{\nu,\mu}$. Hence $\mathbf{Le}^{\nu,\mu} = \lambda_h^{\nu,\mu}\hat{\mathbf{M}}\mathbf{e}^{\nu,\mu}$ holds with

$$\lambda_h^{\nu,\mu} = 4h^{-2}\left(\sin^2\frac{\nu\pi}{2n} + \sin^2\frac{\mu\pi}{2n}\right) \qquad \left(h = \frac{a}{n}\right).$$

Again, $\lambda_h^{\nu,\mu} = \lambda_h^{\mu,\nu}$ holds, and the asymptotic expansion is now

$$\lambda_h^{\nu,\mu} = \frac{\pi^2}{a^2}\left(\nu^2 + \mu^2\right) - \frac{h^2}{12}\frac{\pi^4}{a^4}\left(\nu^4 + \mu^4\right) + \mathcal{O}\big(h^4\big)$$

leading to the same conclusions.

Solution of Exercise 11.9. (a) (11.6) follows from Lemma 6.94 or, respectively, Lemma 8.13.

(b) The resolvent $R(\lambda) := (L - \lambda I)^{-1}$ is continuous outside its singularities as can be seen from $R(\lambda) - R(\mu) = (\lambda - \mu)\,R(\lambda)R(\mu)$. Since the norm is continuous (cf. Exercise 6.3), also $\omega(\lambda)$ continuous except possibly the singularities of R. Let λ be an eigenvalue: $Le = \lambda e$. In the neighbourhood of λ we have $R(\mu)e = \frac{1}{\lambda-\mu}e$, so that $\|R(\mu)\|_{V \leftarrow V'} \to \infty$ for $\mu \to \lambda$. The inverse yields continuity even at λ: $\omega(\mu) \to \omega(\lambda) = 0$.

(c) Combine Lemma 6.94 and Lemma 6.109.

Solution of Exercise 11.17. (a) Let $\dim E(\lambda_0) = 1$. Theorem 11.15 guarantees $\dim E_h(\lambda_h) \geq 1$ for sufficiently small h and the existence of e^h with $e^h \to e$. For an indirect proof assume that $\dim E_h(\lambda_h) \geq 2$ for infinitely many $h \in H$. Hence there is another linearly independent eigenvector $\hat{e}^h$. At least for sufficiently small h the eigenvector $\hat{e}^h$ can be scaled to that $\hat{e}^h \perp e$. According to Theorem 11.16 there is a subsequence so that $\lim \hat{e}^h = \hat{e}$ is an eigenfunction of L. Since $\hat{e} \perp e$, it follows that $\dim E(\lambda_0) \geq 2$ in contradiction to the assumption.

(b) By assumption and part (a) $\dim E(\lambda_0) = \dim E(\lambda_h) = 1$ holds for sufficiently small h. Let e and e^h be normed: $\|e^h\|_V = \|e\|_V = 1$. According to $V = \operatorname{span}\{e\} \oplus e^\perp$ we split e^h into $e^h_\parallel + e^h_\perp$. If $\limsup \|e^h_\perp\|_V > 0$, one proves $\dim E(\lambda_0) \geq 2$ as in the proof of part (a). Hence $\lim e^h_\perp = 0$. $e^h_\parallel = \gamma_h e$ holds with the factor $\gamma_h := (e^h, e)_V$. $1 = \|e^h\|^2_V = \|e^h_\parallel\|^2_V + \|e^h_\perp\|^2_V = |\gamma_h|^2 + \|e^h_\perp\|^2_V$ implies $|\gamma_h| \to 1$. For sufficiently small h we have $|\gamma_h| \geq 1/2$, and $\hat{e}^h := \frac{1}{(e^h,e)_V} e^h = \frac{1}{\gamma_h}(e^h_\parallel + e^h_\perp) = e + \frac{1}{\gamma_h} e^h_\perp$ converges to e.

Solution of Exercise 11.20. The statement follows from Conclusion 8.8.

Solution of Exercise 11.22. Let $\hat{u}^h = P_{V_h} u \in V_h$ be the approximation with $\|\hat{u}^h - u\|_V = d(u, V_h)$ and set $\hat{u}^{h\perp} := u - \hat{u}^h$ (cf. Remark 8.23). Since $(\cdot, v)_0$ is a functional in V', the Riesz isomorphism ensures the existence of $w \in V$ with $(\cdot, v)_0 = (\cdot, w)_V$ (cf. Theorem 6.68). We split w into $w^h + w^{h\perp}$ with $w^h \in V_h$ and $w^{h\perp} \in V_h^\perp$ (orthogonality with respect to $(\cdot,\cdot)_V$). The ansatz for u^h reads

$$u^h = \hat{u}^h + \eta w^h \in V_h.$$

Because

$$\begin{aligned}\left(u^h - u, v\right)_0 &= \left(u^h - u, w\right)_V = \left(\hat{u}^h + \eta w^h - \left(\hat{u}^h + \hat{u}^{h\perp}\right), w^h + w^{h\perp}\right)_V \\ &= \eta \left(w^h, w^h\right)_V - \left(\hat{u}^{h\perp}, w^{h\perp}\right)_V,\end{aligned}$$

the choice $\eta := \frac{\left(\hat{u}^{h\perp}, w^{h\perp}\right)_V}{\left(w^h, w^h\right)_V}$ ensures the side condition $\left(u^h - u, v\right)_0 = 0$. (11.4c) implies $w^{h\perp} \to 0$ in V and $\left(w^h, w^h\right)_V = \|w^h\|^2_V \to \|w\|^2_V > 0$, so that η is well defined for sufficiently small h. Using $\|\hat{u}^{h\perp}\|_V = \|\hat{u}^h - u\|_V = d(u, V_h)$, we can estimate the additional term ηw^h by

$$\frac{\left|\left(\hat{u}^{h\perp}, w^{h\perp}\right)_V\right|}{\left(w^h, w^h\right)_V} \|w^h\|_V \leq \|\hat{u}^{h\perp}\|_V \underbrace{\|w^{h\perp}\|_V / \|w^h\|_V}_{=:\varepsilon_h} = \varepsilon_h \, d(u, V_h)$$

with $\varepsilon_h \to 0$. For sufficiently small h we obtain $\varepsilon_h \leq 1$ and therefore $\|u - u^h\|_V \leq 2d(u, V_h)$.

Solution of Exercise 11.29. (a) Let $u^h \in V_h$ be the solution of (11.21a–c). u^h belongs to $\hat{V}_h$ because of (11.21c). Since (11.21a) is satisfied for all $v \in V_h$, it also holds for all $v \in \hat{V}_h \subset V_h$. Hence $u^h \in \hat{V}_h$ is the solution of $a_\lambda(u^h, v) = (f^{(h)}, v)_0$ for all $v \in \hat{V}_h$.

(b) First we assume (11.21b). Let $u^h \in \hat{V}_h$ satisfy $a_\lambda(u^h, v) = (f^{(h)}, v)_0$ for all $v \in \hat{V}_h$. The equation $a_\lambda(u^h, v) = (f^{(h)}, v)_0$ for $v \in E_h^\star(\lambda_h)$ follows from the definition of $\hat{V}_h$ and (11.21b). From $V_h = \hat{V}_h \oplus E_h^\star(\lambda_h)$ we infer (11.21a). (11.21c) is a consequence of $u^h \in \hat{V}_h$.

(c) If (11.21b) does not hold, (11.21a,b) is not solvable; nevertheless the variational problem in $\hat{V}_h$ has a solution. For this purpose split $f^{(h)}$ into $f_\perp^{(h)} + g^{(h)}$ with $f_\perp^{(h)} \in E_h^\star(\lambda_h)^\perp$ and $g^{(h)} \in E_h^\star(\lambda_h)$. The variation in $\hat{V}_h$ ignores the part $g^{(h)}$ and solves the problem with the right-hand side $f_\perp^{(h)}$ satisfying (11.21b).

Solution of Exercise 11.31. Choose $\hat{u}^h \in V_h$ with $\|u - \hat{u}^h\|_V = d(u, V_h)$.

$$\eta := \left(\hat{u}^h, e^{\star h}\right)_0 = \left(\hat{u}^h - u, e^{\star h}\right)_0 + \left(u, e^{\star h} - e^\star\right)_0$$

holds for all $e^* \in E^*(\lambda_0)$ because of $u \perp e^\star$. We estimate η by

$$|\eta| \le \|u - \hat{u}^h\|_V \|e^{*h}\|_{V'} + |u|_0 |e^{*h} - e^*|_0 \,.$$

Since $\|e^{*h}\|_{V'} = 1$ and $e^* \in E^*(\lambda_0)$ is arbitrary, we obtain

$$|\eta| \le \|u - \hat{u}^h\|_V + |u|_0 \inf_{e^* \in E^*(\lambda_0)} |e^{*h} - e^*|_0 \,.$$

$u^h := \hat{u}^h - \eta e^h$ satisfies $\left(u^h, e^{\star h}\right)_0 = \eta - \eta \left(e^h, e^{\star h}\right)_0 = 0$ and the estimate

$$\begin{aligned} d(u, V_h \cap E_h^\star(\lambda_h)^\perp) \le \left\|u - u^h\right\|_V &\le \|u - \hat{u}^h\|_V + |\eta| \, \|e^h\|_V \le \|u - \hat{u}^h\|_V \\ &+ \|e^h\|_V \left[\|u - \hat{u}^h\|_V \|e^{*h}\|_{V'} + |u|_0 \inf_{e^* \in E^*(\lambda_0)} |e^{*h} - e^*|_0 \right] \\ &= \left(1 + \|e^h\|_V\right) \|u - \hat{u}^h\|_V + \|e^h\|_V |u|_0 \inf_{e^* \in E^*(\lambda_0)} |e^{*h} - e^*|_0. \end{aligned}$$

$\frac{1}{\|e^h\|_V} = \frac{(e^h, e^{\star h})_0}{\|e^h\|_V \|e^{\star h}\|_{V'}} \to \frac{(e, e^\star)_0}{\|e\|_V \|e^\star\|_{V'}} \neq 0$ follows from the convergence of the eigenvectors. Hence $\|e^h\|_V$ is uniformly bounded. Together with $\left\|u - \hat{u}^h\right\|_V = d(u, V_h)$ we obtain

$$d(u, V_h \cap E_h^\star(\lambda_h)^\perp) \le C \left[d(u, V_h) + |u|_0 \inf_{e^* \in E^*(\lambda_0)} |e^{*h} - e^*|_0 \right].$$

Finally, we use $|\cdot|_0 \le C \|\cdot\|_V$.

Solution of Exercise 11.34. (a) Proof of $\|R_h - \check{R}_h\|_{L_h^2(\Omega_h) \leftarrow H_0^1(\Omega)} \le Ch$. We have $R_h - \check{R}_h = \sigma_h^x \sigma_h^y (E_2 - \sigma_h^x \sigma_h^y E_0)$. Corresponding to the grid Ω_h containing the points $\mathbf{x} = (x_i)_{i=1}^n$ we define

$$\Omega^{(h)} := \bigcup_{\mathbf{x} \in \Omega_h} \bigtimes_{j=1}^n \left(x_j - \frac{h}{2}, x_j + \frac{h}{2} \right).$$

Exercise 9.46d shows

$$\|\sigma_h^x \sigma_h^y\|_{L_h^2(\Omega_h) \leftarrow L^2(\Omega^{(h)})} \le 1.$$

Since $\Omega^{(h)}$ may be somewhat larger than Ω, the extensions $E_2 u$ and $E_0 u$ differ in $\Omega^{(h)} \backslash \Omega$. Corollary 6.62 defines the boundary strip ω_h and implies

$$\|(E_2 - \sigma_h^x \sigma_h^y E_0)u\|_{L^2(\omega_h \cup \Omega^{(h)} \backslash \Omega)} \le Ch \|u\|_{H^1(\Omega)}.$$

Exercise 9.46g shows that $\|(E_2 - \sigma_h^x \sigma_h^y E_0)u\|_{L^2(\Omega \backslash \omega_h)} \le Ch\|u\|_{H^1(\Omega)}$ holds in the remaining region $\Omega \backslash \omega_h$.

(b) Proof of $\|I - R_h P_h\|_{L_h^2(\Omega_h) \leftarrow H_h^1(\Omega_h)} \le Ch$. The equality $(R_h P_h u_h)(\mathbf{x}) = u_h(\mathbf{x})$ holds at all far-boundary points. Lemma 9.49 yields the estimate in near-boundary points.

(c) Proof of $\|P_h^* - R_h\|_{L_h^2(\Omega_h) \leftarrow H_0^1(\Omega)} \le Ch$. Let $v_h \in L_h^2(\Omega_h)$ and $u \in H_0^1(\Omega)$. We have

$$\begin{aligned}(v_h, (P_h^* - R_h)u)_{L_h^2(\Omega_h)} &= (P_h v_h, u)_{L^2(\Omega)} - (v_h, R_h u)_{L_h^2(\Omega_h)} \\ &= (\sigma_h^x \sigma_h^y \hat{P}_h v_h, u)_{L^2(\Omega)} - (v_h, \sigma_h^x \sigma_h^y E_0 u)_{L_h^2(\Omega_h)} \\ &= (\hat{P}_h v_h, \sigma_h^x \sigma_h^y u)_{L^2(\Omega)} - (v_h, \sigma_h^x \sigma_h^y E_0 u)_{L_h^2(\Omega_h)}.\end{aligned}$$

The equality

$$\begin{aligned}&(\hat{P}_h v_h, \sigma_h^x \sigma_h^y u)_{L^2(Q_\mathbf{x})} - h^2 v_h(\mathbf{x}) \sigma_h^x \sigma_h^y E_0 u(\mathbf{x}) \\ &= v_h(\mathbf{x}) \int_{Q_\mathbf{x}} (\sigma_h^x \sigma_h^y u)(\mathbf{y}) - (\sigma_h^x \sigma_h^y u)(\mathbf{x})) \mathrm{d}\mathbf{y}\end{aligned}$$

holds for all squares $Q_\mathbf{x} = \left(x - \frac{h}{2}, x + \frac{h}{2}\right) \times \left(y - \frac{h}{2}, y + \frac{h}{2}\right)$ corresponding to far-boundary points $\mathbf{x} = (x, y) \in \Omega_h$. The latter expression can be estimated by $Ch^2 |v_h(\mathbf{x})| \|u\|_{H^1(Q_\mathbf{x})}$. The near-boundary regions are treated by Lemma 9.49 and Corollary 6.62. We obtain

$$\left| (v_h, (P_h^* - R_h)u)_{L_h^2(\Omega_h)} \right| \le Ch \|v_h\|_{L_h^2(\Omega_h)} \|u\|_{H_0^1(\Omega)}$$

for all v_h and u, which is the desired estimate.

