Volker John

Numerical Methods for Partial Differential Equations

Chapter 1
Some Partial Differential Equations From Physics

Remark 1.1. Contents. This chapter introduces some partial differential equations (pde's) from physics to show the importance of this kind of equations and to motivate the application of numerical methods for their solution. □

1.1 The Heat Equation

Remark 1.2. Derivation. The derivation follows (Wladimirow, 1972, p. 39). Let $\boldsymbol{x} = (x_1, x_2, x_3)^T \in \Omega \subset \mathbb{R}^3$, where Ω is a domain, $t \in \mathbb{R}$, and consider the following physical quantities

- $u(t, \boldsymbol{x})$ – temperature at time t and at the point $\boldsymbol{x}$ with unit [K],
- $\rho(t, \boldsymbol{x})$ – density of the considered species with unit $[^{\mathrm{kg}}/_{\mathrm{m}^3}]$,
- $c(t, \boldsymbol{x})$ – specific heat capacity of the species with unit $[^{\mathrm{J}}/_{\mathrm{kg\ K}}] = [^{\mathrm{W\ s}}/_{\mathrm{kg\ K}}]$,
- $k(t, \boldsymbol{x})$ – thermal conductivity of the species with unit $[^{\mathrm{W}}/_{\mathrm{m\ K}}]$,
- $F(t, \boldsymbol{x})$ – intensity of heat sources or sinks with unit $[^{\mathrm{W}}/_{\mathrm{m}^3}]$.

Consider the heat equilibrium in an arbitrary volume $V \subset \Omega$ and in an arbitrary time interval $(t, t + \Delta t)$. First, there are sources or sinks of heat: heat can enter or leave V through the boundary ∂V, or heat can be produced or absorbed in V. Let $\boldsymbol{n}(\boldsymbol{x})$ be the unit outer normal at $\boldsymbol{x} \in \partial V$. Due to Fourier's[1] law, one finds that the heat

$$Q_1 = \int_t^{t+\Delta t} \int_{\partial V} k \frac{\partial u}{\partial \boldsymbol{n}}(t, \boldsymbol{s}) \ d\boldsymbol{s} \ dt = \int_t^{t+\Delta t} \int_{\partial V} (k \nabla u \cdot \boldsymbol{n})(t, \boldsymbol{s}) \ d\boldsymbol{s} \ dt, \ \text{[J]},$$

enters through ∂V into V. One obtains with integration by parts (Gaussian theorem)

$$Q_1 = \int_t^{t+\Delta t} \int_V \nabla \cdot (k \nabla u)(t, \boldsymbol{x}) \ d\boldsymbol{x} \ dt, \ \text{[J]}.$$

[1] Jean Baptiste Joseph Fourier (1768 – 1830)

In addition, the heat

$$Q_2 = \int_t^{t+\mathit{\Delta} t} \int_V F(t, \boldsymbol{x}) \ d\boldsymbol{x} \ dt, \ [\mathrm{W\ s}] = [\mathrm{J}],$$

is produced in V.

Second, a law for the change of the temperature in V has to be derived. Using a Taylor series expansion, on gets that the temperature at $\boldsymbol{x}$ changes in $(t, t + \mathit{\Delta} t)$ by

$$u(t + \mathit{\Delta} t, \boldsymbol{x}) - u(t, \boldsymbol{x}) = \frac{\partial u}{\partial t}(t, \boldsymbol{x})\mathit{\Delta} t + \mathcal{O}((\mathit{\Delta} t)^2).$$

Now, a linear ansatz is utilized, i.e.,

$$u(t + \mathit{\Delta} t, \boldsymbol{x}) - u(t, \boldsymbol{x}) = \frac{\partial u}{\partial t}(t, \boldsymbol{x})\mathit{\Delta} t.$$

With this ansatz, one has that for the change of the temperature in V and for arbitrary sufficiently small $\mathit{\Delta} t$, the heat

$$\begin{aligned} Q_3 &= \int_t^{t+\mathit{\Delta} t} \int_V c\rho \frac{u(t + \mathit{\Delta} t, \boldsymbol{x}) - u(t, \boldsymbol{x})}{\mathit{\Delta} t} \ d\boldsymbol{x} \ dt \\ &= \int_t^{t+\mathit{\Delta} t} \int_V c\rho \frac{\partial u}{\partial t}(t, \boldsymbol{x}) \ d\boldsymbol{x} \ dt, \ [\mathrm{J}], \end{aligned}$$

is needed. This heat has to be equal to the heat sources, i.e., it holds $Q_3 = Q_2 + Q_1$, from what follows that

$$\int_t^{t+\mathit{\Delta} t} \int_V \left[c\rho \frac{\partial u}{\partial t} - \nabla \cdot (k \nabla u) - F \right] (t, \boldsymbol{x}) \ d\boldsymbol{x} \ dt = 0.$$

Since the volume V was chosen to be arbitrary and $\mathit{\Delta} t$ was arbitrary as well, the term in the integral has to vanish. One obtains the so-called heat equation

$$c\rho \frac{\partial u}{\partial t} - \nabla \cdot (k \nabla u) = F \ \text{ in } (0, T) \times \mathit{\Omega}.$$

At this point of modeling one should check if the equation is dimensionally correct. One finds that all terms have the unit $[\mathrm{W}/\mathrm{m}^3]$.

For a homogeneous species, c, ρ, and k are positive constants. Then, the heat equation simplifies to

$$\frac{\partial u}{\partial t} - \varepsilon^2 \Delta u = f \ \text{ in } (0, T) \times \mathit{\Omega}, \tag{1.1}$$

with $\varepsilon^2 = k/(c\rho)$, $[\mathrm{m}^2/\mathrm{s}]$ and $f = F/(c\rho)$, $[\mathrm{K}/\mathrm{s}]$. To obtain a well-posed problem, (1.1) has to be equipped with an initial condition $u(0, \boldsymbol{x})$ and appropriate boundary conditions on $(0, T) \times \partial \mathit{\Omega}$. □

Remark 1.3. Boundary conditions. For the theory and the numerical simulation of partial differential equations, the choice of boundary conditions is of utmost importance. For the heat equation (1.1), one can prescribe the following types of boundary conditions:

- Dirichlet[2] condition: The temperature $u(t, \boldsymbol{x})$ at a part of the boundary is prescribed
$$u = g_1 \text{ on } (0, T) \times \partial\Omega_D$$
with $\partial\Omega_D \subset \partial\Omega$. In the context of the heat equation, the Dirichlet condition is also called essential boundary conditions.
- Neumann[3] condition: The heat flux is prescribed at a part of the boundary
$$-k\frac{\partial u}{\partial \boldsymbol{n}} = g_2 \text{ on } (0, T) \times \partial\Omega_N$$
with $\partial\Omega_N \subset \partial\Omega$. This boundary condition is a so-called natural boundary condition for the heat equation.
- Mixed boundary condition, Robin[4] boundary condition: At the boundary, there is a heat exchange according to Newton's[5] law
$$k\frac{\partial u}{\partial \boldsymbol{n}} + h(u - u_{\text{env}}) = 0 \text{ on } (0, T) \times \partial\Omega_m,$$
with $\partial\Omega_m \subset \partial\Omega$, the heat exchange coefficient h, $[\mathrm{W}/\mathrm{m}^2\mathrm{K}^2]$, and the temperature of the environment u_{env}.

□

Remark 1.4. The stationary case. An important special case is that the temperature is constant in time $u(t, \boldsymbol{x}) = u(\boldsymbol{x})$. Then, one obtains the stationary heat equation
$$-\varepsilon^2 \Delta u = f \quad \text{in } \Omega. \tag{1.2}$$
This equation is called Poisson[6] equation. Its homogeneous form, i.e., with $f(\boldsymbol{x}) = 0$, is called Laplace[7] equation. Solution of the Laplace equation are called harmonic functions. The Poisson equation is the simplest partial differential equation. The most part of this lecture will consider numerical methods for solving this equation. □

Remark 1.5. Another application of the Poisson equation. The stationary distribution of an electric field with charge distribution $f(\boldsymbol{x})$ satisfies also the Poisson equation (1.2). □

[2] Johann Peter Gustav Lejeune Dirichlet (1805 –1859)

[3] Carl Gottfried Neumann (1832 – 1925)

[4] Gustave Robin (1855 – 1897)

[5] Isaac Newton (1642 – 1727)

[6] Siméon Denis Poisson (1781 – 1840)

[7] Pierre Simon Laplace (1749 – 1829)

Remark 1.6. Non-dimensional equations. The mathematical analysis as well as the application of numerical methods relies on equations for functions without physical units, the so-called non-dimensional equations. Let

- L – a characteristic length scale of the problem, [m],
- U – a characteristic temperature scale of the problem, [K],
- T^* – a characteristic time scale of the problem, [s].

If the new coordinates and functions are denoted with a prime, one gets with the transformations

$$\boldsymbol{x}' = \frac{\boldsymbol{x}}{L}, \quad u' = \frac{u}{U}, \quad t' = \frac{t}{T^*}$$

from (1.1) the non-dimensional equation

$$\frac{\partial}{\partial t'}\left(Uu'\right)\frac{\partial t'}{\partial t} - \varepsilon^2 \sum_{i=1}^{3} \frac{\partial}{\partial x_i'}\left(\frac{\partial}{\partial x_i'}\left(Uu'\right)\frac{\partial x_i'}{\partial x_i}\right)\frac{\partial x_i'}{\partial x_i} = f \quad \text{in } \left(0, \frac{T}{T^*}\right) \times \Omega'$$
$$\Longleftrightarrow$$
$$\frac{U}{T^*}\frac{\partial u'}{\partial t'} - \frac{\varepsilon^2 U}{L^2}\sum_{i=1}^{3}\frac{\partial^2 u'}{\partial\left(x_i'\right)^2} = f \quad \text{in } \left(0, \frac{T}{T^*}\right) \times \Omega'.$$

Usually, one denotes the non-dimensional functions like the dimensional functions, leading to

$$\frac{\partial u}{\partial t} - \frac{\varepsilon^2 T^*}{L^2}\Delta u = \frac{T^*}{U} f \quad \text{in } \left(0, \frac{T}{T^*}\right) \times \Omega.$$

For the analysis, one sets $L = 1$ m, $U = 1$ K, and $T^* = 1$ s which yields

$$\frac{\partial u}{\partial t} - \varepsilon^2 \Delta u = f \ \text{ in } (0, T) \times \Omega, \tag{1.3}$$

with a non-dimensional temperature diffusion ε^2 and a non-dimensional right-hand side $f(t, \boldsymbol{x})$.

The same approach can be applied to the stationary equation (1.2) and one gets

$$-\varepsilon^2 \Delta u = f \ \text{ in } \Omega, \tag{1.4}$$

with the non-dimensional temperature diffusion ε^2 and the non-dimensional right-hand side $f(\boldsymbol{x})$. □

Remark 1.7. A standard approach for solving the instationary equation. The heat equation (1.3) is an initial value problem with respect to time and a boundary value problem with respect to space. Numerical methods for solving initial value problems were topic of Numerical Mathematics 2.

A standard approach for solving the instationary problem consists in using a so-called one-step θ-scheme for discretizing the temporal derivative. Consider two consecutive discrete times t_n and t_{n+1} with $\tau = t_{n+1} - t_n$. Then,

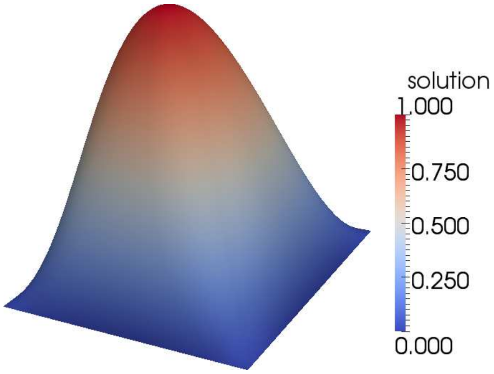

Fig. 1.1 Solution of the two-dimensional example of Example 1.8.

the application of a one-step θ-scheme yields for the solution at t_{n+1}

$$\frac{u_{n+1}-u_n}{\tau} - \theta\varepsilon^2 \Delta u_{n+1} - (1-\theta)\varepsilon^2\Delta u_n = \theta f_{n+1} + (1-\theta) f_n,$$

where the subscript at the functions denotes the time level. This equation is equivalent to

$$u_{n+1} - \tau\theta\varepsilon^2\Delta u_{n+1} = u_n + \tau(1-\theta)\varepsilon^2\Delta u_n + \tau\theta f_{n+1} + \tau(1-\theta)f_n. \quad (1.5)$$

For $\theta = 0$, one obtains the forward Euler scheme, for $\theta = 0.5$ the Crank–Nicolson scheme (trapezoidal rule), and for $\theta = 1$ the backward Euler scheme.

Given u_n, (1.5) is a boundary value problem for u_{n+1}. That means, one has to solve in each discrete time a boundary value problem. For this reason, this lecture will concentrate on the numerical solution of boundary value problems. □

Example 1.8. Demonstrations with the code MooNMD John & Matthies (2004).

- Consider the Poisson equation (1.4) in $\Omega = (0,1)^2$ with $\varepsilon = 1$. The right-hand side and the Dirichlet boundary conditions are chosen such that $u(x,y) = \sin(\pi x)\sin(\pi y)$ is the prescribed solution, see Figure 1.1 Hence, this solution satisfies homogeneous Dirichlet boundary conditions. Denote by $u_h(x,y)$ the computed solution, where h indicates the refinement of a mesh in Ω. The errors obtained on successively refined meshes with the simplest finite element method are presented in Table 1.1.
  One can observe in Table 1.1 that $\|u-u_h\|_{L^2(\Omega)}$ converges with second order and $\|\nabla(u-u_h)\|_{L^2(\Omega)}$ converges with first order. A main topic of the numerical analysis of discretizations for partial differential equations consists in showing that the computed solution converges to the solution

Table 1.1 Example 1.8, two-dimensional example.

h	degrees of freedom	$\|u-u_h\|_{L^2(\Omega)}$	$\|\nabla(u-u_h)\|_{L^2(\Omega)}$
1/4	25	8.522e-2	8.391e-1
1/8	81	2.256e-2	4.318e-1
1/16	289	5.726e-3	2.175e-1
1/32	1089	1.437e-3	1.089e-1
1/64	4225	3.596e-4	5.451e-2
1/128	16641	8.993e-5	2.726e-2
1/256	66049	2.248e-5	1.363e-2
1/512	263169	5.621e-6	6.815e-3

of an appropriate continuous problem in appropriate norms. In addition, to prove a certain order of convergence (in the asymptotic regime) is of interest.

- Consider the Poisson equation (1.4) in $\Omega = (0,1)^3$ with $\varepsilon = 1$ and $f = 0$. At $z = 1$ the temperature profile should be $u(x,y,1) = 16x(1-x)y(1-y)$ and at the opposite wall should be cooled $u(x,y,0) = 0$. At all other walls, there should be an undisturbed temperature flux $\frac{\partial u}{\partial \boldsymbol{n}}(x,y,z) = 0$. A approximation of the solution computed with a finite element method is presented in Figure 1.2.

  The analytical solution is not known in this example (or it maybe hard to compute). It is important for applications that one obtains, e.g., good visualizations of the solution or approximate values for quantities of interest. One knows by the general theory that the computed solution converges to the solution of the continuous problem in appropriate norms and one hopes that the computed solution is already sufficiently close.

□

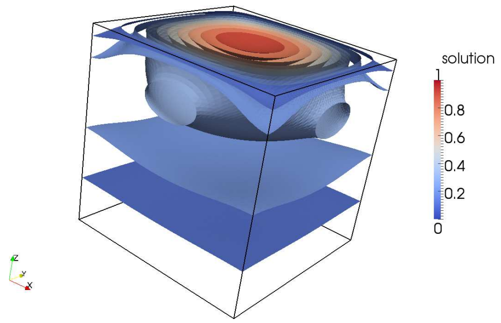

Fig. 1.2 Contour lines of the solution of the three-dimensional example of Example 1.8.

1.2 The Diffusion Equation

Remark 1.9. Derivation. Diffusion is the transport of a species caused by the movement of particles. Instead of Fourier's law, Newton's law for the particle flux through ∂V per time unit is used

$$dQ = -D\nabla u \cdot \boldsymbol{n}\ d\boldsymbol{s}$$

with

- $u(t,\boldsymbol{x})$ – particle density, concentration with unit $[\mathrm{mol}/\mathrm{m}^3]$,
- $D(t,\boldsymbol{x})$ – diffusion coefficient with unit $[\mathrm{m}^2/\mathrm{s}]$.

The derivation of the diffusion equation proceeds in the same way as for the heat equation. It has the form

$$c\frac{\partial u}{\partial t} - \nabla\cdot(D\nabla u) + qu = F \quad \text{in } (0,T)\times\Omega, \tag{1.6}$$

where

- $c(t,\boldsymbol{x})$ – is the porosity of the species, $[\cdot]$,
- $q(t,\boldsymbol{x})$ – is the absorption coefficient of the species with unit $[1/\mathrm{s}]$,
- $F(t,\boldsymbol{x})$ – describes sources and sinks, $[\mathrm{mol}/\mathrm{s\ m}^3]$.

The porosity and the absorption coefficient are positive functions. To obtain a well posed problem, an initial condition and boundary conditions are necessary.

If the concentration is constant in time, one obtains

$$-\nabla\cdot(D\nabla u) + qu = F \quad \text{in } \Omega. \tag{1.7}$$

Hence, the diffusion equation possesses a similar form as the heat equation. □

1.3 The Navier–Stokes Equations

This section was not presented in the course. It is included in the lecture notes for students who are interested in.

Remark 1.10. Generalities. The Navier[8]–Stokes[9] equations are the fundamental equations of fluid dynamics. In this section, a viscous fluid (with internal friction) with constant density (incompressible) will be considered. □

Remark 1.11. Conservation of mass. The first basic principle of the flow of an incompressible fluid is the conservation of mass. Let V be an arbitrary

[8] Claude Louis Marie Henri Navier (1785 – 1836)

[9] George Gabriel Stokes (1819 – 1903)

volume. Then, the change of fluid in V satisfies

$$\underbrace{-\frac{\partial}{\partial t}\int\limits_V \rho\ d\boldsymbol{x}}_{\text{change}} = \underbrace{\int\limits_{\partial V} \rho \boldsymbol{v}\cdot\boldsymbol{n}\ d\boldsymbol{s}}_{\text{flux through the boundary of } V} = \int\limits_V \nabla\cdot(\rho\boldsymbol{v})\ d\boldsymbol{x},$$

where

- $\boldsymbol{v}(t,\boldsymbol{x})$ – velocity $(v_1,v_2,v_3)^T$ at time t and at point $\boldsymbol{x}$ with unit $[{}^{\mathrm{m}}/_{\mathrm{s}}]$,
- ρ – density of the fluid, $[{}^{\mathrm{kg}}/_{\mathrm{m}^3}]$.

Since V is arbitrary, the terms in the volume integrals have to be the same. One gets the so-called continuity equation

$$\frac{\partial\rho}{\partial t} + \nabla\cdot(\rho\boldsymbol{v}) = 0 \ \text{ in } (0,T)\times\Omega.$$

Since ρ is constant, one obtains the first equation of the Navier–Stokes equation, the so-called incompressibility constraint,

$$\nabla\cdot\boldsymbol{v} = 0 \ \text{ in } (0,T)\times\Omega. \tag{1.8}$$

□

Remark 1.12. Conservation of linear momentum. The second equation of the Navier–Stokes equations represents Newton's second law of motion

$$\text{net force} = \text{mass } \times \text{ acceleration.}$$

It states that the rate of change of the linear momentum must be equal to the net force acting on a collection of fluid particles.

The forces acting on an arbitrary volume V are given by

$$F_V = \underbrace{\int\limits_{\partial V} -P\boldsymbol{n}\ d\boldsymbol{s}}_{\text{outer pressure}} + \underbrace{\int\limits_{\partial V} \mathbb{S}'\boldsymbol{n}\ d\boldsymbol{s}}_{\text{friction}} + \underbrace{\int\limits_V \rho\boldsymbol{g}\ d\boldsymbol{x}}_{\text{gravitation}},$$

where

- $S'(t,\boldsymbol{x})$ – stress tensor with unit $[{}^{\mathrm{N}}/_{\mathrm{m}^2}]$,
- $P(t,\boldsymbol{x})$ – the pressure with unit $[{}^{\mathrm{N}}/_{\mathrm{m}^2}]$,
- $\boldsymbol{g}(t,\boldsymbol{x})$ – standard gravity (directed), $[{}^{\mathrm{m}}/_{\mathrm{s}^2}]$.

The pressure possesses a negative sign since it is directed into V, whereas the stress acts outwardly.

The integral on ∂V can be transformed into an integral on V with integration by parts. One obtains the force per unit volume

$$-\nabla P + \nabla\cdot\mathbb{S}' + \rho\boldsymbol{g}.$$

On the basis of physical considerations (Landau & Lifschitz, 1966, p. 53) or (John, 2016, Chapter 2), one uses the following ansatz for the stress tensor

$$\mathbb{S}' = \eta\Big(\nabla \boldsymbol{v} + \nabla \boldsymbol{v}^T - \frac{2}{3}(\nabla \cdot \boldsymbol{v})\mathbb{I}\Big) + \zeta(\nabla \cdot \boldsymbol{v})\mathbb{I},$$

where

- η – first order viscosity of the fluid, [$^{\mathrm{kg}}/_{\mathrm{m\ s}}$],
- ζ – second order viscosity of the fluid, [$^{\mathrm{kg}}/_{\mathrm{m\ s}}$],
- $\mathbb{I}$ – unit tensor.

For Newton's second law of motion one considers the movement of particles with velocity $\boldsymbol{v}(t, \boldsymbol{x}(t))$. One obtains the following equation

$$\begin{aligned}\underbrace{-\nabla P + \nabla \cdot \mathbb{S}' + \rho \boldsymbol{g}}_{\text{force per unit volume}} &= \underbrace{\rho}_{\text{mass per unit volume}} \underbrace{\frac{d\boldsymbol{v}(t, \boldsymbol{x}(t))}{dt}}_{\text{acceleration}} \\ &= \rho\left(\partial_t \boldsymbol{v} + (\boldsymbol{v} \cdot \nabla)\boldsymbol{v}\right).\end{aligned}$$

The second formula was obtained with the chain rule. The detailed form of the second term is

$$(\boldsymbol{v} \cdot \nabla)\boldsymbol{v} = \begin{pmatrix} v_1\partial_x v_1 + v_2\partial_y v_1 + v_3\partial_z v_1 \\ v_1\partial_x v_2 + v_2\partial_y v_2 + v_3\partial_z v_2 \\ v_1\partial_x v_3 + v_2\partial_y v_3 + v_3\partial_z v_3 \end{pmatrix}.$$

If both viscosities are constant, one gets

$$\frac{\partial \boldsymbol{v}}{\partial t} - \nu \Delta \boldsymbol{v} + (\boldsymbol{v} \cdot \nabla)\boldsymbol{v} + \frac{\nabla P}{\rho} = \boldsymbol{g} + \frac{1}{\rho}\left(\frac{\eta}{3} + \zeta\right)\nabla(\nabla \cdot \boldsymbol{v}),$$

where $\nu = \eta/\rho, [m^2/s]$ is the kinematic viscosity. The second term on the right-hand side vanishes because of the incompressibility constraint (1.8).

One obtains the dimensional Navier–Stokes equations

$$\frac{\partial \boldsymbol{v}}{\partial t} - \nu \Delta \boldsymbol{v} + (\boldsymbol{v} \cdot \nabla)\boldsymbol{v} + \frac{\nabla P}{\rho} = \boldsymbol{g}, \quad \nabla \cdot \boldsymbol{v} = 0 \ \text{ in } (0, T) \times \Omega.$$

□

Remark 1.13. Non-dimensional Navier–Stokes equations. The final step in the modeling process is the derivation of non-dimensional equations. Let

- L – a characteristic length scale of the problem, [m],
- U – a characteristic velocity scale of the problem, [$^{\mathrm{m}}/_{\mathrm{s}}$],
- T^* – a characteristic time scale of the problem, [s].

Denoting here the old coordinates with a prime, one obtains with the transformations

$$\boldsymbol{x} = \frac{\boldsymbol{x}'}{L}, \quad \boldsymbol{u} = \frac{\boldsymbol{v}}{U}, \quad t = \frac{t'}{T^*}$$

the non-dimensional equations

$$\frac{L}{UT^*}\partial_t \boldsymbol{u} - \frac{\nu}{UL}\Delta \boldsymbol{u} + (\boldsymbol{u}\cdot\nabla)\boldsymbol{u} + \nabla p = \boldsymbol{f}, \quad \nabla\cdot\boldsymbol{u} = 0 \ \text{ in } (0,T)\times\Omega,$$

with the redefined pressure and the new right-hand side

$$p(t,\boldsymbol{x}) = \frac{P}{\rho U^2}(t,\boldsymbol{x}), \quad \boldsymbol{f}(t,\boldsymbol{x}) = \frac{L\boldsymbol{g}}{U^2}(t,\boldsymbol{x}).$$

This equation has two dimensionless characteristic parameters: the Strouhal[10] number St and the Reynolds [11] number Re

$$St := \frac{L}{UT^*}, \quad Re := \frac{UL}{\nu}.$$

Setting $T^* = L/U$, one obtains the form of the incompressible Navier–Stokes equations which can be found in the literature

$$\begin{aligned} \frac{\partial \boldsymbol{u}}{\partial t} - Re^{-1}\Delta\boldsymbol{u} + (\boldsymbol{u}\cdot\nabla)\boldsymbol{u} + \nabla p &= \boldsymbol{f} \text{ in } (0,T)\times\Omega, \\ \nabla\cdot\boldsymbol{u} &= 0 \ \text{ in } (0,T)\times\Omega. \end{aligned}$$

□

Remark 1.14. About the incompressible Navier–Stokes equations. The Navier–Stokes equations are not yet understood completely. For instance, the existence of an appropriately defined classical solution for $\Omega \subset \mathbb{R}^3$ is not clear. This problem is among the so-called millennium problems of mathematics Fefferman (2000) and its answer is worth one million dollar. Also the numerical methods for solving the Navier–Stokes equations are by far not developed sufficiently well as it is required by many applications, e.g. for turbulent flows in weather prediction. □

Remark 1.15. Slow flows. Am important special case is the case of slow flows which lead to a stationary (independent of time) flow field. In this case, the first term in the in the momentum balance equation vanish. In addition, if the flow is very slow, the nonlinear term can be neglected as well. One gets the so-called Stokes equations

$$\begin{aligned} -Re^{-1}\Delta\boldsymbol{u} + \nabla p &= \boldsymbol{f} \text{ in } \Omega, \\ \nabla\cdot\boldsymbol{u} &= 0 \ \text{ in } \Omega. \end{aligned}$$

□

[10] Čeněk Strouhal (1850 – 1923)

[11] Osborne Reynolds (1842 – 1912)

1.4 Classification of Second Order Partial Differential Equations

Definition 1.16. Quasi-linear and linear second order partial differential equation. Let $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$. A quasi-linear second order partial differential equation defined on Ω has the form

$$\sum_{j,k=1}^{d} a_{jk}(\boldsymbol{x})\partial_j\partial_k u + F\left(\boldsymbol{x}, u, \partial_1 u, \ldots, \partial_d u\right) = 0 \tag{1.9}$$

or in nabla notation

$$\nabla \cdot (A(\boldsymbol{x})\nabla u) + \tilde{F}\left(\boldsymbol{x}, u, \partial_1 u, \ldots, \partial_d u\right) = 0.$$

A linear second order partial differential equation has the form

$$\sum_{j,k=1}^{d} a_{jk}(\boldsymbol{x})\partial_j\partial_k u + \boldsymbol{b}(\boldsymbol{x}) \cdot \nabla u + c(\boldsymbol{x})u = F(\boldsymbol{x}).$$

□

Remark 1.17. The matrix of the second order operator. If $u(\boldsymbol{x})$ is sufficiently regular, then the application of the Theorem of Schwarz[12] yields $\partial_j\partial_k u(\boldsymbol{x}) = \partial_k\partial_j u(\boldsymbol{x})$. It follows that equation (1.9) contains the coefficient $\partial_j\partial_k u(\boldsymbol{x})$ twice, namely in $a_{jk}(\boldsymbol{x})$ and $a_{kj}(\boldsymbol{x})$. For definiteness, one requires that

$$a_{jk}(\boldsymbol{x}) = a_{kj}(\boldsymbol{x}).$$

Now, one can write the coefficient of the second order derivative with the symmetric matrix

$$A(\boldsymbol{x}) = \begin{pmatrix} a_{11}(\boldsymbol{x}) & \cdots & a_{1d}(\boldsymbol{x}) \\ \vdots & \ddots & \vdots \\ a_{d1}(\boldsymbol{x}) & \cdots & a_{dd}(\boldsymbol{x}) \end{pmatrix}.$$

All eigenvalues of this matrix are real and the classification of quasi-linear second order partial differential equations is based on these eigenvalues. □

Definition 1.18. Classification of quasi-linear second order partial differential equation. On a subset $\tilde{\Omega} \subset \Omega$ let α be the number of positive eigenvalues of $A(\boldsymbol{x})$, β be the number of negative eigenvalues, and γ be the multiplicity of the eigenvalue zero. The quasi-linear second order partial differential equation (1.9) is said to be of type (α, β, γ) on $\tilde{\Omega}$. It is called to be

- elliptic on $\tilde{\Omega}$ if it is of type $(d, 0, 0) = (0, d, 0)$,

[12] Hermann Amandus Schwarz (1843 – 1921)

- hyperbolic on $\tilde{\Omega}$, if its type is $(d-1,1,0) = (1,d-1,0)$,
- parabolic on $\tilde{\Omega}$, if it is of type $(d-1,0,1) = (0,d-1,1)$.

In the case of linear partial differential equations, one speaks of a parabolic equation if in addition to the requirement from above it holds that

$$\operatorname{rank}(A(\boldsymbol{x}), \boldsymbol{b}(\boldsymbol{x})) = d$$

in $\tilde{\Omega}$. □

Remark 1.19. Other cases. Definition 1.18 does not cover all possible cases. However, the other cases are only of little interest in practice. □

Example 1.20. Types of second order partial differential equations.