(d) Proof of $\|I - P_h^* P_h\|_{L_h^2(\Omega_h) \leftarrow H_h^1(\Omega_h)} \le Ch$. The assertion follows from the parts (b,c) and $|R_h|_{1 \leftarrow 1} \le C$ in (11.25h), since

$$I - P_h^* P_h = I - R_h P_h - (P_h^* - R_h) P_h.$$

(e) Proof of $\|I - \check{R}_h P_h\|_{L_h^2(\Omega_h) \leftarrow H_h^1(\Omega_h)} \le Ch$. Same argument as in (d).

(f) Proof of $\|I - \check{R}_h^* R_h\|_{L^2(\Omega) \leftarrow H_0^1(\Omega)} \le Ch$. First we assume $\Omega = \mathbb{R}^n$, so that the extensions E_0 and E_2 become the identity. We abbreviate $L^2(\Omega)$ and $H^1(\Omega)$ by L^2 and H^1. $R_h : L^2 \to L_h^2$ describes the averaging. The following constant extension $\hat{P}_h : L_h^2 \to L^2$ in (9.53a) defines the map $\Pi := \hat{P}_h R_h$.

Assertion A: $\Pi : L^2 \to L^2$ is the orthogonal projection. Proof: $R_h \hat{P}_h : L_h^2 \to L_h^2$ is the identity.

Assertion B: $(v_h, u_h)_{L_h^2} = (\hat{P}_h v_h, \hat{P}_h u_h)_{L^2}$. Proof: $h^2 v_h(\mathbf{x}) u_h(\mathbf{x})$ coincides with the integral of the constant extension on $Q_{\mathbf{x}} := \left(x - \frac{h}{2}, x + \frac{h}{2}\right) \times \left(y - \frac{h}{2}, y + \frac{h}{2}\right)$. Summation over all grid points $\mathbf{x} = (x, y)$ shows the assertion.

Assertion C: $\|I - \Pi\|_{L^2 \leftarrow H^1} \leq Ch$. The proof follows from (A.9) for $T := Q_{\mathbf{x}}$.

Assertion D: According to Exercise 9.46c,g and part C, $\|I - \sigma\Pi\|_{L^2 \leftarrow H^1} \leq Ch$ follows from $I - \Pi\sigma = (I - \sigma) + \sigma(I - \Pi)$.

Let $v \in L^2$ and $u \in H^1$ be arbitrary. The assertion is proved if

$$\left|(v, (I - \check{R}_h^* R_h) u)_{L^2}\right| \leq Ch \|v\|_{L^2} \|u\|_{H^1}$$

can be shown. We have $(v, (I - \check{R}_h^* R_h) u)_{L^2} = (v, u)_{L^2} - (v, \check{R}_h^* R_h u)_{L^2}$, where

$$\begin{aligned}(v, \check{R}_h^* R_h u)_{L^2} &= (\check{R}_h v, R_h u)_{L_h^2} \underset{(B)}{=} (\hat{P}_h R_h \sigma v, \hat{P}_h R_h u)_{L^2} \\ &= (\Pi\sigma v, \Pi u)_{L^2} \underset{(A)}{=} (\sigma v, \Pi u)_{L^2} \underset{\text{Exercise 9.46a,b}}{=} (v, \sigma\Pi u)_{L^2}.\end{aligned}$$

By Assertion D, the difference between the above expression and $(v, u)_{L^2}$ is estimated by $Ch\|v\|_{L^2}\|u\|_{H^1}$.

The proved estimate also holds for functions with support in Ω. Because of the extension operators deviations may appear close to the boundary. These can be estimated with the help of Lemma 9.49 and Corollary 6.62.

Solution of Exercise 11.36. (a) $|u|_1^2 = (\Lambda u, u)_0$ and $|u_h|_1^2 = (\Lambda_h u_h, u_h)_0$ are the definitions of these norms.

(b) Λ has the square root $\Lambda^{1/2}$ with

$$|u|_1^2 = (\Lambda u, u)_0 = \left(\Lambda^{1/2} u, \Lambda^{1/2} u\right)_0 = |\Lambda^{1/2} u|_0^2.$$

The scalar product in V is $(x, y)_1 = \left(\Lambda^{1/2} x, \Lambda^{1/2} y\right)_0$. The Riesz isomorphism $J : V \to V'$ is given by $J = \Lambda$ since

$$\begin{aligned}\left(J^{-1}\varphi, v\right)_V &= \varphi(v) = (\varphi, v)_0 = (\Lambda^{-1/2}\varphi, \Lambda^{1/2} v)_0 \\ &= (\Lambda^{1/2}\Lambda^{-1}\varphi, \Lambda^{1/2} v)_0 = \left(\Lambda^{-1}\varphi, v\right)_1.\end{aligned}$$

The isomorphism implies

$$|w|_{-1}^2 = |\Lambda^{-1} w|_1^2 \underset{(a)}{=} \left(\Lambda\Lambda^{-1} w, \Lambda^{-1} w\right)_0 = \left(w, \Lambda^{-1} w\right)_0 = \left(\Lambda^{-1} w, w\right)_0 = (v, w)_0.$$

The statement $|w_h|_{-1}^2 = (v_h, w_h)_0$ is completely analogous.

(c) Since $H^2(\Omega) \cap H_0^1(\Omega)$ is dense in $H_0^1(\Omega)$, the last statement of the exercise follows from the consistency condition (11.25a).

Solution of Exercise 11.43. (a) $\dim E(\lambda_0) = 1$ implies $\lim_{h\to 0} \dim E_h(\lambda_h) = 1$. For a proof repeat the solution of Exercise 11.17 and replace Theorem 11.40 and Theorem 11.16 by Theorem 11.41.

(b) If $\dim E(\lambda_0) = 1$, the scaling in Exercise 11.17b produces discrete eigenvectors $\hat{e}^h$ with $\lim \hat{e}^h = e$. The proof is unchanged.

Solution of Exercise 11.49. Each pair (ν, μ) is associated with a square $[\nu-1, \nu]\times[\mu-1, \mu]$. All squares with $\nu^2+\mu^2 \le \Lambda/\pi^2$ lie in the quarter circle

$$\left\{(x, y) : 0 \le x \le \sqrt{\Lambda}/\pi,\ 0 \le y \le \sqrt{\Lambda/\pi^2 - x^2}\right\}$$

which has the area $\frac{1}{4}\pi(\Lambda/\pi^2) = \frac{\Lambda}{4\pi}$.

Exercises of Chapter 12

Solution of Exercise 12.3. (a) Obviously, the sums $m_i + m'_j$ are independent of k.

(b) For $q = 1$, $m_1 = m'_1 = 1$, the function $L^P_{11}(\boldsymbol{\xi}; \mathbf{x})$ corresponding to (1.16) coincides with $\langle \mathbf{A}(\mathbf{x})\boldsymbol{\xi}, \boldsymbol{\xi}\rangle = \sum_{i,j=1}^n a_{ij}(\mathbf{x})\xi_i\xi_j$ from Definition 1.14a and (5.4a).

(c) In the case of the $q \times q$ system (1.21) (there with n instead of q) the orders are $k_{ij} = 1$ with $m_i = 1$, $m'_j = 0$. The operator L_{ij} $(1 \le i, j \le q)$ is $\sum_{k=1}^n (A_k)_{ij} \frac{\partial}{\partial x_k}$ so that $L^P_{ij}(\boldsymbol{\xi}; \mathbf{x}) = \sum_{k=1}^n (A_k(\mathbf{x}))_{ij}\xi_k$ and $L^P(\boldsymbol{\xi}; \mathbf{x}) = \sum_{i=k}^n \xi_k A_k(\mathbf{x})$.

Solution of Exercise 12.4. The operator $(L_{ij})_{1\le i,j\le 3}$ has the components

$$L_{ij} = \mu\delta_{ij}\Delta + (\lambda+\mu)\frac{\partial^2}{\partial x_i \partial x_j} \qquad (\delta_{ij} \text{ Kronecker symbol}),$$

so that $L^P_{ij}(\boldsymbol{\xi}; \mathbf{x}) = \mu\delta_{ij}|\boldsymbol{\xi}|^2 + (\lambda+\mu)\xi_i\xi_j$. The determinant of

$$\begin{bmatrix} (\lambda+\mu)\xi_1^2 + \mu|\boldsymbol{\xi}|^2 & (\lambda+\mu)\xi_1\xi_2 & (\lambda+\mu)\xi_1\xi_3 \\ (\lambda+\mu)\xi_1\xi_2 & (\lambda+\mu)\xi_2^2 + \mu|\boldsymbol{\xi}|^2 & (\lambda+\mu)\xi_2\xi_3 \\ (\lambda+\mu)\xi_1\xi_3 & (\lambda+\mu)\xi_2\xi_3 & (\lambda+\mu)\xi_3^2 + \mu|\boldsymbol{\xi}|^2 \end{bmatrix}$$

is equal to $\mu^2(2\mu+\lambda)|\boldsymbol{\xi}|^6$.

Solution of Exercise 12.7. (a) By assumption (12.10a,b) $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ are bounded; let C_a and C_b be the bounds. Then $|c(u,z)| \le C_c \|u\|_X \|z\|_X$ holds with $C_c := \frac{1}{2}C_a + \sqrt{\frac{1}{4}C_a^2 + C_b^2}$, hence $c(\cdot,\cdot)$ is continuous. Here we use $\|u\|_X^2 = \|v\|_V^2 + \|w\|_W^2$ for $u = \binom{v}{w}$.

(b) Adding the equations in (12.11), we obtain (12.12a,b). On the other hand, inserting $y = 0$ resp. $x = 0$ in (12.12a,b), both equations in (12.11) follow.

Solution of Exercise 12.11. (i) Let $V = V_0 \oplus V_\perp$ be the decomposition (12.21a) in orthogonal spaces. $(v,w)_V = (J_V w)(v)$ holds by definition of the Riesz isomorphism $J_V : V \to V'$. $(v,w)_V = 0$ for all $v \in V_\perp$ only holds for $w \in (V_\perp)^\perp = V_0$. This property shows that $J_V w \in V_0'$ and vice versa. Hence $J_V : V_0 \to V_0'$ and $J_V : V_\perp \to V_\perp'$ are bijective. This proves part (b) of the exercise and shows that $V' = V_0' \oplus V_\perp'$ is a direct sum.

(ii) According to Conclusion 6.69 V' is a Hilbert space with the scalar product $(\varphi,\psi)_{V'} = (J_V^{-1}\varphi, J_V^{-1}\psi)_V$. Let $\varphi \in V_0'$ and $\psi \in V_\perp'$. Part (i) ensures the existence of $v_0 \in V_0$ and $v_\perp \in V_\perp$ with $\varphi = J_V v_0$ and $\psi = J_V v_\perp$. This implies $(\varphi,\psi)_{V'} = (J_V^{-1}\varphi, J_V^{-1}\psi)_V = (v_0, v_\perp)_V = 0$ and proves the orthogonality in $V' = V_0' \oplus V_\perp'$ and therefore also part (c) of the exercise.

Solution of Exercise 12.13. (12.24a) is equivalent to

$$\sup_{x_0 \in V_0,\ \|x_0\|_V = 1} |a(v_0, x_0)| \geq \alpha \|v_0\|_V \quad \text{for all } x_0 \in V_0 \text{ with } \|x_0\|_V = 1.$$

Since the inequality is scaling invariant, it holds for all $v_0 \in V_0$. In the case of $b(\cdot,\cdot)$ one uses analogous arguments.

Solution of Exercise 12.16. Consider the system $C\binom{v_0+v_\perp}{w} = \binom{f_{10}+f_{1\perp}}{f_2}$ where $v_0 \in V_0$ etc. Elimination as in the proof of Theorem 12.12 yields the following statements:

(i) $v_\perp = B^{*-1} f_2$, i.e., $\|v_\perp\|_V \leq \beta \|f_2\|_{W'}$ with

$$\beta := \|B^{-1}\|_{W \leftarrow V_\perp'} = \|B^{*-1}\|_{V_\perp \leftarrow W'};$$

(ii) $v_0 = A_{00}^{-1}(f_{10} - A_{0\perp} v_\perp)$, i.e., $\|v_0\|_V \leq \alpha\,(\|f_{10}\|_{V'} + C_a \beta \|f_2\|_{W'})$ holds with $\alpha := \|A_{00}^{-1}\|_{V_0 \leftarrow V_0'}$ and $C_a := \|A\|_{V' \leftarrow V'}$.

(iii) $w = B^{-1}(f_{1\perp} - A_{\perp 0} v_0 - A_{\perp\perp} v_\perp)$, i.e., $\|w\|_W \leq \beta(\|f_{1\perp}\|_{V'} + C_a \|v\|_V)$.

Therefore

$$\|\tbinom{v_0+v_\perp}{w}\|_X^2 = \|v_\perp\|_V^2 + \|v_0\|_V^2 + \|w\|_W^2$$

can be estimated by $K^2(\|f_1\|_{V'}^2 + \|f_2\|_{W'}^2)$, where K only depends on α, β, C_a.

So far we have proved that (12.23) implies the existence of $C^{-1} \in L(X', X)$. On the other hand, $\binom{v_0}{w} = C^{-1}\binom{f_{10}}{0}$ contains the components $v_0 = A_{00}^{-1} f_{10}$, while $\binom{0}{w} = C^{-1}\binom{f_{1\perp}}{0}$ holds with $w = B^{-1} f_{1\perp}$ so that (12.23) follows.

The proof shows that the boundedness of A can be reduced to the boundedness of $A_{\perp 0}$ and $A_{0\perp}$.

Solution of Exercise 12.22. The domain Ω_ε lies in $[-1,1]\times[0,1]$. Let $\Omega_{\varepsilon,k}$ ($k \in \mathbb{Z}$) be the domain shifted by $2k$: $\Omega_{\varepsilon,k} = \{(x+2k, y) : (x,y) \in \Omega_\varepsilon\}$. Let Ω be the interior of $\bigcup_{k\in\mathbb{Z}} \overline{\Omega_{1/(2+|k|),k}}$. Clearly Ω is contained in the strip $\mathbb{R}\times(0,1)$. For an indirect proof we assume that $\|w\|_{L^2(\Omega)} \leq \|\nabla w\|_{\mathbf{H}^{-1}(\Omega)} / \beta$. We set $w(x,y) = 1$ for $2k < x < 2k + 2K$, $w(x,y) = -1$ for $2k - 2K' < x < 2k$ and $w = 0$ otherwise. Here $K' \approx K$ is chosen such that $w \in L_0^2(\Omega)$. Analogously to the proof

in Example 12.21 one obtains for $\Omega^+ := \Omega \cap (2k, 2k+2K) \times (0,1)$ with the boundary $\kappa = \{2k+2K\} \times (0,1)$ on the right side that

$$K \le \int_{\Omega_\varepsilon^+} |w(\mathbf{x})|^2 \, \mathrm{d}\mathbf{x} = \int_{\Omega_\varepsilon^+} w \operatorname{div} u \mathrm{d}\mathbf{x} = \int_{\Omega_\varepsilon^+} \operatorname{div} u \mathrm{d}\mathbf{x} = \int_{\gamma \cup \kappa} \langle u, \mathbf{n} \rangle \, \mathrm{d}\Gamma$$
$$= \int_\kappa u_1 \mathrm{d}\Gamma - \int_\gamma u_1 \mathrm{d}\Gamma = \int_0^1 u_1(2k+2K, y) \mathrm{d}y - \int_0^{1/(2+|k|)} u_1(0,y) \mathrm{d}y.$$

Since $\|u_1\|_{L^2(0,1)} \le C \|u_1\|_{H^{1/2}(0,1)} \le C' \|u\|_{H^1(\Omega \cap (2k+2K,\infty) \times (0,1))} \to 0$ for $K \to \infty$, the inequality $\left| \int_\kappa u_1 \mathrm{d}\Gamma \right| \le \frac{K}{3}$ follows for sufficiently large K, while $\left| \int_\gamma u_1 \mathrm{d}\Gamma \right| \le \frac{K}{3}$ as in Example 12.21 for sufficiently large k. This produces the contradiction $K \le \frac{2K}{3}$.

Solution of Exercise 12.41. Let ◸ be the combination of two triangles in the left lower corner of the square $\Omega = (0,1) \times (0,1)$. Let u_i^+ $(i = 1, 2)$ be the piecewise linear functions in the upper triangle and u_i^- another one in the lower triangle. The zero boundary condition implies $\partial u_i^+ / \partial y = 0$ and $\partial u_i^- / \partial x = 0$. In addition the divergence yields $\partial u_1^+ / \partial x = -\partial u_2^+ / \partial y = 0$ and $\partial u_2^- / \partial y = -\partial u_1^- / \partial x = 0$. Together with the zero boundary values $u_1^+ = 0$ and $u_2^- = 0$ follows. Continuity also implies $u_1^- = 0$ and $u_2^+ = 0$. Therefore the situation repeats for the triangles ◸ in $(0,h) \times (h, 2h)$, so that $u_1 = u_2 = 0$ in $(0,h) \times (0,1)$. Analogously $u_1 = u_2 = 0$ holds in $(h, 2h) \times (0,1)$ etc.

In the bilinear case the zero boundary values imply $u_i(x,y) = \gamma_i xy$ $(i = 1, 2)$ in $(0,h) \times (0,h)$. Vanishing divergence leads to $\gamma_1 y = \gamma_2 x$ implying $\gamma_1 = \gamma_2 = 0$ and $u = 0$ in $(0,h) \times (0,h)$. As above the above result can be used to prove $u = 0$ for the neighbouring squares.

Solution of Exercise 12.27. $f_2 = 0$ implies that the solution v^h belongs to $V_{0,h}$. Equation (12.30) determines the solution $v^h \in V_{0,h}$ uniquely. For any $x \in V_{0,h}$ one can replace $a(v^h, x) = f_1(x)$ by $a(v^h, x) + b(w^h, x) = f_1(x)$ with arbitrary $w^h \in W_h$ since by definition $b(w^h, x) = 0$.

Solution of Exercise 12.32. This is the statement of Theorem 8.16.

Solution of Exercise 12.34. $\phi : T \to \tilde{T}$ is a linear transformation so that the functional determinant $\det \phi'$ is constant. The substitution $\tilde{\mathbf{x}} = \phi(\mathbf{x})$ yields

$$\int_{\tilde{T}} \tilde{u}_{\tilde{T}}(\tilde{\mathbf{x}}) \mathrm{d}\tilde{\mathbf{x}} = \int_T \tilde{u}_{\tilde{T}}(\phi(\mathbf{x})) \, |\det \phi'| \, \mathrm{d}\mathbf{x}.$$

The choice $\tilde{u} = 1$ shows that $|\det \phi'| = \operatorname{area}(\tilde{T}) / \operatorname{area}(T) = 2 \operatorname{area}(\tilde{T})$. (12.38) implies $\tilde{u}_{\tilde{T}}(\phi(\mathbf{x})) = u_T(\mathbf{x})$. The value of $\int_T u_T(\mathbf{x}) \mathrm{d}\mathbf{x}$ is $1/120$.

References

1. Adams, R.A.: Sobolev Spaces. Academic Press, New York (1975)
2. Afendikova, N.G.: История метода Галеркина и его ролъ М.В. Келдыша [The history of Galerkin's method and its role in M.V. Keldysh's work]. Preprint 77, Keldyš Institute of Appl. Math. (2014). http://library.keldysh.ru/preprint.asp?id=2014-77
3. Agmon, S.: The coerciveness problem for integro-differential forms. J. Analyse Math. **6**, 183–223 (1958)
4. Agmon, S., Douglis, A., Nirenberg, L.: Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions II. Comm. Pure Appl. Math. **17**, 35–92 (1964)
5. Alinhac, S.: Hyperbolic Partial Differential Equations. Springer, Berlin (2009)
6. Apel, T., Nicaise, S.: The finite element method with anisotropic mesh grading for elliptic problems in domains with corners and edges. Math. Methods Appl. Sci. **21**, 519–549 (1998)
7. Armentano, M.G., Durán, R.G.: Mass-lumping or not mass-lumping for eigenvalue problems. Numer. Methods Partial Differential Equations **19**, 653–664 (2003)
8. Arnold, D.N., Brezzi, F., Fortin, M.: A stable finite element for the Stokes equations. Calcolo **21**, 337–344 (1984)
9. Arzelà, C.: Sulle funzioni di linee [On functions of lines]. Memorie della Reale Accademia delle Scienze dell'Istituto di Bologna, Ser. 5 **5**, 55–74 (1895)
10. Ascoli, G.: Le curve limiti di una varietà data di curve [The limit curve of a given family of curves]. Atti della R. Accad. Dei Lincei Memorie della Cl. Sci. Fis. Mat. Nat. **18**(3), 521–586 (1884)
11. Aubin, J.P.: Behaviour of the error of the approximate solution of boundary value problems for linear operators by Galerkin's and finite difference methods. Ann. Scuola Norm. Sup. Pisa **21**, 599–637 (1967)
12. Auzinger, W.: A quantitative discrete H^2-regularity estimate for the Shortley–Weller scheme in convex domains. Numer. Math. **52**, 523–537 (1988)
13. Aziz, A.K. (ed.): The Mathematical Foundation of the Finite Element Method with Applications to Partial Differential Equations. Academic Press, New York (1972). (Maryland, June 1972)
14. Babuška, I.: The theory of small changes in the domain of existence in the theory of partial differential equations and its applications. In: Babuška and Zlámal [21], pp. 13–26
15. Babuška, I.: Error-bounds for finite element method. Numer. Math. **16**, 322–333 (1971)
16. Babuška, I., Aziz, A.K.: Survey lectures on the mathematical foundations of the finite element method. In: Aziz [13], pp. 3–363
17. Babuška, I., Osborn, J.E.: Estimates for the errors in eigenvalue and eigenvector approximation by Galerkin methods, with particular attention to the case of multiple eigenvalues. SIAM J. Numer. Anal. **24**, 1249–1276 (1987)

W. Hackbusch, *Elliptic Differential Equations*, Springer Series
in Computational Mathematics 18, DOI 10.1007/978-3-662-54961-2

18. Babuška, I., Osborn, J.E.: Finite element-Galerkin approximation of the eigenvalues and eigenvectors of selfadjoint problems. Math. Comp. **52**(186), 275–297 (1989)
19. Babuška, I., Osborn, J.E.: Eigenvalue problems. In: Ciarlet and Lions [68], pp. 641–792
20. Babuška, I., Rosenzweig, H.B.: A finite element scheme for domains with corners. Numer. Math. **20**, 1–21 (1972)
21. Babuška, I., Zlámal, M. (eds.): Differential Equations and Their Applications. Czech. Academy of Sciences (1963) (also Acad. Press, New York 1963). (Prague, Sept. 1962)
22. Bank, R.E.: A software package for solving elliptic partial differential equations – users' guide 6.0. SIAM, Philadelphia (1990)
23. Bank, R.E.: Hierarchical bases and the finite element method. Acta Numerica **5**, 1–43 (1996)
24. Bank, R.E., Deotte, C.: Adventures in adaptivity. Comput. Vis. Sci. **18**, 79–91 (2017)
25. Bank, R.E., Dupont, T.F., Yserentant, H.: The hierarchical basis multigrid method. Math. Comp. **52**, 427–458 (1988)
26. Bank, R.E., Scott, L.R.: On the conditioning of finite element equations with highly refined meshes. SIAM J. Numer. Anal. **26**, 1383–1394 (1989)
27. Bank, R.E., Yserentant, H.: On the H^1-stability of the L_2-projection onto finite element spaces. Numer. Math. **126**, 361–381 (2014)
28. Baranov, V.B. et al.: Georgiĭ Ivanovich Petrov (Obituary). Sov. Phys. Usp. **30**, 289–291 (1987)
29. Bartels, S.: Numerical Approximation of Partial Differential Equations. Springer, Cham (2016)
30. Bebendorf, M.: A note on the Poincaré inequality for convex domains. Z. Anal. Anwendungen **22**, 751–756 (2003)
31. Bebendorf, M.: Efficient inversion of the Galerkin matrix of general second order elliptic operators with non-smooth coefficients. Math. Comp. **74**, 1179–1199 (2005)
32. Bebendorf, M.: Low-rank approximation of elliptic boundary value problems with high-contrast coefficients. SIAM J. Math. Anal. **48**, 932–949 (2016)
33. Bebendorf, M., Hackbusch, W.: Existence of $\mathcal{H}$-matrix approximants to the inverse FE-matrix of elliptic operators with L^∞-coefficients. Numer. Math. **95**, 1–28 (2003)
34. Becker, R., Rannacher, R.: An optimal control approach to a posteriori error estimation in finite element methods. Acta Math. **10**, 1–102 (2001)
35. Bellman, R.: Adaptive control processes – a guided tour. Princeton University Press, New Jersey (1961)
36. Bermúdez de Castro, A., Valli, A. (eds.): Computational Electromagnetism, *Lect. Notes Math.*, Vol. 2148. Springer, Berlin (2015). (Cetraro, June 2014)
37. Bernardi, C., Maday, Y., Patera, A.T.: Domain decomposition by the mortar element method. In: Kaper and Garbey [167], pp. 269–286.
38. Bers, L., Bochner, S., John, F. (eds.): Contributions to the Theory of Partial Differential Equations, *Princeton Math. Series in Annals of Mathematics Studies*, Vol. 33. Princeton University Press, New Jersey (1954). (Harriman, NY, 1952)
39. Birkhoff, G., Schulz, M.H., Varga, R.S.: Piecewise Hermite interpolation in one and two variables with applications to partial differential equations. Numer. Math. **11**, 232–256 (1968)
40. Blanchard, P., Brüning, E.: Direkte Methoden der Variationsrechnung [Direct Methods of the Variational Calculus]. Springer, Wien (1982)
41. Blum, H., Dobrowolski, K.: On finite element methods for elliptic equations on domains with corners. Computing **28**, 53–63 (1982)
42. Blum, H., Rannacher, R.: On the boundary value problem of the biharmonic operator on domains with angular corners. Math. Meth. Appl. Sci. **2**, 556–581 (1980)
43. Boccardo, L., Croce, G.: Elliptic Partial Differential Equations. Existence and Regularity of Distributional Solutions. Walter de Gruyter, Berlin (2014)
44. Bochev, P.B., Gunzburger, M.D.: Least-Squares Finite Element Methods. Springer, New York (2009)
45. Braess, D.: Finite Elements: Theory, Fast Solvers, and Applications in Elasticity Theory, 3rd ed. Cambridge Univ. Press, 2007. (German edition: Finite Elemente: Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie, 5th ed. Springer, Berlin 2013)
46. Braess, D., Dahmen, W.: Stability estimates of the mortar finite element method for 3-dimensional problems. East-West J. Numer. Math. **6**, 249–264 (1998)