- For the Poisson equation (1.4) one has $a_{ii} = -\varepsilon^2 < 0$ and $a_{ij} = 0$ for $i \neq j$. It follows that all eigenvalues of A are negative and the Poisson equation is an elliptic partial differential equation. The same reasoning can be applied to the stationary diffusion equation (1.7).
- In the heat equation (1.3) there is besides the spatial derivatives also the temporal derivative. The derivative in time has to be taken into account in the definition of the matrix A. Since this derivative is only of first order, one obtains in A a zero row and a zero column. One has, e.g., $a_{ii} = -\varepsilon^2 < 0, i = 2, \ldots, d+1$, $a_{11} = 0$, and $a_{ij} = 0$ for $i \neq j$. It follows that one eigenvalue is zero and the others have the same sign. The vector of the first order term has the form $\boldsymbol{b} = (1, 0, \ldots, 0)^T \in \mathbb{R}^{d+1}$, where the one comes from $\partial_t u(t, \boldsymbol{x})$. Now, one can see immediately that $(A, \boldsymbol{b})$ possesses full column rank. Hence, (1.3) is a parabolic partial differential equation.
- An example for a hyperbolic partial differential equation is the wave equation

$$\partial_{tt} u - \varepsilon^2 \Delta u = f \quad \text{in } (0,T) \times \Omega.$$

□

1.5 Literature

Remark 1.21. Some books about the topic of this class. Books about finite difference methods are

- Samarskij (1984), classic book, the English version is Samarskii (2001)
- LeVeque (2007)

Much more books can be found about finite element methods

- Ciarlet (2002), classic text,
- Strang & Fix (2008), classic text,
- Braess (2001), very popular book in Germany, English version available,

- Brenner & Scott (2008), rather abstract treatment, from the point of view of functional analysis,
- Ern & Guermond (2004), modern comprehensive book,
- Grossmann & Roos (2007)
- Šolín (2006), written by somebody who worked a lot in the implementation of the methods,
- Goering *et al.* (2010), introductory text, good for beginners,
- Deuflhard & Weiser (2012), strong emphasis on adaptive methods
- Dziuk (2010).

These lists are not complete.

These lectures notes are based in some parts on lecture notes from Sergej Rjasanow (Saarbrücken) and Manfred Dobrowolski (Würzburg). □

Chapter 2

Finite Difference Methods for Elliptic Equations

Remark 2.1. Model problem. The model problem in this chapter is the Poisson equation with Dirichlet boundary conditions

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ u &= g \quad \text{on } \partial\Omega, \end{aligned} \tag{2.1}$$

where $\Omega \subset \mathbb{R}^2$. This chapter follows in wide parts Samarskij (1984). □

2.1 Basics on Finite Differences

Remark 2.2. Grid. This section considers the one-dimensional case. Consider the interval $[0, 1]$ that is decomposed by an equidistant grid

$$\begin{aligned} x_i &= ih, \quad i = 0, \ldots, n, \quad h = 1/n, \text{ – nodes,} \\ \omega_h &= \{x_i \ : \ i = 0, \ldots, n\} \text{ – grid.} \end{aligned}$$

□

Definition 2.3. Grid function. A vector $\underline{u}_h = (u_0, \ldots, u_n)^T \in \mathbb{R}^{n+1}$ that assigns every grid point a function value is called grid function. □

Definition 2.4. Finite differences. Let $v(x)$ be a sufficiently smooth function and denote by $v_i = v(x_i)$, where x_i are the nodes of the grid. The following quotients are called

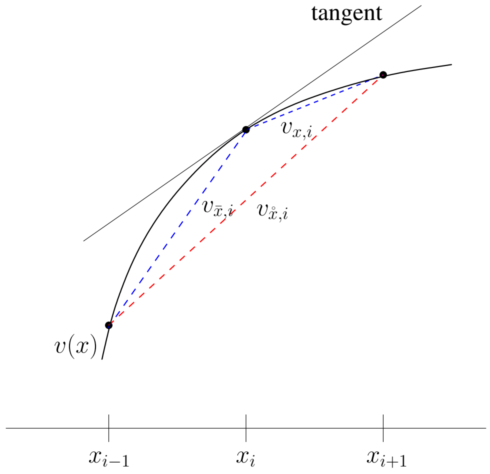

Fig. 2.1 Illustration of the finite differences.

$$
\begin{aligned}
v_{x,i} &= \frac{v_{i+1}-v_i}{h} \text{ – forward difference,}\\
v_{\bar{x},i} &= \frac{v_i-v_{i-1}}{h} \text{ – backward difference,}\\
v_{\mathring{x},i} &= \frac{v_{i+1}-v_{i-1}}{2h} \text{ – central difference,}\\
v_{\bar{x}x,i} &= \frac{v_{i+1}-2v_i+v_{i-1}}{h^2} \text{ – second order difference,}
\end{aligned}
$$

see Figure 2.1. □

Remark 2.5. Some properties of the finite differences. It is (*exercise*)

$$v_{\mathring{x},i} = \frac{1}{2}(v_{x,i}+v_{\bar{x},i}), \quad v_{\bar{x}x,i} = (v_{\bar{x},i})_{x,i}.$$

Using the Taylor series expansion for $v(x)$ at the node x_i, one gets (*exercise*)

$$
\begin{aligned}
v_{x,i} &= v'(x_i) + \frac{1}{2}hv''(x_i) + \mathcal{O}\left(h^2\right),\\
v_{\bar{x},i} &= v'(x_i) - \frac{1}{2}hv''(x_i) + \mathcal{O}\left(h^2\right),\\
v_{\mathring{x},i} &= v'(x_i) + \mathcal{O}\left(h^2\right),\\
v_{\bar{x}x,i} &= v''(x_i) + \mathcal{O}\left(h^2\right).
\end{aligned}
$$

□

Definition 2.6. Consistent difference operator. Let L be a differential operator. The difference operator $L_h \ : \ \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ is called consistent with L of order k if

$$\max_{0\le i\le n} |(Lu)(x_i) - (L_h u_h)_i| = \|Lu - L_h u_h\|_{\infty,\omega_h} = \mathcal{O}\left(h^k\right)$$

for all sufficiently smooth functions $u(x)$. □

Example 2.7. Consistency orders. The order of consistency measures the quality of approximation of L by L_h.

The difference operators $v_{x,i}, v_{\overline{x},i}, v_{\mathring{x},i}$ are consistent to $L = \frac{d}{dx}$ with order $1, 1$, and 2, respectively. The operator $v_{\overline{x}x,i}$ is consistent of second order to $L = \frac{d^2}{dx^2}$, see Remark 2.5. □

Example 2.8. Approximation of a more complicated differential operator by difference operators. Consider the differential operator

$$Lu = \frac{d}{dx}\left(k(x)\frac{du}{dx}\right),$$

where $k(x)$ is assumed to be continuously differentiable. Define the difference operator L_h as follows

$$\begin{aligned}(L_h u_h)_i = (a u_{\overline{x},i})_{x,i} &= \frac{1}{h}\Big(a(x_{i+1})u_{\overline{x},i}(x_{i+1}) - a(x_i)u_{\overline{x},i}(x_i)\Big) \\ &= \frac{1}{h}\left(a_{i+1}\frac{u_{i+1}-u_i}{h} - a_i\frac{u_i - u_{i-1}}{h}\right), \qquad (2.2)\end{aligned}$$

where a is a grid function that has to be determined appropriately. One gets with the product rule

$$(Lu)_i = k'(x_i)(u')_i + k(x_i)(u'')_i$$

and with a Taylor series expansion for u_{i-1}, u_{i+1}, which is inserted in (2.2),

$$(L_h u_h)_i = \frac{a_{i+1}-a_i}{h}(u')_i + \frac{a_{i+1}+a_i}{2}(u'')_i + \frac{h(a_{i+1}-a_i)}{6}(u''')_i + \mathcal{O}\left(h^2\right).$$

Thus, the difference of the differential operator and the difference operator is

$$\begin{aligned}(Lu)_i - (L_h u_h)_i = &\left(k'(x_i) - \frac{a_{i+1}-a_i}{h}\right)(u')_i + \left(k(x_i) - \frac{a_{i+1}+a_i}{2}\right)(u'')_i \\ &-\frac{h(a_{i+1}-a_i)}{6}(u''')_i + \mathcal{O}\left(h^2\right). \qquad (2.3)\end{aligned}$$

In order to define L_h such that it is consistent of second order to L, one has to satisfy the following two conditions

$$\frac{a_{i+1}-a_i}{h} = k'(x_i) + \mathcal{O}\left(h^2\right), \quad \frac{a_{i+1}+a_i}{2} = k(x_i) + \mathcal{O}\left(h^2\right).$$

From the first requirement, it follows that $a_{i+1} - a_i = \mathcal{O}(h)$. Hence, the third term in the consistency error equation (2.3) is of order $\mathcal{O}\left(h^2\right)$. Possible choices for the grid function are (*exercise*)

$$a_i = \frac{k_i + k_{i-1}}{2}, \quad a_i = k\left(x_i - \frac{h}{2}\right), \quad a_i = \left(k_i k_{i-1}\right)^{1/2}.$$

Note that the 'natural' choice, $a_i = k_i$, leads only to first order consistency. (*exercise*) □

2.2 Finite Difference Approximation of the Laplacian in Two Dimensions

Remark 2.9. The five point stencil. The Laplacian in two dimensions is defined by

$$\Delta u(\boldsymbol{x}) = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = \partial_{xx} u + \partial_{yy} u = u_{xx} + u_{yy}, \quad \boldsymbol{x} = (x, y).$$

The simplest approximation uses for both second order derivatives the second order differences. One obtains the so-called five point stencil and the approximation

$$\Delta u \approx \Lambda u = u_{\overline{x}x} + u_{\overline{y}y} = \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{h_x^2} + \frac{u_{i,j+1} - 2u_{ij} + u_{i,j-1}}{h_y^2}, \tag{2.4}$$

see Figure 2.2. From the consistency order of the second order difference, it follows immediately that Λu approximates the Laplacian of order $\mathcal{O}\left(h_x^2 + h_y^2\right)$. □

Remark 2.10. The five point stencil on curvilinear boundaries. There is a difficulty if the five point stencil is used in domains with curvilinear boundaries. The approximation of the second derivative requires three function values in each coordinate direction

$$\begin{gathered}(x - h_x^-, y), (x, y), (x + h_x^+, y),\\ (x, y - h_y^-), (x, y), (x, y + h_y^+),\end{gathered}$$

see Figure 2.3. A guideline of defining the approximation is that the five point stencil is recovered in the case $h_x^- = h_x^+$. A possible approximation of this type is

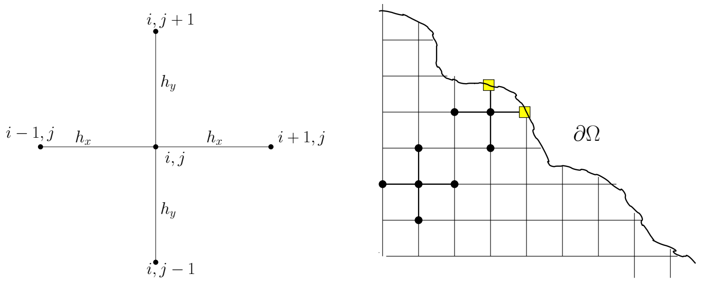

Fig. 2.2 Five point stencils.

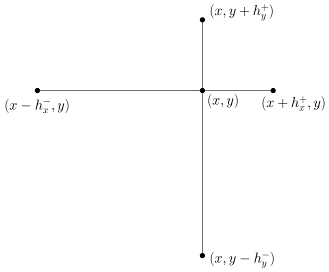

Fig. 2.3 Sketch to Remark 2.10.

$$\frac{\partial^2 u}{\partial x^2} \approx \frac{1}{\overline{h}_x}\left(\frac{u(x+h_x^+,y)-u(x,y)}{h_x^+} - \frac{u(x,y)-u(x-h_x^-,y)}{h_x^-}\right) \tag{2.5}$$

with $\overline{h}_x = (h_x^+ + h_x^-)/2$. Using a Taylor series expansion, one finds that the error of this approximation is

$$\begin{aligned}&\frac{\partial^2 u}{\partial x^2} - \frac{1}{\overline{h}_x}\left(\frac{u(x+h_x^+,y)-u(x,y)}{h_x^+} - \frac{u(x,y)-u(x-h_x^-,y)}{h_x^-}\right)\\ &\qquad = -\frac{1}{3}(h_x^+ - h_x^-)\frac{\partial^3 u}{\partial x^3} + \mathcal{O}\left(\overline{h}_x^2\right).\end{aligned}$$

For $h_x^+ \neq h_x^-$, this approximation is of first order.

A different way consists in using

$$\frac{\partial^2 u}{\partial x^2} \approx \frac{1}{\tilde{h}_x}\left(\frac{u(x+h_x^+,y)-u(x,y)}{h_x^+} - \frac{u(x,y)-u(x-h_x^-,y)}{h_x^-}\right)$$

with $\tilde{h}_x = \max\{h_x^+, h_x^-\}$. However, this approximation possesses only the order zero, i.e., there is actually no approximation.

Altogether, there is a loss of order of consistency at curvilinear boundaries. □

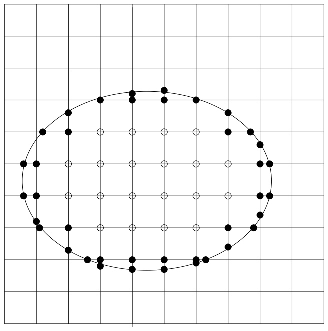

Fig. 2.4 Different types of nodes in the grid.

Example 2.11. The Dirichlet problem. Consider the Poisson equation that is equipped with Dirichlet boundary conditions (2.1). First, $\mathbb{R}^2$ is decomposed by a grid with rectangular mesh cells $x_i = ih_x, y_j = jh_y$, $h_x, h_y > 0$, $i, j \in \mathbb{Z}$. Denote by

$w_h^\circ = \{\circ\}$ inner nodes, five point stencil does not contain any boundary node,
$w_h^* = \{*\}$ inner nodes that are close to the boundary, five point stencil contains boundary nodes,
$\gamma_h = \{*\}$ boundary nodes,
$\omega_h = w_h^\circ \cup w_h^*$ inner nodes,
$\omega_h \cup \gamma_h$ grid,

see Figure 2.4.

The finite difference approximation of problem (2.1) that will be studied in the following consists in finding a mesh function $u(\boldsymbol{x})$ such that

$$\begin{aligned} -\Lambda u(\boldsymbol{x}) &= \phi(\boldsymbol{x}) \ \boldsymbol{x} \in w_h^\circ, \\ -\Lambda^* u(\boldsymbol{x}) &= \phi(\boldsymbol{x}) \ \boldsymbol{x} \in w_h^*, \\ u(\boldsymbol{x}) &= g(\boldsymbol{x}) \ \boldsymbol{x} \in \gamma_h, \end{aligned} \tag{2.6}$$

where $\phi(\boldsymbol{x})$ is a grid function that approximates $f(\boldsymbol{x})$ and Λ^* is an approximation of the Laplacian for nodes that are close to the boundary, e.g., defined by (2.5). The discrete problem is a large sparse linear system of equations. The most important questions are:

- Which properties possesses the solution of (2.6)?
- Converges the solution of (2.6) to the solution of the Poisson problem and if yes, with which order?

□

2.3 The Discrete Maximum Principle for a Finite Difference Approximation

Remark 2.12. Contents of this section. Solutions of the Laplace equation, i.e., of (2.1) with $f(\boldsymbol{x}) = 0$, fulfill so-called maximum principles. This section shows, that the finite difference approximation of an operator, where the five point stencil of the Laplacian is a special case, satisfies a discrete analog of one of the maximum principles. □

Theorem 2.13. Maximum principles for harmonic functions. *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain and $u \in C^2(\Omega) \cap C(\overline{\Omega})$ be harmonic in Ω, i.e., $u(\boldsymbol{x})$ solves the Laplace equation $-\Delta u = 0$ in Ω.*

- *Weak maximum principle. It holds*

$$\max_{\boldsymbol{x} \in \overline{\Omega}} u(\boldsymbol{x}) = \max_{\boldsymbol{x} \in \partial\Omega} u(\boldsymbol{x}).$$

  That means, $u(\boldsymbol{x})$ takes its maximal value at the boundary.
- *Strong maximum principle. If Ω is connected and if the maximum is taken in Ω (note that Ω is open), i.e., $u(\boldsymbol{x}_0) = \max_{\boldsymbol{x} \in \overline{\Omega}} u(\boldsymbol{x})$ for a point $\boldsymbol{x}_0 \in \Omega$, then $u(\boldsymbol{x})$ is constant*

$$u(\boldsymbol{x}) = \max_{\boldsymbol{x} \in \overline{\Omega}} u(\boldsymbol{x}) = u(\boldsymbol{x}_0) \quad \forall\ \boldsymbol{x} \in \overline{\Omega}.$$

Proof. See the literature, e.g., (Evans, 2010, p. 27, Theorem 4) or the course on the theory of partial differential equations. ■

Remark 2.14. Interpretation of the maximum principle.

- The Laplace equation models the temperature distribution of a heated body without heat sources in Ω. Then, the weak maximum principle just states that the temperature in the interior of the body cannot be higher than the highest temperature at the boundary.
- There are maximum principles also for more complicated operators than the Laplacian, e.g., see Evans (2010).
- Since the solution of the partial differential equation will be only approximated by a discretization like a finite difference method, one has to expect that basic physical properties are satisfied by the numerical solution also only approximately. However, in applications, it is often very important that such properties are satisfied exactly.

□

Remark 2.15. The difference equation. In this section, a difference equation of the form

$$a(\boldsymbol{x})u(\boldsymbol{x}) = \sum_{\boldsymbol{y} \in S(\boldsymbol{x})} b(\boldsymbol{x}, \boldsymbol{y})u(\boldsymbol{y}) + F(\boldsymbol{x}),\ \ \boldsymbol{x} \in \omega_h \cup \gamma_h, \tag{2.7}$$

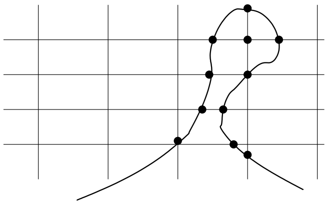

Fig. 2.5 Grid that is not allowed in Section 2.3.

will be considered. In (2.7), for each node $\boldsymbol{x}$, the set $S(\boldsymbol{x})$ is the set of all nodes on which the sum has to be performed, but $\boldsymbol{x} \notin S(\boldsymbol{x})$. That means, $a(\boldsymbol{x})$ describes the contribution of the finite difference scheme of a node $\boldsymbol{x}$ to itself and $b(\boldsymbol{x}, \boldsymbol{y})$ describes the contributions from the neighbors.

It will be assumed that the grid ω_h of inner nodes is connected, i.e., for all $\boldsymbol{x}_a, \boldsymbol{x}_e \in \omega_h$ exist $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \omega_h$ with $\boldsymbol{x}_1 \in S(\boldsymbol{x}_a), \boldsymbol{x}_2 \in S(\boldsymbol{x}_1), \ldots, \boldsymbol{x}_e \in S(\boldsymbol{x}_m)$. E.g., the situation depicted in Figure 2.5 is not allowed.

It will be assumed that the coefficients $a(\boldsymbol{x})$ and $b(\boldsymbol{x}, \boldsymbol{y})$ satisfy the following conditions:

$$
\begin{aligned}
&a(\boldsymbol{x}) > 0,\ b(\boldsymbol{x}, \boldsymbol{y}) > 0,\ \forall\ \boldsymbol{x} \in \omega_h, \forall\ \boldsymbol{y} \in S(\boldsymbol{x}),\\
&a(\boldsymbol{x}) = 1,\ b(\boldsymbol{x}, \boldsymbol{y}) = 0\ \forall\ \boldsymbol{x} \in \gamma_h \text{ (Dirichlet boundary condition).}
\end{aligned}
$$

The values of the Dirichlet boundary condition are incorporated in (2.7) in the function $F(\boldsymbol{x})$. □

Example 2.16. Five point stencil for approximating the Laplacian. Inserting the approximation of the Laplacian with the five point stencil (2.4) for $\boldsymbol{x} = (x, y) \in \omega_h^\circ$ in scheme (2.7) gives

$$
\begin{aligned}
\frac{2(h_x^2 + h_y^2)}{h_x^2 h_y^2} u(x, y) = \Bigg[&\frac{1}{h_x^2} u(x + h_x, y) + \frac{1}{h_x^2} u(x - h_x, y) \\
&+ \frac{1}{h_y^2} u(x, y + h_y) + \frac{1}{h_y^2} u(x, y - h_y) \Bigg] + \phi(x, y).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
a(\boldsymbol{x}) &= \frac{2(h_x^2 + h_y^2)}{h_x^2 h_y^2}, \\
b(\boldsymbol{x}, \boldsymbol{y}) &\in \{h_x^{-2}, h_y^{-2}\}, \\
S(\boldsymbol{x}) &= \{(x - h_x, y), (x + h_x, y), (x, y - h_y), (x, y + h_y)\}.
\end{aligned}
$$

For inner nodes that are close to the boundary, only the one-dimensional case (2.5) will be considered for simplicity. Let $x + h_x^+ \in \gamma_h$, then it follows by inserting (2.5) in (2.7)

$$\frac{1}{\overline{h}_x}\left(\frac{1}{h_x^+}+\frac{1}{h_x^-}\right)u(x,y)=\frac{u(x-h_x^-,y)}{\overline{h}_x h_x^-}+\underbrace{\frac{u(x+h_x^+,y)}{\overline{h}_x h_x^+}}_{\text{on } \gamma_h \to F(x)}+\phi(x), \tag{2.8}$$

such that $a(x)=\frac{1}{\overline{h}_x}\left(\frac{1}{h_x^+}+\frac{1}{h_x^-}\right)$, $b(x,y)=\frac{1}{\overline{h}_x h_x^-}$, and $S(x)=\{(x-h_x^-,y)\}$. □

Remark 2.17. Reformulation of the difference scheme. Scheme (2.7) can be reformulated in the form

$$d(\boldsymbol{x})u(\boldsymbol{x})=\sum_{\boldsymbol{y}\in S(\boldsymbol{x})} b(\boldsymbol{x},\boldsymbol{y})\big(u(\boldsymbol{y})-u(\boldsymbol{x})\big)+F(\boldsymbol{x}) \tag{2.9}$$

with $d(\boldsymbol{x})=a(\boldsymbol{x})-\sum_{\boldsymbol{y}\in S(\boldsymbol{x})} b(\boldsymbol{x},\boldsymbol{y})$. □

Example 2.18. Five point stencil for approximating the Laplacian. Using the five point stencil for approximating the Laplacian, form (2.9) of the scheme is obtained with

$$d(\boldsymbol{x})=\frac{2(h_x^2+h_y^2)}{h_x^2h_y^2}-\frac{2}{h_x^2}-\frac{2}{h_y^2}=0 \tag{2.10}$$

for $\boldsymbol{x}\in\omega_h^\circ$.

The coefficients $a(\boldsymbol{x})$ and $b(\boldsymbol{x},\boldsymbol{y})$ are the weights of the finite difference stencil for approximating the Laplacian. A minimal condition for consistency is that this approximation vanishes for constant functions since the derivatives of constant functions vanish. It follows that also for the nodes $\boldsymbol{x}\in\omega_h^*$, it is $a(\boldsymbol{x})=\sum_{\boldsymbol{y}\in S(\boldsymbol{x})} b(\boldsymbol{x},\boldsymbol{y})$, compare (2.8). However, as it could be seen in Example 2.16, in this case the contributions from the neighbors on γ_h are included in the scheme (2.7) in $F(\boldsymbol{x})$. Hence, one obtains for nodes that are close to the boundary

$$d(\boldsymbol{x})=\underbrace{\sum_{\boldsymbol{y}\in S(\boldsymbol{x})} b(\boldsymbol{x},\boldsymbol{y})}_{=a(\boldsymbol{x})}-\sum_{\boldsymbol{y}\in S(\boldsymbol{x}),\boldsymbol{y}\notin\gamma_h} b(\boldsymbol{x},\boldsymbol{y})=\sum_{\boldsymbol{y}\in S(\boldsymbol{x}),\boldsymbol{y}\in\gamma_h} b(\boldsymbol{x},\boldsymbol{y}). \tag{2.11}$$

In the one-dimensional case, one has, by the definition of $\overline{h}_x$ and with $h_x^-=h_x\geq h_x^+$,

$$\begin{aligned} d(x) &= \frac{1}{\overline{h}_x}\left(\frac{1}{h_x^+}+\frac{1}{h_x^-}\right)-\frac{1}{\overline{h}_x h_x^-}=\frac{1}{\overline{h}_x h_x^+}=\frac{2}{h_xh_x^++h_x^+h_x^+}\\ &\geq \frac{2}{h_xh_x+h_xh_x}=\frac{1}{h_xh_x}>0. \end{aligned}$$

□

Lemma 2.19. Discrete maximum principle (DMP) for inner nodes. *Let $u(\boldsymbol{x}) \neq const$ on ω_h and $d(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \omega_h$. Then, it follows from*

$$L_h u(\boldsymbol{x}) := d(\boldsymbol{x})u(\boldsymbol{x}) - \sum_{\boldsymbol{y}\in S(\boldsymbol{x})} b(\boldsymbol{x},\boldsymbol{y})\big(u(\boldsymbol{y}) - u(\boldsymbol{x})\big) \leq 0 \tag{2.12}$$

(or $L_h u(\boldsymbol{x}) \geq 0$, respectively) on ω_h that $u(\boldsymbol{x})$ does not possess a positive maximum (or negative minimum, respectively) on ω_h.

Proof. The proof is performed by contradiction. Let $L_h u(\boldsymbol{x}) \leq 0$ for all $\boldsymbol{x} \in \omega_h$ and assume that $u(\boldsymbol{x})$ has a positive maximum on ω_h at $\overline{\boldsymbol{x}}$, i.e., $u(\overline{\boldsymbol{x}}) = \max_{\boldsymbol{x}\in\omega_h} u(\boldsymbol{x}) > 0$.

For the node $\overline{\boldsymbol{x}}$, it holds that

$$L_h u(\overline{\boldsymbol{x}}) = \underbrace{d(\overline{\boldsymbol{x}})}_{\geq 0}\underbrace{u(\overline{\boldsymbol{x}})}_{>0} - \sum_{\boldsymbol{y}\in S(\overline{\boldsymbol{x}})} \underbrace{b(\overline{\boldsymbol{x}},\boldsymbol{y})}_{>0} \underbrace{\big(u(\boldsymbol{y}) - u(\overline{\boldsymbol{x}})\big)}_{\leq 0 \text{ by definition of } \overline{\boldsymbol{x}}} \geq d(\overline{\boldsymbol{x}})u(\overline{\boldsymbol{x}}) \geq 0. \tag{2.13}$$

Hence, it follows that $L_h u(\overline{\boldsymbol{x}}) = 0$ and, in particular, that all terms of $L_h u(\overline{\boldsymbol{x}})$ have to vanish. For the first term, it follows that $d(\overline{\boldsymbol{x}}) = 0$. For the terms in the sum to vanish, it must hold

$$u(\boldsymbol{y}) = u(\overline{\boldsymbol{x}}) \quad \forall\, \boldsymbol{y} \in S(\overline{\boldsymbol{x}}). \tag{2.14}$$

From the assumption $u(\boldsymbol{x}) \neq const$, it follows that there exists a node $\hat{\boldsymbol{x}} \in \omega_h$ with $u(\overline{\boldsymbol{x}}) > u(\hat{\boldsymbol{x}})$. Because the grid is connected, there is a path $\overline{\boldsymbol{x}}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_m, \hat{\boldsymbol{x}}$ such that, using (2.14) for all nodes of this path,

$$\begin{aligned}
&\boldsymbol{x}_1 \in S(\overline{\boldsymbol{x}}),\ \ u(\boldsymbol{x}_1) = u(\overline{\boldsymbol{x}}),\\
&\boldsymbol{x}_2 \in S(\boldsymbol{x}_1),\ u(\boldsymbol{x}_2) = u(\boldsymbol{x}_1) = u(\overline{\boldsymbol{x}}),\\
&\ldots\\
&\hat{\boldsymbol{x}} \in S(\boldsymbol{x}_m),\ u(\boldsymbol{x}_m) = u(\boldsymbol{x}_{m-1}) = \ldots = u(\overline{\boldsymbol{x}}) > u(\hat{\boldsymbol{x}}).
\end{aligned}$$

The last inequality is a contradiction to (2.14) for $\boldsymbol{x}_m$. ■

Corollary 2.20. DMP for boundary value problem. *Let $u(\boldsymbol{x}) \geq 0$ for $\boldsymbol{x} \in \gamma_h$ and $L_h u(\boldsymbol{x}) \leq 0$ (or $L_h u(\boldsymbol{x}) \geq 0$, respectively) on ω_h. Then, the grid function $u(\boldsymbol{x})$ is non-positive (or non-negative, respectively) for all $\boldsymbol{x} \in \omega_h \cup \gamma_h$.*

Proof. Let $L_h u(\boldsymbol{x}) \leq 0$ on $\omega_h \cup \gamma_h$. Assume that there is a node $\overline{\boldsymbol{x}} \in \omega_h$ with $u(\overline{\boldsymbol{x}}) > 0$. Then, the grid function has either a positive maximum on ω_h, which is a contradiction to the DMP for the inner nodes, Lemma 2.19, or $u(\boldsymbol{x})$ has to be constant, i.e., $u(\boldsymbol{x}) = u(\overline{\boldsymbol{x}}) > 0$ for all $\boldsymbol{x} \in \omega_h$. For the second case, consider a boundary-connected inner node $\boldsymbol{x} \in \omega_h^*$. Using the same calculations as in (2.13) and taking into account that the values of u at the boundary are non-positive, one obtains

$$\begin{aligned}
L_h u(\boldsymbol{x}) = {} & \underbrace{d(\boldsymbol{x})}_{\geq 0}\underbrace{u(\boldsymbol{x})}_{>0} - \sum_{\boldsymbol{y}\in S(\boldsymbol{x}),\boldsymbol{y}\notin\gamma_h} \underbrace{b(\boldsymbol{x},\boldsymbol{y})}_{>0}\underbrace{(u(\boldsymbol{y}) - u(\boldsymbol{x}))}_{=0}\\
& - \sum_{\boldsymbol{y}\in S(\boldsymbol{x}),\boldsymbol{y}\in\gamma_h} \underbrace{b(\boldsymbol{x},\boldsymbol{y})}_{>0}\underbrace{(u(\boldsymbol{y}) - u(\boldsymbol{x}))}_{<0} > 0,
\end{aligned}$$

which is a contradiction to the assumption on L_h. ■

Corollary 2.21. Unique solution of the discrete Laplace equation with homogeneous Dirichlet boundary conditions. *The discrete Laplace* *equation $L_h u(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \in \omega_h \cup \gamma_h$ possesses only the trivial solution $u(\boldsymbol{x}) = 0$.*

Proof. The statement of the corollary follows by applying Corollary 2.20 and its analog for the non-positivity of the grid function if $u(\boldsymbol{x}) \le 0$ for $\boldsymbol{x} \in \gamma_h$ and $L_h u(\boldsymbol{x}) \le 0$ on ω_h. ■

Corollary 2.22. Comparison lemma. *Let*

$$\begin{aligned} L_h u(\boldsymbol{x}) &= f(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in \omega_h; \quad & u(\boldsymbol{x}) &= g(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in \gamma_h, \\ L_h \overline{u}(\boldsymbol{x}) &= \overline{f}(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in \omega_h; \quad & \overline{u}(\boldsymbol{x}) &= \overline{g}(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in \gamma_h, \end{aligned}$$

with $|f(\boldsymbol{x})| \le \overline{f}(\boldsymbol{x})$ and $|g(\boldsymbol{x})| \le \overline{g}(\boldsymbol{x})$. Then, it is $|u(\boldsymbol{x})| \le \overline{u}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \omega_h \cup \gamma_h$. The function $\overline{u}(\boldsymbol{x})$ is called majorizing function.

Proof. Exercise. ■

Remark 2.23. Remainder of this section. The remaining corollaries presented in this section will be applied in the stability proof in Section 2.4. In this proof, the homogeneous problem (right-hand side vanishes) and the problem with homogeneous Dirichlet boundary conditions will be analyzed separately. □

Corollary 2.24. Homogeneous problem. *For the solution of the problem*

$$\begin{aligned} L_h u(\boldsymbol{x}) &= 0, \qquad \boldsymbol{x} \in \omega_h, \\ u(\boldsymbol{x}) &= g(\boldsymbol{x}), \ \boldsymbol{x} \in \gamma_h, \end{aligned}$$

with $d(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \omega_h^\circ$, it holds that

$$\|u\|_{l^\infty(\omega_h \cup \gamma_h)} \le \|g\|_{l^\infty(\gamma_h)} .$$

Proof. Consider the problem

$$\begin{aligned} L_h \overline{u}(\boldsymbol{x}) &= 0, & \boldsymbol{x} &\in \omega_h, \\ \overline{u}(\boldsymbol{x}) &= \overline{g}(\boldsymbol{x}) = const = \|g\|_{l^\infty(\gamma_h)}, & \boldsymbol{x} &\in \gamma_h. \end{aligned}$$

It will be shown that $\overline{u}(\boldsymbol{x}) = \|g\|_{l^\infty(\gamma_h)} = const$ by inserting this function in the problem.[1] For inner nodes that are not close to the boundary, it holds that

$$L_h \overline{u}(\boldsymbol{x}) = \underbrace{d(\boldsymbol{x})}_{=0,\ (2.10)} \ \overline{u}(\boldsymbol{x}) - \sum_{\boldsymbol{y} \in S(\boldsymbol{x})} b(\boldsymbol{x}, \boldsymbol{y}) \underbrace{\left(\overline{u}(\boldsymbol{y}) - \overline{u}(\boldsymbol{x})\right)}_{=0} = 0.$$

With the same arguments as in Example 2.18, one can derive the representation (2.11) for inner nodes that are close to the boundary. Inserting (2.11) in (2.12) and using in addition $\overline{u}(\boldsymbol{x}) = \overline{u}(\boldsymbol{y})$ yields

[1] The corresponding continuous problem is $-\Delta u = 0$ in Ω, $u = const = \|g\|_{l^\infty(\gamma_h)}$ on $\partial\Omega$. It is clear that $u = \|g\|_{l^\infty(\gamma_h)}$ is the solution of this problem. It is shown that the discrete analog holds, too.

$$L_h\overline{u}(\boldsymbol{x}) = d(\boldsymbol{x})\overline{u}(\boldsymbol{x}) - \sum_{\boldsymbol{y}\in S(\boldsymbol{x})} b(\boldsymbol{x},\boldsymbol{y}) \underbrace{\left(\overline{u}(\boldsymbol{y}) - \overline{u}(\boldsymbol{x})\right)}_{=0} = \sum_{\boldsymbol{y}\in S(\boldsymbol{x}),\boldsymbol{y}\in\gamma_h} b(\boldsymbol{x},\boldsymbol{y})\overline{u}(\boldsymbol{x})$$
$$= \sum_{\boldsymbol{y}\in S(\boldsymbol{x}),\boldsymbol{y}\in\gamma_h} b(\boldsymbol{x},\boldsymbol{y})\overline{u}(\boldsymbol{y}).$$

This expression is exactly the contribution of the nodes on γ_h that is included in $F(\boldsymbol{x})$ in scheme (2.7), see also Example 2.16. That means, the finite difference equation is also satisfied by the nodes that are close to the boundary.

Now, the application of Corollary 2.22 gives $\overline{u}(\boldsymbol{x}) \ge |u(\boldsymbol{x})|$ for all $\boldsymbol{x} \in \omega_h \cup \gamma_h$, such that

$$\|u\|_{l^\infty(\omega_h\cup\gamma_h)} \le \overline{u}(\boldsymbol{x}) = \|g\|_{l^\infty(\gamma_h)}\,,$$

which is the statement of the corollary. ■

Corollary 2.25. Problem with homogeneous boundary condition. *For the solution of the problem*

$$\begin{aligned} L_h u(\boldsymbol{x}) &= f(\boldsymbol{x}),\ \boldsymbol{x}\in\omega_h,\\ u(\boldsymbol{x}) &= 0, \qquad \boldsymbol{x}\in\gamma_h, \end{aligned}$$

with $d(\boldsymbol{x}) > 0$ *for all* $\boldsymbol{x}\in\omega_h$*, it is*

$$\|u\|_{l^\infty(\omega_h\cup\gamma_h)} \le \left\|D^{-1}f\right\|_{l^\infty(\omega_h)}$$

with $D = diag(d(\boldsymbol{x}))$ *for* $\boldsymbol{x}\in\omega_h$.

Proof. Consider the grid function

$$\overline{f}(\boldsymbol{x}) = |f(\boldsymbol{x})| \ge f(\boldsymbol{x}) \ \forall\ \boldsymbol{x}\in\omega_h.$$

From the discrete maximum principle, it follows that the solution of the problem

$$\begin{aligned} L_h\overline{u}(\boldsymbol{x}) &= \overline{f}(\boldsymbol{x}),\ \boldsymbol{x}\in\omega_h,\\ \overline{u}(\boldsymbol{x}) &= 0, \qquad \boldsymbol{x}\in\gamma_h, \end{aligned}$$

is non-negative, i.e., it holds $\overline{u}(\boldsymbol{x}) \ge 0$ for $\boldsymbol{x}\in\omega_h\cup\gamma_h$. Define the node $\overline{\boldsymbol{x}}$ by the condition

$$\overline{u}(\overline{\boldsymbol{x}}) = \|\overline{u}\|_{l^\infty(\omega_h\cup\gamma_h)}\,.$$

In $\overline{\boldsymbol{x}}$, it is

$$L_h\overline{u}(\overline{\boldsymbol{x}}) = d(\overline{\boldsymbol{x}})\overline{u}(\overline{\boldsymbol{x}}) - \sum_{\boldsymbol{y}\in S(\overline{\boldsymbol{x}})} \underbrace{b(\overline{\boldsymbol{x}},\boldsymbol{y})}_{>0} \underbrace{\left(\overline{u}(\boldsymbol{y}) - \overline{u}(\overline{\boldsymbol{x}})\right)}_{\le 0} = |f(\overline{\boldsymbol{x}})|\,,$$

from what follows that

$$\overline{u}(\overline{\boldsymbol{x}}) \le \frac{|f(\overline{\boldsymbol{x}})|}{d(\overline{\boldsymbol{x}})} \le \max_{\boldsymbol{x}\in\omega_h}\frac{|f(\boldsymbol{x})|}{d(\boldsymbol{x})} = \max_{\boldsymbol{x}\in\omega_h}\left|\frac{f(\boldsymbol{x})}{d(\boldsymbol{x})}\right| = \left\|D^{-1}f\right\|_{l^\infty(\omega_h)}.$$

Since $u(\boldsymbol{x}) \le \overline{u}(\overline{\boldsymbol{x}})$ for all $\boldsymbol{x}\in\omega_h\cup\gamma_h$ because of Corollary 2.22, the statement of the corollary is proved. ■

Corollary 2.26. Problem with homogeneous boundary condition and inhomogeneous right-hand side close to the boundary. *Consider*

$$\begin{aligned} L_h u(\boldsymbol{x}) &= f(\boldsymbol{x}),\ \boldsymbol{x}\in\omega_h,\\ u(\boldsymbol{x}) &= 0, \qquad \boldsymbol{x}\in\gamma_h, \end{aligned}$$

with $f(\boldsymbol{x})=0$ for all $\boldsymbol{x}\in\omega_h^\circ$. With respect to the finite difference scheme, it will be assumed that $d(\boldsymbol{x})=0$ for all $\boldsymbol{x}\in\omega_h^\circ$, and $d(\boldsymbol{x})>0$ for all $\boldsymbol{x}\in\omega_h^*$. Then, the following estimate is valid

$$\|u\|_{l^\infty(\omega_h\cup\gamma_h)} \le \left\|D^+ f\right\|_{l^\infty(\omega_h)}$$

with $D^+ = diag(0, d(\boldsymbol{x})^{-1})$. The zero entries appear for $\boldsymbol{x}\in\omega_h^\circ$ and the entries $d(\boldsymbol{x})^{-1}$ for $\boldsymbol{x}\in\omega_h^*$.

Proof. Let $\overline{f}(\boldsymbol{x}) = |f(\boldsymbol{x})|$, $\boldsymbol{x}\in\omega_h$, and $\overline{g}(\boldsymbol{x}) = 0, \boldsymbol{x}\in\gamma_h$. The solution $\overline{u}(\boldsymbol{x})$ is non-negative, $\overline{u}(\boldsymbol{x})\ge 0$ for all $\boldsymbol{x}\in\omega_h\cup\gamma_h$, see the DMP for the boundary value problem, Corollary 2.25. Define $\overline{\boldsymbol{x}}$ by

$$\overline{u}(\overline{\boldsymbol{x}}) = \|\overline{u}\|_{l^\infty(\omega_h\cup\gamma_h)}\,.$$

One can choose $\overline{\boldsymbol{x}}\in\omega_h^*$, because if $\overline{\boldsymbol{x}}\in\omega_h^\circ$, then it holds that

$$\underbrace{d(\overline{\boldsymbol{x}})}_{=0}\,\overline{u}(\overline{\boldsymbol{x}}) - \sum_{\boldsymbol{y}\in S(\overline{\boldsymbol{x}})} \underbrace{b(\overline{\boldsymbol{x}},\boldsymbol{y})}_{>0}\underbrace{\left(\overline{u}(\boldsymbol{y})-\overline{u}(\overline{\boldsymbol{x}})\right)}_{\le 0} = \overline{f}(\overline{\boldsymbol{x}}) = 0,$$

i.e., $\overline{u}(\overline{\boldsymbol{x}}) = \overline{u}(\boldsymbol{y})$ for all $\boldsymbol{y}\in S(\overline{\boldsymbol{x}})$. Let $\hat{\boldsymbol{x}}\in\omega_h^*$ and $\overline{\boldsymbol{x}},\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m,\hat{\boldsymbol{x}}$ be a connection with $\boldsymbol{x}_i\notin\omega_h^*$, $i=1,\ldots,m$. For $\boldsymbol{x}_m$, it holds analogously that

$$\overline{u}(\boldsymbol{x}_m) = \|\overline{u}\|_{l^\infty(\omega_h\cup\gamma_h)} = \overline{u}(\boldsymbol{y})\ \forall\ \boldsymbol{y}\in S(\boldsymbol{x}_m).$$

Hence, it follow in particular that $\overline{u}(\hat{\boldsymbol{x}}) = \|\overline{u}\|_{l^\infty(\omega_h\cup\gamma_h)}$ such that one can choose $\overline{\boldsymbol{x}}=\hat{\boldsymbol{x}}$. It follows that

$$\underbrace{d(\hat{\boldsymbol{x}})}_{>0}\quad\underbrace{\overline{u}(\hat{\boldsymbol{x}})}_{=\|\overline{u}\|_{l^\infty(\omega_h\cup\gamma_h)}}\quad - \sum_{\boldsymbol{y}\in S(\hat{\boldsymbol{x}})} \underbrace{b(\hat{\boldsymbol{x}},\boldsymbol{y})}_{>0}\underbrace{\left(\overline{u}(\boldsymbol{y})-\overline{u}(\hat{\boldsymbol{x}})\right)}_{\le 0} = \overline{f}(\hat{\boldsymbol{x}}).$$

Since all terms in the sum over $\boldsymbol{y}\in\omega_h$ are non-negative, it follows, using also Corollary 2.22, that

$$\|u\|_{l^\infty(\omega_h\cup\gamma_h)} \le \|\overline{u}\|_{l^\infty(\omega_h\cup\gamma_h)} \le \frac{\overline{f}(\hat{\boldsymbol{x}})}{d(\hat{\boldsymbol{x}})} \le \max_{\boldsymbol{x}\in\omega_h^*}\frac{\overline{f}(\boldsymbol{x})}{d(\boldsymbol{x})} \le \left\|D^+ f\right\|_{l^\infty(\omega_h)}\,.$$

■

2.4 Stability and Convergence of the Finite Difference Approximation of the Poisson Problem with Dirichlet Boundary Conditions

Remark 2.27. Decomposition of the solution. A short form to write (2.6) is

$$L_h u(\boldsymbol{x}) = f(\boldsymbol{x}),\ \boldsymbol{x}\in\omega_h,\quad u(\boldsymbol{x}) = g(\boldsymbol{x}),\ \boldsymbol{x}\in\gamma_h.$$

The solution of (2.6) can be decomposed into

$$u(\boldsymbol{x}) = u_1(\boldsymbol{x}) + u_2(\boldsymbol{x}),$$

with

$$\begin{aligned} &L_h u_1(\boldsymbol{x}) = f(\boldsymbol{x}),\ \boldsymbol{x} \in \omega_h, \quad u_1(\boldsymbol{x}) = 0,\ \boldsymbol{x} \in \gamma_h \text{ (homogeneous boundary cond.)},\\ &L_h u_2(\boldsymbol{x}) = 0,\ \boldsymbol{x} \in \omega_h, \quad u_2(\boldsymbol{x}) = g(\boldsymbol{x}),\ \boldsymbol{x} \in \gamma_h \text{ (homogeneous right-hand side)}. \end{aligned}$$

□

Stability with Respect to the Boundary Condition

Remark 2.28. Stability with respect to the boundary condition. From Corollary 2.24, it follows that

$$\|u_2\|_{l^\infty(\omega_h)} \le \|g\|_{l^\infty(\gamma_h)}\,. \tag{2.15}$$

□

Stability with Respect to the Right-Hand Side

Remark 2.29. Decomposition of the right-hand side. The right-hand side will be decomposed into

$$f(\boldsymbol{x}) = f^\circ(\boldsymbol{x}) + f^*(\boldsymbol{x})$$

with

$$f^\circ(\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}), & \boldsymbol{x} \in \omega_h^\circ, \\ 0, & \boldsymbol{x} \in \omega_h^*, \end{cases} \qquad f^*(\boldsymbol{x}) = f(\boldsymbol{x}) - f^\circ(\boldsymbol{x}).$$

Since the considered finite difference scheme is linear, also the function $u_1(\boldsymbol{x})$ can be decomposed into

$$u_1(\boldsymbol{x}) = u_1^\circ(\boldsymbol{x}) + u_1^*(\boldsymbol{x})$$

with

$$\begin{aligned} L_h u_1^\circ(\boldsymbol{x}) &= f^\circ(\boldsymbol{x}),\ \boldsymbol{x} \in \omega_h, \quad u_1^\circ(\boldsymbol{x}) = 0,\ \boldsymbol{x} \in \gamma_h,\\ L_h u_1^*(\boldsymbol{x}) &= f^*(\boldsymbol{x}),\ \boldsymbol{x} \in \omega_h, \quad u_1^*(\boldsymbol{x}) = 0,\ \boldsymbol{x} \in \gamma_h. \end{aligned}$$

□

Remark 2.30. Estimate for the inner nodes. Let $B((0,0),R)$ be a circle with center $(0,0)$ and radius R, which is chosen such that $R \ge \|\boldsymbol{x}\|_2$ for all $\boldsymbol{x} \in \Omega$. Consider the function

$$\overline{u}(\boldsymbol{x}) = \alpha\left(R^2 - x^2 - y^2\right) \quad \text{with } \alpha > 0,$$

that takes only non-negative values for $(x, y) \in \Omega$. Applying the definition of the five point stencil, it follows that

$$\begin{aligned}
\Lambda\overline{u}(\boldsymbol{x}) &= -\alpha\Lambda(x^2+y^2-R^2) \\
&= -\alpha\left(\frac{(x+h_x)^2-2x^2+(x-h_x)^2}{h_x^2}+\frac{(y+h_y)^2-2y^2+(y-h_y)^2}{h_y^2}\right) \\
&= -4\alpha =: -\overline{f}(\boldsymbol{x}),\ \boldsymbol{x}\in\omega_h^\circ,
\end{aligned}$$

and

$$\begin{aligned}
\Lambda^*\overline{u}(\boldsymbol{x}) &= -\alpha\left[\frac{1}{\overline{h}_x}\left(\frac{(x+h_x^+)^2-x^2}{h_x^+}-\frac{x^2-(x-h_x^-)^2}{h_x^-}\right)\right. \\
&\left.\quad+\frac{1}{\overline{h}_y}\left(\frac{(y+h_y^+)^2-y^2}{h_y^+}-\frac{y^2-(y-h_y^-)^2}{h_y^-}\right)\right] \\
&= -\alpha\left(\frac{h_x^++h_x^-}{\overline{h}_x}+\frac{h_y^++h_y^-}{\overline{h}_y}\right) =: -\overline{f}(\boldsymbol{x}),\ \boldsymbol{x}\in\omega_h^*.
\end{aligned}$$

Hence, $\overline{u}(\boldsymbol{x})$ is the solution of the problem

$$\begin{aligned}
L_h\overline{u}(\boldsymbol{x}) &= \overline{f}(\boldsymbol{x}), & \boldsymbol{x}&\in\omega_h, \\
\overline{u}(\boldsymbol{x}) &= \alpha\left(R^2-x^2-y^2\right)\ge 0, & \boldsymbol{x}&\in\gamma_h.
\end{aligned}$$

It is $\overline{u}(\boldsymbol{x}) \ge 0$ for all $\boldsymbol{x} \in \gamma_h$. Choosing $\alpha = \frac{1}{4}\|f^\circ\|_{l^\infty(\omega_h)}$, one obtains

$$\begin{aligned}
&\overline{f}(\boldsymbol{x}) = 4\alpha = \|f^\circ\|_{l^\infty(\omega_h)} \ge |f^\circ(\boldsymbol{x})|\,,\ \boldsymbol{x}\in\omega_h^\circ, \\
&\overline{f}(\boldsymbol{x}) \ge 0 = |f^\circ(\boldsymbol{x})|\,,\ \boldsymbol{x}\in\omega_h^*.
\end{aligned}$$

Now, Corollary 2.22 (Comparison Lemma) can be applied, which leads to

$$\|u_1^\circ\|_{l^\infty(\omega_h)} \le \|\overline{u}\|_{l^\infty(\omega_h)} \le \alpha R^2 = \frac{R^2}{4}\|f^\circ\|_{l^\infty(\omega_h)}\,. \tag{2.16}$$

One gets the last lower or equal estimate because $(0, 0)$ does not need to belong to Ω or ω_h. □

Remark 2.31. Estimate for the nodes that are close to the boundary. Corollary 2.26 can be applied to estimate $u_1^*(\boldsymbol{x})$. For $\boldsymbol{x} \in \omega_h^\circ$, it is $d(\boldsymbol{x}) = 0$, see Example 2.18. For $\boldsymbol{x} \in \omega_h^*$, one has

$$d(\boldsymbol{x}) = a(\boldsymbol{x}) - \sum_{\boldsymbol{y}\in S(\boldsymbol{x}),\boldsymbol{y}\notin\gamma_h} b(\boldsymbol{x},\boldsymbol{y}) \ge \frac{1}{h^2}$$

with $h = \max\{h_x, h_y\}$, since all terms in the sum are of the form

$$\frac{1}{\overline{h}_x h_x^+},\ \frac{1}{\overline{h}_x h_x^-},\ \frac{1}{\overline{h}_y h_y^+},\ \frac{1}{\overline{h}_y h_y^-},$$

see Example 2.18. One obtains

$$\|u_1^*\|_{l^\infty(\omega_h)} \le \left\|D^+ f^*\right\|_{l^\infty(\omega_h)} \le h^2 \|f^*\|_{l^\infty(\omega_h)}. \tag{2.17}$$

□

Lemma 2.32. Stability estimate *The solution of the discrete Dirichlet problem* (2.6) *satisfies*

$$\|u\|_{l^\infty(\omega_h \cup \gamma_h)} \le \|g\|_{l^\infty(\gamma_h)} + \frac{R^2}{4} \|\phi\|_{l^\infty(\omega_h^\circ)} + h^2 \|\phi\|_{l^\infty(\omega_h^*)} \tag{2.18}$$

with $R \ge \|\boldsymbol{x}\|_2$ *for all* $\boldsymbol{x} \in \Omega$ *and* $h = \max\{h_x, h_y\}$, *i.e., the solution* $u(\boldsymbol{x})$ *can be bounded in the norm* $\|\cdot\|_{l^\infty(\omega_h \cup \gamma_h)}$ *by the data of the problem.*

Proof. The statement of the lemma is obtained by combining the estimates (2.15), (2.16), and (2.17). ■

Convergence

Theorem 2.33. Convergence. *Let* $u(\boldsymbol{x})$ *be the solution of the Poisson equation* (2.1) *and* $u_h(\boldsymbol{x})$ *be the finite difference approximation given by the solution of* (2.6). *Then, it is*

$$\|u - u_h\|_{l^\infty(\omega_h \cup \gamma_h)} \le Ch^2$$

with $h = \max\{h_x, h_y\}$.

Proof. The error in the node (x_i, y_j) is defined by $e_{ij} = u(x_i, y_j) - u_h(x_i, y_j)$. With

$$-\Lambda u(x_i, y_j) = -\Delta u(x_i, y_j) + \mathcal{O}\left(h^2\right) = f(x_i, y_j) + \mathcal{O}\left(h^2\right),$$

one obtains by subtracting the finite difference equation, the following problem for the error

$$\begin{aligned} -\Lambda e(\boldsymbol{x}) &= \psi(\boldsymbol{x}),\ \boldsymbol{x} \in w_h^\circ,\ \psi(\boldsymbol{x}) = \mathcal{O}\left(h^2\right), \\ -\Lambda^* e(\boldsymbol{x}) &= \psi(\boldsymbol{x}),\ \boldsymbol{x} \in w_h^*,\ \psi(\boldsymbol{x}) = \mathcal{O}(1), \\ e(\boldsymbol{x}) &= 0, \qquad \boldsymbol{x} \in \gamma_h, \end{aligned}$$

where $\psi(\boldsymbol{x})$ is the consistency error, see Section 2.2. Applying the stability estimate (2.18) to this problem, one obtains immediately

$$\|e\|_{l^\infty(\omega_h \cup \gamma_h)} \le \frac{R^2}{4} \|\psi\|_{l^\infty(\omega_h^\circ)} + h^2 \|\psi\|_{l^\infty(\omega_h^*)} = \mathcal{O}\left(h^2\right).$$

■

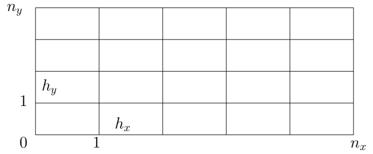

Fig. 2.6 Grid for the Dirichlet problem in the rectangular domain.

2.5 An Efficient Solver for the Dirichlet Problem in the Rectangle

Remark 2.34. Contents of this section. This section considers the Poisson equation (2.1) in the special case $\Omega = (0, l_x) \times (0, l_y)$. In this case, a modification of the difference stencil in a neighborhood of the boundary of the domain is not needed. The convergence of the finite difference approximation was already established in Theorem 2.33. Applying this approximation results in a large linear system of equations $A\boldsymbol{u} = \underline{f}$ which has to be solved. This section discusses some properties of the matrix A and it presents an approach for solving this system in the case of a rectangular domain in an almost optimal way.

A number of result obtained here will be need also in Section 2.6. □

Remark 2.35. The considered problem and its approximation. The considered continuous problem consists in solving

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega = (0, l_x) \times (0, l_y), \\ u &= g \text{ on } \partial\Omega, \end{aligned}$$

and the corresponding discrete problem in solving

$$\begin{aligned} -\Lambda u(\boldsymbol{x}) &= \phi(\boldsymbol{x}),\ \boldsymbol{x} \in \omega_h, \\ u(\boldsymbol{x}) &= g(\boldsymbol{x}),\ \boldsymbol{x} \in \gamma_h, \end{aligned}$$

where the discrete Laplacian is of the form (for simplicity of notation, the subscript h is omitted)

$$\Lambda u = \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{h_x^2} + \frac{u_{i,j+1} - 2u_{ij} + u_{i,j-1}}{h_y^2} =: \Lambda_x u + \Lambda_y u, \quad (2.19)$$

with $h_x = l_x/n_x, h_y = l_y/n_y$, $i = 0, \ldots, n_x, j = 0, \ldots, n_y$, see Figure 2.6. □

Remark 2.36. The linear system of equations. The difference scheme (2.19) is equivalent to a linear system of equations $A\underline{u} = \underline{f}$.

For assembling the matrix and the right-hand side of the system, usually a lexicographical enumeration of the nodes of the grid is used. The nodes are called enumerated lexicographically if the node (i_1, j_1) has a smaller number than the node (i_2, j_2), if for the corresponding coordinates, it is

$$y_1 < y_2 \ \text{ or } (y_1 = y_2) \wedge (x_1 < x_2).$$

Using this lexicographical enumeration of the nodes, one obtains for the inner nodes a system of the form

$$\begin{aligned}
A &= \text{BlockTriDiag}(C, B, C) \in \mathbb{R}^{(n_x-1)(n_y-1)\times(n_x-1)(n_y-1)},\\
B &= \text{TriDiag}\left(-\frac{1}{h_x^2}, \frac{2}{h_x^2} + \frac{2}{h_y^2}, -\frac{1}{h_x^2}\right) \in \mathbb{R}^{(n_x-1)\times(n_x-1)},\\
C &= \text{Diag}\left(-\frac{1}{h_y^2}\right) \in \mathbb{R}^{(n_x-1)\times(n_x-1)},\\
\underline{f} &= \begin{cases}
\phi(\boldsymbol{x}), & \boldsymbol{x} \in \omega_h^\circ,\\
\phi(\boldsymbol{x}) + \dfrac{g(x \pm h_x, y)}{h_x^2}, & \boldsymbol{x} \in \omega_h^*, \text{ close to lower}\\
 & \text{or upper boundary,}\\
\phi(\boldsymbol{x}) + \dfrac{g(x, y \pm h_y)}{h_y^2}, & \boldsymbol{x} \in \omega_h^*, , \text{ close to left}\\
 & \text{or right boundary,}\\
\phi(\boldsymbol{x}) + \dfrac{g(x \pm h_x, y)}{h_x^2} + \dfrac{g(x, yx \pm h_y)}{h_y^2}, & \boldsymbol{x} \in \omega_h^*, \text{ corner of inner nodes.}
\end{cases}
\end{aligned}$$

In this approach, the known Dirichlet boundary values are already substituted into the system and they appear in the right-hand side vector. The matrices B and C possess some modifications for nodes that have a neighbor on the boundary.

The linear system of equations has the following properties:

- high dimension: $N = (n_x - 1)(n_y - 1) \sim 10^3 \cdots 10^7$,
- sparse: per row and column of the matrix there are only 3, 4, or 5 non-zero entries,
- symmetric: hence, all eigenvalues are real,
- positive definite: all eigenvalues are positive. It holds that

$$\begin{aligned}
\lambda_{\min} &= \lambda_{(1,1)} \sim \pi^2\left(\frac{1}{l_x^2} + \frac{1}{l_y^2}\right) = \mathcal{O}\left(1\right),\\
\lambda_{\max} &= \lambda_{(n_x-1,n_y-1)} \sim \pi^2\left(\frac{1}{h_x^2} + \frac{1}{h_y^2}\right) = \mathcal{O}\left(h^{-2}\right),
\end{aligned} \tag{2.20}$$

  with $h = \max\{h_x, h_y\}$, see Remark 2.37 below.
- high condition number: For the spectral condition number of a symmetric and positive definite matrix, it is

$$\kappa_2(A) = \frac{\lambda_{\max}}{\lambda_{\min}} = \mathcal{O}\left(h^{-2}\right).$$

Since the dimension of the matrix is large and the matrix is sparse, iterative solvers are an appropriate approach for solving the linear system of equations. The main costs for iterative solvers are the matrix-vector multiplications (often one per iteration). The cost of one matrix-vector multiplication is for sparse matrices proportional to the number of unknowns. Hence, an optimal solver with respect to the number of floating point operations is given if the number of operations for solving the linear system of equations is proportional to the number of unknowns. It is known that the number of iterations of many iterative solvers depends on the condition number of the matrix:

- *(damped) Jacobi method, SOR, SSOR.* The number of iteration is proportional to $\kappa_2(A)$. That means, if the grid is refined once, $h \to h/2$, then the number of unknowns is increased by around the factor 4 in two dimensions and also the number of iterations increases by a factor of around 4. Altogether, for one refinement step, the total costs increase by a factor of around 16.
- *(preconditioned) conjugate gradient (PCG) method.* The number of iterations is proportional to $\sqrt{\kappa_2(A)}$, see the corresponding theorem from the class Numerical Mathematics II. Hence, the total costs increase by a factor of around 8 if the grid is refined once.
- *multigrid methods.* For multigrid methods, the number of iterations on each grid is bounded by a constant that is independent of the grid. Hence, the total costs are proportional to the number of unknowns and these methods are optimal. However, the implementation of multigrid methods is involved.