47. Braess, D., Schöberl, J.: Equilibrated residual error estimator for edge elements. Math. Comp. **77**(262), 651–672 (2008)
48. Bramble, J.H.: A second order finite element analog of the first biharmonic boundary value problem. Numer. Math. **9**, 236–249 (1966)
49. Bramble, J.H., Hubbard, B.E.: A theorem on error estimation for finite difference analogues of the Dirichlet problem for elliptic equations. Contributions to Diff. Eq. **2**, 319–340 (1963)
50. Bramble, J.H., Hubbard, B.E.: Approximation of solutions of mixed boundary value problems for Poisson's equation by finite differences. J. Assoc. Comput. Mach. **12**, 114–123 (1965)
51. Bramble, J.H., Schatz, A.H.: Higher order local accuracy by averaging in the finite element method. Math. Comp. **31**, 94–111 (1977)
52. Brenner, S.C., Scott, L.R.: The Mathematical Theory of Finite Element Methods, 3rd ed. Springer, New York (2008)
53. Brezzi, F.: On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multiplies. RAIRO Anal. Numer. **8**, 129–151 (1974)
54. Brezzi, F., Buffa, A., Corsaro, S., Murli, A. (eds.): Numerical Mathematics and Advanced Applications. Proceedings of Enumath 2001 the 4th European Conference, Ischia. Springer, Mailand (2003). (Ischia, July 2001)
55. Brezzi, F., Fortin, M.: Mixed and Hybrid Finite Element Methods. Springer, New York (1991)
56. Brezzi, F., Marini, D., Süli, E.: Residual-free bubbles for advection-diffusion problems: the general error analysis. Numer. Math. **85**, 31–47 (2000)
57. Brezzi, F., Pitkäranta, J.: On the stability of the finite element approximations to the Stokes equations. In: Hackbusch [131], pp. 11–29.
58. Bube, K.P., Strikwerda, J.C.: Interior regularity estimates for elliptic systems of difference equations. SIAM J. Numer. Anal. **20**, 653–670 (1983)
59. Bubnov, I. G.: Отзыв о работе Проф. С.П. Тимошенко 'Об устойчивости упругих систем' [Report on the work 'On the stability of elastic systems' by Prof. S.P. Timoshenko]. Сборник Институт а инженеров путей сообщения императора Алексцандра I [Collection of the Institute of Transportation Engineering of Emperor Alexander I], **LXXXI**(7), 33–36 (1913)[1] – Also in: Избранные труды / Иван Григорьевич Бубнов [Selected works / Ivan Grigorevich Bubnov]. Sudprom GIZ, Leningrad (1956), pages 136–139
60. Bungartz, H.J., Griebel, M.: Sparse grids. Acta Numerica **13**, 147–269 (2004)
61. Cabannes, H., Temam, R. (eds.): Proc. 3rd Int. Conf. on the Numerical Methods in Fluid Mechanics, Vol. I, *Lect. Notes Phys.*, Vol. 18. Springer, Berlin (1973). (Paris, July 1972)
62. Canuto, C., Nochetto, R.H., Stevenson, R., Verani, M.: Convergence and optimality of hp-AFEM. Numer. Math. **135**, 1073–1119 (2017)
63. Canuto, C., Nochetto, R.H., Stevenson, R., Verani, M.: On p-robust saturation for hp-AFEM. Comput. Math. Appl. (2017). To appear
64. Carrero, J., Cockburn, B., Schötzau, D.: Hybridized globally divergencefree LDG methods. Part I: the Stokes problem. Math. Comp. **75**(254), 533–563 (2006)
65. Céa, J.: Approximation variationelle des problèmes aux limites [Variational approximation of boundary-value problems]. Ann. Inst. Fourier **14**, 345–444 (1964)
66. Chatelin, F.: Spectral Approximation of Linear Operators. Academic Press, New York (1983)
67. Ciarlet, P.G.: The Finite Element Method for Elliptic Problems. North-Holland, Amsterdam (1978)
68. Ciarlet, P.G., Lions, J.L. (eds.): Handbook of Numerical Analysis, Vol. 2. Elsevier (1991)
69. Ciarlet, P.G., Lions, J.L. (eds.): Handbook of Numerical Analysis, Vol. IV. Elsevier (1996)
70. Ciarlet, P.G., Lions, J.L. (eds.): Handbook of Numerical Analysis, Vol. VI. Elsevier (1998)

[1] This article belongs to a collection of four reports: Отзывы профессоров Кирпичева, Белзецкого, Бубнова и Колосова о сочинениях профессора Тимошенко, удостоенных премии Д.И. Журавского [Reviews of Prof. Kirpichev, Belzetskiy, Bubnov, and Kolosov about the works of Professor Timoshenko, awarded the D. I. Zhuravsky Prize], pages 1–40

71. Ciarlet, P.G., Raviart, P.A.: A mixed finite element method for the biharmonic equation. In: de Boor [82], pp. 125–145
72. Clément, P.: Approximation by finite element functions using local regularization. RAIRO Anal. Numer. **9, R-2**, 77–84 (1975)
73. Cockburn, B., Karniadakis, G.E., Shu, C.W. (eds.): Discontinuous Galerkin Methods – Theory, Computation and Applications, *Lect. Notes Comput. Sci. Eng.*, Vol. 11. Springer, Berlin (2000). (Newport, May 1999)
74. Collatz, L.: The Numerical Treatment of Differential Equations. Springer, Berlin (1966)
75. Cosner, C.: On the definition of ellipticity for systems of partial differential equations. J. Math. Anal. Appl. **158**, 80–93 (1991)
76. Courant, R.: Variational methods for the solution of problems of equilibrium and vibrations. Bull. Amer. Math. Soc. **49**, 1–23 (1943)
77. Courant, R., Friedrichs, K.O., Lewy, H.: Über die partiellen Differentialgleichungen der mathematischen Physik [On partial differential equations of mathematical physics]. Math. Ann. **100**, 32–74 (1928). (English translation in Appendix C of de Moura–Kubrusly [83])
78. Courant, R., Hilbert, D.: Methods of Mathematical Physics, Vol. 1. 2nd reprint. Wiley-VCH Verlag, Weinberg (2009). (German original: Methoden der mathematischen Physik, 4th ed. Springer, Berlin, 1993)
79. Crouzeix, M., Raviart, P.A.: Conforming and nonconforming finite element methods for solving the stationary Stokes equations I. RAIRO Anal. Numer. **7**, 33–75 (1973)
80. Crouzeix, M., Thomée, V.: The stability in L_p and W_p^1 of the L_2-projection onto finite element function spaces. Math. Comp. **48**(178), 521–532 (1987)
81. Dahmen, W., Faermann, B., Graham, I.G., Hackbusch, W., Sauter, S.A.: Inverse inequalities on non-quasiuniform meshes and applications to the mortar element method. Math. Comp. **73**, 1107–1138 (2003)
82. de Boor, C. (ed.): Mathematical Aspects of Finite Elements in Partial Differential Equations. Academic Press, New York (1974). (Madison, April 1974)
83. de Moura, C.A., Kubrusly, C.S. (eds.): The Courant-Friedrichs-Lewy (CFL) Condition. Springer, New York (2013)
84. Di Pietro, D.A., Ern, A.: Mathematical Aspects of Discontinuous Galerkin Methods. Springer, Berlin (2011)
85. Diening, L., Kreuzer, C., Stevenson, R.: Instance optimality of the adaptive maximum strategy. Found. Comput. Math. **16**, 33–68 (2016)
86. Dieudonné, J.: Foundations of Modern Analysis. Academic Press, New York (1969)
87. Dirac, P.A.M.: The Principles of Quantum Mechanics, 1st ed. Oxford University Press, Oxford (1930)
88. Dolejší, V., Feistauer, M.: Discontinuous Galerkin Method: Analysis and Applications to Compressible Flow, *SSCM*, Vol. 48. Springer, Cham (2015)
89. Dolzmann, G., Müller, S.: Estimates for Green's matrices of elliptic of systems by L^p theory. Manuscripta Math. **88**, 261–273 (1995)
90. Dörfler, W.: A convergent adaptive algorithm for Poisson's equation. SIAM J. Numer. Anal. **33**, 1106–1124 (1996)
91. Dryja, M.: Priori estimates in W_2^2 in a convex domain for systems of difference equations. USSR Comput. Math. Math. Phys.**12**(6), 291–300 (1972)
92. Dupaigne, L.: Stable Solutions of Elliptic Partial Differential Equations. Chapman & Hall/CRC, London (2011)
93. Dziuk, G.: Theorie und Numerik partieller Differentialgleichungen [Theory and Numerics of Partial Differential Equations]. Walter de Gruyter, Berlin (2010)
94. Egger, H., Rüde, U., Wohlmuth, B.I.: Energy-corrected finite element methods for corner singularities. SIAM J. Numer. Anal. **52**, 171–193 (2014)
95. Emmrich, E., Grigorieff, R.D.: Supraconvergence of a finite differrence scheme for elliptic boundary value problems of the third kind in fractional order Sobolev spaces. Comput. Methods Appl. Math. **6**, 154–177 (2006)
96. Eriksson, K., Estep, D., Hansbo, P., Johnson, C.: Introduction to adaptive methods for differential equations. Acta Numerica **4**, 105–158 (1995)

97. Euler, L.: Principia motus fluidorum [Principles of fluid motion]. Novi Commentarii Academiae Scientiarum Imperialis Petropolitanae **6**, 271–311 (1761)
98. Faustmann, M.: Approximation inverser Finite Elemente- und Randelementematrizen mittels hierarchischer Matrizen [Approximation of inverse finite-element and boundary-element matrices by hierarchical matrices]. Dissertation, Technische Universität Wien (2015)
99. Faustmann, M., Melenk, J.M., Praetorius, D.: $\mathcal{H}$-matrix approximability of the inverses of FEM matrices. Numer. Math. **131**, 615–642 (2015)
100. Fitzgibbon, W., Hoppe, R.H.W., Périaux, J., Pironneau, O., Vassilevski, P. (eds.): Advances in Numerical Mathematics. Institute of Numerical Mathematics RAS, Moscow (2006). (Moscow, Sept. 2005)
101. Flemisch, B., Melenk, J.M., Wohlmuth, B.I.: Mortar methods with curved interfaces. Appl. Numer. Math. **54**, 339–361 (2005)
102. Fletcher, C.A.J.: Computational Galerkin Methods. Springer, New York (1984)
103. Folland, G.: How to integrate a polynomial over a sphere. Amer. Math. Monthly **108**, 446–448 (2001)
104. Fortin, M.: Approximation des fonctions à divergence nulle par la méthode des éléments finis [Approximation of divergence-free functions by the finite-element method]. In: Cabannes and Temam [61], pp. 60–68
105. Friedman, A.: Partial Differential Equations of Parabolic Type. Prentice-Hall, Englewood Cliffs (1964). Reprint by R.E. Krieger Pub. Co., 1983
106. Galerkin, B.G.: Стержни и пластинки - ряды в некоторых вопросах упругого равновесия стержней и пластинок [Beams and plates - a number of problems of the elastic equilibrium of beams and plates]. Vestnik Inzh. **19**, 897–908 (1915).
107. Galligani, I., Magenes, E. (eds.): Mathematical Aspects of Finite Element Methods, *Lect. Notes Math.*, Vol. 606. Springer, Berlin (1977). (Rom, Dec. 1975)
108. Gantmacher, F.R.: The Theory of Matrices, Vol 1. AMS, Providence, Rhode Island (2000) (German edition: Matrizenrechnung, Band I. Deutscher Verlag der Wissenschaften, Berlin 1958)
109. Garcke, J.: Sparse grids in a nutshell. In: Garcke and Griebel [110], pp. 57–80.
110. Garcke, J., Griebel, M. (eds.): Sparse Grids and Applications, *Lect. Notes Comput. Sci. Eng.*, Vol. 88. Springer, Berlin (2013). (Bonn, May 2011)
111. Gårding, L.: Dirichlet's problem for linear elliptic partial differential equations. Math. Scand. **1**, 55–72 (1953)
112. Gaul, L.: From Newton's Principia via Lord Rayleigh's theory of sound to finite elements. In: Stein [271], pp. 385–398
113. Gerschgorin, S.A.: Über die Abgrenzung der Eigenwerte einer Matrix [On bounds of the eigenvalues of a matrix]. Bulletin de l'Académie des Sciences de l'URSS. Classe des sciences mathématiques et naturelles **6**, 749–754 (1931). Also in Varga [295, Appendix A]
114. Giani, S., Grubišić, L., Międlar, A., Ovall, J.S.: Robust error estimates for approximations of non-self-adjoint eigenvalue problems. Numer. Math. **133**, 471–495 (2016)
115. Gilbarg, D., Trudinger, N.S.: Elliptic Partial Differential Equations of Second Order. Springer, Berlin (1983)
116. Giles, M.B., Süli, E.: Adjoint methods for PDEs: *a posteriori* error analysis and postprocessing by duality. Acta Numerica **11**, 145–236 (2002)
117. Girault, V., Raviart, P.A.: Finite Element Methods for Navier-Stokes Equations, *SCM*, Vol. 5. Springer, Berlin (1986)
118. Giroire, J., Nédélec, J.C.: Numerical solution of an exterior Neumann problem using a double layer potential. Math. Comp. **32**(144), 973–990 (1978)
119. Gladwell, L., Wait, R. (eds.): A survey of numerical methods for partial differential equations. Clarendon Press, Oxford (1979)
120. Grasedyck, L., Greff, I., Sauter, S.A.: The AL basis for the solution of elliptic problems in heterogeneous media. SIAM J. Multiscale Model. Simul. **10**, 245–258 (2012)
121. Green, G.: An assay on the application of mathematical analysis in the theories of electricity and magnetism. Nottingham (1828)