□

Remark 2.37. An eigenvalue problem. The derivation of an alternative direct solver is based on the eigenvalues and eigenvectors of the discrete Laplacian. It is possible to computed these quantities only in special situations, e.g., if the Poisson problem with Dirichlet boundary conditions is considered, the domain is rectangular, and the Laplacian is approximated with the five point stencil.

Consider the following eigenvalue problem

$$\begin{aligned} -\Lambda v(\boldsymbol{x}) &= \lambda v(\boldsymbol{x}), \ \boldsymbol{x} \in \omega_h, \\ v(\boldsymbol{x}) &= 0, \qquad\quad \boldsymbol{x} \in \gamma_h. \end{aligned}$$

Denote the node $\boldsymbol{x} = (x_i, y_j)$ by $\boldsymbol{x}_{ij}$ and grid functions in a similar way. The solution of this problem is sought in product form (separation of variables)

$$v_{ij}^{(\boldsymbol{k})} = v_i^{(k_x),x} v_j^{(k_y),y}, \quad \boldsymbol{k} = (k_x, k_y)^T.$$

It is

$$\Lambda v_{ij}^{(\boldsymbol{k})} = \left(\Lambda_x v_i^{(k_x),x}\right) v_j^{(k_y),y} + v_i^{(k_x),x} \left(\Lambda_y v_j^{(k_y),y}\right) = -\lambda_{\boldsymbol{k}} v_i^{(k_x),x} v_j^{(k_y),y},$$

where $i = 0, \ldots, n_x$, $j = 0, \ldots, n_y$ refers to the nodes and $k_x = 1, \ldots, n_x - 1$, $k_y = 1, \ldots, n_y - 1$ refers to the eigenvalues. Note that the number of eigenvalues is equal to the number of inner nodes, i.e., it is $(n_x - 1)(n_y - 1)$. In this ansatz, also a splitting of the eigenvalues in a contribution from the x coordinate and a contribution from the y coordinate is included. From the boundary condition, it follows that

$$v_0^{(k_x),x} = v_{n_x}^{(k_x),x} = v_0^{(k_y),y} = v_{n_y}^{(k_y),y} = 0.$$

Dividing by $v_i^{(k_x),x} v_j^{(k_y),y}$ and rearranging terms, the eigenvalue problem can be split

$$\frac{\Lambda_x v_i^{(k_x),x}}{v_i^{(k_x),x}} + \lambda_{k_x}^{(x)} = -\frac{\Lambda_y v_j^{(k_y),y}}{v_j^{(k_y),y}} - \lambda_{k_y}^{(y)}$$

with $\lambda_{\boldsymbol{k}} = \lambda_{k_x}^{(x)} + \lambda_{k_y}^{(y)}$. Both sides of this equation have to be constant since one of them depends only on i, i.e., on x, and the other one only on j, i.e., on y. The splitting of $\lambda_{\boldsymbol{k}}$ can be chosen such that the constant is zero. Then, one gets

$$\Lambda_x v_i^{(k_x),x} + \lambda_{k_x}^{(x)} v_i^{(k_x),x} = 0, \quad \Lambda_y v_j^{(k_y),y} + \lambda_{k_y}^{(y)} v_j^{(k_y),y} = 0.$$

The solution of these eigenvalue problems is known (exercise)

$$\begin{aligned} v_i^{(k_x),x} &= \sqrt{\frac{2}{l_x}} \sin\left(\frac{k_x \pi i}{n_x}\right), \quad \lambda_{k_x}^{(x)} = \frac{4}{h_x^2} \sin^2\left(\frac{k_x \pi}{2n_x}\right), \\ v_j^{(k_y),y} &= \sqrt{\frac{2}{l_y}} \sin\left(\frac{k_y \pi j}{n_y}\right), \quad \lambda_{k_y}^{(y)} = \frac{4}{h_y^2} \sin^2\left(\frac{k_y \pi}{2n_y}\right). \end{aligned}$$

It follows that the solution of the full eigenvalue problem is

$$\begin{aligned} v_{ij}^{(\boldsymbol{k})} &= \frac{2}{\sqrt{l_x l_y}} \sin\left(\frac{k_x \pi i}{n_x}\right) \sin\left(\frac{k_y \pi j}{n_y}\right), \\ \lambda_{\boldsymbol{k}} &= \frac{4}{h_x^2} \sin^2\left(\frac{k_x \pi}{2n_x}\right) + \frac{4}{h_y^2} \sin^2\left(\frac{k_y \pi}{2n_y}\right), \end{aligned} \tag{2.21}$$

with $i = 0, \ldots, n_x, j = 0, \ldots, n_y$ and $k_x = 1, \ldots, n_x - 1, k_y = 1, \ldots, n_y - 1$. Using a Taylor series expansion, one obtains now the asymptotic behavior of the eigenvalues as given in (2.20). Note that because of the splitting of the eigenvalues into the directional contributions, the number of individual terms for computing the eigenvalues is only proportional to $(n_x + n_y)$. □

Remark 2.38. On the eigenvectors, weighted Euclidean inner product. Since the matrix corresponding to Λ is symmetric, the eigenvectors are orthogonal with respect to the Euclidean vector product. They become orthonormal with respect to the weighted Euclidean vector product

$$\langle u, v\rangle = h_x h_y \sum_{\boldsymbol{x}\in\omega_h\cup\gamma_h} u(\boldsymbol{x})v(\boldsymbol{x}) = h_x h_y \sum_{i=0}^{n_x}\sum_{j=0}^{n_y} u_{ij}v_{ij}, \tag{2.22}$$

with

$$h_x = \frac{l_x}{n_x}, h_y = \frac{l_y}{n_y},$$

i.e., then it is

$$\langle v^{(\boldsymbol{k})}, v^{(\boldsymbol{m})}\rangle = \delta_{\boldsymbol{k},\boldsymbol{m}}.$$

This property can be checked by using the relation

$$\sum_{i=0}^{n} \sin^2\left(\frac{i\pi}{n}\right) = \frac{n}{2}, \quad n > 1.$$

The norm induced by the weighted Euclidean vector product is given by

$$\|v\|_h = \langle v, v\rangle^{1/2} = \left(h_x h_y \sum_{i=0}^{n_x}\sum_{j=0}^{n_y} v_{ij}^2\right)^{1/2}. \tag{2.23}$$

The weights are such that this norm can be bounded for constants independently of the mesh, i.e.,

$$\|1\|_h = \left(h_x h_y (n_x+1)(n_y+1)\right)^{1/2} = \left(l_x l_y \frac{n_x+1}{n_x}\frac{n_y+1}{n_y}\right)^{1/2} \le 2\left(l_x l_y\right)^{1/2}. \tag{2.24}$$

□

Remark 2.39. Solver based on the eigenvalues and eigenvectors. One uses the ansatz

$$\phi(\boldsymbol{x}) = \sum_{\boldsymbol{k}} f_{\boldsymbol{k}} v^{(\boldsymbol{k})}(\boldsymbol{x}) \tag{2.25}$$

with the Fourier coefficients

$$f_{\boldsymbol{k}} = \langle f, v^{(\boldsymbol{k})}\rangle = \frac{2h_x h_y}{\sqrt{l_x l_y}} \sum_{i=0}^{n_x}\sum_{j=0}^{n_y} f_{ij} \sin\left(\frac{k_x\pi i}{n_x}\right)\sin\left(\frac{k_y\pi j}{n_y}\right), \quad \boldsymbol{k} = (k_x, k_y),$$

with $f_{ij} = f(\boldsymbol{x}_{ij})$. The solution $u(\boldsymbol{x})$ of (2.19) is sought as a linear combination of the eigenfunctions

$$u(\boldsymbol{x}) = \sum_{\boldsymbol{k}} u_{\boldsymbol{k}} v^{(\boldsymbol{k})}(\boldsymbol{x})$$

with unknown coefficients $u_{\boldsymbol{k}}$. With this ansatz, one obtains for the finite difference operator

$$\Lambda u = \sum_{\boldsymbol{k}} u_{\boldsymbol{k}} \Lambda v^{(\boldsymbol{k})} = \sum_{\boldsymbol{k}} u_{\boldsymbol{k}} \lambda_{\boldsymbol{k}} v^{(\boldsymbol{k})}.$$

Since the eigenfunctions form a basis of the space of the grid functions, a comparison of the coefficients with the right-hand side (2.25) gives

$$-u_{\boldsymbol{k}}\lambda_{\boldsymbol{k}} = f_{\boldsymbol{k}} \quad \Longleftrightarrow \quad u_{\boldsymbol{k}} = -\frac{f_{\boldsymbol{k}}}{\lambda_{\boldsymbol{k}}}$$

or, for each component, using (2.21),

$$u_{ij} = -\sum_{\boldsymbol{k}} \frac{f_{\boldsymbol{k}}}{\lambda_{\boldsymbol{k}}} v_{ij}^{(\boldsymbol{k})} = -\frac{2h_x h_y}{\sqrt{l_x l_y}} \sum_{k_x=1}^{n_x-1} \sum_{k_y=1}^{n_y-1} \frac{f_{\boldsymbol{k}}}{\lambda_{\boldsymbol{k}}} \sin\left(\frac{k_x \pi i}{n_x}\right) \sin\left(\frac{k_y \pi j}{n_y}\right),$$

$i = 0, \ldots, n_x$, $j = 0, \ldots, n_y$.

It is possible to implement this approach with the Fast Fourier Transform (FFT) with

$$\mathcal{O}\left(n_x n_y \log_2 n_x + n_x n_y \log_2 n_y\right) = \mathcal{O}\left(N \log_2 N\right), \ N = (n_x - 1)(n_y - 1),$$

operations. Hence, this method is almost, up to a logarithmic factor, optimal. □

2.6 A Higher Order Discretizations

Remark 2.40. Contents. The five point stencil is a second order discretization of the Laplacian. In this section, a discretization of higher order will be studied. In these studies, only the case of a rectangular domain $\Omega = (0, l_x) \times (0, l_y)$ and Dirichlet boundary conditions will be considered. □

Remark 2.41. Derivation of a fourth order approximation. Let $u(\boldsymbol{x})$ be the solution of the Poisson equation (2.1) and assume that $u(\boldsymbol{x})$ is sufficiently smooth. It is

$$Lu(\boldsymbol{x}) = \Delta u(\boldsymbol{x}) = L_x u(\boldsymbol{x}) + L_y u(\boldsymbol{x}), \quad L_\alpha u := \frac{\partial^2 u}{\partial x_\alpha^2}.$$

Let the five point stencil be represented by the following operator

$$\Lambda u = \Lambda_x u + \Lambda_y u.$$

Applying a Taylor series expansion, one finds that

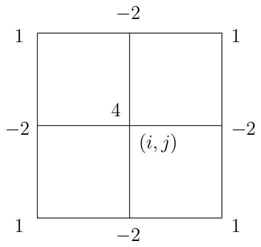

Fig. 2.7 The nine point stencil.

$$\Lambda u - \Delta u = \frac{h_x^2}{12} L_x^2 u + \frac{h_y^2}{12} L_y^2 u + \mathcal{O}\left(h^4\right). \tag{2.26}$$

From the equation $-Lu = f$, it follows with differentiation that

$$L_x^2 u = -L_x f - L_x L_y u, \quad L_y^2 u = -L_y f - L_y L_x u.$$

Inserting these expressions in (2.26) gives

$$\Lambda u - \Delta u = -\frac{h_x^2}{12} L_x f - \frac{h_y^2}{12} L_y f - \frac{h_x^2 + h_y^2}{12} L_x L_y u + \mathcal{O}\left(h^4\right). \tag{2.27}$$

The operator $L_x L_y = \frac{\partial^4}{\partial x^2 \partial y^2}$ can be approximated as follows

$$L_x L_y u \approx \Lambda_x \Lambda_y u = u_{\overline{x}x\overline{y}y}.$$

The difference operator in this approximation requires nine points, see Figure 2.7,

$$\Lambda_x \Lambda_y u = \frac{1}{h_x^2 h_y^2}\Big(u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1} - 2u_{i+1,j} + 4u_{ij}$$
$$-2u_{i-1,j} + u_{i+1,j-1} - 2u_{i,j-1} + u_{i-1,j-1}\Big).$$

Therefore it is called nine point stencil.

One checks, as usual by using a Taylor series expansion, that this approximation is of second order

$$L_x L_y u - \Lambda_x \Lambda_y u = \mathcal{O}\left(h^2\right).$$

Inserting this expansion in (2.27) and using the partial differential equation shows that the difference equation

$$-\left(\Lambda + \frac{h_x^2 + h_y^2}{12} \Lambda_x \Lambda_y\right) u = \left(f + \frac{h_x^2}{12} L_x f + \frac{h_y^2}{12} L_y f\right)$$

is a fourth order approximation of the differential equation (2.1). In addition, one can replace the derivatives of $f(\boldsymbol{x})$ also by finite differences

$$L_x f = \Lambda_x f + \mathcal{O}\left(h_x^2\right), \quad L_y f = \Lambda_y f + \mathcal{O}\left(h_y^2\right).$$

Finally, one obtains a finite difference equation $-\Lambda' u = \phi$ with

$$\Lambda' = \Lambda_x + \Lambda_y + \frac{h_x^2 + h_y^2}{12}\Lambda_x\Lambda_y, \quad \phi = f + \frac{h_x^2}{12}\Lambda_x f + \frac{h_y^2}{12}\Lambda_y f.$$

□

Remark 2.42. On the convergence of the fourth order approximation. The finite difference problem with the higher order approximation property can be written with the help of the second order differences. Since the convergence proof is based on the five point stencil, the following lemma considers this stencil. It will be proved that one can estimate the values of the grid function by the second order differences. This result will be used in the convergence proof for the fourth order approximation. □

Lemma 2.43. Stability estimate. *Let*

$$\omega_h = \{(ih_x, jh_y) \ : \ i = 1, \ldots, n_x - 1, j = 1, \ldots, n_y - 1\},$$

and let y be a grid function on $\omega_h \cup \gamma_h$ with $y(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \in \gamma_h$. Then, the following estimate holds

$$\|y\|_{l^\infty(\omega_h \cup \gamma_h)} \le M \|\Lambda y\|_h,$$

with the mesh-independent constant $M = \frac{\max\{l_x^2, l_y^2\}}{2\sqrt{l_x l_y}}$, Λ is the matrix obtained by using the five point stencil $\Lambda = \Lambda_x + \Lambda_y$ for approximating the second derivatives, and the norm on the right-hand side is defined in (2.23)*.*

Proof. Let $\{v_{ij}^{\boldsymbol{k}}\}$, $\boldsymbol{k} = (k_x, k_y)$, be the orthonormal basis with

$$v_{ij}^{\boldsymbol{k}} = \frac{2}{\sqrt{l_x l_y}} \sin\left(\frac{k_x \pi i}{n_x}\right) \sin\left(\frac{k_y \pi j}{n_y}\right),$$

which was derived in Remark 2.37. Then, there is a unique representation of the grid function $y = \sum_{\boldsymbol{k}} y_{\boldsymbol{k}} v^{\boldsymbol{k}}$ and it holds with (2.22)

$$\Lambda y = \sum_{\boldsymbol{k}} y_{\boldsymbol{k}} \lambda_{\boldsymbol{k}} v^{\boldsymbol{k}}, \quad \|\Lambda y\|_h^2 = \sum_{\boldsymbol{k}} y_{\boldsymbol{k}}^2 \lambda_{\boldsymbol{k}}^2. \tag{2.28}$$

It follows for $\boldsymbol{x} \in \omega_h$, because of $|\sin(x)| \le 1$ for all $x \in \mathbb{R}$, that

$$|y(\boldsymbol{x})| = \left|\sum_{\boldsymbol{k}} y_{\boldsymbol{k}} v^{\boldsymbol{k}}(\boldsymbol{x})\right| \le \sum_{\boldsymbol{k}} |y_{\boldsymbol{k}}| \left|v^{\boldsymbol{k}}(\boldsymbol{x})\right| \le \frac{2}{\sqrt{l_x l_y}} \sum_{\boldsymbol{k}} |y_{\boldsymbol{k}}|.$$

Using this estimate, applying the Cauchy–Schwarz inequality for sums, and utilizing (2.28) gives

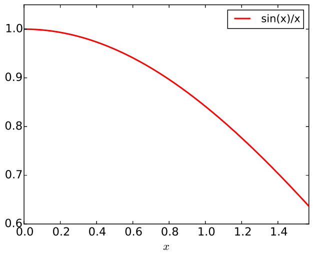

Fig. 2.8 The function $\sin(\phi)/\phi$.

$$
\begin{aligned}
|y(\boldsymbol{x})|^2 &\le \frac{4}{l_x l_y}\left(\sum_{\boldsymbol{k}} |y_{\boldsymbol{k}}|\right)^2 = \frac{4}{l_x l_y}\left(\sum_{\boldsymbol{k}} |\lambda_{\boldsymbol{k}} y_{\boldsymbol{k}}| \frac{1}{\lambda_{\boldsymbol{k}}}\right)^2 \\
&\le \frac{4}{l_x l_y} \sum_{\boldsymbol{k}} \lambda_{\boldsymbol{k}}^2 y_{\boldsymbol{k}}^2 \sum_{\boldsymbol{k}} \frac{1}{\lambda_{\boldsymbol{k}}^2} = \frac{4}{l_x l_y} \|Ay\|_h^2 \sum_{\boldsymbol{k}} \frac{1}{\lambda_{\boldsymbol{k}}^2}. \qquad (2.29)
\end{aligned}
$$

Now, one has to estimate the last sum. It is already known that

$$
\lambda_{\boldsymbol{k}} = \frac{4}{h_x^2}\sin^2\left(\frac{k_x \pi}{2n_x}\right) + \frac{4}{h_y^2}\sin^2\left(\frac{k_y \pi}{2n_y}\right), \quad k_x = 1,\ldots,n_x-1,\ k_y = 1,\ldots,n_y-1.
$$

Setting $l = \max\{l_x, l_y\}$ and $h_\alpha = l_\alpha/n_\alpha, \phi_\alpha = \frac{k_\alpha \pi}{2n_\alpha} \in (0,\pi/2)$, $\alpha \in \{x,y\}$, leads to

$$
\lambda_{\boldsymbol{k}} = \frac{k_x^2\pi^2}{l_x^2}\left(\frac{\sin\phi_x}{\phi_x}\right)^2 + \frac{k_y^2\pi^2}{l_y^2}\left(\frac{\sin\phi_y}{\phi_y}\right)^2 \ge 4\left(\frac{k_x^2}{l_x^2} + \frac{k_y^2}{l_y^2}\right) \ge \frac{4}{l^2}\left(k_x^2 + k_y^2\right).
$$

In performing this estimate, it was used that the function $\sin(\phi)/\phi$ is monotonically decreasing on $(0,\pi/2)$, see Figure 2.8, and that

$$
\frac{\sin\phi}{\phi} \ge \frac{\sin(\pi/2)}{\pi/2} = \frac{2}{\pi} \quad \forall\ \phi \in (0,\pi/2).
$$

The estimate will be continued by constructing a function that majorizes $\left(k_x^2 + k_y^2\right)^{-2}$ and that can be easily integrated. Let $G = \{(x,y)\ :\ x > 0, y > 0, x^2 + y^2 > 1\}$ be the first quadrant of the complex plane without the part that belongs to the unit circle, see Figure 2.9. The function $\left(k_x^2 + k_y^2\right)^{-2}$ has its smallest value in the square $[k_x - 1, k_x] \times [k_y - 1, k_y]$ in the point (k_x, k_y). Using the lower estimate of $\lambda_{\boldsymbol{k}}$, one obtains

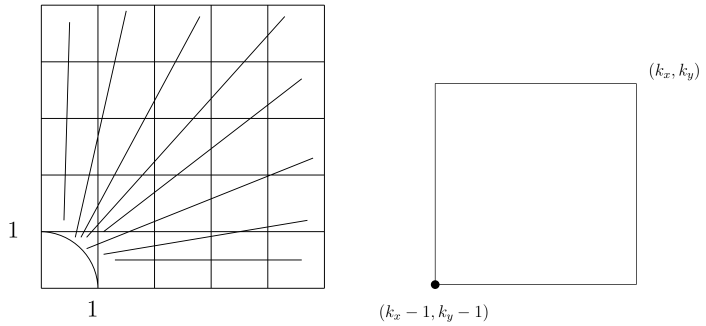

Fig. 2.9 Illustration to the proof of Lemma 2.43.

$$\begin{aligned}
\sum_{\boldsymbol{k},\boldsymbol{k}\neq(1,1)} \frac{1}{\lambda_{\boldsymbol{k}}^2} &\le \frac{l^4}{16} \sum_{\boldsymbol{k},\boldsymbol{k}\neq(1,1)} \left(k_x^2+k_y^2\right)^{-2} \\
&= \frac{l^4}{16} \sum_{\boldsymbol{k},\boldsymbol{k}\neq(1,1)} \underbrace{\left(k_x^2+k_y^2\right)^{-2}}_{\text{smallest value in square}} \underbrace{\int_{k_x-1}^{k_x}\int_{k_y-1}^{k_y} dxdy}_{=1} \\
&= \frac{l^4}{16} \sum_{\boldsymbol{k},\boldsymbol{k}\neq(1,1)} \int_{k_x-1}^{k_x}\int_{k_y-1}^{k_y} \left(k_x^2+k_y^2\right)^{-2}\ dxdy \\
&\le \frac{l^4}{16} \sum_{\boldsymbol{k},\boldsymbol{k}\neq(1,1)} \int_{k_x-1}^{k_x}\int_{k_y-1}^{k_y} \left(x^2+y^2\right)^{-2}\ dxdy \\
&\le \frac{l^4}{16} \int_G \left(x^2+y^2\right)^{-2}\ dxdy \\
&\overset{\text{polar coord.}}{=} \frac{l^4}{16} \int_1^\infty \int_0^{\pi/2} \frac{\rho}{\rho^4}\ d\phi d\rho = \frac{l^4}{16}\frac{\pi}{2}\left(-\frac{\rho^{-2}}{2}\Big|_{\rho=1}^{\rho=\infty}\right) = \frac{\pi l^4}{64}.
\end{aligned}$$

For performing this computation, one has to exclude $\rho \to 0$.

For $\lambda_{(1,1)}$, it is

$$\begin{aligned}
\lambda_{(1,1)} &= \frac{4}{h_x^2}\sin^2\left(\frac{\pi}{2n_x}\right) + \frac{4}{h_y^2}\sin^2\left(\frac{\pi}{2n_y}\right) = \frac{4}{h_x^2}\sin^2\left(\frac{h_x\pi}{2l_x}\right) + \frac{4}{h_y^2}\sin^2\left(\frac{h_y\pi}{2l_y}\right) \\
&= \frac{\pi^2}{l_x^2}\left(\frac{2l_x}{h_x\pi}\right)^2\sin^2\left(\frac{h_x\pi}{2l_x}\right) + \frac{\pi^2}{l_y^2}\left(\frac{2l_y}{h_y\pi}\right)^2\sin^2\left(\frac{h_y\pi}{2l_y}\right) \\
&\ge \frac{\pi^2}{l_x^2}\frac{8}{\pi^2} + \frac{\pi^2}{l_y^2}\frac{8}{\pi^2} \ge \frac{16}{l^2}. \qquad (2.30)
\end{aligned}$$

For this estimate, the following relations and the monotonicity of $\sin(x)/x$, see Figure 2.8, were used

$$h_\alpha \le \frac{l_\alpha}{2}, \quad \phi_\alpha = \frac{h_\alpha\pi}{2l_\alpha} \le \frac{\pi}{4}, \quad \left(\frac{\sin\phi_\alpha}{\phi_\alpha}\right)^2 \ge \left(\frac{\sin(\pi/4)}{\pi/4}\right)^2 = \frac{8}{\pi^2}.$$

Collecting all estimates gives

$$\sum_{\boldsymbol{k}} \frac{1}{\lambda_{\boldsymbol{k}}^2} = \lambda_{(1,1)}^{-2} + \sum_{\boldsymbol{k},\boldsymbol{k}\neq(1,1)} \frac{1}{\lambda_{\boldsymbol{k}}^2} \le \frac{l^4}{256} + \frac{\pi l^4}{64} \le \frac{l^4}{16}.$$

Inserting this estimate in (2.29), the final bound has the form

$$\|y\|_{l^\infty(\omega_h \cup \gamma_h)} \le \frac{2}{\sqrt{l_x l_y}} \|Ay\|_h \frac{l^2}{4} =: M \|Ay\|_h .$$

■

Theorem 2.44. Convergence of the higher order finite difference scheme. *Let $\Omega = (0,l_x) \times (0,l_y)$. The finite difference scheme*

$$\begin{aligned} -\Lambda' u(\boldsymbol{x}) &= \phi(\boldsymbol{x}),\ \boldsymbol{x} \in \omega_h, \\ u(\boldsymbol{x}) &= g(\boldsymbol{x}),\ \boldsymbol{x} \in \gamma_h, \end{aligned}$$

with

$$\Lambda' = \Lambda_x + \Lambda_y + \frac{h_x^2 + h_y^2}{12} \Lambda_x \Lambda_y, \quad \phi = f + \frac{h_x^2}{12}\Lambda_x f + \frac{h_y^2}{12}\Lambda_y f,$$

converges of fourth order.

Proof. Analogously as in the proof of Theorem 2.33, one finds that the following equation holds for the error $e = u(x_i, y_j) - u_{ij}$:

$$\begin{aligned} -\Lambda' e(\boldsymbol{x}) &= \psi(\boldsymbol{x}),\ \psi = \mathcal{O}\left(h^4\right), \boldsymbol{x} \in \omega_h, \\ e(\boldsymbol{x}) &= 0, \qquad \boldsymbol{x} \in \gamma_h. \end{aligned}$$

Let Ω_h be the vector space of grid functions, which are non-zero only in the interior, i.e., at the nodes from ω_h, and which vanish on γ_h. Let $A_\alpha y = -\Lambda_\alpha y$, $y \in \Omega_h$, $\alpha \in \{x, y\}$. The operators $A_\alpha : \Omega_h \to \Omega_h$ are linear and they have the following properties:

- They are symmetric and positive definite, i.e., $A_\alpha = A_\alpha^* > 0$, where A_α^* is the adjoint (transposed) of A_α, and $(A_\alpha u, v) = (u, A_\alpha v)$, $\forall\ u, v \in \Omega_h$.
- They are elliptic, i.e., $(A_\alpha u, u) \ge \lambda_1^{(\alpha)}(u,u)$, $\forall u \in \Omega_h$, with

  $$\lambda_1^{(\alpha)} = \frac{4}{h_\alpha^2} \sin^2\left(\frac{\pi h_\alpha}{2 l_\alpha}\right) \ge \frac{8}{l_\alpha^2},$$

  see (2.30).
- They are bounded, i.e., it holds $(A_\alpha u, u) \le \lambda_{n_\alpha - 1}^{(\alpha)}(u,u)$ with

  $$\lambda_{n_\alpha-1}^{(\alpha)} = \frac{4}{h_\alpha^2} \sin^2\left(\frac{k_\alpha \pi}{2 n_\alpha}\right) \le \frac{4}{h_\alpha^2} \quad \Longrightarrow \quad (A_\alpha u, u) \le \frac{4}{h_\alpha^2}(u,u), \qquad (2.31)$$

  and $\|A_\alpha\|_2 \le 4/h_\alpha^2$, since the spectral norm of a symmetric positive definite matrix equals the largest eigenvalue.
- They are commutative, i.e., $A_x A_y = A_y A_x$.
- It holds $A_x A_y = (A_x A_y)^*$.

The error equation on ω_h is given by

$$A_x e + A_y e - (\kappa_x + \kappa_y) A_x A_y e = A' e = \psi \quad \text{with} \quad \kappa_\alpha = \frac{h_\alpha^2}{12}. \qquad (2.32)$$

Using the boundedness of the operators, one finds with (2.31) for all $v \in \Omega_h$ that

$$\begin{aligned}(\kappa_x A_x A_y v + \kappa_y A_x A_y v, v) &= ((\kappa_x A_x)\, A_y v, v) + ((\kappa_y A_y)\, A_x v, v) \\ &\le \frac{h_x^2}{12}\frac{4}{h_x^2}\,(A_y v, v) + \frac{h_y^2}{12}\frac{4}{h_y^2}\,(A_x v, v) \\ &= \frac{1}{3}\,((A_x + A_y)\, v, v)\,.\end{aligned}$$

Now, it follows for all $v \in \Omega_h$ that

$$\begin{aligned}(A' v, v) &= ((A_x + A_y)\, v, v) - (\kappa_x A_x A_y v + \kappa_y A_x A_y v, v) \\ &\ge \frac{2}{3}\,((A_x + A_y)\, v, v) \ge 0.\end{aligned}$$

The matrices on both sides of this inequality are symmetric and because the matrix on the lower estimate is positive definite, also the matrix at the upper estimate is positive definite. The matrices commute since the order of applying the finite differences in x and y direction does not matter. Using these properties, one gets (*exercise*)

$$\left\| \frac{2}{3}\,(A_x + A_y)\, e \right\|_h \le \left\| A' e \right\|_h = \| \psi \|_h\,, \tag{2.33}$$

where the last equality follows from (2.32). The application of Lemma 2.43 to the error, (2.33), (2.32), and (2.24) yields

$$\begin{aligned}\|e\|_{l^\infty(\omega_h \cup \gamma_h)} &\le \frac{l^2}{2\sqrt{l_x l_y}}\, \|(\Lambda_x + \Lambda_y)\, e\|_h \le \frac{3l^2}{4\sqrt{l_x l_y}}\, \left\|A' e\right\|_h = \frac{3l^2}{4\sqrt{l_x l_y}}\, \|\psi\|_h \\ &\le \frac{3l^2}{4\sqrt{l_x l_y}}\, (h_x h_y (n_x+1)(n_y+1))^{1/2}\, \|\psi\|_{l^\infty(\omega_h \cup \gamma_h)} \\ &= \frac{3l^2}{4}\left(\frac{n_x+1}{n_x}\frac{n_y+1}{n_y}\right)^{1/2} \|\psi\|_{l^\infty(\omega_h \cup \gamma_h)} = \mathcal{O}\left(h^4\right).\end{aligned}$$

■

Remark 2.45. On the discrete maximum principle. Reformulation of the finite difference scheme $-\Lambda' u = \phi$ in the form studied for the discrete maximum principle gives for the node (i, j)

$$\begin{aligned}a(\boldsymbol{x})u(\boldsymbol{x}) &= \sum_{\boldsymbol{y} \in S(\boldsymbol{x})} b(\boldsymbol{x}, \boldsymbol{y})u(\boldsymbol{y}) + \phi(\boldsymbol{x}), \\ a(\boldsymbol{x}) &= \frac{2}{h_x^2} + \frac{2}{h_y^2} - \frac{1}{12}\left(h_x^2 + h_y^2\right)\frac{4}{h_x^2 h_y^2} = \frac{5}{3}\left(\frac{1}{h_x^2} + \frac{1}{h_y^2}\right) > 0, \\ b(\boldsymbol{x}, \boldsymbol{y}) &= \frac{1}{h_x^2} - \frac{1}{12}\left(h_x^2 + h_y^2\right)\frac{2}{h_x^2 h_y^2} = \frac{1}{6}\left(\frac{5}{h_x^2} - \frac{1}{h_y^2}\right),\ i \pm 1, j, \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(left, right node)} \\ b(\boldsymbol{x}, \boldsymbol{y}) &= \frac{1}{6}\left(-\frac{1}{h_x^2} + \frac{5}{h_y^2}\right),\ i, j \pm 1,\ \text{(bottom, top node)} \\ b(\boldsymbol{x}, \boldsymbol{y}) &= \frac{1}{12}\left(\frac{1}{h_x^2} + \frac{1}{h_y^2}\right),\ i \pm 1, j \pm 1,\ \text{(other neighbors)}.\end{aligned}$$

Hence, the assumptions for the discrete maximum principle, see Remark 2.15, are satisfied only if

$$\frac{1}{\sqrt{5}} < \frac{h_x}{h_y} < \sqrt{5}.$$

Consequently, the ratio of the grid widths has to be bounded and it has to be of order one. In this case, one speaks of an isotropic grid. □

2.7 Summary

Remark 2.46. Summary.

- Finite difference methods are the simplest approach for discretizing partial differential equations. The derivatives are just approximated by difference quotients.
- They are very popular in the engineering community.
- One large drawback are the difficulties in approximating domains that are not of tensor-product type. However, in the engineering communities, a number of strategies have been developed to deal with this issue in practice.
- Another drawback arises from the point of view of numerical analysis. The numerical analysis of finite difference methods is mainly based on Taylor series expansions. For this tool to be applicable, one has to assume a high regularity of the solution. These assumptions are generally not realistic.
- In Numerical Mathematics, one considers often other schemes then finite difference methods. However, there are problems, where finite difference methods can compete with other discretizations.

□

Chapter 3
Introduction to Sobolev Spaces

Remark 3.1. Contents. Sobolev spaces are the basis of the theory of weak or variational forms of partial differential equations. A very popular approach for discretizing partial differential equations, the finite element method, is based on variational forms. In this chapter, a short introduction into Sobolev spaces will be given. Recommended literature are the books Adams (1975); Adams & Fournier (2003), and Evans (2010). □

3.1 Elementary Inequalities

Lemma 3.2. Inequality for strictly monotonically increasing function. *Let $f : \mathbb{R}_+ \cup \{0\} \to \mathbb{R}$ be a continuous and strictly monotonically increasing function with $f(0) = 0$ and $f(x) \to \infty$ for $x \to \infty$. Then, for all $a, b \in \mathbb{R}_+ \cup \{0\}$, it is*

$$ab \le \int_0^a f(x)\ dx + \int_0^b f^{-1}(y)\ dy,$$

where $f^{-1}(y)$ is the inverse of $f(x)$.

Proof. Since $f(x)$ is strictly monotonically increasing, the inverse function exists.

The proof is based on a geometric argument, see Figure 3.1.

Consider the interval $(0, a)$ on the x-axis and the interval $(0, b)$ on the y-axis. Then, the area of the corresponding rectangle is given by ab, $\int_0^a f(x)\ dx$ is the area below the curve, and $\int_0^b f^{-1}(y)\ dy$ is the area between the positive y-axis and the curve. From Figure 3.1, the inequality follows immediately. The equal sign holds only iff $f(a) = b$. ■

Remark 3.3. Young's[1] inequality. Young's inequality

$$ab \le \frac{\varepsilon}{2}a^2 + \frac{1}{2\varepsilon}b^2 \quad \forall\, a, b \in \mathbb{R}_+ \cup \{0\}, \varepsilon \in \mathbb{R}_+, \tag{3.1}$$

[1] William Henry Young (1863 – 1942)

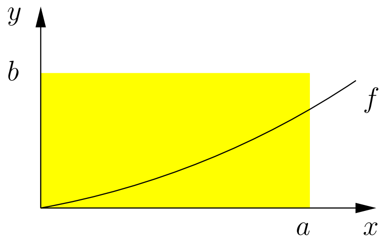

Fig. 3.1 Sketch to the proof of Lemma 3.2.

follows from Lemma 3.2 with $f(x) = \varepsilon x$, $f^{-1}(y) = \varepsilon^{-1}y$. It is also possible to derive this inequality from the binomial theorem. For proving the generalized Young inequality

$$ab \le \frac{\varepsilon^p}{p}a^p + \frac{1}{q\varepsilon^q}b^q, \quad \forall\, a,b \in \mathbb{R}_+ \cup \{0\}, \varepsilon \in \mathbb{R}_+, \tag{3.2}$$

with $p^{-1}+q^{-1} = 1, p,q \in (1,\infty)$, one chooses $f(x) = x^{p-1}$, $f^{-1}(y) = y^{1/(p-1)}$ and applies Lemma 3.2 with intervals where the upper bounds are given by εa and $\varepsilon^{-1}b$. □

Remark 3.4. Cauchy–Schwarz inequality.

- The Cauchy[2]–Schwarz[3] inequality (for vectors, for sums)

$$\left|(\underline{x},\underline{y})\right| \le \|\underline{x}\|_2 \left\|\underline{y}\right\|_2 \ \forall\, \underline{x},\underline{y} \in \mathbb{R}^n, \tag{3.3}$$

  where $(\cdot,\cdot)$ is the Euclidean product and $\|\cdot\|_2$ the Euclidean norm, is well known.
- One can prove this inequality with the help of Young's inequality. First, it is clear that the Cauchy–Schwarz inequality is correct if one of the vectors is the zero vector. Now, let $\underline{x},\underline{y}$ with $\|\underline{x}\|_2 = \left\|\underline{y}\right\|_2 = 1$. One obtains with the triangle inequality and Young's inequality (3.1)

$$\left|(\underline{x},\underline{y})\right| = \left|\sum_{i=1}^n x_i y_i\right| \le \sum_{i=1}^n |x_i|\,|y_i| \le \frac{1}{2}\sum_{i=1}^n |x_i|^2 + \frac{1}{2}\sum_{i=1}^n |y_i|^2 = 1.$$

  Hence, the Cauchy–Schwarz inequality is correct for $\underline{x},\underline{y}$. Last, one considers arbitrary vectors $\underline{\tilde{x}} \ne \underline{0}, \underline{\tilde{y}} \ne \underline{0}$. Now, one can utilize the homogeneity of the Cauchy–Schwarz inequality. From the validity of the Cauchy–Schwarz inequality for $\underline{x}$ and $\underline{y}$, one obtains by a scaling argument

[2] Augustin Louis Cauchy (1789 – 1857)

[3] Hermann Amandus Schwarz (1843 – 1921)

$$\big|(\underbrace{\|\underline{\tilde{x}}\|_2^{-1}\,\underline{\tilde{x}}}_{\underline{x}},\underbrace{\|\underline{\tilde{y}}\|_2^{-1}\,\underline{\tilde{y}}}_{\underline{y}})\big| \le 1$$

Both vectors $\underline{x}, \underline{y}$ have the Euclidean norm 1, hence

$$\frac{1}{\|\underline{\tilde{x}}\|_2 \|\underline{\tilde{y}}\|_2}\big|(\underline{\tilde{x}},\underline{\tilde{y}})\big| \le 1 \quad \Longleftrightarrow \quad \big|(\underline{\tilde{x}},\underline{\tilde{y}})\big| \le \|\underline{\tilde{x}}\|_2 \|\underline{\tilde{y}}\|_2 .$$

- The generalized Cauchy–Schwarz inequality or Hölder's[4] inequality

$$\big|(\underline{x},\underline{y})\big| \le \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} \left(\sum_{i=1}^n |y_i|^q\right)^{1/q}$$

with $p^{-1} + q^{-1} = 1, p, q \in (1,\infty)$, can be proved in the same way with the help of the generalized Young inequality.

□

Definition 3.5. *Lebesgue spaces.* The space of functions that are Lebesgue[5] integrable on Ω to the power of $p \in [1,\infty)$ is denoted by

$$L^p(\Omega) = \left\{ f \ : \ \int_\Omega |f(\boldsymbol{x})|^p \ d\boldsymbol{x} < \infty \right\},$$

which is equipped with the norm

$$\|f\|_{L^p(\Omega)} = \left(\int_\Omega |f(\boldsymbol{x})|^p \ d\boldsymbol{x} \right)^{1/p} .$$

For $p = \infty$, this space is given by

$$L^\infty(\Omega) = \{ f \ : \ |f(\boldsymbol{x})| < \infty \text{ almost everywhere in } \Omega \}$$

with the norm

$$\|f\|_{L^\infty(\Omega)} = \text{ess sup}_{\boldsymbol{x}\in\Omega} |f(\boldsymbol{x})|.$$

□

Lemma 3.6. Hölder's inequality. *Let $p^{-1} + q^{-1} = 1, p, q \in [1,\infty]$. If $u \in L^p(\Omega)$ and $v \in L^q(\Omega)$, then it is $uv \in L^1(\Omega)$ and it holds that*

$$\|uv\|_{L^1(\Omega)} \le \|u\|_{L^p(\Omega)} \|v\|_{L^q(\Omega)} . \tag{3.4}$$

If $p = q = 2$, then this inequality is also known as Cauchy–Schwarz inequality

[4] Otto Hölder (1859 – 1937)

[5] Henri Lebesgue (1875 – 1941)

$$\|uv\|_{L^1(\Omega)} \le \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} . \tag{3.5}$$

Proof. *i)* $p, q \in (1, \infty)$. First, one has to show that $|uv(\boldsymbol{x})|$ can be estimated from above by an integrable function. Setting in the generalized Young inequality (3.2) $\varepsilon = 1$, $a = |u(\boldsymbol{x})|$, and $b = |v(\boldsymbol{x})|$ gives

$$|u(\boldsymbol{x})v(\boldsymbol{x})| \le \frac{1}{p} |u(\boldsymbol{x})|^p + \frac{1}{q} |v(\boldsymbol{x})|^q . \tag{3.6}$$

Since the right-hand side of this inequality is integrable, by assumption, it follows that $uv \in L^1(\Omega)$. Integrating (3.6), Hölder's inequality is proved for the case $\|u\|_{L^p(\Omega)} = \|v\|_{L^q(\Omega)} = 1$

$$\int_\Omega |u(\boldsymbol{x})v(\boldsymbol{x})| \ d\boldsymbol{x} \le \frac{1}{p} \int_\Omega |u(\boldsymbol{x})|^p \ d\boldsymbol{x} + \frac{1}{q} \int_\Omega |v(\boldsymbol{x})|^q \ d\boldsymbol{x} = 1.$$

The general inequality follows, for the case that both functions do not vanish almost everywhere, with the same homogeneity argument as used for proving the Cauchy–Schwarz inequality of sums. In the case that one of the functions vanishes almost everywhere, (3.4) is trivially satisfied.

ii) $p = 1, q = \infty$. It is

$$\int_\Omega |u(\boldsymbol{x})v(\boldsymbol{x})| \ d\boldsymbol{x} \le \int_\Omega |u(\boldsymbol{x})| \operatorname{ess\,sup}_{\boldsymbol{x}\in\Omega} |v(\boldsymbol{x})| \ d\boldsymbol{x} = \|u\|_{L^1(\Omega)} \|v\|_{L^\infty(\Omega)} .$$

■

3.2 Weak Derivative and Distributions

Remark 3.7. Contents. This section introduces a generalization of the derivative which is needed for the definition of weak or variational problems. For an introduction to the topic of this section, e.g., see Haroske & Triebel (2008)

Let $\Omega \subset \mathbb{R}^d$ be a domain with boundary $\Gamma = \partial\Omega$, $d \in \mathbb{N}$, $\Omega \neq \emptyset$. A domain is always an open set. □

Definition 3.8. The space $C_0^\infty(\Omega)$. The space of infinitely often differentiable real functions with compact (closed and bounded) support in Ω is denoted by $C_0^\infty(\Omega)$

$$C_0^\infty(\Omega) = \{v \ : \ v \in C^\infty(\Omega), \ \operatorname{supp}(v) \subset \Omega\},$$

where

$$\operatorname{supp}(v) = \overline{\{\boldsymbol{x} \in \Omega \ : \ v(\boldsymbol{x}) \neq 0\}}.$$

In particular, functions from $C_0^\infty(\Omega)$ vanish in a neighborhood of the boundary. □

Definition 3.9. Convergence in $C_0^\infty(\Omega)$. The sequence $\{\phi_n(\boldsymbol{x})\}_{n=1}^\infty$, $\phi_n \in C_0^\infty(\Omega)$ for all n, is said to convergence to the zero functions if and only if

a) $\exists K \subset \Omega, K$ compact with $\operatorname{supp}(\phi_n) \subset K$ for all n,

b) $D^{\boldsymbol{\alpha}}\phi_n(\boldsymbol{x}) \to 0$ for $n \to \infty$ on K for all multi-indices $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$, $|\boldsymbol{\alpha}| = \alpha_1 + \ldots + \alpha_d$.

It is

$$\lim_{n\to\infty} \phi_n = \phi \quad \Longleftrightarrow \quad \lim_{n\to\infty} (\phi_n - \phi) = 0.$$

□

Definition 3.10. Weak derivative. Let $f, F \in L^1_{\text{loc}}(\Omega)$. A function u belongs to $L^1_{\text{loc}}(\Omega)$ if for each compact subset $\Omega' \subset \Omega$, it holds

$$\int_{\Omega'} |u(\boldsymbol{x})| \ d\boldsymbol{x} < \infty.$$

If for all functions $g \in C_0^\infty(\Omega)$, it holds that

$$\int_\Omega F(\boldsymbol{x})g(\boldsymbol{x}) \ d\boldsymbol{x} = (-1)^{|\boldsymbol{\alpha}|} \int_\Omega f(\boldsymbol{x}) D^{\boldsymbol{\alpha}} g(\boldsymbol{x}) \ d\boldsymbol{x},$$

then $F(\boldsymbol{x})$ is called weak derivative of $f(\boldsymbol{x})$ with respect to the multi-index $\boldsymbol{\alpha}$. □

Remark 3.11. On the weak derivative.

- The notion 'weak' is used in mathematics if something holds for all appropriate test elements (test functions).
- One uses the same notations for the derivative as in the classical case : $F(\boldsymbol{x}) = D^{\boldsymbol{\alpha}} f(\boldsymbol{x})$.
- If $f(\boldsymbol{x})$ is classically differentiable on Ω, then the classical derivative is also the weak derivative.
- The assumptions on $f(\boldsymbol{x})$ and $F(\boldsymbol{x})$ are such that the integrals in the definition of the weak derivative are well defined. In particular, since the test functions vanish in a neighborhood of the boundary, the behavior of $f(\boldsymbol{x})$ and $F(\boldsymbol{x})$ if $\boldsymbol{x}$ approaches the boundary is not of importance.
- The main aspect of the weak derivative is due to the fact that the (Lebesgue) integral is not influenced from the values of the functions on a set of (Lebesgue) measure zero. Hence, the weak derivative is defined only up to a set of measure zero. It follows that $f(\boldsymbol{x})$ might be not classically differentiable on a set of measure zero, e.g., in a point, but it can still be weakly differentiable.
- The weak derivative is uniquely determined, in the sense described above.

□

Example 3.12. Weak derivative. The weak derivative of the function $f(x) = |x|$ is

$$f'(x) = \begin{cases} -1 & x < 0, \\ 0 & x = 0, \\ 1 & x > 0. \end{cases}$$

In $x = 0$, one can use also any other real number. The proof of this statement follows directly from the definition and it is left as an exercise. □

Definition 3.13. Distribution. A continuous linear functional defined on $C_0^\infty(\Omega)$ is called distribution. The set of all distributions is denoted by $(C_0^\infty(\Omega))'$.

Let $u \in C_0^\infty(\Omega)$ and $\psi \in (C_0^\infty(\Omega))'$, then the following notations are used for the application of the distribution to the function

$$\psi(u) = \langle \psi, u \rangle \in \mathbb{R}.$$

□

Remark 3.14. On distributions. Distributions are a generalization of functions. They assign each function from $C_0^\infty(\Omega)$ a real number. □

Example 3.15. Regular distribution. Let $u \in L^1_{\mathrm{loc}}(\Omega)$. Then, a distribution is defined by

$$\int_\Omega u(\boldsymbol{x})\phi(\boldsymbol{x})\ d\boldsymbol{x} = \langle \psi, \phi \rangle \quad \forall\ \phi \in C_0^\infty(\Omega).$$

This distribution will be identified with $u \in L^1_{\mathrm{loc}}(\Omega)$.

Distributions with such an integral representation are called regular, otherwise they are called singular. □

Example 3.16. Dirac distribution. Let $\boldsymbol{\xi} \in \Omega$ be fixed, then

$$\langle \delta_{\boldsymbol{\xi}}, \phi \rangle = \phi(\boldsymbol{\xi}) \quad \forall\ \phi \in C_0^\infty(\Omega)$$

defines a singular distribution, the so-called Dirac[6] distribution or δ-distribution. It is denoted by $\delta_{\boldsymbol{\xi}} = \delta(\boldsymbol{x} - \boldsymbol{\xi})$. □

Definition 3.17. *Derivatives of distributions.* Let $\phi \in (C_0^\infty(\Omega))'$ be a distribution. The distribution $\psi \in (C_0^\infty(\Omega))'$ is called derivative in the sense of distributions or distributional derivative of ϕ if

$$\langle \psi, u \rangle = (-1)^{|\boldsymbol{\alpha}|} \langle \phi, D^{\boldsymbol{\alpha}} u \rangle \quad \forall\ u \in C_0^\infty(\Omega),$$

$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$, $\alpha_j \geq 0, j = 1, \ldots, d$, $|\boldsymbol{\alpha}| = \alpha_1 + \ldots + \alpha_d$. □

Remark 3.18. On derivatives of distributions.

- Each distribution has derivatives in the sense of distributions of arbitrary order.
- If the derivative in the sense of distributions $D^{\boldsymbol{\alpha}} u(\boldsymbol{x})$ with $u \in L^1_{\mathrm{loc}}(\Omega)$ is a regular distribution, then also the weak derivative of $u(\boldsymbol{x})$ exists and both derivatives are identified.

□

[6] Paul Adrien Maurice Dirac (1902 – 1984)

3.3 Lebesgue Spaces and Sobolev Spaces

Remark 3.19. On the spaces $L^p(\Omega)$. These spaces were introduced in Definition 3.5.

- The elements of $L^p(\Omega)$ are, strictly speaking, equivalence classes of functions that are different only on a set of Lebesgue measure zero.
- The spaces $L^p(\Omega)$ are Banach[7] spaces (complete normed spaces). A space X is complete, if each so-called Cauchy sequence $\{u_n\}_{n=0}^\infty \in X$, i.e., for all $\varepsilon > 0$ there is an index $n_0(\varepsilon)$ such that for all $i, j > n_0(\varepsilon)$
$$\|u_i - u_j\|_X < \varepsilon,$$
converges and the limit is an element of X.
- The space $L^2(\Omega)$ becomes a Hilbert[8] spaces with the inner product
$$(f, g) = \int_\Omega f(\boldsymbol{x})g(\boldsymbol{x})\ d\boldsymbol{x}, \quad \|f\|_{L^2} = (f, f)^{1/2}, \quad f, g \in L^2(\Omega).$$
- The dual space of a space X is the space of all bounded linear functionals defined on X. Let Ω be a domain with sufficiently smooth boundary Γ and consider the Lebesgue space $L^p(\Omega)$, $p \in [1, \infty]$, then
$$\begin{aligned} (L^p(\Omega))' &= L^q(\Omega) \text{ with } p, q \in (1, \infty), \ \frac{1}{p} + \frac{1}{q} = 1, \\ \left(L^1(\Omega)\right)' &= L^\infty(\Omega), \\ (L^\infty(\Omega))' &\neq L^1(\Omega), \end{aligned}$$
where the prime symbolizes the dual space. The spaces $L^1(\Omega)$, $L^\infty(\Omega)$ are not reflexive, i.e., the dual space of the dual space is not the original space again.

□

Definition 3.20. Sobolev[9] spaces. Let $k \in \mathbb{N} \cup \{0\}$ and $p \in [1, \infty]$, then the Sobolev space $W^{k,p}(\Omega)$ is defined by
$$W^{k,p}(\Omega) := \{u \in L^p(\Omega) \ : \ D^{\boldsymbol{\alpha}} u \in L^p(\Omega) \ \forall \ \boldsymbol{\alpha} \text{ with } |\boldsymbol{\alpha}| \le k\}.$$

This space is equipped with the norm
$$\|u\|_{W^{k,p}(\Omega)} := \sum_{|\boldsymbol{\alpha}| \le k} \|D^{\boldsymbol{\alpha}} u\|_{L^p(\Omega)} . \tag{3.7}$$

□

[7] Stefan Banach (1892 – 1945)

[8] David Hilbert (1862 – 1943)

[9] Sergei Lvovich Sobolev (1908 – 1989)

Remark 3.21. On the spaces $W^{k,p}(\Omega)$.

- Definition 3.20 has the following meaning. From $u \in L^p(\Omega)$, $p \in [1,\infty)$, it follows in particular that $u \in L^1_{\mathrm{loc}}(\Omega)$, such that u defines (represents) a distribution. Then, all derivatives $D^{\boldsymbol{\alpha}} u$ exist in the sense of distributions. The statement $D^{\boldsymbol{\alpha}} u \in L^p(\Omega)$ means that the distribution $D^{\boldsymbol{\alpha}} u \in (C_0^\infty(\Omega))'$ can be represented by a function from $L^p(\Omega)$.
- One can add elements from $W^{k,p}(\Omega)$ and one can multiply them with real numbers. The result is again a function from $W^{k,p}(\Omega)$. With this property, the space $W^{k,p}(\Omega)$ becomes a vector space (linear space). It is straightforward to check that (3.7) is a norm. (*exercise*)
- It is $D^{\boldsymbol{\alpha}} u(\boldsymbol{x}) = u(\boldsymbol{x})$ for $\boldsymbol{\alpha} = (0,\ldots,0)$ and $W^{0,p}(\Omega) = L^p(\Omega)$.
- The spaces $W^{k,p}(\Omega)$ are Banach spaces.
- Sobolev spaces have for $p \in [1,\infty)$ a countable basis $\{\varphi_n(\boldsymbol{x})\}_{n=1}^\infty$ (Schauder[10] basis), i.e., each element $u(\boldsymbol{x})$ can be written in the form

$$u(\boldsymbol{x}) = \sum_{n=1}^\infty u_n \varphi_n(\boldsymbol{x}), \quad u_n \in \mathbb{R},\ n = 1,2,\ldots\ .$$

- Sobolev spaces are reflexive for $p \in (1,\infty)$.
- The subspace $C^\infty(\Omega) \cap W^{k,p}(\Omega)$ is dense in $W^{k,p}(\Omega)$, see (Gilbarg & Trudinger, 1983, p. 154). Under a certain condition on the smoothness of the boundary of a bounded domain Ω, one can show that $C_0^\infty(\Omega)$ is dense in $W^{k,p}(\Omega)$, $p \in [1,\infty)$, with respect to the norm (3.7), e.g., (Adams, 1975, Thm. 3.18). With this property, one can characterize the Sobolev spaces $W^{k,p}(\Omega)$ as completion of the functions from $C_0^\infty(\Omega)$ with respect to the norm (3.7). It follows that $C^k(\overline{\Omega})$ is dense in $W^{k,p}(\Omega)$, $p \in [1,\infty)$.
- The Sobolev space $H^k(\Omega) = W^{k,2}(\Omega)$ is a Hilbert space with the inner product

$$(u,v)_{H^k(\Omega)} = \sum_{|\boldsymbol{\alpha}|\le k} \int_\Omega D^{\boldsymbol{\alpha}} u(\boldsymbol{x}) D^{\boldsymbol{\alpha}} v(\boldsymbol{x})\ d\boldsymbol{x}$$

and the induced norm $\|u\|_{H^k(\Omega)} = (u,u)_{H^k(\Omega)}^{1/2}$.

□

Definition 3.22. The space $W_0^{k,p}(\Omega)$. The Sobolev space $W_0^{k,p}(\Omega)$ is defined as the completion of $C_0^\infty(\Omega)$ in the norm of $W^{k,p}(\Omega)$

$$W_0^{k,p}(\Omega) = \overline{C_0^\infty(\Omega)}^{\|\cdot\|_{W^{k,p}(\Omega)}}.$$

□

[10] Juliusz Pawel Schauder (1899 – 1943)

3.4 The Trace of a Function from a Sobolev Space

Remark 3.23. Motivation. This class considers boundary value problems for partial differential equations. In the theory of weak or variational solutions, the solution of the partial differential equation is searched in an appropriate Sobolev space. Then, for the boundary value problem, this solution has to satisfy the boundary condition. However, since the boundary of a domain is a manifold of dimension $(d-1)$, and consequently it has Lebesgue measure zero, one has to clarify how a function from a Sobolev space is defined on this manifold. This definition will be presented in this section. □

Definition 3.24. Lipschitz boundary, Lipschitz domain, (Grisvard, 1985, Def. 1.2.1.1). Let Ω be a bounded domain in $\mathbb{R}^d$, then Ω is called Lipschitz[11] domain, respectively the boundary Γ of Ω is called Lipschitz boundary, if for every $\boldsymbol{x} \in \Gamma$ there exists a neighborhood U of $\boldsymbol{x}$ in $\mathbb{R}^d$ and new orthogonal coordinates $(y_1, \ldots, y_d)$ such that

1) U is a hypercube in the new coordinates

$$U = \{(y_1, \ldots, y_d) \ : \ -a_i < y_i < a_i, \ i = 1, \ldots, d\}.$$

2) There exists a Lipschitz continuous function ϕ, defined in

$$U' = \{(y_1, \ldots, y_{d-1}) \ : \ -a_i < y_i < a_i, \ i = 1, \ldots, d-1\},$$

such that

$$\begin{aligned} |\phi(\boldsymbol{y}')| &\le \frac{a_n}{2} \text{ for every } \boldsymbol{y}' = (y_1, \ldots, y_{d-1}) \in U', \\ \Omega \cap U &= \{\boldsymbol{y} = (\boldsymbol{y}', y_n) \in V \ : \ y_n < \phi(\boldsymbol{y}')\}, \\ \Gamma \cap U &= \{\boldsymbol{y} = (\boldsymbol{y}', y_n) \in V \ : \ y_n = \phi(\boldsymbol{y}')\}. \end{aligned}$$

□

Remark 3.25. Lipschitz boundary.

- In a neighborhood of $\boldsymbol{y}$, Ω is below the graph of ϕ und the boundary Γ is the graph of ϕ.
- The domain Ω is not on both sides of the boundary at any point of Γ.
- The outer normal vector is defined almost everywhere at the boundary and it is almost everywhere continuous.

□

Example 3.26. On Lipschitz domains.

- Domains with Lipschitz boundary are, for example, balls or polygonal domains in two dimensions where the domain is always on one side of the boundary.

[11] Rudolf Otto Sigismund Lipschitz (1832 – 1903)

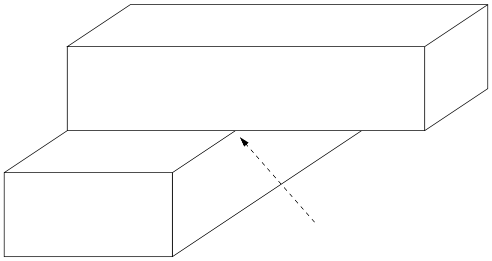

Fig. 3.2 Polyhedral domain in three dimensions that is not Lipschitz continuous (at the corner where the arrow points to).

- A domain that is not a Lipschitz domain is a circle with a slit
$$\Omega = \{(x,y)\ :\ x^2+y^2<1\} \setminus \{(x,y)\ :\ x \ge 0, y=0\}.$$
At the slit, the domain is on both sides of the boundary.
- In three dimension, a polyhedral domain is not not necessarily a Lipschitz domain. For instance, if the domain is build of two bricks that are laying on each other like in Figure 3.2, then the boundary is not Lipschitz continuous where the edge of one brick meets the edge of the other brick.