122. Griffiths, D.F.: Finite elements for incompressible flow. Math. Meth. Appl. Sci. **1**, 16–31 (1979)
123. Grivard, P.: Elliptic Problems in Nonsmooth Domains. Pitman, Boston (1985)
124. Großmann, C., Roos, H.G., Stynes: Numerical Treatment of Partial Differential Equations. Springer, Berlin (2007) (German edition: Großmann, C., Roos: Numerische Behandlung partieller Differentialgleichungen. Springer Vieweg, Wiesbaden 2005)
125. Grüter, M., Widman, K.O.: The Green function for uniformly elliptic equations. Manuscripta Math. **37**, 303–342 (1982)
126. Gustafson, K., Abe, T.: The third boundary condition – was it Robin's? Math. Intelligencer **20**(1), 63–71 (1998)
127. Gustafson, K., Abe, T.: (Victor) Gustave Robin (1855–1897). Math. Intelligencer **20**(2), 47–53 (1998)
128. Gustafson, K., Hartman, R.: Divergence-free bases for finite element schemes in hydrodynamics. SIAM J. Numer. Anal. **20**, 697–721 (1983)
129. Hackbusch, W.: On the regularity of difference schemes. Ark. Mat. **19**, 71–95 (1981)
130. Hackbusch, W.: On the regularity of difference schemes, part II: regularity estimates for linear and nonlinear problems. Ark. Mat. **21**, 3–28 (1983)
131. Hackbusch, W. (ed.): Efficient solvers for elliptic systems, *Notes on Numerical Fluid Mechanics*, Vol. 10. Friedr. Vieweg & Sohn, Braunschweig (1984). (Kiel, Jan. 1984)
132. Hackbusch, W.: Local defect correction method and domain decomposition techniques. Comput. Suppl. **5**, 89–113 (1984)
133. Hackbusch, W.: On first and second order box schemes. Computing **41**, 277–296 (1989)
134. Hackbusch, W.: Elliptic Differential Equations: Theory and Numerical Treatment, *SSCM*, Vol. 18. Springer, Berlin (1992)
135. Hackbusch, W. (ed.): Parallel Algorithms for Partial Differential Equations, *Notes on Numerical Fluid Mechanics*, Vol. 31. Friedr. Vieweg & Sohn, Braunschweig (1991). (Kiel, Jan. 1990)
136. Hackbusch, W.: Integral Equations: Theory and Numerical Treatment, *ISNM*, Vol. 128. Birkhäuser, Basel (1995). (German edition: Integralgleichungen: Theorie und Numerik, 2nd ed. Teubner, Stuttgart 1997)
137. Hackbusch, W.: Multi-Grid Methods and Applications, *SCM*, Vol. 4. Springer, Berlin (2003)
138. Hackbusch, W.: Tensor Spaces and Numerical Tensor Calculus, *SSCM*, Vol. 42. Springer, Berlin (2012)
139. Hackbusch, W.: The Concept of Stability in Numerical Mathematics, *SSCM*, Vol. 45. Springer, Berlin (2014)
140. Hackbusch, W.: Hierarchical Matrices: Algorithms and Analysis, *SSCM*, Vol. 49. Springer, Berlin (2015). (German edition: Hierarchische Matrizen: Algorithmen und Analysis. Springer, Berlin 2009)
141. Hackbusch, W.: Solution of linear systems in high spatial dimensions. Comput. Vis. Sci. **17**, 111–118 (2015)
142. Hackbusch, W.: Iterative Solution of Large Sparse Systems of Equations, 2nd ed. Springer, Berlin (2016). (German edition: Iterative Lösung großer schwachbesetzter Gleichungssysteme, 2nd ed. Teubner, Stuttgart 1993)
143. Hackbusch, W.: Theorie und Numerik elliptischer Differentialgleichungen [Theory and Numerics of Elliptic Differential Equations], 4th ed. Springer Spektrum, Wiesbaden (2017)
144. Hackbusch, W., Sauter, S.A.: Adaptive composite finite elements for the solution of PDEs containing non-uniformly distributed micro-scales. Mat. Model. **8**(9), 31–43 (1996)
145. Hackbusch, W., Sauter, S.A.: Composite finite elements for the approximation of PDEs on domains with complicated micro-structures. Numer. Math. **75**, 447–472 (1997)
146. Hackbusch, W., Sauter, S.A.: Composite finite elements for problems containing small geometric details, part II: implementation and numerical results. Comput. Vis. Sci. **1**, 15–25 (1997)
147. Hackbusch, W., Sauter, S.A.: A new finite element approach for problems containing small geometric details. Arch. math. (Brno) **34**, 105–117 (1998)

148. Hadamard, J.: Le problème de Cauchy et les équations aux dérivées partielles linéaires hyperboliques [The Cauchy Problem and Linear Hyperbolic Partial Differential Equations]. Hermann, Paris (1932)
149. Harnack, A.: Die Grundlagen der Theorie des logarithmischen Potentiales und der eindeutigen Potentialfunktion in der Ebene [Foundations of the Theory of the Logarithmic Potential and the Unique Potential Function in the Plane]. Teubner, Leipzig (1887)
150. Hellwig, G.: Partial Differential Equations, 2nd ed. Springer Fachmedien, Wiesbaden (1977). (German original: Partielle Differentialgleichungen. Teubner, Stuttgart 1960)
151. Heuveline, V., Rannacher, R.: Adaptive FEM for eigenvalue problems. In: Brezzi et al. [54], pp. 713–722
152. Heuveline, V., Rannacher, R.: Adaptive FEM for eigenvalue problems with application in hydrodynamic stability analysis. In: Fitzgibbon et al. [100], pp. 109–140
153. Hiptmair, R.: Finite elements in computational electromagnetism. Acta Numerica **11**, 237–339 (2002)
154. Hiptmair, R.: Maxwell's equations: continuous and discrete. In: Bermúdez de Castro and Valli [36], pp. 1–58.
155. Hopf, E.: Elementare Bemerkungen über die Lösungen partieller Differentialgleichungen zweiter Ordnung vom elliptischen Typus [Elementary remarks on the solution of elliptic partial differential equations of second order]. Sitzungsber. Preuss. Akad. Wiss. **19**, 147–152 (1927)
156. Houston, P., Schwab, C., Süli, E.: Discontinuous hp-finite element methods for advection-diffusion-reaction problems. SIAM J. Numer. Anal. **39**, 2133–2163 (2002)
157. Hsiao, G.C., Wendland, W.L.: Boundary Integral Equations. Springer, Berlin (2008)
158. Hüeber, S., Wohlmuth, B.I.: Mortar methods for contact problems. In: Wriggers and Nackenhorst [310], pp. 39–47
159. Hurwitz, A., Courant, R.: Vorlesungen über allgemeine Funktionentheorie und elliptische Funktionen [Lectures on General Function Theory and Elliptic Functions], 1st ed. Springer, Berlin (1922)
160. John, F.: Partial Differential Equations, 3rd ed. Springer, New York (1978)
161. John, V.: Finite Element Methods for Incompressible Flow Problems. Springer, Berlin (2016)
162. Jost, J.: Partial Differential Equations. 3rd ed. Springer Science+Business Media New York (2013). (German edition: Partielle Differentialgleichungen. Springer, Berlin 1998)
163. Jovanovič, B., Süli, E.: Analysis of finite difference schemes for linear partial differential equations with generalized solutions, *SSCM*, Vol. 46. Springer, London (2014)
164. Jung, M., Langer, U.: Methode der finiten Elemente für Ingenieure [Finite-Element Method of Engineers], 2nd ed. Springer Vieweg, Wiesbaden (2013)
165. Kadlec, J.: О регулярности решения задачи Пуассона на области с границей выруклой области [On the regularity of the solution of the Poisson problem on a domain with boundary locally similar to the boundary of a convex open set]. Czech. Math. J. **14**, 386–393 (1964).
166. Kanschat, G.: Discontinuous Galerkin Methods for Viscous Incompressible Flows. Advances in Numerical Mathematics. Deutscher Universitätsverlag, Wiesbaden (2007)
167. Kaper, H.G., Garbey, M. (eds.): Asymptotic and Numerical Methods for Partial Differential Equations with Critical Parameters. Springer, Dordrecht (1993). (Beaune, May 1992)
168. Kato, T.: Perturbation Theory of Linear Operators. Springer, Berlin (1995)
169. Kellogg, O.D.: Foundations of Potential Theory. Advance Publ., London (1929)
170. Kellogg, R.B., Osborn, J.E.: A regularity result for the Stokes problem in a convex polygon. J. Funct. Anal. **21**, 397–431 (1976)
171. Kim, C., Lazarov, R.D., Pasciak, J.E., Vassilevski, P.: Multiplier spaces for the mortar finite element method in three dimensions. SIAM J. Numer. Anal. **39**, 519–538 (2001)
172. Knabner, P., Angermann, L.: Numerical Methods for Elliptic and Parabolic Partial Differential Equations. Springer, New York (2003) (German edition: Numerik partieller Differentialgleichungen. Springer, Berlin 2000)
173. Kondrat'ev, V.A.: Boundary value problems for elliptic equations in domains with conical or angular points. Trans. Moscow Math. Society **16**, 227–313 (1967)

174. Kornhuber, R., Peterseim, D., Yserentant, H.: An analysis of a class of variational multiscale methods based on subspace decomposition. arXiv (2017)
175. Kornhuber, R., Yserentant, H.: Numerical homogenization of elliptic multiscale problems by subspace decomposition. Multiscale Model. Simul. **14**, 1017–1036 (2016)
176. Kress, R.: Linear Integral Equations, 2nd ed. Springer, New York (1997)
177. Kreuzer, C., Siebert, K.G.: Decay rates of adaptive finite elements with Dörfler marking. Numer. Math. **117**, 679–716 (2011)
178. Ladyženskaja, O.A.: The Mathematical Theory of Viscous Incompressible Flow. Gordon and Breach Science Publishers, New York (1963). (2nd ed.: 1969; first Russian edition: 1961)
179. Ladyženskaja, O.A.: Funktionalanalytische Untersuchungen der Navier-Stokesschen Gleichungen [Functional Analytic Studies of the Navier-Stokes Equations]. Akademie-Verlag, Berlin (1965)
180. Ladyženskaja, O.A., Ural'ceva, N.N.: Linear and Quasilinear Elliptic Equations. Academic Press, Orlando (1968)
181. Lapin, A.V.: Study of the $W_2^{(2)}$ convergence of difference schemes for quasilinear elliptic equations. USSR Comput. Math. Math. Phys. **14**(6), 140–149 (1974)
182. Larson, M.G.: A posteriori and a priori error analysis for finite element approximations of self-adjoint elliptic eigenvalue problems. SIAM J. Numer. Anal. **38**, 608–625 (2001)
183. Larsson, S., Thomée, V.: Partial Differential Equations with Numerical Methods, *Texts in Applied Mathematics*, Vol. 45. Springer, Berlin (2003)
184. Lax, P.D., Milgram, A.N.: Parabolic equations. In: Bers et al. [38], pp. 167–190
185. Layton, W., Morley, T.D.: On central difference approximations to general second order elliptic equations. Linear Algebra Appl. **97**, 65–75 (1987)
186. Le Borne, S.: Preconditioned nullspace method for the two-dimensional Oseen problem. SIAM J. Sci. Comput. **31**, 2494–2509 (2009)
187. Le Borne, S., Ovall, J.S.: Rapid error reduction for block Gauss–Seidel based on p-hierarchical basis. Numer. Linear Algebra Appl. **20**, 743–760 (2013)
188. Lebesgue, H.: Sur des cas d'impossibilité du problème de Dirichlet ordinaire [On the non-solvability of the Dirichlet problem]. C. R. des séances de la Soc. Math. de France **41** (1913). Also in: Œuvres Scientifiques, Vol. 4. Université de Genève, 1973, page 131
189. Leis, R.: Vorlesungen über partielle Differentialgleichungen zweiter Ordnung [Lectures on Partial Differential Equations of Second Order]. Bibliographisches Institut, Mannheim (1967)
190. Leissa, A.W.: The historical bases of the Rayleigh and Ritz methods. J. Sound Vibration **287**, 961–978 (2005)
191. Lesaint, P., Zlámal, M.: Superconvergence of the gradient of finite element solutions. RAIRO Anal. Numer. **13**, 139–166 (1979)
192. Levendorskii, S.Z.: Asymptotic Distribution of Eigenvalues of Differential Operators. Springer, Dordrecht (1990)
193. Liesen, J., Mehrmann, V.: Linear Algebra. Springer, Cham (2015). (German edition: Lineare Algebra, 2nd ed. Springer Spektrum, Wiesbaden 2015)
194. Lions, J.L., Magenes, E.: Non-Homogeneous Boundary Value Problems and Applications, Vol. I. Springer, Berlin (1972)
195. Lions, J.L., Magenes, E.: Non-Homogeneous Boundary Value Problems and Applications, Vol. II. Springer, Berlin (1972)
196. Lions, J.L., Magenes, E.: Non-Homogeneous Boundary Value Problems and Applications, Vol. III. Springer, Berlin (1973)
197. Louis, A.: Acceleration of convergence for finite element solutions of the Poisson equation. Numer. Math. **33**, 43–53 (1979)
198. Maeß, G.: Vorlesungen über numerische Mathematik, I. Lineare Algebra [Lectures on Numerical Mathematics, I. Linear Algebra]. Birkhäuser, Basel (1985)
199. Marchuk, G., Shaidurov, V.V.: Difference methods and their extrapolation. Libreria Editrice Universitaria Levrotto & Bella, Torino (1983)
200. Marion, M., Temam, R.: Navier-Stokes Equations: Theory and Approximation. In: Ciarlet and Lions [70], pp. 503–688

201. Meis, T., Marcowitz, U.: Numerical Solution of Partial Differential Equations. Springer New York (1981). (German edition: Numerische Behandlung partieller Differentialgleichungen. Springer, Berlin 1978)
202. Meißner, E. (ed.): Verhandlungen des 2. Internationalen Kongresses für Technische Mechanik [Proceedings of the 2nd International Congress on Technical Mechanics], Zürich 1926. Verlag Orell Füssli, Zürich and Leipzig (1927). (Zürich, Sept. 1926)
203. Meister, A., Struckmeier, J. (eds.): Hyperbolic Partial Differential Equations: Theory, Numerics and Applications. Vieweg + Teubner Verlag, Wiesbaden (2002)
204. Melenk, J.M.: hp-Finite Element Methods for Singular Perturbations, *Lect. Notes Math.*, Vol. 1796. Springer, Berlin (2002)
205. Miranda, C.: Partial Differential Equations of Elliptic Type. Springer, Berlin (1970)
206. Mizohata, S.: The Theory of Partial Differential Equations. University Press, Cambridge (1973)
207. Monk, P.: A mixed finite element method for the biharmonic equation. SIAM J. Numer. Anal. **24**, 737–749 (1987)
208. Nédélec, J.C.: Mixed finite elements in R^3. Numer. Math. **35**, 315–341 (1980)
209. Nédélec, J.C.: A new family of mixed finite elements in R^3. Numer. Math. **50**, 57–81 (1986)
210. Nečas, J.: Sur la coercivité des formes sesqui-linéares elliptiques [On the coercivity of elliptic sesquilinear forms]. Rev. Roumaine Math. Pures Appl. **9**, 47–69 (1964)
211. Nečas, J.: Equations aux dérivées partielles [Partial Differential Equations]. Presse de l'Université de Montréal (1965)
212. Nitsche, J.: Ein Kriterium für die Quasi-Optimalität des Ritzschen Verfahrens [A criterion on the quasi-optimality of the Ritz method]. Numer. Math. **11**, 346–348 (1968)
213. Oganesjan, L.A., Ruchovec, L.A.: Study of the rate of convergence of variational difference schemes for second-order elliptic equations in a two-dimensional field with a smooth boundary. USSR Comput. Math. Math. Phys.**9**(5), 158–183 (1969)
214. Oganesjan, L.A., Ruchovec, L.A.: Вариационно-разностные методы решения эллиптичестих уравнеий [Variational difference method for solving elliptic equations]. Izdatel'stvo Akademii Nauk Armjanskoj SSR, Jerevan (1979)
215. Ovall, J.S.: The Laplacian and mean and extreme values. Amer. Math. Monthly **123**, 287–291 (2016)
216. Parlett, B.N.: The Rayleigh quotient iteration and some generalizations for nonnormal matrices. Math. Comp. **28**(127), 679–693 (1974)
217. Payne, L.E., Weinberger, H.F.: An optimal Poincaré inequality for convex domains. Arch. Rat. Mech. Anal. **5**, 286–292 (1960)
218. Pazy, A.: Semigroups of Linear Operators and Applications to Partial Differential Equations. Springer, New York (1983)
219. Pereyra, V., Proskurowski, W., Widlund, O.B.: High order fast Laplace solvers for the Dirichlet problem on general regions. Math. Comp. **31**, 1–16 (1977)
220. Peterseim, D.: The Composite Mini Element: A Mixed FEM for the Stokes Equations on Complicated Domains. Doctoral thesis, Universität Zürich, Institut für Mathematik (2007)
221. Peterseim, D., Sauter, S.A.: The composite mini element - Coarse mesh computation of Stokes flows on complicated domains. SIAM J. Numer. Anal. **46**, 3181–3206 (2008)
222. Peterseim, D., Sauter, S.A.: Finite element methods for the Stokes problem on complicated domains. Comput. Methods Appl. Mech. Engrg. **200**, 2611–2623 (2011)
223. Peterseim, D., Sauter, S.A.: Composite finite elements for elliptic interface problems. Math. Comp. **83**(290), 2657–2674 (2014)
224. Petrov, G.I.: Применение метода галеркина к задаче об устойчивости течения вязкой жидкости [Application of the Galerkin method and the problem of flow stability of a viscous liquid]. J. Appl. Math. Mech. **4**, 3–12 (1940).
225. Petrovsky, I.G.: Lectures on Partial Differential Equations. Interscience Publishers, New York (1954)
226. Pflaumann, E., Unger, H.: Funktionalanalysis I. Bibliographisches Institut, Mannheim (1968)

227. Pizzetti, P.: Sulla media dei valori che una funzione dei punti dello spazio assume alla superficie di una sfera [On the mean value which a spatial function takes on a sphere]. Rend. Reale Accad. Lincei, Ser. 5^a, 1^o Sem. **18**, 182–185 (1909)
228. Prager, W., Synge, J.L.: Approximations in elasticity based on the concept of function spaces. Quart. Appl. Math. **5**, 241–269 (1947)
229. Preusser, T., Rumpf, M., Sauter, S.A., Schwen, O.: 3D composite finite elements for elliptic boundary value problems with discontinuous coefficients. SIAM J. Sci. Comput. **33**, 2115–2143 (2011)
230. Quarteroni, A., Sacco, R., Saleri, F.: Numerical Mathematics, 2nd ed. Springer Berlin (2007). (German edition: Numerische Mathematik 1. Springer, Berlin 2002)
231. Rannacher, R.: Approximation of simply supported plates and the Babuška paradox. ZAMM **59**, T73–T76 (1979)
232. Raviart, P.A., Thomas, J.M.: A mixed finite element method for second order elliptic problems. In: Galligani and Magenes [107], pp. 292–315
233. Raviart, P.A., Thomas, J.M.: Primal hybrid finite element methods for 2nd order elliptic equations. Math. Comp. **31**(138), 391–413 (1977)
234. Rayleigh, J.W.: The Theory of Sound, Vol. 1. Macmillan & Co, London (1877)
235. Rektorys, K.: Variational Methods in Mathematics, Science and Engineering. Reidel Publ., Dordrecht (1977)
236. Richardson, L.F., Gaunt, A.: The deferred approach to the limit. Philosophical Transactions of the Royal Society of London, Series A **226**, 299–361 (1927)
237. Richtmyer, R.D., Morton, K.W.: Difference Methods for Initial-value Problems, 2nd ed. John Wiley & Sons, New York (1967). Reprint by Krieger Publ., Malabar, Florida, 1994
238. Riesz, F.: Über lineare Funktionalgleichungen [On linear functional equations]. Acta Math. **41**, 71–98 (1916)
239. Riesz, F., Sz.-Nagy, B.: Functional Analysis. Dover Publ. Inc, New York (1990)
240. Ritz, W.: Theorie der Transversalschwingungen einer quadratischen Platte mit freien Rändern [Theory of transversal vibrations of a square plate with free boundaries]. Ann. Phys. **38**, 737–786 (1809)
241. Ritz, W.: Über eine neue Methode zur Lösung gewisser Randwertaufgaben [On a new method for solving certain boundary-value problems]. Göttinger Nachr. Math.-Phys. Klasse pp. 236–248 (1908)
242. Ritz, W.: Über eine neue Methode zur Lösung gewisser Variationsprobleme der mathematischen Physik [On a new method for solving certain variational problems of mathematical physics]. J. Reine Angew. Math. **135**, 1–61 (1909)
243. Rivara, M.C.: Algorithms for refining triangular grids suitable for adaptive and multigrid techniques. Int. J. Num. Meth. Engng. **20**, 745–756 (1984)
244. Rivara, M.C.: New longest-edge algorithms for the refinement and/or improvement of unstructured triangulations. Int. J. Num. Meth. Engng. **40**, 3313–3324 (1997)
245. Rivière, B.: Discontinuous Galerkin Methods for Solving Elliptic and Parabolic Equations: Theory and Implementation. SIAM, Philadelphia (2008)
246. Roos, H.G., Stynes, M., Tobiska, L.: Numerical Methods for Singularly Perturbed Differential Equations: Convection-Diffusion and Flow Problems, *SSCM*, Vol. 24. Springer, Berlin (1996)
247. Salamon, D.A.: Funktionentheorie [Function Theory]. Springer, Basel (2012)
248. Samarskii, A.A.: The Theory of Difference Schemes. Marcel Dekker, Ney York (2001). (German edition: Theorie der Differenzenverfahren. Akademische Verlagsgesellschaft Geest & Portig K.-G., Leipzig 1984)
249. Sauter, S.A.: hp-Finite elements for elliptic eigenvalue problems: error estimates which are explicit with respect to λ, h, and p. SIAM J. Numer. Anal. **48**, 95–108 (2010)
250. Sauter, S.A., Schwab, C.: Boundary Element Methods, *SSCM*, Vol. 39. Springer, Berlin (2011). (German edition: Randelementmethoden. Teubner, Stuttgart 2004)
251. Sauter, S.A., Warnke, R.: Extension operators and approximation on domains containing small geometric details. East-West J. Numer. Math. **7**, 61–77 (1999)

252. Sauter, S.A., Warnke, R.: Composite finite elements for elliptic boundary value problems with discontinuous coefficients. Computing **77**, 29–55 (2006)
253. Schatz, A.H.: A weak discrete maximum principle and stability in the finite element method in L^∞ on plane polygonal domains. Math. Comp. **34**, 77–91 (1980)
254. Schatz, A.H., Wahlbin, L.B.: Maximum norm estimates in the finite element method on plane polygonal domains, part 1. Math. Comp. **32**(141), 73–109 (1978)
255. Schatz, A.H., Wahlbin, L.B.: Maximum norm estinates in the finite element method on plane polygonal domains, part 2, refinements. Math. Comp. **33**, 465–492 (1979)
256. Schatz, A.H., Wahlbin, L.B.: On the quasi-optimality in L_∞ of the $\mathring{H}^1$-projection into finite element spaces. Math. Comp. **38**(157), 1–22 (1982)
257. Schatz, A.H., Wahlbin, L.B.: On the finite element method for singularly perturbed reaction-diffusion problems in two and one dimensions. Math. Comp. **40**(161), 47–89 (1983)
258. Schauder, J.: Über lineare elliptische Differentialgleichungen zweiter Ordnung [On linear elliptic differential equations of second order]. Math. Z. **38**, 257–282 (1934)
259. Schellbach, K.: Probleme der Variationsrechnung [Problems of variational calculus]. J. Reine Angew. Math. **41**, 293–363 (1851)
260. Schulze, B.W., Wildenhain, G.: Methoden der Potentialtheorie für elliptische Differentialgleichungen beliebiger Ordnung [Methods of Potential Theory of Elliptic Differential Equations of Arbitrary Order]. Springer, Basel (1977)
261. Schwab, C.: p- and hp-Finite Element Methods. Oxford University Press, Oxford (1998)
262. Schwarz, H.R.: Finite element methods. Academic Press, London (1988). (German original: Methode der finiten Elemente, 3rd ed. Teubner, Stuttgart 1991)
263. Scott, L.R.: Interpolated boundary conditions in the finite element method. SIAM J. Numer. Anal. **12**, 404–427 (1975)
264. Shortley, G.H., Weller, R.: Numerical solution of Laplace's equation. J. Appl. Phys. **9**, 334–348 (1938)
265. Smolyak, S.A.: Quadrature and interpolation formulas for tensor products of certain classes of functions. Sov. Math. Dokl. **4**, 240–243 (1963)
266. Sobolev, S.L.: Some Applications of Functional Analysis in Mathematical Physics, 3rd. ed. AMS Providence, Rhode Island (1991). (German edition: Einige Anwendungen der Funktionalanalysis auf Gleichungen der mathematischen Physik. Akademie-Verlag, Berlin 1964)
267. Stein, E.: An appreciation of Erich Trefftz. Comput. Assist. Mech. Eng. Sci. **4**, 301–304 (1997)
268. Stein, E.: History of the finite element method - mathematics meets mechanics - part I: engineering developments. In: Stein [271], pp. 399–442
269. Stein, E.: History of the finite element method - mathematics meets mechanics - part II: mathematical foundation of primal FEM for elastic deformations, error analysis and adaptivity. In: Stein [271], pp. 443–478
270. Stein, E.: The origins of mechanical conservation principles and variational calculus in the 17th century. In: Stein [271], pp. 3–22
271. Stein, E. (ed.): The History of Theoretical, Material and Computational Mechanics - Mathematics Meets Mechanics and Engineering. Springer, Berlin (2014)
272. Steinbach, O.: Numerical Approximation Methods for Elliptic Boundary Value Problems: Finite and Boundary Elements. Springer, New York (2008) – Deutsche Ausgabe: Teubner Wiesbaden (2003)
273. Steklov, V.A.: Sur les problèmes fondamentaux de la physique mathematique [On fundamental problems of mathematical physics]. Annales scientifiques de l'École Normale Supérieure 3e série **19**, 191–259 and 455–490 (1902)
274. Stoer, J.: Einführung in die Numerische Mathematik I [Introduction to Numerical Mathematics], 5th ed. Springer, Berlin (1989)
275. Strang, G.: Variational crimes in the finite element method. In: Aziz [13], pp. 689–710.
276. Strang, G., Fix, G.: An Analysis of the Finite Element Method. North Oxford Academic, Oxford (1973)
277. Stummel, F.: Diskrete Konvergenz linearer Operatoren I [Discrete convergence of linear operators I]. Math. Ann. **190**, 557–597 (1970)