□

Theorem 3.27. Trace theorem. *Let $\Omega \subset \mathbb{R}^d$, $d \ge 2$, with a Lipschitz boundary. Then, there is exactly one linear and continuous operator γ : $W^{1,p}(\Omega) \to L^p(\Gamma)$, $p \in [1,\infty)$, that gives for functions $u \in C(\overline{\Omega}) \cap W^{1,p}(\Omega)$ the classical boundary values*
$$\gamma u(\boldsymbol{x}) = u(\boldsymbol{x}),\ \boldsymbol{x} \in \Gamma,\ \forall\, u \in C(\overline{\Omega}) \cap W^{1,p}(\Omega),$$
i.e., $\gamma u(\boldsymbol{x}) = u(\boldsymbol{x})|_{\boldsymbol{x}\in\Gamma}$.

Proof. The proof can be found in the literature, e.g., in Adams (1975); Adams & Fournier (2003). ■

Remark 3.28. On the trace.
- The operator γ is called trace or trace operator.
- By definition of the trace, one gets for $u \in C(\overline{\Omega})$ the classical boundary values. By the density of $C^\infty(\overline{\Omega}) \subset C(\overline{\Omega})$ in $W^{1,p}(\Omega)$ for domains with smooth boundary that for all $u \in W^{1,p}(\Omega)$ there is a sequence $\{u_n\}_{n=1}^\infty \in C^\infty(\overline{\Omega})$ with $u_n \to u$ in $W^{1,p}(\Omega)$. Then, the trace of u is defined to be $\gamma u = \lim_{n\to\infty}(\gamma u_n)$.
- Since a linear and continuous operator is bounded, there is a constant $C > 0$ with
$$\|\gamma u\|_{L^p(\Gamma)} \le C\,\|u\|_{W^{1,p}(\Omega)}\ \ \forall\, u \in W^{1,p}(\Omega)$$

or

$$\|\gamma\|_{\mathcal{L}(W^{1,p}(\Omega),L^p(\Gamma))} \le C.$$

- It is

$$\begin{aligned} \gamma u(\boldsymbol{x}) &= 0 \quad \forall\, u \in W_0^{1,p}(\Omega), \\ \gamma D^{\boldsymbol{\alpha}} u(\boldsymbol{x}) &= 0 \quad \forall\, u \in W_0^{k,p}(\Omega), |\boldsymbol{\alpha}| \le k-1. \end{aligned} \tag{3.8}$$

□

3.5 Sobolev Spaces with Non-Integer and Negative Exponents

Remark 3.29. Motivation. Sobolev spaces with non-integer and negative exponents are important in the theory of variational formulations of partial differential equations.

Let $\Omega \subset \mathbb{R}^d$ be a domain and $p \in (1,\infty)$ with $p^{-1} + q^{-1} = 1$. □

Definition 3.30. The space $W^{-k,q}(\Omega)$. The space $W^{-k,q}(\Omega), k \in \mathbb{N} \cup \{0\}$, contains distributions that are defined on $W^{k,p}(\Omega)$

$$W^{-k,q}(\Omega) = \left\{\varphi \in (C_0^\infty(\Omega))' \ : \ \|\varphi\|_{W^{-k,q}(\Omega)} < \infty\right\}$$

with

$$\|\varphi\|_{W^{-k,q}(\Omega)} = \sup_{u \in C_0^\infty(\Omega), u \ne 0} \frac{\langle \varphi, u\rangle}{\|u\|_{W^{k,p}(\Omega)}}.$$

□

Remark 3.31. On the spaces $W^{-k,p}(\Omega)$.

- It is $W^{-k,q}(\Omega) = \left[W_0^{k,p}(\Omega)\right]'$, i.e., $W^{-k,q}(\Omega)$ can be identified with the dual space of $W_0^{k,p}(\Omega)$. In particular, it is $H^{-1}(\Omega) = \left(H_0^1(\Omega)\right)'$.
- It is

$$\ldots \subset W^{2,p}(\Omega) \subset W^{1,p}(\Omega) \subset L^p(\Omega) \subset W^{-1,q}(\Omega) \subset W^{-2,q}(\Omega) \ldots$$

□

Definition 3.32. Sobolev–Slobodeckij space. Let $s \in \mathbb{R}$, then the Sobolev–Slobodeckij[12] or Sobolev space $H^s(\Omega)$ is defined as follows:

- $s \in \mathbb{Z}$. $H^s(\Omega) = W^{s,2}(\Omega)$.
- $s > 0$ with $s = k + \sigma$, $k \in \mathbb{N} \cup \{0\}$, $\sigma \in (0,1)$. The space $H^s(\Omega)$ contains all functions u for which the following norm is finite:

[12] L. N. Slobodeckij

$$\|u\|^2_{H^s(\Omega)} = \|u\|^2_{H^k(\Omega)} + |u|^2_{k+\sigma} ,$$

with

$$(u,v)_{H^s(\Omega)} = (u,v)_{H^k} + (u,v)_{k+\sigma}, \quad |u|^2_{k+\sigma} = (u,u)_{k+\sigma},$$

and

$$(u,v)_{k+\sigma} = \sum_{|\boldsymbol{\alpha}|=k} \int_\Omega \int_\Omega \frac{\left(D^{\boldsymbol{\alpha}} u(\boldsymbol{x}) - D^{\boldsymbol{\alpha}} u(\boldsymbol{y})\right)\left(D^{\boldsymbol{\alpha}} v(\boldsymbol{x}) - D^{\boldsymbol{\alpha}} v(\boldsymbol{y})\right)}{\|\boldsymbol{x}-\boldsymbol{y}\|_2^{d+2\sigma}} \, d\boldsymbol{x} d\boldsymbol{y},$$

- $s < 0$. $H^s(\Omega) = \left[H_0^{-s}(\Omega)\right]'$ with $H_0^{-s}(\Omega) = \overline{C_0^\infty(\Omega)}^{\|\cdot\|_{H^{-s}(\Omega)}}$.

□

3.6 Theorem on Equivalent Norms

Definition 3.33. Equivalent norms. Two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ on the linear space X are said to be equivalent if there are constants C_1 and C_2 such that

$$C_1 \|u\|_1 \le \|u\|_2 \le C_2 \|u\|_1 \;\; \forall\, u \in X.$$

□

Remark 3.34. On equivalent norms.

- Many important properties, like continuity or convergence, do not change if an equivalent norm is considered.
- In finite-dimensional spaces, all norms are equivalent.

□

Theorem 3.35. Equivalent norms in $W^{k,p}(\Omega)$ (Smirnow, 1967, § 114, Satz 3). *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary Γ, $p \in [1,\infty]$, and $k \in \mathbb{N}$. Let $\{f_i\}_{i=1}^l$ be a system of functionals with the following properties:*

1) $f_i \; : \; W^{k,p}(\Omega) \to \mathbb{R}_+ \cup \{0\}$ is a seminorm,

2) boundedness: $\exists C_i > 0$ with $0 \le f_i(v) \le C_i \|v\|_{W^{k,p}(\Omega)}$, $\forall\, v \in W^{k,p}(\Omega)$,

3) f_i is a norm on the polynomials of degree $k-1$, i.e., if for $v \in P_{k-1} = \left\{\sum_{|\boldsymbol{\alpha}|\le k-1} C_{\boldsymbol{\alpha}} x^{\boldsymbol{\alpha}}\right\}$, it holds that $f_i(v) = 0$, $i = 1,\ldots,l$, then it is $v \equiv 0$.

Then, the norm $\|\cdot\|_{W^{k,p}(\Omega)}$ defined in (3.7) *and the norm*

$$\|u\|'_{W^{k,p}(\Omega)} := \left(\sum_{i=1}^{l} f_i^p(u) + |u|^p_{W^{k,p}(\Omega)} \right)^{1/p} \quad \textit{with}$$

$$|u|_{W^{k,p}(\Omega)} = \left(\sum_{|\boldsymbol{\alpha}|=k} \int_\Omega |D^{\boldsymbol{\alpha}} u(\boldsymbol{x})|^p \ d\boldsymbol{x} \right)^{1/p}$$

are equivalent.

Remark 3.36. On seminorms. For a seminorm $f_i(\cdot)$, one cannot conclude from $f_i(v) = 0$ that $v = 0$. The third assumptions however states, that this conclusion can be drawn for all polynomials up to a certain degree. □

Example 3.37. Equivalent norms in Sobolev spaces.

- The following norms are equivalent to the standard norm (3.7) in $W^{1,p}(\Omega)$:

$$\text{a) } \|u\|'_{W^{1,p}(\Omega)} = \left(\left| \int_\Omega u \ d\boldsymbol{x} \right|^p + |u|^p_{W^{1,p}(\Omega)} \right)^{1/p},$$

$$\text{b) } \|u\|'_{W^{1,p}(\Omega)} = \left(\left| \int_\Gamma u \ d\boldsymbol{s} \right|^p + |u|^p_{W^{1,p}(\Omega)} \right)^{1/p},$$

$$\text{c) } \|u\|'_{W^{1,p}(\Omega)} = \left(\int_\Gamma |u|^p \ d\boldsymbol{s} + |u|^p_{W^{1,p}(\Omega)} \right)^{1/p}.$$

- In $W^{k,p}(\Omega)$, it is

$$\|u\|'_{W^{k,p}(\Omega)} = \left(\sum_{i=0}^{k-1} \int_\Gamma \left| \frac{\partial^i u}{\partial \boldsymbol{n}^i} \right|^p \ d\boldsymbol{s} + |u|^p_{W^{k,p}(\Omega)} \right)^{1/p}$$

  equivalent to the standard norm. Here, $\boldsymbol{n}$ denotes the outer normal on Γ with $\|\boldsymbol{n}\|_2 = 1$.

- In the case $W_0^{k,p}(\Omega)$, one does not need the regularity of the boundary. It is

$$\|u\|'_{W_0^{k,p}(\Omega)} = |u|_{W^{k,p}(\Omega)},$$

  i.e., in the spaces $W_0^{k,p}(\Omega)$ the standard seminorm is equivalent to the standard norm.

  In particular, it is for $u \in H_0^1(\Omega)$ $(k = 1, p = 2)$

$$C_1 \|u\|_{H^1(\Omega)} \le \|\nabla u\|_{L^2(\Omega)} \le C_2 \|u\|_{H^1(\Omega)}.$$

  It follows that there is a constant $C > 0$ such that

$$\|u\|_{L^2(\Omega)} \le C \|\nabla u\|_{L^2(\Omega)} \quad \forall\, u \in H_0^1(\Omega). \tag{3.9}$$

□

3.7 Some Inequalities in Sobolev Spaces

Remark 3.38. Motivation. This section presents a generalization of the last part of Example 3.37. It will be shown that for inequalities of type (3.9), it is not necessary that the trace vanishes on the complete boundary.

Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary Γ and let $\Gamma_1 \subset \Gamma$ with $\text{meas}_{\mathbb{R}^{d-1}}(\Gamma_1) = \int_{\Gamma_1} d\boldsymbol{s} > 0$.

One considers the space

$$\begin{aligned}
&V_0 = \left\{ v \in W^{1,p}(\Omega) \ : \ v|_{\Gamma_1} = 0 \right\} \subset W^{1,p}(\Omega) \text{ if } \Gamma_1 \subsetneq \Gamma, \\
&V_0 = W_0^{1,p}(\Omega) \text{ if } \Gamma_1 = \Gamma,
\end{aligned}$$

with $p \in [1, \infty)$. □

Lemma 3.39. Friedrichs[13] inequality, Poincaré[14] inequality, Poincaré–Friedrichs inequality. *Let $p \in [1,\infty)$ and $\text{meas}_{\mathbb{R}^{d-1}}(\Gamma_1) > 0$. Then, it is for all $u \in V_0$*

$$\int_\Omega |u(\boldsymbol{x})|^p \ d\boldsymbol{x} \le C_P \int_\Omega \|\nabla u(\boldsymbol{x})\|_2^p \ d\boldsymbol{x}, \tag{3.10}$$

where $\|\cdot\|_2$ is the Euclidean vector norm.

Proof. The inequality will be proved with the theorem on equivalent norms, Theorem 3.35. Let $f_1(u) \ : \ W^{1,p}(\Omega) \to \mathbb{R}_+ \cup \{0\}$ with

$$f_1(u) = \left(\int_{\Gamma_1} |u(\boldsymbol{s})|^p \ d\boldsymbol{s} \right)^{1/p}.$$

This functional has the following properties:

1) $f_1(u)$ is a seminorm.
2) It is bounded, since

$$\begin{aligned}
0 \le f_1(u) &= \left(\int_{\Gamma_1} |u(\boldsymbol{s})|^p \ d\boldsymbol{s} \right)^{1/p} \le \left(\int_{\Gamma} |u(\boldsymbol{s})|^p \ d\boldsymbol{s} \right)^{1/p} \\
&= \|u\|_{L^p(\Gamma)} = \|\gamma u\|_{L^p(\Gamma)} \le C \|u\|_{W^{1,p}(\Omega)}.
\end{aligned}$$

The last estimate follows from the continuity of the trace operator.

3) Let $v \in P_0$, i.e., v is a constant. Then, one obtains from

$$0 = f_1(v) = \left(\int_{\Gamma_1} |v(\boldsymbol{s})|^p \ d\boldsymbol{s} \right)^{1/p} = |v| \left(\text{meas}_{\mathbb{R}^{d-1}}(\Gamma_1) \right)^{1/p},$$

that $|v| = 0$.

Hence, all assumptions of Theorem 3.35 are satisfied. That means, there are two constants C_1 and C_2 with

[13] Kurt Otto Friedrichs (1901 – 1982)

[14] Henri Poincaré (1854 – 1912)

$$\underbrace{C_1\left(\int_{\Gamma_1} |u(\boldsymbol{s})|^p \ d\boldsymbol{s} + \int_\Omega \|\nabla u(\boldsymbol{x})\|_2^p \ d\boldsymbol{x}\right)^{1/p}}_{\|u\|'_{W^{1,p}(\Omega)}} \le \|u\|_{W^{1,p}(\Omega)} \le C_2 \|u\|'_{W^{1,p}(\Omega)}$$

for all $u \in W^{1,p}(\Omega)$. In particular, it follows that

$$\int_\Omega |u(\boldsymbol{x})|^p \ d\boldsymbol{x} + \int_\Omega \|\nabla u(\boldsymbol{x})\|_2^p \ d\boldsymbol{x} \le C_2^p \left(\int_{\Gamma_1} |u(\boldsymbol{s})|^p \ d\boldsymbol{s} + \int_\Omega \|\nabla u(\boldsymbol{x})\|_2^p \ d\boldsymbol{x}\right)$$

or, neglecting the non-negative term on the left-hand side,

$$\int_\Omega |u(\boldsymbol{x})|^p \ d\boldsymbol{x} \le C_P \left(\int_{\Gamma_1} |u(\boldsymbol{s})|^p \ d\boldsymbol{s} + \int_\Omega \|\nabla u(\boldsymbol{x})\|_2^p \ d\boldsymbol{x}\right)$$

with $C_P = C_2^p$. Since $u \in V_0$ vanishes on Γ_1, the statement of the lemma is proved. ■

Remark 3.40. On the Poincaré–Friedrichs inequality. In the space V_0 becomes $|\cdot|_{W^{1,p}}$ a norm that is equivalent to $\|\cdot\|_{W^{1,p}(\Omega)}$. The classical Poincaré–Friedrichs inequality is given for $\Gamma_1 = \Gamma$ and $p = 2$

$$\|u\|_{L^2(\Omega)} \le C_P \|\nabla u\|_{L^2(\Omega)} \ \forall\, u \in H_0^1(\Omega),$$

where the constant depends only on the diameter of the domain Ω, e.g., see (Galdi, 2011, Theorem II.5.1). □

3.8 The Gaussian Theorem

Remark 3.41. Motivation. The Gaussian theorem is the generalization of the integration by parts from calculus. This operation is very important for the theory of weak or variational solutions of partial differential equations. One has to study, under which conditions on the regularity of the domain and of the functions it is well defined. □

Theorem 3.42. Gaussian theorem. *Let $\Omega \subset \mathbb{R}^d, d \ge 2$, be a bounded domain with Lipschitz boundary Γ. Then, the following identity holds for all $u \in W^{1,1}(\Omega)$*

$$\int_\Omega \partial_i u(\boldsymbol{x}) \ d\boldsymbol{x} = \int_\Gamma u(\boldsymbol{s})\boldsymbol{n}_i(\boldsymbol{s}) \ d\boldsymbol{s}, \tag{3.11}$$

where $\boldsymbol{n}$ is the unit outer normal vector on Γ.

Proof. It is referred to the literature. ■

Corollary 3.43. Vector field. *Let the conditions of Theorem 3.42 on the domain Ω be satisfied and let $\boldsymbol{u} \in \left(W^{1,1}(\Omega)\right)^d$ be a vector field. Then, it is*

$$\int_\Omega \nabla \cdot \boldsymbol{u}(\boldsymbol{x}) \ d\boldsymbol{x} = \int_\Gamma \boldsymbol{u}(\boldsymbol{s}) \cdot \boldsymbol{n}(\boldsymbol{s}) \ d\boldsymbol{s}.$$

Proof. The statement follows by adding (3.11) from $i = 1$ to $i = d$. ■

Corollary 3.44. Integration by parts. *Let the conditions of Theorem 3.42 on the domain Ω be satisfied. Consider $u \in W^{1,p}(\Omega)$ and $v \in W^{1,q}(\Omega)$ with $p \in (1,\infty)$ and $\frac{1}{p} + \frac{1}{q} = 1$. Then, it is*

$$\int_\Omega \partial_i u(\boldsymbol{x}) v(\boldsymbol{x}) \ d\boldsymbol{x} = \int_\Gamma u(\boldsymbol{s}) v(\boldsymbol{s}) \boldsymbol{n}_i(\boldsymbol{s}) \ d\boldsymbol{s} - \int_\Omega u(\boldsymbol{x}) \partial_i v(\boldsymbol{x}) \ d\boldsymbol{x}.$$

Proof. *exercise.* ■

Corollary 3.45. First Green[15]'s formula. *Let the conditions of Theorem 3.42 on the domain Ω be satisfied, then it is*

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

for all $u \in H^2(\Omega)$ and $v \in H^1(\Omega)$.

Proof. From the definition of the Sobolev spaces, it follows that the integrals are well defined. Now, the proof follows the proof of Corollary 3.44, where one has to sum over the components and one has to take $\partial_i v$ instead of v. ■

Remark 3.46. On the first Green's formula. The first Green's formula is the formula of integrating by parts once. The boundary integral can be equivalently written in the form

$$\int_\Gamma \nabla u(\boldsymbol{s}) \cdot \boldsymbol{n}(\boldsymbol{s}) v(\boldsymbol{s}) \ d\boldsymbol{s}.$$

The formula of integrating by parts twice is called second Green's formula. □

Corollary 3.47. Second Green's formula. *Let the conditions of Theorem 3.42 on the domain Ω be satisfied, then one has*

$$\int_\Omega \big(\Delta u(\boldsymbol{x}) v(\boldsymbol{x}) - \Delta v(\boldsymbol{x}) u(\boldsymbol{x})\big) \ d\boldsymbol{x} = \int_\Gamma \left(\frac{\partial u}{\partial \boldsymbol{n}}(\boldsymbol{s}) v(\boldsymbol{s}) - \frac{\partial v}{\partial \boldsymbol{n}}(\boldsymbol{s}) u(\boldsymbol{s}) \right) \ d\boldsymbol{s}$$

for all $u, v \in H^2(\Omega)$.

3.9 Sobolev Imbedding Theorems

Remark 3.48. Motivation. This section studies the question which (Sobolev) spaces are subspaces of other Sobolev spaces. With this property, called

[15] Georg Green (1793 – 1841)

imbedding, it is possible to estimate the norm of a function in the subspace by the norm in the larger space, compare (3.12). □

Lemma 3.49. Imbedding of Sobolev spaces with same integration power p and different orders of the derivative. *Let $\Omega \subset \mathbb{R}^d$ be a domain, $p \in [1,\infty]$, and $k \le m$, then it is $W^{m,p}(\Omega) \subset W^{k,p}(\Omega)$.*

Proof. The statement of this lemma follows directly from the definition of Sobolev spaces, see Definition 3.20. ■

Lemma 3.50. Imbedding of Sobolev spaces with the same order of the derivative k and different integration powers. *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain, $k \ge 0$, and $p, q \in [1,\infty]$ with $q > p$. Then, it is $W^{k,q}(\Omega) \subset W^{k,p}(\Omega)$.*

Proof. *exercise.* ■

Remark 3.51. Imbedding of Sobolev spaces with the same order of the derivative k and the same integration power p in imbedded domains. Let $\Omega \subset \mathbb{R}^d$ be a domain with sufficiently smooth boundary Γ, $k \ge 0$, and $p \in [1,\infty]$. Then, there is a map $E\ :\ W^{k,p}(\Omega) \to W^{k,p}(\mathbb{R}^d)$, the so-called (simple) extension, with

- $Ev|_\Omega = v$,
- $\|Ev\|_{W^{k,p}(\mathbb{R}^d)} \le C\,\|v\|_{W^{k,p}(\Omega)}$, with $C > 0$ independent of v,

e.g., see (Adams, 1975, Chapter IV) for details. Likewise, the natural restriction $e\ :\ W^{k,p}(\mathbb{R}^d) \to W^{k,p}(\Omega)$ can be defined and it is $\|ev\|_{W^{k,p}(\Omega)} \le \|v\|_{W^{k,p}(\mathbb{R}^d)}$. □

Theorem 3.52. A Sobolev inequality. *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary Γ, $k \ge 0$, and $p \in [1,\infty)$ with*

$$\begin{array}{ll} k \ge d & \text{for } p = 1, \\ k > d/p & \text{for } p > 1. \end{array}$$

Then, there is a constant C such that for all $u \in W^{k,p}(\Omega)$, it follows that $u \in C_B(\Omega)$, where

$$C_B(\Omega) = \{v \in C(\Omega)\ :\ v \text{ is bounded}\},$$

and it is

$$\|u\|_{C_B(\Omega)} = \|u\|_{L^\infty(\Omega)} \le C\,\|u\|_{W^{k,p}(\Omega)}\,. \tag{3.12}$$

Proof. See literature, e.g., Adams (1975); Adams & Fournier (2003). ■

Remark 3.53. On the Sobolev inequality.

- The Sobolev inequality states that each function with sufficiently many weak derivatives (the number depends on the dimension of Ω and the integration power) can be considered as a continuous and bounded function in Ω, i.e., there is such a representative in the equivalence class where this function belongs to. One says that $W^{k,p}(\Omega)$ is imbedded in $C_B(\Omega)$.

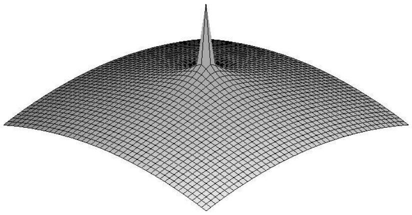

Fig. 3.3 The function $f(\boldsymbol{x})$ of Example 3.55 for $d = 2$.

- It is
$$C\left(\overline{\Omega}\right) \subsetneq C_B(\Omega) \subsetneq C(\Omega).$$
Consider $\Omega = (0,1)$ and $f_1(x) = 1/x$ and $f_2(x) = \sin(1/x)$. Then, $f_1 \in C(\Omega)$, $f_1 \notin C_B(\Omega)$ and $f_2 \in C_B(\Omega)$, $f_2 \notin C(\overline{\Omega})$.
- Of course, it is possible to apply this theorem to weak derivatives of functions. Then, one obtains imbeddings like $W^{k,p}(\Omega) \to C_B^s(\Omega)$ for $(k-s)p > d, p > 1$. A comprehensive overview on imbeddings can be found in Adams (1975); Adams & Fournier (2003).

□

Example 3.54. $H^1(\Omega)$ in one dimension. Let $d = 1$ and Ω be a bounded interval. Then, each function from $H^1(\Omega)$ $(k = 1, p = 2)$ is continuous and bounded in Ω. □

Example 3.55. $H^1(\Omega)$ in higher dimensions. The functions from $H^1(\Omega)$ are in general not continuous for $d \geq 2$. This property will be shown with the following example.

Let $\Omega = \{\boldsymbol{x} \in \mathbb{R}^d \ : \ \|\boldsymbol{x}\|_2 < 1/2\}$ and $f(\boldsymbol{x}) = \ln|\ln\|\boldsymbol{x}\|_2|$, see Figure 3.3. For $\|\boldsymbol{x}\|_2 < 1/2$, it is $|\ln\|\boldsymbol{x}\|_2| = -\ln\|\boldsymbol{x}\|_2$ and one gets for $\boldsymbol{x} \neq \boldsymbol{0}$

$$\partial_i f(\boldsymbol{x}) = -\frac{1}{\ln\|\boldsymbol{x}\|_2}\frac{1}{\|\boldsymbol{x}\|_2}\frac{x_i}{\|\boldsymbol{x}\|_2} = -\frac{x_i}{\|\boldsymbol{x}\|_2^2\ln\|\boldsymbol{x}\|_2}.$$

For $p \leq d$, one obtains

$$\left|\frac{\partial f}{\partial x_i}(\boldsymbol{x})\right|^p = \underbrace{\left|\frac{x_i}{\|\boldsymbol{x}\|_2}\right|^p}_{\leq 1}\underbrace{\left|\frac{1}{\|\boldsymbol{x}\|_2\ln\|\boldsymbol{x}\|_2}\right|^p}_{\geq e} \leq \left|\frac{1}{\|\boldsymbol{x}\|_2\ln\|\boldsymbol{x}\|_2}\right|^d.$$

The estimate of the second factor can be obtained, e.g., with a discussion of the curve. Using now spherical coordinates, $\rho = e^{-t}$ and S^{d-1} is the unit sphere, yields

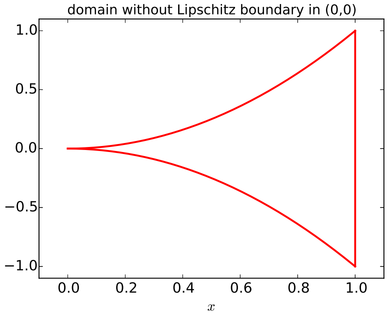

Fig. 3.4 Domain of Example 3.56.

$$\int_\Omega |\partial_i f(\boldsymbol{x})|^p \, d\boldsymbol{x} \le \int_\Omega \frac{d\boldsymbol{x}}{\|\boldsymbol{x}\|_2^d \left|\ln \|\boldsymbol{x}\|_2\right|^d} = \int_{S^{d-1}} \int_0^{1/2} \frac{\rho^{d-1}}{\rho^d \left|\ln \rho\right|^d} \, d\rho d\omega$$

$$= \text{meas}\left(S^{d-1}\right) \int_0^{1/2} \frac{d\rho}{\rho \left|\ln \rho\right|^d} = -\text{meas}\left(S^{d-1}\right) \int_\infty^{\ln 2} \frac{dt}{t^d} < \infty,$$

because of $d \ge 2$.

It follows that $\partial_i f \in L^p(\Omega)$ with $p \le d$. Analogously, one proves that $f \in L^p(\Omega)$ with $p \le d$. Altogether, one has $f \in W^{1,p}(\Omega)$ with $p \le d$. However, it is $f \notin L^\infty(\Omega)$ and consequently $f \notin C_B(\Omega)$. This example shows that the condition $k > d/p$ for $p > 1$ is sharp.

In particular, it was proved for $p = 2$ that from $f \in H^1(\Omega)$ in general it does not follow that $f \in C(\Omega)$. □

Example 3.56. The assumption of a Lipschitz boundary. Also the assumption that Ω is a Lipschitz domain is of importance.

Consider $\Omega = \{(x,y) \in \mathbb{R}^2 \ : \ 0 < x < 1, \ |y| < x^r, r > 1\}$, see Figure 3.4 for $r = 2$. The boundary is not Lipschitz in $(0,0)$.

For $u(x,y) = x^{-\varepsilon/p}$ with $0 < \varepsilon \le r$, it is

$$\partial_x u = x^{-\varepsilon/p-1}\left(-\frac{\varepsilon}{p}\right) = C(\varepsilon,p)x^{-\varepsilon/p-1}, \ \partial_y u = 0.$$

Using the same notation for the constant, which might take different values at different occasions, it follows that

$$\begin{aligned}\sum_{|\boldsymbol{\alpha}|=1} \int_{\Omega} |D^{\boldsymbol{\alpha}} u(x,y)|^p \; dxdy &= C(\varepsilon,p) \int_{\Omega} x^{-\varepsilon-p} \; dxdy \\ &= C(\varepsilon,p) \int_0^1 x^{-\varepsilon-p} \left(\int_{-x^r}^{x^r} \; dy \right) dx \\ &= C(\varepsilon,p) \int_0^1 x^{-\varepsilon-p+r} \; dx.\end{aligned}$$

This value is finite for $-\varepsilon - p + r > -1$ or for $p < 1 + r - \varepsilon$, respectively. If one chooses $r \geq \varepsilon > 0$, then it is $u \in W^{1,p}(\Omega)$. But for $\varepsilon > 0$, the function $u(\boldsymbol{x})$ is not bounded in Ω, i.e., $u \notin L^{\infty}(\Omega)$ and consequently $u \notin C_B(\Omega)$.

The unbounded values of the function are compensated in the integration by the fact that the neighborhood of the singular point $(0,0)$ possesses a small measure. □

Chapter 4
The Ritz Method and the Galerkin Method

Remark 4.1. Contents. This chapter studies variational or weak formulations of boundary value problems of partial differential equations in Hilbert spaces. The existence and uniqueness of an appropriately defined weak solution will be discussed. The approximation of this solution with the help of finite-dimensional spaces is called Ritz method or Galerkin method. Some basic properties of this method will be proved.

In this chapter, a Hilbert space V will be considered with inner product $a(\cdot,\cdot)\ :\ V\times V\to\mathbb{R}$ and norm $\|v\|_V = a(v,v)^{1/2}$. □

4.1 The Theorems of Riesz and Lax–Milgram

Theorem 4.2. Representation theorem of Riesz[1]**.** *Let $f\in V'$ be a continuous and linear functional, then there is a uniquely determined $u\in V$ with*

$$a(u,v) = f(v) \quad \forall\, v\in V. \tag{4.1}$$

In addition, u is the unique solution of the variational problem

$$F(v) = \frac{1}{2}a(v,v) - f(v) \to \min\ \ \forall\, v\in V. \tag{4.2}$$

Proof. First, the existence of a solution u of the variational problem will be proved. Since f is continuous, it holds

$$|f(v)| \le C\,\|v\|_V \quad \forall\, v\in V,$$

from what follows that

$$F(v) \ge \frac{1}{2}\,\|v\|_V^2 - C\,\|v\|_V \ge -\frac{1}{2}C^2,$$

[1] Frigyes Riesz (1880 – 1956)

where in the last estimate the necessary criterion for a local minimum of the expression of the first estimate,

$$\frac{2}{2}\left\|v\right\|_V - C = 0 \quad \Longleftrightarrow \quad \left\|v\right\|_V = C,$$

is used. Hence, the function $F(\cdot)$ is bounded from below and

$$\kappa = \inf_{v\in V} F(v)$$

exists.

Let $\{v_k\}_{k\in\mathbb{N}}$ be a sequence with $F(v_k) \to \kappa$ for $k \to \infty$. A straightforward calculation (parallelogram identity in Hilbert spaces) gives

$$\left\|v_k - v_l\right\|_V^2 + \left\|v_k + v_l\right\|_V^2 = 2\left\|v_k\right\|_V^2 + 2\left\|v_l\right\|_V^2 .$$

Using the linearity of $f(\cdot)$ and $\kappa \le F(v)$ for all $v \in V$, one obtains

$$\begin{aligned}
&\left\|v_k - v_l\right\|_V^2 \\
&= 2\left\|v_k\right\|_V^2 + 2\left\|v_l\right\|_V^2 - 4\left\|\frac{v_k + v_l}{2}\right\|_V^2 - 4f(v_k) - 4f(v_l) + 8f\left(\frac{v_k + v_l}{2}\right) \\
&= 4F(v_k) + 4F(v_l) - 8F\left(\frac{v_k + v_l}{2}\right) \\
&\le 4F(v_k) + 4F(v_l) - 8\kappa \to 0
\end{aligned}$$

for $k, l \to \infty$. Hence, $\{v_k\}_{k\in\mathbb{N}}$ is a Cauchy sequence. Because V is a complete space, there exists a limit u of this sequence with $u \in V$. Because $F(\cdot)$ is continuous, it is $F(u) = \kappa$ and u is a solution of the variational problem.

In the next step, it will be shown that each solution of the variational problem (4.2) is also a solution of (4.1). It is for arbitrary $v \in V$

$$\begin{aligned}
\Phi(\varepsilon) &= F(u + \varepsilon v) = \frac{1}{2}a(u + \varepsilon v, u + \varepsilon v) - f(u + \varepsilon v) \\
&= \frac{1}{2}a(u, u) + \varepsilon a(u, v) + \frac{\varepsilon^2}{2}a(v, v) - f(u) - \varepsilon f(v).
\end{aligned}$$

If u is a minimum of the variational problem, then the function $\Phi(\varepsilon)$ has in particular a local minimum at $\varepsilon = 0$. The necessary condition for a local minimum leads to

$$0 = \Phi'(0) = a(u, v) - f(v) \quad \text{for all } v \in V.$$

Finally, the uniqueness of the solution will be proved. It is sufficient to prove the uniqueness of the solution of the equation (4.1). If the solution of (4.1) is unique, then the existence of two solutions of the variational problem (4.2) would be a contradiction to the fact proved in the previous step. Let u_1 and u_2 be two solutions of the equation (4.1). Computing the difference of both equations gives

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

This equation holds, in particular, for $v = u_1 - u_2$. Hence, $\left\|u_1 - u_2\right\|_V = 0$, such that $u_1 = u_2$. ■

Definition 4.3. Bounded bilinear form, coercive bilinear form, V-elliptic bilinear form. Let $b(\cdot,\cdot) : V \times V \to \mathbb{R}$ be a bilinear form on the Banach space V. Then, it is bounded if

$$|b(u, v)| \le M \left\|u\right\|_V \left\|v\right\|_V \quad \forall\, u, v \in V, M > 0, \tag{4.3}$$

where the constant M is independent of u and v. The bilinear form is coercive or V-elliptic if

$$b(u,u) \geq m \left\|u\right\|_V^2 \quad \forall\, u \in V, m > 0, \tag{4.4}$$

where the constant m is independent of u. □

Remark 4.4. Application to an inner product. Let V be a Hilbert space. Then, the inner product $a(\cdot,\cdot)$ is a bounded and coercive bilinear form, since by the Cauchy–Schwarz inequality

$$|a(u,v)| \leq \left\|u\right\|_V \left\|v\right\|_V \quad \forall\, u, v \in V,$$

and obviously $a(u,u) = \left\|u\right\|_V^2$. Hence, the constants can be chosen to be $M = 1$ and $m = 1$.

Next, the representation theorem of Riesz will be generalized to the case of coercive and bounded bilinear forms. □

Theorem 4.5. Theorem of Lax[2]–Milgram[3]. *Let $b(\cdot,\cdot)\ :\ V \times V \to \mathbb{R}$ be a bounded and coercive bilinear form on the Hilbert space V. Then, for each bounded linear functional $f \in V'$ there is exactly one $u \in V$ with*

$$b(u,v) = f(v) \quad \forall\, v \in V. \tag{4.5}$$

Proof. One defines operators $T, T'\ :\ V \to V$ by

$$a(Tu,v) = b(u,v)\ \forall\, v \in V, \quad a(T'u,v) = b(v,u)\ \forall\, v \in V. \tag{4.6}$$

These operators are linear, e.g., using that $b(\cdot,\cdot)$ is a bilinear form, one gets

$$a\left(T(\alpha_1 u_1 + \alpha_2 u_2), v\right) = \alpha_1 b(u_1,v) + \alpha_2 b(u_2,v) = a\left(\alpha_1 Tu_1 + \alpha_2 Tu_2, v\right) \quad \forall\, v \in V.$$

Because this relation holds for all $v \in V$, it is $T(\alpha_1 u_1 + \alpha_2 u_2) = \alpha_1 Tu_1 + \alpha_2 Tu_2$. Since $b(u,\cdot)$ and $b(\cdot,u)$ are continuous linear functionals on V, it follows from Theorem 4.2 that the elements Tu and $T'u$ exist and they are defined uniquely. Because the operators satisfy the relation

$$a(Tu,v) = b(u,v) = a(T'v,u) = a(u,T'v), \tag{4.7}$$

T' is called adjoint operator of T. Setting $v = Tu$ in (4.6) and using the boundedness of $b(\cdot,\cdot)$ yields

$$\left\|Tu\right\|_V^2 = a(Tu,Tu) = b(u,Tu) \leq M \left\|u\right\|_V \left\|Tu\right\|_V \implies \left\|Tu\right\|_V \leq M \left\|u\right\|_V$$

for all $u \in V$. Hence, T is bounded. Since T is linear, it follows that T is continuous. Using the same argument, one shows that T' is also bounded and continuous.

Define the bilinear form

$$d(u,v) := a(TT'u,v) = a(T'u,T'v) \quad \forall\, u,v \in V, \tag{4.8}$$

where (4.7) was used. Hence, this bilinear form is symmetric. Using the coercivity of $b(\cdot,\cdot)$, the Cauchy–Schwarz inequality, the definition of $\left\|\cdot\right\|_V$, and (4.8) gives

[2] Peter Lax, born 1926

[3] Arthur Norton Milgram (1912 – 1961)

$$m^2 \|v\|_V^4 \le b(v,v)^2 = a(T'v,v)^2 \le \|v\|_V^2 \left\|T'v\right\|_V^2 = \|v\|_V^2 \, a(T'v,T'v) = \|v\|_V^2 \, d(v,v).$$

Applying now the boundedness of $a(\cdot,\cdot)$ and of T' yields

$$m^2 \|v\|_V^2 \le d(v,v) = a(T'v,T'v) = \left\|T'v\right\|_V^2 \le M \|v\|_V^2 . \tag{4.9}$$

Hence, $d(\cdot,\cdot)$ is also coercive and, since it is symmetric, it defines an inner product on V. From (4.9), one has that the norm induced by $d(v,v)^{1/2}$ is equivalent to the norm $\|v\|_V$. From Theorem 4.2, it follows that there is a exactly one $w \in V$ with

$$d(w,v) = f(v) \quad \forall\, v \in V.$$

Now, inserting $u = T'w$ in $b(\cdot,\cdot)$ gives with (4.6)

$$b(T'w,v) = a(TT'w,v) = d(w,v) = f(v) \quad \forall\, v \in V,$$

hence $u = T'w$ is a solution of (4.5).

The uniqueness of the solution is proved analogously as in the symmetric case. ■

4.2 Weak Formulation of Boundary Value Problems

Remark 4.6. Model problem. Consider the Poisson equation with homogeneous Dirichlet boundary conditions

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega \subset \mathbb{R}^d, \\ u &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{4.10}$$

□

Definition 4.7. Weak formulation of (4.10)**.** Let $f \in L^2(\Omega)$. A weak formulation of (4.10) consists in finding $u \in V = H_0^1(\Omega)$ such that

$$a(u,v) = (f,v) \quad \forall\, v \in V \tag{4.11}$$

with

$$a(u,v) = (\nabla u, \nabla v) = \int_\Omega \nabla u(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) \; d\boldsymbol{x}$$

and $(\cdot,\cdot)$ is the inner product in $L^2(\Omega)$. □

Remark 4.8. On the weak formulation.

- The weak formulation is also called variational formulation.
- As usual in mathematics, 'weak' means that something holds for all appropriately chosen test functions.
- Formally, one obtains the weak formulation by multiplying the strong form of the equation (4.10) with the test function, by integrating the equation on Ω, and applying integration by parts. Because of the Dirichlet boundary condition, one can use as test space $H_0^1(\Omega)$ and therefore the integral on the boundary vanishes.

- The ansatz space for the solution and the test space are defined such that the arising integrals are well defined.
- The weak formulation reduces the necessary regularity assumptions for the solution by the integration and the transfer of derivatives to the test function. Whereas the solution of (4.10) has to be in $C^2(\Omega) \cap C(\overline{\Omega})$, the solution of (4.11) has to be only in $H_0^1(\Omega)$. The latter assumption is much more realistic for problems coming from applications.
- The regularity assumption on the right-hand side can be relaxed to $f \in H^{-1}(\Omega)$. Then, the right-hand side of the weak formulation has the form

  $$f(v) = \langle f, v \rangle_{H^{-1}(\Omega), H_0^1(\Omega)},$$

  where the symbol $\langle \cdot, \cdot, \rangle_{H^{-1}(\Omega), H_0^1(\Omega)}$ denotes the dual pairing of the spaces $H_0^1(\Omega)$ and $H^{-1}(\Omega)$.

□

Theorem 4.9. Existence and uniqueness of the weak solution. *Let $f \in L^2(\Omega)$. There is exactly one solution of* (4.11).

Proof. Because of the Poincaré inequality (3.10), there is a constant C with

$$\|v\|_{L^2(\Omega)} \le C \|\nabla v\|_{L^2(\Omega)} \quad \forall\, v \in H_0^1(\Omega).$$

It follows for $v \in H_0^1(\Omega) \subset H^1(\Omega)$ that

$$\begin{aligned}\|v\|_{H^1(\Omega)} &= \left(\|v\|_{L^2(\Omega)}^2 + \|\nabla v\|_{L^2(\Omega)}^2\right)^{1/2} \le \left(C \|\nabla v\|_{L^2(\Omega)}^2 + \|\nabla v\|_{L^2(\Omega)}^2\right)^{1/2} \\ &\le C \|\nabla v\|_{L^2(\Omega)} \le C \|v\|_{H^1(\Omega)}.\end{aligned}$$

Hence, $a(\cdot,\cdot)$ is an inner product on $H_0^1(\Omega)$ with the induced norm

$$\|v\|_{H_0^1(\Omega)} = a(v,v)^{1/2},$$

which is equivalent to the norm $\|\cdot\|_{H^1(\Omega)}$.

Define for $f \in L^2(\Omega)$ the linear functional

$$\tilde{f}(v) := \int_\Omega f(\boldsymbol{x}) v(\boldsymbol{x})\, d\boldsymbol{x} \quad \forall\, v \in H_0^1(\Omega).$$

Using the Cauchy–Schwarz inequality (3.5) and the Poincaré inequality (3.10) shows that this functional is continuous on $H_0^1(\Omega)$

$$\left|\tilde{f}(v)\right| = |(f,v)| \le \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \le C \|f\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} = C \|f\|_{L^2(\Omega)} \|v\|_{H_0^1(\Omega)}.$$

Applying the representation theorem of Riesz, Theorem 4.2, gives the existence and uniqueness of the weak solution of (4.11). In addition, $u(\boldsymbol{x})$ solves the variational problem

$$F(v) = \frac{1}{2}\|\nabla v\|_2^2 - \int_\Omega f(\boldsymbol{x}) v(\boldsymbol{x})\, d\boldsymbol{x} \to \min \quad \text{for all } v \in H_0^1(\Omega).$$

■

Example 4.10. A more general elliptic problem. Consider the problem

$$\begin{aligned} -\nabla \cdot (A(\boldsymbol{x})\nabla u) + c(\boldsymbol{x})u &= f \text{ in } \Omega \subset \mathbb{R}^d, \\ u &= 0 \text{ on } \partial\Omega, \end{aligned} \tag{4.12}$$

with $A(\boldsymbol{x}) \in \mathbb{R}^{d\times d}$ for each point $\boldsymbol{x} \in \Omega$. It will be assumed that the coefficients $a_{ij}(\boldsymbol{x})$ and $c(\boldsymbol{x}) \ge 0$ are bounded, $f \in L^2(\Omega)$, and that the matrix (tensor) $A(\boldsymbol{x})$ is for all $\boldsymbol{x} \in \Omega$ uniformly elliptic, i.e., there are positive constants m and M independent of $\boldsymbol{x}$ such that

$$m \left\|\underline{y}\right\|_2^2 \le \underline{y}^T A(\boldsymbol{x}) \underline{y} \le M \left\|\underline{y}\right\|_2^2 \quad \forall\, \underline{y} \in \mathbb{R}^d,\ \forall\, \boldsymbol{x} \in \Omega.$$

The weak form of (4.12) is obtained in the usual way by multiplying (4.12) with test functions $v \in H_0^1(\Omega)$, integrating on Ω, and applying integration by parts: Find $u \in H_0^1(\Omega)$, such that

$$a(u,v) = (f,v) \quad \forall\, v \in H_0^1(\Omega)$$

with

$$a(u,v) = \int_\Omega \left(\nabla u(\boldsymbol{x})^T A(\boldsymbol{x}) \nabla v(\boldsymbol{x}) + c(\boldsymbol{x}) u(\boldsymbol{x}) v(\boldsymbol{x})\right) \, d\boldsymbol{x}.$$

This bilinear form is bounded (*exercise*). The coercivity of the bilinear form is proved by using the uniform ellipticity of $A(\boldsymbol{x})$ and the non-negativity of $c(\boldsymbol{x})$:

$$\begin{aligned} a(u,u) &= \int_\Omega \nabla u(\boldsymbol{x})^T A(\boldsymbol{x}) \nabla u(\boldsymbol{x}) + c(\boldsymbol{x}) u(\boldsymbol{x}) u(\boldsymbol{x}) \, d\boldsymbol{x} \\ &\geq \int_\Omega m \nabla u(\boldsymbol{x})^T \nabla u(\boldsymbol{x}) \, d\boldsymbol{x} = m \left\|u\right\|^2_{H_0^1(\Omega)} . \end{aligned}$$

Applying the Theorem of Lax–Milgram, Theorem 4.5, gives the existence and uniqueness of a weak solution of (4.12).

If the tensor is not symmetric, $a_{ij}(\boldsymbol{x}) \neq a_{ji}(\boldsymbol{x})$ for one pair i, j, then the solution cannot be characterized as the solution of a variational problem. □

4.3 The Ritz Method and the Galerkin Method

Remark 4.11. Idea of the Ritz method. Let V be a Hilbert space with the inner product $a(\cdot,\cdot)$. Consider the problem

$$F(v) = \frac{1}{2} a(v,v) - f(v) \to \min, \tag{4.13}$$

where $f \; : \; V \to \mathbb{R}$ is a bounded linear functional. As already proved in Theorem 4.2, there is a unique solution $u \in V$ of this variational problem which is also the unique solution of the equation

$$a(u,v) = f(v) \quad \forall\, v \in V. \tag{4.14}$$

For approximating the solution of (4.13) or (4.14) with a numerical method, it will be assumed that V has a countable orthonormal basis (Schauder basis). Then, there are finite-dimensional subspaces $V_1, V_2, \ldots \subset V$ with $\dim V_k = k$, which have the following property: for each $u \in V$ and each $\varepsilon > 0$ there is a $K \in \mathbb{N}$ and a $u_k \in V_k$ with

$$\left\|u - u_k\right\|_V \leq \varepsilon \quad \forall\, k \geq K. \tag{4.15}$$

Note that it is not required that there holds an inclusion of the form $V_k \subset V_{k+1}$.

The Ritz approximation of (4.13) and (4.14) is defined by: Find $u_k \in V_k$ with

$$a(u_k, v_k) = f(v_k) \quad \forall\, v_k \in V_k. \tag{4.16}$$

□

Lemma 4.12. Existence and uniqueness of a solution of (4.16)**.** *There exists exactly one solution of* (4.16)*.*

Proof. Finite-dimensional subspaces of Hilbert spaces are Hilbert spaces as well. For this reason, one can apply the representation theorem of Riesz, Theorem 4.2, to (4.16) which gives the statement of the lemma. In addition, the solution of (4.16) solves a minimization problem on V_k. ■

Lemma 4.13. Best approximation property. *The solution of* (4.16) *is the best approximation of u in V_k, i.e., it is*

$$\|u-u_k\|_V = \inf_{v_k \in V_k} \|u-v_k\|_V . \tag{4.17}$$

Proof. Since $V_k \subset V$, one can use the test functions from V_k in the weak equation (4.14). Then, the difference of (4.14) and (4.16) gives the orthogonality, the so-called Galerkin orthogonality,

$$a(u-u_k, v_k) = 0 \quad \forall\, v_k \in V_k. \tag{4.18}$$

Hence, the error $u-u_k$ is orthogonal to the space V_k: $u-u_k \perp V_k$. That means, u_k is the orthogonal projection of u onto V_k with respect of the inner product of V.

Let now $w_k \in V_k$ be an arbitrary element, then it follows with the Galerkin orthogonality (4.18) and the Cauchy–Schwarz inequality that

$$\begin{aligned}\|u-u_k\|_V^2 &= a(u-u_k, u-u_k) = a(u-u_k, u-\underbrace{(u_k-w_k)}_{v_k}) = a(u-u_k, u-v_k)\\ &\le \|u-u_k\|_V \|u-v_k\|_V .\end{aligned}$$

Since $w_k \in V_k$ was arbitrary, also $v_k \in V_k$ is arbitrary. If $\|u-u_k\|_V > 0$, division by $\|u-u_k\|_V$ gives the statement of the lemma, since the error cannot be smaller than the best approximation error. If $\|u-u_k\|_V = 0$, the statement of the lemma is trivially true. ■

Theorem 4.14. Convergence of the Ritz approximation. *The Ritz approximation converges*

$$\lim_{k\to\infty} \|u-u_k\|_V = 0.$$

Proof. The best approximation property (4.17) and property (4.15) give

$$\|u-u_k\|_V = \inf_{v_k \in V_k} \|u-v_k\|_V \le \varepsilon$$

for each $\varepsilon > 0$ and $k \ge K(\varepsilon)$. Hence, the convergence is proved. ■

Remark 4.15. Formulation of the Ritz method as linear system of equations. One can use an arbitrary basis $\{\phi_i\}_{i=1}^k$ of V_k for the computation of u_k. First of all, the equation for the Ritz approximation (4.16) is satisfied for all $v_k \in V_k$ if and only if it is satisfied for each basis function ϕ_i. This statement follows from the linearity of both sides of the equation with respect to the test function and from the fact that each function $v_k \in V_k$ can be represented as linear combination of the basis functions. Let $v_k = \sum_{i=i}^k \alpha_i \phi_i$, then from (4.16), it follows that

$$a(u_k, v_k) = \sum_{k=1}^k \alpha_i a(u_k, \phi_i) = \sum_{k=1}^k \alpha_i f(\phi_i) = f(v_k).$$

This equation is satisfied if $a(u_k, \phi_i) = f(\phi_i)$, $i = 1, \ldots, k$. On the other hand, if (4.16) holds then it holds in particular for each basis function ϕ_i.

Now, one uses as ansatz for the solution also a linear combination of the basis functions

$$u_k = \sum_{j=1}^k u^j \phi_j$$

with unknown coefficients $u^j \in \mathbb{R}$. Using as test functions the basis functions yields

$$\sum_{j=1}^k a(u^j \phi_j, \phi_i) = \sum_{j=1}^k a(\phi_j, \phi_i) u^j = f(\phi_i), \quad i = 1, \ldots, k.$$

This equation is equivalent to the linear system of equations $A\underline{u} = \underline{f}$, where

$$A = (a_{ij})_{i,j=1}^k = a(\phi_j, \phi_i)_{i,j=1}^k$$

is called stiffness matrix. Note that the order of the indices is different for the entries of the matrix and the arguments of the inner product. The right-hand side is a vector of length k with the entries $f_i = f(\phi_i)$, $i = 1, \ldots, k$.

Using the one-to-one mapping between the coefficient vector $(v^1, \ldots, v^k)^T$ and the element $v_k = \sum_{i=1}^k v^i \phi_i$, one can show that the matrix A is symmetric and positive definite (*exercise*)

$$\begin{aligned} A = A^T &\Longleftrightarrow a(v,w) = a(w,v) \quad \forall\, v, w \in V_k, \\ \underline{x}^T A \underline{x} > 0 \text{ for } \underline{x} \neq \underline{0} &\Longleftrightarrow a(v,v) > 0 \quad \forall\, v \in V_k, v \neq 0. \end{aligned}$$

□

Remark 4.16. The case of a bounded and coercive bilinear form. If $b(\cdot,\cdot)$ is bounded and coercive, but not symmetric, it is possible to approximate the solution of (4.5) with the same idea as for the Ritz method. In this case, it is called Galerkin method. The discrete problem consists in finding $u_k \in V_k$ such that

$$b(u_k, v_k) = f(v_k) \quad \forall\, v_k \in V_k. \tag{4.19}$$

□

Lemma 4.17. Existence and uniqueness of a solution of (4.19)**.** *There is exactly one solution of* (4.19).

Proof. The statement of the lemma follows directly from the Theorem of Lax–Milgram, Theorem 4.5. ■

Remark 4.18. On the discrete solution. The discrete solution is not the orthogonal projection into V_k in the case of a bounded and coercive bilinear form, which is not the inner product of V. □

Lemma 4.19. Lemma of Cea[4], error estimate. *Let* $b: V \times V \to \mathbb{R}$ *be a bounded and coercive bilinear form on the Hilbert space* V *and let* $f \in V'$ *be a bounded linear functional. Let* u *be the solution of* (4.5) *and* u_k *be the solution of* (4.19)*, then the following error estimate holds*

$$\|u - u_k\|_V \le \frac{M}{m} \inf_{v_k \in V_k} \|u - v_k\|_V\,, \tag{4.20}$$

where the constants M *and* m *are given in* (4.3) *and* (4.4)*.*

Proof. Considering the difference of the continuous equation (4.5) and the discrete equation (4.19), one obtains the error equation

$$b(u - u_k, v_k) = 0 \quad \forall\, v_k \in V_k,$$

i.e., Galerkin orthogonality holds. With (4.4), the Galerkin orthogonality, and (4.3), it follows that

$$\begin{aligned}\|u - u_k\|_V^2 &\le \frac{1}{m} b(u - u_k, u - u_k) = \frac{1}{m} b(u - u_k, u - v_k)\\ &\le \frac{M}{m} \|u - u_k\|_V \|u - v_k\|_V\,, \quad \forall\, v_k \in V_k,\end{aligned}$$

from what the statement of the lemma follows immediately. ■

Remark 4.20. On the best approximation error. It follows from estimate (4.20) that the error is bounded by a multiple of the best approximation error, where the factor depends on properties of the bilinear form $b(\cdot,\cdot)$. Thus, concerning error estimates for concrete finite-dimensional spaces, the study of the best approximation error will be of importance. □

Remark 4.21. The corresponding linear system of equations. The corresponding linear system of equations is derived analogously to the symmetric case. The system matrix is still positive definite but not symmetric. □

Remark 4.22. Choice of the basis. The most important issue of the Ritz and Galerkin method is the choice of the spaces V_k, or more concretely, the choice of an appropriate basis $\{\phi_i\}_{i=1}^k$ that spans the space V_k. From the point of view of numerics, there are the requirements that:

- it should be possible to compute the entries a_{ij} of the stiffness matrix efficiently,
- and that the matrix A should be sparse.

□

[4] Jean Cea, born 1932

Chapter 5
Finite Element Methods

5.1 Finite Element Spaces

Remark 5.1. Mesh cells, faces, edges, vertices. A mesh cell K is a compact polyhedron in $\mathbb{R}^d$, $d \in \{2,3\}$, whose interior is not empty. The boundary ∂K of K consists of m-dimensional linear manifolds (points, pieces of straight lines, pieces of planes), $0 \le m \le d-1$, which are called m-faces. The 0-faces are the vertices of the mesh cell, the 1-faces are the edges, and the $(d-1)$-faces are just called faces. □

Remark 5.2. Finite-dimensional spaces defined on K. Let $s \in \mathbb{N}$. Finite element methods use finite-dimensional spaces $P(K) \subset C^s(K)$ that are defined on K. In general, $P(K)$ consists of polynomials. The dimension of $P(K)$ will be denoted by $\dim P(K) = N_K$. □

*Example 5.3. The space $P(K) = P_1(K)$.*The space consisting of linear polynomials on a mesh cell K is denoted by $P_1(K)$:

$$P_1(K) = \left\{ a_0 + \sum_{i=1}^{d} a_i x_i \ : \ \boldsymbol{x} = (x_1, \ldots, x_d)^T \in K \right\}.$$

There are $d+1$ unknown coefficients a_i, $i = 0, \ldots, d$, such that $\dim P_1(K) = N_K = d+1$. □

Remark 5.4. Linear functionals defined on $P(K)$, nodal functionals. For the definition of finite elements, linear functional that are defined on $P(K)$ are of importance. These functionals are called nodal functionals.

Consider linear and continuous functionals $\Phi_{K,1}, \ldots, \Phi_{K,N_K} \ : C^s(K) \to \mathbb{R}$ which are linearly independent. There are different types of functionals that can be utilized in finite element methods:

- point values: $\Phi(v) = v(\boldsymbol{x})$, $\boldsymbol{x} \in K$,
- point values of a first partial derivative: $\Phi(v) = \partial_i v(\boldsymbol{x})$, $\boldsymbol{x} \in K$,

- point values of the normal derivative on a face E of K: $\Phi(v) = \nabla v(\boldsymbol{x}) \cdot \boldsymbol{n}_E$, $\boldsymbol{n}_E$ is the outward pointing unit normal vector on E,
- integral mean values on K: $\Phi(v) = \frac{1}{|K|} \int_K v(\boldsymbol{x}) \; d\boldsymbol{x}$,
- integral mean values on faces E: $\Phi(v) = \frac{1}{|E|} \int_E v(\boldsymbol{s}) \; d\boldsymbol{s}$.

The smoothness parameter s has to be chosen in such a way that the functionals $\Phi_{K,1}, \ldots, \Phi_{K,N_K}$ are continuous. If, e.g., a functional requires the evaluation of a partial derivative or a normal derivative, then one has to choose at least $s = 1$. For the other functionals given above, $s = 0$ is sufficient. □

Definition 5.5. Unisolvence of $P(K)$ with respect to the functionals $\Phi_{K,1}, \ldots, \Phi_{K,N_K}$. The space $P(K)$ is called unisolvent with respect to the functionals $\Phi_{K,1}, \ldots, \Phi_{K,N_K}$ if there is for each $\underline{a} \in \mathbb{R}^{N_K}$, $\underline{a} = (a_1, \ldots, a_{N_K})^T$, exactly one $p \in P(K)$ with

$$\Phi_{K,i}(p) = a_i, \quad 1 \le i \le N_K.$$

□

Remark 5.6. Local basis. Unisolvence means that for each vector $\underline{a} \in \mathbb{R}^{N_K}$, $\underline{a} = (a_1, \ldots, a_{N_K})^T$, there is exactly one element in $P(K)$ such that a_i is the image of the i-th functional, $i = 1, \ldots, N_K$.

Choosing in particular the Cartesian[1] unit vectors for $\underline{a}$, then it follows from the unisolvence that a set $\{\phi_{K,i}\}_{i=1}^{N_K}$ exists with $\phi_{K,i} \in P(K)$ and

$$\Phi_{K,i}(\phi_{K,j}) = \delta_{ij}, \quad i,j = 1, \ldots, N_K.$$

Consequently, the set $\{\phi_{K,i}\}_{i=1}^{N_K}$ forms a basis of $P(K)$. This basis is called local basis. □

Remark 5.7. Transform of an arbitrary basis to the local basis. If an arbitrary basis $\{p_i\}_{i=1}^{N_K}$ of $P(K)$ is known, then the local basis can be computed by solving a linear system of equations. To this end, represent the local basis in terms of the known basis

$$\phi_{K,j} = \sum_{k=1}^{N_K} c_{jk} p_k, \quad c_{jk} \in \mathbb{R}, \ j = 1, \ldots, N_K,$$

with unknown coefficients c_{jk}. Applying the definition of the local basis leads to the linear system of equations

$$\Phi_{K,i}(\phi_{K,j}) = \sum_{k=1}^{N_K} c_{jk} a_{ik} = \delta_{ij}, \quad i,j = 1, \ldots, N_K, \quad a_{ik} = \Phi_{K,i}(p_k).$$

Because of the unisolvence, the matrix $A = (a_{ij})$ is non-singular and the coefficients c_{jk} are determined uniquely. □

[1] René Descartes (1596 – 1650)

Example 5.8. Local basis for the space of linear functions on the reference triangle. Consider the reference triangle $\hat{K}$ with the vertices $(0,0)$, $(1,0)$, and $(0,1)$. A linear space on $\hat{K}$ is spanned by the functions $1, \hat{x}, \hat{y}$. Let the functionals be defined by the values of the functions in the vertices of the reference triangle. Then, the given basis is not a local basis because the function 1 does not vanish at the vertices.