278. Stummel, F.: The generalized patch test. SIAM J. Numer. Anal. **16**, 449–471 (1979)
279. Süli, E.: Convergence of finite volume schemes for Poisson's equation on non-uniform meshes. SIAM J. Numer. Anal. **28**, 1419–1430 (1991)
280. Sylvester, J.J.: A demonstration of the theorem that every homogeneous quadratic polynomial is reducible by real orthogonal substitutions to the form of a sum of positive and negative squares. The London, Edinburgh and Dublin Philos. Mag. and J. of Science (Ser. 4) **4**, 138–142 (1852)
281. Temam, R.: Navier-Stokes Equations and Nonlinear Functional Analysis, 2nd ed. SIAM, Philadelphia (1995)
282. Temam, R.: Navier-Stokes Equations and Theory and Numerical Analysis. AMS Chelsea Publ. (2001)
283. Thatcher, R.W.: On the finite element method for unbounded regions. SIAM J. Numer. Anal. **15**, 466–477 (1978)
284. Thomasset, F.: Implementation of Finite Element Methods for Navier-Stokes Equations. Springer, New York (1981)
285. Thomée, V.: Discrete interior Schauder estimates for elliptic difference operators. SIAM J. Numer. Anal. **5**, 626–645 (1968)
286. Thomée, V.: High order local approximations to derivatives in the finite element method. Math. Comp. **31**, 652–660 (1977)
287. Thomée, V.: Galerkin Finite Element Methods for Parabolic Problems, *SSCM*, Vol. 25. Springer, Berlin (1997)
288. Thomée, V.: From finite difference to finite elements - a short history of numerical analysis of partial differential equations. J. Comput. Appl. Math. **128**, 1–54 (2001)
289. Thomée, V., Westergren, B.: Elliptic difference equations and interior regularity. Numer. Math. **11**, 196–210 (1968)
290. Trangenstein, J.A.: Numerical Solution of Hyperbolic Partial Differential Equations. Cambridge University Press, Cambridge (2007)
291. Trangenstein, J.A.: Numerical Solution of Elliptic and Parabolic Partial Differential Equations. Cambridge University Press, Cambridge (2013)
292. Trefftz, E.: Ein Gegenstück zum Ritzschen Verfahren [A counterpart of the Ritz method]. In: Meißner [202], pp. 131–138
293. Trefftz, E.: Konvergenz und Fehlerabschätzung beim Ritzschen Verfahren [Convergence and error estimation of Ritz' method]. Math. Ann. **100**, 503–521 (1928)
294. van Linde, H.J.: High-order finite-difference methods for Poisson's equation. Math. Comp. **28**(126), 369–391 (1974)
295. Varga, R.S.: Geršgorin and his Circles. Springer, Berlin (2004)
296. Velte, W.: Direkte Methoden der Variationsrechnung [Direct Methods of the Variational Calculus]. Teubner, Stuttgart (1976)
297. Verfürth, R.: A Review of a posteriori Error Estimation and Adaptive Mesh-refinement Techniques. J. Wiley and Teubner, Stuttgart (1996)
298. Wahlbin, L.B.: Maximum norm error estimates in the finite element method with isoparametric quadratic elements and numerical integration. RAIRO Anal. Numer. **12**, 173–262 (1978)
299. Wahlbin, L.B.: On the sharpness of certain local estimates for $\mathring{H}^1$ projections into finite element spaces: influence of a reentrant corner. Math. Comp. **42**(165), 1–8 (1984)
300. Wahlbin, L.B.: Superconvergence in Galerkin Finite Element Methods, *Lect. Notes Math.*, Vol. 1605. Springer, Berlin (1995)
301. Walter, W.: Differential and Integral Inequalities. Springer, Berlin (1970)
302. Walter, W.: Einführung in die Potentialtheorie [Introduction into Potential Theory]. BI-Hochschultaschenbuch, Mannheim (1971)
303. Wathen, A.J.: Realistic eigenvalue bounds for the Galerkin mass matrix. IMA J. Numer. Anal. **7**, 449–457 (1987)
304. Weyl, H.: Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung) [The asymptotic distribution of eigenvalues of linear partial differential equations (with an application to the theory of black-body radiation)]. Math. Ann. **71**, 441–479 (1912)

305. Weymuth, M., Sauter, S.A.: An adaptive local (AL) basis for elliptic problems with complicated discontinuous coefficients. Proc. Appl. Math. Mech. (PAMM) **15**, 605–606 (2015)
306. Wienholtz, E., Kalf, H., Kriecherbauer, T.: Elliptische Differentialgleichungen zweiter Ordnung [Elliptic Differential Equations of Second Order]. Springer, Berlin (2009)
307. Witsch, K.: Numerische Quadratur bei Projektionsverfahren [Numerical quadrature in projection methods]. Numer. Math. **30**, 185–206 (1978)
308. Wloka, J.: Partial Differential Equations. Cambridge Univ. Press, Cambridge (1987). (German original: Partielle Differentialgleichungen: Sobolevräume und Randwertaufgaben. Teubner, Stuttgart 1982)
309. Wohlmuth, B.: Discretization methods and iterative solvers based on domain decomposition. Habilitationsschrift, Universität Augsburg (1999)
310. Wriggers, P., Nackenhorst, U. (eds.): Analysis and Simulation of Contact Problems, *Lect. Notes Appl. Comput. Mech.*, Vol. 27. Springer, Berlin (2006) (Loccum, July 2005)
311. Ye, X., Hall, C.A.: A discrete divergence-free basis for finite element methods. Numer. Algorithms **16**, 365–380 (1997)
312. Yosida, K.: Functional Analysis. Springer, New York (1968)
313. Yserentant, H.: On the multi-level splitting of finite element spaces. Numer. Math. **49**, 379–412 (1986)
314. Yserentant, H.: A short theory of the Rayleigh-Ritz method. Comput. Methods Appl. Math. **13**, 495–502 (2013)
315. Zeidler, E. (ed.): Oxford Users' Guide to Mathematics. Oxford Univ. Press, Oxford 2004. (German edition: Springer-Taschenbuch der Mathematik, 3rd ed. Springer Spektrum, Wiesbaden 2013)
316. Zenger, C.: Sparse grids. In: Hackbusch [135], pp. 241–251
317. Zenger, C., Gietl, H.: Improved difference schemes for the Dirichlet problem of Poisson's equation in the neighbourhood of corners. Numer. Math. **30**, 315–332 (1978)
318. Zhou, G., Rannacher, R.: Pointwise superconvergence of the streamline diffusion finite-element method. Numer. Methods Partial Differential Equations **12**, 123–145 (1996)
319. Zienkiewicz, O.O.: The Finite Element Method. 3rd ed. McGraw-Hill, London (1977). (German edition: Methode der finiten Elemente. Carl Hanser, München 1984)
320. Zienkiewicz, O.O.: Origins, milestones and directions of the finite element method – a personal view. In: Ciarlet and Lions [69], pp. 3–67
321. Zlámal, M.: Discretisation and error estimates for elliptic boundary value problems of the fourth order. SIAM J. Numer. Anal. **4**, 626–639 (1967)
322. Zlámal, M.: On the finite element method. Numer. Math. **12**, 394–409 (1968)
323. Zlámal, M.: Some superconvergence results in the finite element method. In: Galligani and Magenes [107], pp. 353–362
324. Zlámal, M.: Superconvergence and reduced integration in the finite element method. Math. Comp. **32**(143), 663–685 (1978)

List of authors involved in the references from above, but not placed as first author.

Abe, T. [126, 127]
Angermann, L. [172]
Aziz, A.K. [16]
Bochner, S. [38]
Brezzi, F. [8]
Brüning, E. [40]
Buffa, A. [54]
Cockburn, B. [64]
Corsaro, S. [54]
Courant, R. [159]
Croce, G. [43]
Dahmen, W. [46]
Deotte, C. [24]
Dobrowolski, K. [41]
Douglis, A. [4]
Dupont, T.F. [25]
Durán, R.G. [7]
Ern, A. [84]
Estep, D. [96]
Faermann, B. [81]
Feistauer, M. [88]
Fix, G. [276]
Fortin, M. [8, 55]
Friedrichs, K.O. [77]
Garbey, M. [167]
Gaunt, A. [236]
Gietl, H. [317]
Graham, I.G. [81]
Greff, I. [120]
Griebel, M. [60, 110]
Grigorieff, R.D. [95]
Grubišić, L. [114]
Gunzburger, M.D. [44]
Hackbusch, W. [33, 81]
Hall, C.A. [311]
Hansbo, P. [96]
Hartman, R. [128]
Hilbert, D. [78]
Hoppe, R.H.W. [100]
Hubbard, B.E. [49, 50]
John, F. [38]
Johnson, C. [96]
Kalf, H. [306]
Karniadakis, G.E. [73]
Kreuzer, C. [85]
Kriecherbauer, T. [306]
Kubrusly, C.S. [83]
Langer, U. [164]
Lazarov, R.D. [171]
Lewy, H. [77]
Lions, J.L. [68, 69, 70]
Maday, Y. [37]
Magenes, E. [107, 194–196]
Marcowitz, U. [201]
Marini, D. [56]
Mehrmann, V. [193]
Melenk, J.M. [99, 101]
Międlar, A. [114]
Milgram, A.N. [184]
Morley, T.D. [185]
Morton, K.W. [237]
Müller, S. [89]
Murli, A. [54]
Nackenhorst, U. [310]
Nédélec, J.C. [118]
Nicaise, S. [6]
Nirenberg, L. [4]
Nochetto, R.H. [62, 63]
Osborn, J.E. [17, 18, 19, 170]
Ovall, J.S. [114, 187]
Pasciak, J.E. [171]
Patera, A. [37]
Périaux, J. [100]
Peterseim, D. [174]
Pieper, G.W. [167]
Pironneau, O. [100]
Pitkäranta, J. [57]
Praetorius, D. [99]
Proskurowski, W. [219]
Rannacher, R. [34, 42, 151, 152, 318]
Raviart, P.A. [71, 79, 117]
Roos, H.G. [124]
Rosenzweig, H.B. [20]
Ruchovec, L.A. [213, 214]
Rüde, U. [94]
Rumpf, M. [229]
Sacco, R. [230]
Saleri, F. [230]
Sauter, S.A. [81, 120, 144–147, 221–223, 229, 305]
Schatz, A.H. [51]
Schöberl, J. [47]
Schötzau, D. [64]
Schulz, M.H. [39]
Schwab, C. [156, 250]
Schwen, O. [229]
Scott, L.R. [26, 52]
Shaidurov, V.V. [199]
Shu, C.W. [73]
Siebert, K.G. [177]
Stevenson, R. [62, 63, 85]
Strikwerda, J.C. [58]
Struckmeier, J. [203]
Stynes, M. [124, 246]
Süli, E. [56, 116, 156, 163]
Synge, J.L. [228]
Sz.-Nagy, B. [239]
Temam, R. [61, 200]
Thomas, J.M. [232, 233]
Thomée, V. [80, 183]
Tobiska, L. [246]
Trudinger, N.S. [115]
Unger, H. [226]
Ural'ceva, N.N. [180]
Valli, A. [36]
Varga, R.S. [39]
Vassilevski, P. [100, 171]
Verani, M. [62, 63]
Wahlbin, L.B. [254–257]
Wait, R. [119]
Warnke, R. [251, 252]
Weinberger, H.F. [217]
Weller, R. [264]
Wendland, W.L. [157]
Westergren, B. [289]
Widlund, O.B. [219]
Widman, K.O. [125]
Wildenhain, G. [260]
Wohlmuth, B.I. [94, 101, 158]
Yserentant, H. [25, 27, 174, 175]
Zlámal, M. [21, 191]

List of Symbols and Abbreviations

Symbols

$\S$	Chapter, Section, Subsection, etc.
$\#X$	Cardinality of a set X (number of elements).
$\oplus$	Direct sum; see Lemma 6.15.
∇	Gradient
$\partial^+, \partial^-, \partial^0, \ldots$	Difference operators; see §4.1.
$\partial_n^{\pm}$	Difference operator in normal direction; see (4.44).
$\partial_x^\alpha = \partial^\alpha / \partial x^\alpha$	Partial derivative of order $\lvert\alpha\rvert$ with respect to x.
$\partial/\partial n$	Normal derivative; see page 16.
$\bullet^{\mathsf{H}}$	Hermitian transposition of a matrix or a vector.
$\bullet^{\mathsf{T}}$	Transposition of a matrix or a vector.
$\bullet^{-\mathsf{T}}$	Transposition of an inverse matrix.
$\bullet^{\perp}$	Orthogonal space; see §6.1.4.
$\cdot\lvert_\omega, \cdot\lvert_\Gamma$	Restriction of a function $\cdot$ to a (smaller) domain ω, Γ etc.
$\lvert\cdot\rvert$	Absolute value of a real or complex number.
$\lvert\cdot\rvert$	Euclidean norm in $\mathbb{R}^n$; see §2.2 and (4.34).
$\lvert\nu\rvert$	Length of a multi-index ν; see (3.11a).
$\lvert\cdot\rvert_0$	Short notation of the L^2-Norm $\lVert\cdot\rVert_{L^2}$; see (6.8).
$\lvert\cdot\rvert_k$, $\lvert\cdot\rvert_s$	Short notation of the Sobolev norms $\lVert\cdot\rVert_{H^k}$, $\lVert\cdot\rVert_{H^s}$; see (6.13) and (6.22b).
$\lvert\cdot\rvert_k^\wedge$, $\lvert\cdot\rvert_s^\wedge$	Sobolev norms defined by Fourier transform; see (6.20) and (6.21b).
$\lvert\cdot\rvert_{k,0}$	Norm on $H_0^k(\Omega)$; see (6.15).
$\lVert\cdot\rVert_\infty$	Maximum norm of vectors, row-sum norm of matrices (see (4.32)), Supremum norm of functions; see Example 6.1b.
$\lVert\cdot\rVert_2$	Euclidean norm of vectors (see (4.34)) and spectral norm of matrices (see 4.28c).
$\lVert\cdot\rVert_h$	Euclidean norm scaled by $h^{n/2}$; see (8.87b).
$\lVert\cdot\rVert_{L_h^2}$, $\lVert\cdot\rVert_{H_h^{\pm 1}}$	Norms of grid functions on Q_h; see §9.3.1.

W. Hackbusch, *Elliptic Differential Equations*, Springer Series
in Computational Mathematics 18, DOI 10.1007/978-3-662-54961-2

$\|\cdot\|_P$	Norm in $\mathbb{R}^N$; see Theorem 8.76.
$\|\cdot\|_{Y\leftarrow X}$	Operator norm; see (6.3).
$\|\cdot\|_{i\leftarrow j}$	Norm of operators defined on grid functions; see §9.3.1.
$\langle\cdot,\cdot\rangle$	Euclidean scalar product in $\mathbb{R}^n$; see page 17.
$\langle\cdot,\cdot\rangle_{X\times X'}$	Dual form; see §6.3.1.
$(\cdot,\cdot)_X$	Scalar product of a Hilbert space X; see §6.1.4.
$(\cdot,\cdot)_0$	Scalar product of $L^2(\Omega)$; see (6.7).
$[a,b]$, (a,b), $[a,b)$	Closed, open and half-open interval.
$[\cdot]_E$	Difference of the right- and left-sided limits; see §8.7.1.3.
$\widehat{\dots}$	Fourier transform of $\dots$; identical to $\mathcal{F}(\dots)$.
$A>B$, $A\geq B,\dots$	Elementwise inequalities of matrices; see §4.3.
$\subset$, $\supset$	These signs include the case of equal sets.
$\subsetneq$, $\supsetneq$	Strict inclusions.
$\subset\subset$	Compact inclusion; see (6.9).
$T',\dots$	Dual map corresponding to T; see §6.3.
$X',\dots$	Dual space corresponding to X; see §6.3.
$\int\dots\mathrm{d}\Gamma$	Surface integral

Greek Letters

γ	Trace of a function, e.g., the restriction to the boundary; see (6.24).
γ	Internal boundary in §10.1.1.
$\gamma(\cdot,\cdot)$	Fundamental solution; see (2.10).
Γ	$\Gamma=\partial\Omega$ is the boundary of the domain Ω; see page 14.
Γ_h	Set of grid points on the boundary; see (4.8b).
$\Gamma(\cdot)$	Gamma function.
$\delta(\cdot)$	Dirac distribution; see page 16.
δ_{ij}	Kronecker symbol.
Δ	Laplace operator; see (2.1a).
Δ_h	Five-point formula (4.10).
η_T	Residual corresponding to the triangle T; see (8.85).
$\boldsymbol{\xi}$	Argument $\boldsymbol{\xi}\in\mathbb{R}^n$ of a Fourier-transformed function; see (6.18).
$\rho(\cdot)$	Spectral radius of a matrix; see (4.27).
$\sigma(\cdot)$	Spectrum of a matrix; see page 52.
σ_h^x, σ_h^y	Averaging operators in (9.38).
φ	Often boundary values; see (2.1b).
φ_h	Boundary-value part of the right-hand side of the discrete system; see (4.57).
ω_n	Surface measure of the n-dimensional unit sphere; see (2.4b).
ω_E	Set of triangles neighboured to the edge E; see §8.7.1.3.
ω_T	Set of triangles neighboured to T; see §8.7.1.3.
$\omega(\lambda)$, $\omega_h(\lambda)$	Quantities in (11.5b,c).

$\omega^*(\lambda)$, $\omega_h^*(\lambda)$	Analogous quantities of the adjoint problem.
Ω	Domain of the boundary-value problem; see (2.1a).
Ω_h, $\overline{\Omega_h}$	Set of grid points in Ω and $\overline{\Omega}$, respectively; see (4.8a), (4.8c).

Latin Letters

$a(\cdot,\cdot)$	Bilinear or sesquilinear form; see §6.5.
$a_\lambda(\cdot,\cdot)$	$a_\lambda(u,v) = a(u,v) - \lambda(u,v)_0$; see (11.5a).
$a_N(\cdot,\cdot)$	Restriction of $a(\cdot,\cdot)$ to $V_N \times V_N$; see (8.10).
A, $A(x)$	Matrix or matrix function of the coefficients of the principal part; see (5.2).
A	Operator associated to $a(\cdot,\cdot)$; see Lemma 6.91.
$A, B, \ldots$	Matrices.
B, $B_j, \ldots$	Differential operators defined on the boundary; see (5.21b).
b_i	Basis functions of the subspace V_N for $i = 1, \ldots, N$; see (8.5).
$c(\cdot,\cdot)$	Bilinear form in the case of saddle-point problems; see §8.9.1.
$C(D)$, $C^0(D)$	Set of continuous functions on D; see page 14.
$C^k(D)$	Set of k-times continuously differentiable functions on D for $k \in \mathbb{N}_0 \cup \{\infty\}$; see page 14.
$C^{k+\lambda}(D)$, $C^{k,1}(D)$	Set of Hölder- or Lipschitz-continuously differentiable functions; see Definition 3.14.
$C_0^\infty(D)$	Set of C^∞ functions with compact support; see (6.9).
C^t, $C^{k,1}$	Set of domains with correspondingly smooth boundary; see Definition 6.51.
C, const	Existing but not specified constant in estimates.
C_E	Positive ellipticity constant in (6.44) and (6.48).
C_I	Constant of the inverse inequality; see (8.92).
C_K	Constant of the coercivity inequality (6.48).
C_S	Upper bound of the bilinear form; see (8.2).
cond, cond_2	Condition (spectral condition) of a matrix; see Remark 5.45, Theorem 8.83.
$\cosh(\cdots)$	Hyperbolic cosine, $\cosh(x) = (\exp(x) + \exp(-x))/2$.
$\mathrm{dist}(\cdot,\cdot)$	Distance in $\mathbb{R}^n$ with respect to the Euclidean norm.
$d(u, V_N)$	Distance of the function u from the subspace V_N; see (8.21).
D^α	Differential operator of order α; see (3.11b).
D_h	Difference operator; see Remark 4.7.
e, e^h	Eigenfunction or eigenvector in §11.
e^*, e^{*h}	Eigenfunction or eigenvector of the adjoint problem in §11.
E_0, E_2	Extension operator; see §9.3.2.
$E(\lambda)$	Eigenspace; see Definition 11.1.
$E^*(\lambda)$	Eigenspace of the adjoint problem; see Definition 11.1.
$E_h(\lambda)$, $E_h^*(\lambda)$	Eigenspaces of the discrete problems; see §11.2.1.
$\mathcal{E}$	Set of all edges of the triangulation $\mathcal{T}$; see §8.7.1.3

$\mathcal{E}(T)$	Set of edges of the triangle T; see §8.7.1.3
f	Often the right-hand side of the differential equation; see (3.1a).
f_h	Often the right-hand side of the difference equation.
$\mathcal{F}$, $\mathcal{F}^{-1}$	Fourier transform and Fourier back-transform; see §6.2.3.
$\mathcal{F}_n$	n-dimensional Fourier transform.
$\mathbf{f}$	Right-hand side $\mathbf{f} \in \mathbb{R}^N$ of the discrete Galerkin system (8.9).
$g(\cdot,\cdot)$	Green's function; see Definition 3.6.
$g_h(\cdot,\cdot)$	Discrete Green's function; see (4.40b).
$\underline{G(A)}$	Graph of a matrix A; see Definition 4.9.
$\overline{G(A)}$	Transitive extension of $G(A)$; see page 51.
h	Step size of a difference method; see page 4.1. Maximal element size of a finite-element method; see §8.5.3.
H	Set of grid widths with 0 as accumulation point; see Definition 4.46 and page 332.
$H^k(\Omega)$, $H_0^k(\Omega)$, $H^s(\Omega),\ldots$	Sobolev spaces; see §6.2.
$H(\mathrm{div})$	Hilbert space defined in §8.9.1.
i	Imaginary unit.
I	Unit matrix or identity map.
I	Index set, e.g., in $\mathbb{R}^I$; see Notation 4.4.
$J(\cdot)$	Quadratic functional to be minimised ; see (7.13).
J_X	Riesz isomorphism in $L(X,X')$; see Conclusion 6.69.
k	Often the order of differentiation.
$\mathbb{K}$	Either the field $\mathbb{R}$ or $\mathbb{C}$.
$K_R(x)$	Open ball $\{y \in X : \|y-x\| < R\}$.
$\ker(A)$	Kernel $\{x \in X : Ax = 0\}$ of a linear map $A : X \to Y$.
L	Differential operator.
L_N	Operator associated to a bilinear form a_N; see §8.2.2.6.
$\mathbf{L}$	Matrix of the discrete Galerkin system of equations (8.9).
$L(X,Y)$	Set of linear and continuous maps from X to Y; see page 121.
$L^2(\Omega)$	Set of quadratically integrable functions on Ω; see §6.2.1.
$L^\infty(\Omega)$	Set of bounded functions on Ω; see §6.1.3.
$\underline{\lim}$	Limes inferior; smallest accumulation point.
$\limsup$, $\overline{\lim}$	Limes superior; largest accumulation point.
$\log$	Natural logarithm.
$\mathbf{M}$	Mass matrix; see (8.91).
n	Often dimension of the space $\mathbb{R}^n$ containing Ω.
n	Number of grid points per direction in §4.2.
$\mathbf{n}$	Normal direction; see page 16
$\mathbb{N}$	Set of natural numbers $\{1,2,\ldots.\}$
$\mathbb{N}_0$	$\mathbb{N} \cup \{0\} = \{0,1,2,\ldots.\}$
$\mathcal{N}$	Set of all corner points of the triangulation $\mathcal{T}$; see §8.7.1.3
$\mathcal{N}(T)$	Set of corner points of the triangle T; see §8.7.1.3
O	Zero matrix.
$\mathcal{O}(\cdot)$	Landau symbol; $f(x) = \mathcal{O}(g(x))$ holds with respect to the limit process $x \to x_0$ if $\limsup_{x\to x_0} \lvert f(x)/g(x)\rvert < \infty$.

$o(\cdot)$	Landau symbol; $f(x)=o(g(x))$ holds with respect to the limit process $x \to x_0$ if $\lim_{x\to x_0} \lvert f(x)/g(x)\rvert = 0$.
P	Isomorphism from $\mathbb{R}^N$ to V_N; see (8.6).
$P_h,\ \hat{P}_h$	Prolongation operators for grid functions; see (9.53a,b).
q_h	Grid function; often the right-hand side of the discrete system; see (4.13a,b).
$\mathbb{Q}$	Set of rational numbers.
Q_h	Regular grid of step size h in $\mathbb{R}^n$; see §9.3.1.
Q_N	Orthogonal projection onto the subspace V_N; see §8.2.2.5.
$\mathbb{R}$	Set of real numbers.
$\mathbb{R}^I$	Vector space of the vectors $(x_i)_{i\in I}$ with $x_i \in \mathbb{R}$; see Notation 4.4.
$\mathbb{R}^n_+$	Half space; see (6.23).
R_h	Restriction $u \to u_h$ of a function u onto the grid $\overline{\Omega}_h$ or Ω_h; see (4.46) and (9.39a).
$\tilde{R}_h$	Averaging a function f onto the grid Ω_h; see (4.47).
$\check{R}_h$	Restriction $f \to f_h$ of a function f onto the grid Ω_h; see (9.39b).
$\mathrm{range}(A)$	The set $\{Ax : x \in X\}$ of a map $A : X \to Y$.
$S_N,\ S_h$	Ritz projection onto V_N or V_h; see §8.3.3.
$s(\mathbf{x},\mathbf{y})$	Singularity function; see §2.2.
$\mathrm{span}\{\ldots\}$	Subspace spanned by $\{\ldots\}$.
$\sinh(\cdots)$	Hyperbolic sine, $\sinh(x) = (\exp(x) - \exp(-x))/2$.
$\mathrm{supp}(\cdot)$	Support of a function; see (6.9).
T	Reference triangle; see Figure 8.4 on page 205.
$\mathcal{T}$	Triangulation; see Definition 8.36.
T_i	Finite elements in $\mathcal{T}$; see Definition 8.36.
trace	Trace of a matrix; see Exercise 5.6.
u	Often solution of the boundary-value problem.
$\mathbf{u}$	Solution $\mathbf{u} \in \mathbb{R}^N$ of the discrete Galerkin system (8.9).
u^N	Ritz–Galerkin solution in V_N; see (8.4).
u_h	Grid function; often solution of the discrete system; see (4.6a).
U, V	Spaces of the Gelfand triple (see (6.36)); often $U = L^2(\Omega)$ and $V = H^1(\Omega)$ or $H^1_0(\Omega)$.
U_N	Galerkin subspace V_N endowed with the norm of U; see §8.2.2.1.
V_h	Finite-element subspace with grid size h; see (8.45).
V_N	Galerkin subspace of dimension N; see (8.3).
$(x,y),\ (x,y,z)$	Independent variables of functions in Ω for $n = 2$ and $n = 3$.
$\mathbf{x}=(x_1,\ldots,x_n)$	Independent variables of functions in Ω for general n.
X, Y, Z	Banach spaces; see §6.1.1.
$X', \ldots$	Dual space.
$\mathbb{Z}$	Set of integers.
$Z(\lambda),\ Z_h(\lambda)$	Solution operators in (11.8a,b).

Abbreviations

AFEM	Adaptive finite-element method.
BEM	Boundary-element method; see page 3.6.
DGFEM	Discontinuous Galerkin method; see §8.9.10.2.
FEM	Finite-element method; §8.4.
hp-FEM	Finite-element method with elements of variable grid size and variable polynomial degree; see §8.7.3.5.
SUPG	streamline upwind Petrov–Galerkin; see footnote on page 323.

Index

W. Hackbusch, *Elliptic Differential Equations*, Springer Series in Computational Mathematics 18, DOI 10.1007/978-3-662-54961-2