Consider first the vertex $(0,0)$. A linear basis function $a\hat{x}+b\hat{y}+c$ that has the value 1 in $(0,0)$ and that vanishes in the other vertices has to satisfy the following set of equations

$$\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

The solution is $a=-1, b=-1, c=1$. The two other basis functions of the local basis are $\hat{x}$ and $\hat{y}$, such that the local basis has the form $\{1-\hat{x}-\hat{y}, \hat{x}, \hat{y}\}$. □

Remark 5.9. Triangulation, grid, mesh, grid cell. For the definition of global finite element spaces, a decomposition of the domain Ω into polyhedra K is needed. This decomposition is called triangulation $\mathcal{T}^h$ and the polyhedra K are called mesh cells. The union of the polyhedra is called grid or mesh.

A triangulation is called admissible, see the definition in (Ciarlet, 1978, p. 38, p. 51), if:

- It holds $\overline{\Omega} = \cup_{K\in\mathcal{T}^h} K$.
- Each mesh cell $K \in \mathcal{T}^h$ is closed and the interior $\overset{\circ}{K}$ is non-empty.
- For distinct mesh cells K_1 and K_2 there holds $\overset{\circ}{K}_1 \cap \overset{\circ}{K}_2 = \emptyset$.
- For each $K \in \mathcal{T}^h$, the boundary ∂K is Lipschitz continuous.
- The intersection of two mesh cells is either empty or a common m-face, $m \in \{0, \ldots, d-1\}$.

□

Remark 5.10. Global and local functionals. Let

$$\Phi_1, \ldots, \Phi_N \ : \ \{v \in L^\infty(\Omega) \ : \ v|_K \in P(K)\} \to \mathbb{R}$$

be continuous linear functionals of the same types as given in Remark 5.4, where for each K, $v|_K \in P(K)$ has to be understood in the sense that the polynomial in K is extended continuously to the boundary of K. The restriction of the functionals to $C^s(K)$ defines a set of local functionals $\Phi_{K,1}, \ldots, \Phi_{K,N_K}$, where it is assumed that the local functionals are unisolvent on $P(K)$. The union of all mesh cells K_j, for which there is a $p \in P(K_j)$ with $\Phi_i(p) \neq 0$, will be denoted by ω_i. □

Example 5.11. On subdomains ω_i. Consider the two-dimensional case and let Φ_i be defined as nodal value of a function in $\boldsymbol{x} \in K$. If $\boldsymbol{x} \in \overset{\circ}{K}$, then $\omega_i = K$. In the case that $\boldsymbol{x}$ is on a face of K but not in a vertex, then ω_i is the union

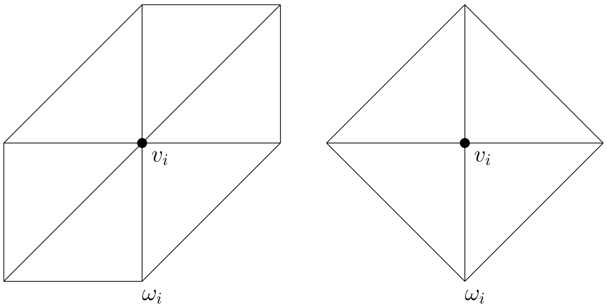

Fig. 5.1 Subdomains ω_i.

of K and the other mesh cell whose boundary contains this face. Last, if $\boldsymbol{x}$ is a vertex of K, then ω_i is the union of all mesh cells that possess this vertex, see Figure 5.1. □

Definition 5.12. Finite element space, global basis. A function $v(\boldsymbol{x})$ defined on Ω with $v|_K \in P(K)$ for all $K \in \mathcal{T}^h$ is called continuous with respect to a global functional Φ_i defined in Remark 5.10 if

$$\Phi_i(v|_{K_1}) = \Phi_i(v|_{K_2}), \quad \forall\, K_1, K_2 \in \omega_i.$$

The space

$$S = \Big\{ v \in L^\infty(\Omega) \ : \ v|_K \in P(K) \text{ and } v \text{ is continuous with respect to } \Phi_i, i = 1, \ldots, N \Big\}$$

is called finite element space.

The global basis $\{\phi_j\}_{j=1}^N$ of S is defined by the condition

$$\phi_j \in S, \quad \Phi_i(\phi_j) = \delta_{ij}, \quad i, j = 1, \ldots, N.$$

□

Example 5.13. Piecewise linear global basis function. Figure 5.2 shows a piecewise linear global basis function in two dimensions. Because of its form, such a function is called hat function. □

Remark 5.14. On global basis functions. A global basis function coincides on each mesh cell with a local basis function. This property implies the uniqueness of the global basis functions.

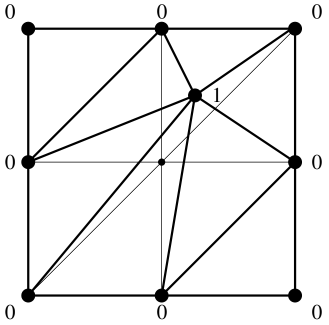

Fig. 5.2 Piecewise linear global basis function (boldface lines), hat function.

Whether the continuity with respect to $\{\Phi_i\}_{i=1}^N$ implies the continuity of the finite element functions depends on the functionals that define the finite element space. □

Definition 5.15. Parametric finite elements. Let $\hat{K}$ be a reference mesh cell with the local space $\hat{P}(\hat{K})$, the local functionals $\hat{\Phi}_1, \ldots, \hat{\Phi}_{\hat{N}}$, and a class of bijective mappings $\{F_K \ : \ \hat{K} \to K\}$. A finite element space is called a parametric finite element space if:

- The images $\{K\}$ of $\{F_K\}$ form the set of mesh cells.
- The local spaces are given by

$$P(K) = \left\{p \ : \ p = \hat{p} \circ F_K^{-1}, \hat{p} \in \hat{P}(\hat{K})\right\}. \tag{5.1}$$

- The local functionals are defined by

$$\Phi_{K,i}(v(\boldsymbol{x})) = \hat{\Phi}_i\left(\hat{v}(\hat{\boldsymbol{x}})\right) = \hat{\Phi}_i\left(v(F_K(\hat{\boldsymbol{x}}))\right), \tag{5.2}$$

  where $\hat{\boldsymbol{x}} = (\hat{x}_1, \ldots, \hat{x}_d)^T$ are the coordinates of the reference mesh cell and it holds $\boldsymbol{x} = F_K(\hat{\boldsymbol{x}})$, $\hat{v} = v \circ F_K$.

□

Remark 5.16. Motivations for using parametric finite elements. Definition 5.12 of finite elements spaces is very general. For instance, different types of mesh cells are allowed. However, as well the finite element theory as the implementation of finite element methods become much simpler if only parametric finite elements are considered. □

5.2 Finite Elements on Simplices

Definition 5.17. ***d*-simplex.** A d-simplex $K \subset \mathbb{R}^d$ is the convex hull of $(d+1)$ points $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{d+1} \in \mathbb{R}^d$ which form the vertices of K. □

Remark 5.18. On d-simplices. It will be always assumed that the simplex is not degenerated, i.e., its d-dimensional measure is positive. This property is equivalent to the non-singularity of the matrix (*exercise*)

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1,d+1} \\ a_{21} & a_{22} & \ldots & a_{2,d+1} \\ \vdots & \vdots & \ddots & \vdots \\ a_{d1} & a_{d2} & \ldots & a_{d,d+1} \\ 1 & 1 & \ldots & 1 \end{pmatrix},$$

where $\boldsymbol{a}_i = (a_{1i}, a_{2i}, \ldots, a_{di})^T$, $i = 1, \ldots, d+1$.

For $d = 2$, the simplices are the triangles and for $d = 3$ they are the tetrahedra. □

Definition 5.19. Barycentric coordinates. Since K is the convex hull of the points $\{\boldsymbol{a}_i\}_{i=1}^{d+1}$, the parametrization of K with a convex combination of the vertices reads as follows

$$K = \left\{ \boldsymbol{x} \in \mathbb{R}^d \ : \ \boldsymbol{x} = \sum_{i=1}^{d+1} \lambda_i \boldsymbol{a}_i,\ 0 \le \lambda_i \le 1,\ \sum_{i=1}^{d+1} \lambda_i = 1 \right\}.$$

The coefficients $\lambda_1, \ldots, \lambda_{d+1}$ are called barycentric coordinates of $\boldsymbol{x} \in K$. □

Remark 5.20. On barycentric coordinates.

- From the definition, it follows that the barycentric coordinates are the solution of the linear system of equations

  $$\sum_{i=1}^{d+1} a_{ji}\lambda_i = x_j, \quad 1 \le j \le d, \quad \sum_{i=1}^{d+1} \lambda_i = 1.$$

  Since the system matrix is non-singular, see Remark 5.18, the barycentric coordinates are determined uniquely.
- The barycentric coordinates of the vertex $\boldsymbol{a}_i$, $i = 1, \ldots, d+1$, of the simplex are $\lambda_i = 1$ and $\lambda_j = 0$ if $i \ne j$. Since $\lambda_i(\boldsymbol{a}_j) = \delta_{ij}$, the barycentric coordinate λ_i can be identified with the linear function that has the value 1 in the vertex $\boldsymbol{a}_i$ and that vanishes in all other vertices $\boldsymbol{a}_j$ with $j \ne i$.
- The barycenter of the simplex is given by

  $$S_K = \frac{1}{d+1} \sum_{i=1}^{d+1} \boldsymbol{a}_i = \sum_{i=1}^{d+1} \frac{1}{d+1} \boldsymbol{a}_i.$$

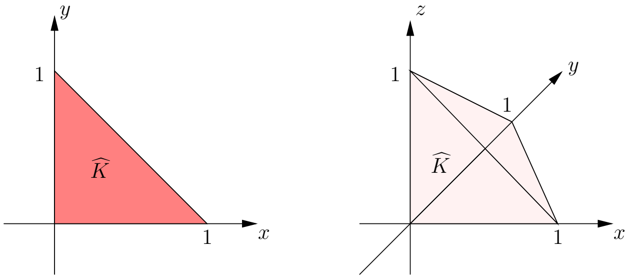

Fig. 5.3 The unit simplices in two and three dimensions.

Hence, its barycentric coordinates are $\lambda_i = 1/(d+1)$, $i = 1, \ldots, d+1$. □

Remark 5.21. Simplicial reference mesh cells. A commonly used reference mesh cell for triangles and tetrahedra is the unit simplex

$$\hat{K} = \left\{ \hat{\boldsymbol{x}} \in \mathbb{R}^d \ : \ \sum_{i=1}^{d} \hat{x}_i \le 1, \ \hat{x}_i \ge 0, \ i = 1, \ldots, d \right\},$$

see Figure 5.3. The class $\{F_K\}$ of admissible mappings are the bijective affine mappings

$$F_K \hat{\boldsymbol{x}} = B_K \hat{\boldsymbol{x}} + \boldsymbol{b}, \qquad B_K \in \mathbb{R}^{d \times d}, \ \det(B_K) \neq 0, \ \boldsymbol{b} \in \mathbb{R}^d. \tag{5.3}$$

The images of these mappings generate the set of the non-degenerated simplices $\{K\} \subset \mathbb{R}^d$. □

Definition 5.22. Affine family of simplicial finite elements. Given a simplicial reference mesh cell $\hat{K}$, affine mappings $\{F_K\}$, and an unisolvent set of functionals on $\hat{K}$. Using (5.1) and (5.2), one obtains a local finite element space on each non-degenerated simplex. The set of these local spaces is called affine family of simplicial finite elements. □

Definition 5.23. Polynomial space P_k. Let $\boldsymbol{x} = (x_1, \ldots, x_d)^T$, $k \in \mathbb{N} \cup \{0\}$, and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)^T$. Then, the polynomial space P_k is given by

$$P_k = \operatorname{span} \left\{ \prod_{i=1}^{d} x_i^{\alpha_i} = \boldsymbol{x}^{\boldsymbol{\alpha}} \ : \ \alpha_i \in \mathbb{N} \cup \{0\} \ \text{ for } \ i = 1, \ldots, d, \ \sum_{i=1}^{d} \alpha_i \le k \right\}.$$

□

Remark 5.24. Lagrangian[2] finite elements. In many examples given below, the linear functionals on the reference mesh cell $\hat{K}$ are the values of the

[2] Joseph-Louis de Lagrange (1736 – 1813)

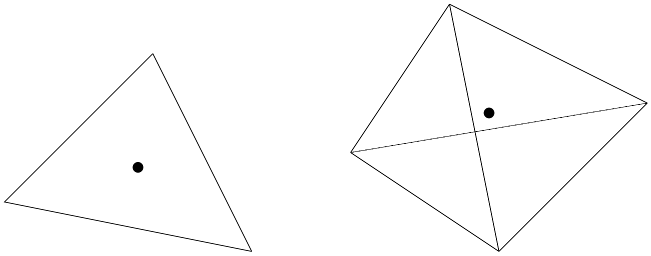

Fig. 5.4 The finite element $P_0(K)$.

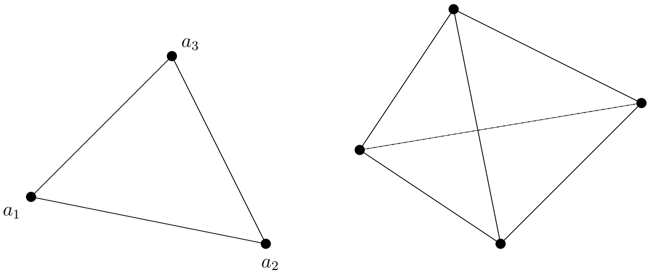

Fig. 5.5 The finite element $P_1(K)$.

polynomials with the same barycentric coordinates as on the general mesh cell K. Finite elements whose linear functionals are values of the polynomials on certain points in K are called Lagrangian finite elements. □

Example 5.25. P_0 : piecewise constant finite element. The piecewise constant finite element space consists of discontinuous functions. The linear functional is the value of the polynomial in the barycenter of the mesh cell, see Figure 5.4. It is $\dim P_0(K) = 1$. □

Example 5.26. P_1 : conforming piecewise linear finite element. This finite element space is a subspace of $C(\overline{\Omega})$. The linear functionals are the values of the function in the vertices of the mesh cells, see Figure 5.5. It follows that $\dim P_1(K) = d + 1$.

The local basis for the functionals $\{\Phi_i(v) = v(\boldsymbol{a}_i),\ i = 1, \ldots, d+1\}$ is $\{\lambda_i\}_{i=1}^{d+1}$ since $\Phi_i(\lambda_j) = \delta_{ij}$, compare Remark 5.20. Since a local basis exists, the functionals are unisolvent with respect to the polynomial space $P_1(K)$.

Now, it will be shown that the corresponding finite element space consists of continuous functions. Let K_1, K_2 be two mesh cells with the common face E and let $v \in P_1(= S)$. The restriction of v_{K_1} on E is a linear function on E as well as the restriction of v_{K_2} on E. It has to be shown that both linear functions are identical. A linear function on the $(d-1)$-dimensional face E is

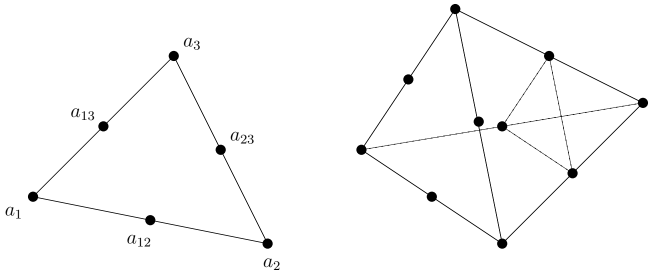

Fig. 5.6 The finite element $P_2(K)$.

uniquely determined with d linearly independent functionals that are defined on E. These functionals can be chosen to be the values of the function in the d vertices of E. The functionals in S are continuous by the definition of S. Thus, it must hold that both restrictions on E have the same values in the vertices of E. Hence, it is $v_{K_1}|_E = v_{K_2}|_E$ and the functions from P_1 are continuous. □

Example 5.27. P_2 : conforming piecewise quadratic finite element. This finite element space is also a subspace of $C(\overline{\Omega})$. It consists of piecewise quadratic functions. The functionals are the values of the functions in the $d+1$ vertices of the mesh cell and the values of the functions in the centers of the edges, see Figure 5.6. Since each vertex is connected to each other vertex, there are $\sum_{i=1}^{d} i = d(d+1)/2$ edges. Hence, it follows that $\dim P_2(K) = (d+1)(d+2)/2$.

The part of the local basis that belongs to the functionals $\{\Phi_i(v) = v(\boldsymbol{a}_i),$ $i = 1, \ldots, d+1\}$, is given by

$$\{\phi_i(\lambda) = \lambda_i(2\lambda_i - 1), \quad i = 1, \ldots, d+1\}.$$

Denote the center of the edges between the vertices $\boldsymbol{a}_i$ and $\boldsymbol{a}_j$ by $\boldsymbol{a}_{ij}$. The corresponding part of the local basis is given by

$$\{\phi_{ij} = 4\lambda_i\lambda_j, \quad i,j = 1, \ldots, d+1,\ i < j\}.$$

The unisolvence follows from the fact that there exists a local basis. The continuity of the corresponding finite element space is shown in the same way as for the P_1 finite element. The restriction of a quadratic function defined in a mesh cell to a face E is a quadratic function on that face. Hence, the function on E is determined uniquely with $d(d+1)/2$ linearly independent functionals on E.

The functions ϕ_{ij} are called in two dimensions edge bubble functions. □

Example 5.28. P_3 : conforming piecewise cubic finite element. This finite element space consists of continuous piecewise cubic functions. It is a subspace of $C(\overline{\Omega})$. The functionals in a mesh cell K are defined to be the values in

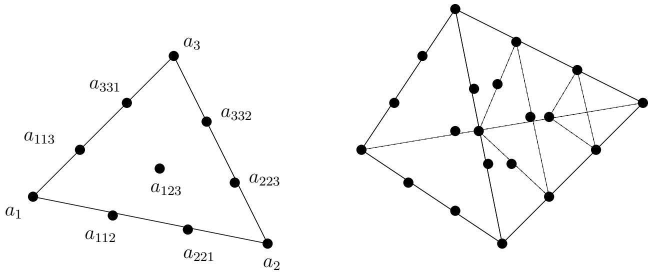

Fig. 5.7 The finite element $P_3(K)$.

the vertices ($(d+1)$ values), two values on each edge (dividing the edge in three parts of equal length) ($2\sum_{i=1}^{d} i = d(d+1)$ values), and the values in the barycenter of the 2-faces of K, see Figure 5.7. Each 2-face of K is defined by three vertices. If one considers for each vertex all possible pairs with other vertices, then each 2-face is counted three times. Hence, there are $(d+1)(d-1)d/6$ 2-faces. The dimension of $P_3(K)$ is given by

$$\dim P_3(K) = (d+1) + d(d+1) + \frac{(d-1)d(d+1)}{6} = \frac{(d+1)(d+2)(d+3)}{6}.$$

For the functionals

$$\left\{\begin{array}{ll} \Phi_i(v) = v(\boldsymbol{a}_i),\ i = 1,\ldots,d+1, & \text{(vertex)}, \\ \Phi_{iij}(v) = v(\boldsymbol{a}_{iij}),\ i,j = 1,\ldots,d+1, i \neq j, & \text{(point on edge)}, \\ \Phi_{ijk}(v) = v(\boldsymbol{a}_{ijk}),\ i = 1,\ldots,d+1, i<j<k, & \text{(point on 2-face)} \end{array}\right\},$$

the local basis is given by

$$\left\{\begin{array}{l} \phi_i(\lambda) = \frac{1}{2}\lambda_i(3\lambda_i - 1)(3\lambda_i - 2), \quad \phi_{iij}(\lambda) = \frac{9}{2}\lambda_i\lambda_j(3\lambda_i - 1), \\ \phi_{ijk}(\lambda) = 27\lambda_i\lambda_j\lambda_k \end{array}\right\}.$$

In two dimensions, the function $\phi_{ijk}(\lambda)$ is called cell bubble function. □

Example 5.29. Cubic Hermite[3] *element.* The finite element space is a subspace of $C(\overline{\Omega})$, its dimension is $(d+1)(d+2)(d+3)/6$ and the functionals are the values of the function in the vertices of the mesh cell ($(d+1)$ values), the value of the barycenter at the 2-faces of K ($(d+1)(d-1)d/6$ values), and the partial derivatives at the vertices ($d(d+1)$ values), see Figure 5.8. The

[3] Charles Hermite (1822 – 1901)

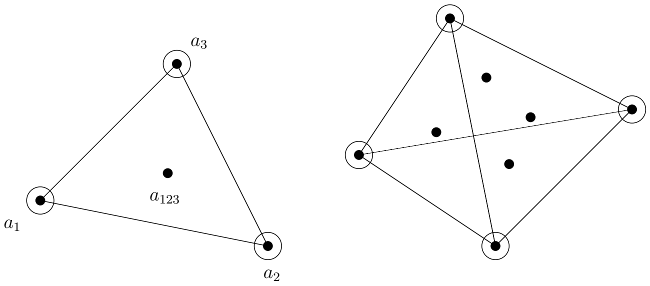

Fig. 5.8 The cubic Hermite element.

dimension is the same as for the P_3 element. Hence, the local polynomials can be defined to be cubic.

This finite element does not define an affine family in the strict sense, because partial derivatives on the reference cell are mapped to directional derivatives on the physical cell. Concretely, the functionals for the partial derivatives $\hat{\Phi}_i(\hat{v}) = \partial_i \hat{v}(\mathbf{0})$ on the reference cell are mapped to the functionals $\Phi_i(v) = \partial_{\boldsymbol{t}_i} v(\boldsymbol{a})$, where $\boldsymbol{a} = F_K(\mathbf{0})$ and $\boldsymbol{t}_i$ are the directions of edges which are adjacent to $\boldsymbol{a}$, i.e., $\boldsymbol{a}$ is an end point of this edge. This property suffices to control all first derivatives. One has to take care of this property in the implementation of this finite element.

Because of this property, one can use the derivatives in the direction of the edges as functionals

$$\begin{aligned} \Phi_i(v) &= v(\boldsymbol{a}_i), && \text{(vertices)}\\ \Phi_{ij}(v) &= \nabla v(\boldsymbol{a}_i)\cdot(\boldsymbol{a}_j - \boldsymbol{a}_i),\ i,j = 1,\ldots,d-1, i \neq j, && \text{(directional deriv.)}\\ \Phi_{ijk}(v) &= v(\boldsymbol{a}_{ijk}),\ i<j<k, && \text{(2-faces)} \end{aligned}$$

with the corresponding local basis

$$\begin{aligned} \phi_i(\lambda) &= -2\lambda_i^3 + 3\lambda_i^2 - 7\lambda_i \textstyle\sum_{j<k,j\neq i,k\neq i} \lambda_j\lambda_k,\\ \phi_{ij}(\lambda) &= \lambda_i\lambda_j(2\lambda_i - \lambda_j - 1),\\ \phi_{ijk}(\lambda) &= 27\lambda_i\lambda_j\lambda_k. \end{aligned}$$

The proof of the unisolvence can be found in the literature.

Here, the continuity of the functions will be shown only for $d = 2$. Let K_1, K_2 be two mesh cells with the common edge E and the unit tangential vector $\boldsymbol{t}$. Let V_1, V_2 be the end points of E. The restrictions $v|_{K_1}, v|_{K_2}$ to E satisfy four conditions

$$v|_{K_1}(V_i) = v|_{K_2}(V_i), \quad \partial_{\boldsymbol{t}} v|_{K_1}(V_i) = \partial_{\boldsymbol{t}} v|_{K_2}(V_i),\ i = 1,2.$$

Since both restrictions are cubic polynomials and four conditions have to be satisfied, their values coincide on E.

The cubic Hermite finite element possesses an advantage in comparison with the P_3 finite element. For $d = 2$, it holds for a regular triangulation $\mathcal{T}^h$ that

$$\#(K) \approx 2\#(V), \quad \#(E) \approx 2\#(V),$$

where $\#(\cdot)$ denotes the number of triangles, nodes, and edges, respectively. Hence, the dimension of P_3 is approximately $\#(V) + 2\#(E) + \#(K) \approx 7\#(V)$, whereas the dimension of the cubic Hermite element is approximately $3\#(V)+\#(K) \approx 5\#(V)$. This difference comes from the fact that both spaces are different proper subspaces of the space of all continuous piecewise cubic functions. The elements of both spaces are continuous functions, but for the functions of the cubic Hermite finite element, in addition, the first derivatives are continuous at the nodes. That means, these two spaces are different finite element spaces whose degree of the local polynomial space is the same (cubic). One can see at this example the importance of the functionals for the definition of the global finite element space. □

Example 5.30. P_1^{nc} : non-conforming linear finite element, Crouzeix–Raviart finite element, Crouzeix & Raviart (1973). This finite element consists of piecewise linear but discontinuous functions. The functionals are given by the values of the functions in the barycenters of the faces such that $\dim P_1^{\mathrm{nc}}(K) = (d+1)$. It follows from the definition of the finite element space, Definition 5.12, that the functions from P_1^{nc} are continuous in the barycenter of the faces

$$P_1^{\mathrm{nc}} = \big\{v \in L^2(\Omega) \ : \ v|_K \in P_1(K), \ v(\boldsymbol{x}) \text{ is continuous at the barycenter of all faces}\big\}. \tag{5.4}$$

Equivalently, the functionals can be defined to be the integral mean values on the faces and then the global space is defined to be

$$P_1^{\mathrm{nc}} = \left\{ v \in L^2(\Omega) \ : \ v|_K \in P_1(K), \right. \left. \int_E v|_K \, d\boldsymbol{s} = \int_E v|_{K'} \, d\boldsymbol{s} \ \forall\, E \in \mathcal{E}(K) \cap \mathcal{E}(K') \right\}, \tag{5.5}$$

where $\mathcal{E}(K)$ is the set of all $(d-1)$-dimensional faces of K.

For the description of this finite element, one defines the functionals by

$$\Phi_i(v) = v(\boldsymbol{a}_{i-1,i+1}) \text{ for } d = 2, \quad \Phi_i(v) = v(\boldsymbol{a}_{i-2,i-1,i+1}) \text{ for } d = 3,$$

where the points are the barycenters of the faces with the vertices that correspond to the indices, see Figure 5.9. This system is unisolvent with the local basis

$$\phi_i(\lambda) = 1 - d\lambda_i, \quad i = 1, \ldots, d+1.$$

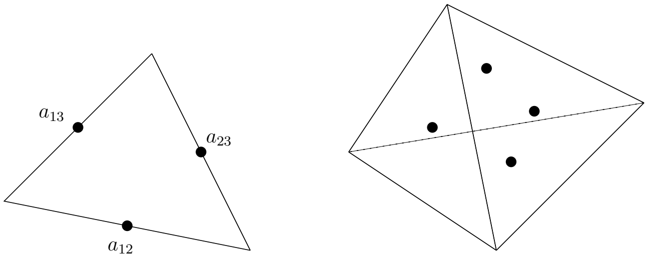

Fig. 5.9 The finite element $P_1^{\mathrm{nc}}(K)$.

□

5.3 Finite Elements on Parallelepipeds and Quadrilaterals

Remark 5.31. Reference mesh cells, reference map to parallelepipeds. One can find in the literature two reference cells: the unit cube $[0,1]^d$ and the large unit cube $[-1,1]^d$. It does not matter which reference cell is chosen. Here, the large unit cube will be used: $\hat{K} = [-1,1]^d$. The class of admissible reference maps $\{F_K\}$ to parallelepipeds consists of bijective affine mappings of the form

$$F_K \hat{\boldsymbol{x}} = B_K \hat{\boldsymbol{x}} + \boldsymbol{b}, \quad B_K \in \mathbb{R}^{d\times d},\ \boldsymbol{b} \in \mathbb{R}^d.$$

If B_K is a diagonal matrix, then $\hat{K}$ is mapped to d-rectangles.

The class of mesh cells that is obtained in this way is not sufficient to triangulate general domains. If one wants to use more general mesh cells than parallelepipeds, then the class of admissible reference maps has to be enlarged, see Remark 5.40. □

Definition 5.32. Polynomial space Q_k. Let $\boldsymbol{x} = (x_1,\ldots,x_d)^T$ and denote by $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_d)^T$ a multi-index. Then, the polynomial space Q_k is given by

$$Q_k = \operatorname{span}\left\{\prod_{i=1}^d x_i^{\alpha_i} = \boldsymbol{x}^{\boldsymbol{\alpha}} \ : \ 0 \le \alpha_i \le k \ \text{ for } \ i = 1,\ldots,d\right\}.$$

□

Example 5.33. Q_1 vs. P_1. The space Q_1 consists of all polynomials that are d-linear. Let $d = 2$, then it is

$$Q_1 = \operatorname{span}\{1, x, y, xy\},$$

whereas

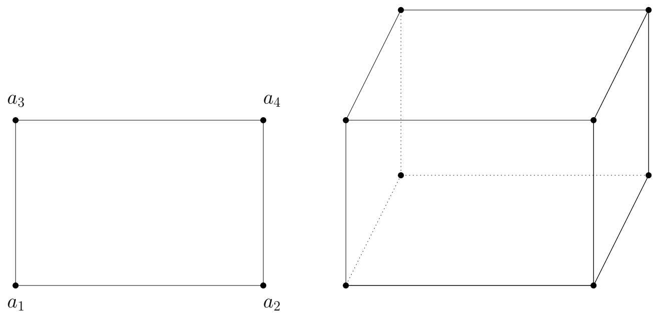

Fig. 5.10 The finite element $Q_1(K)$.

$$P_1 = \operatorname{span}\{1, x, y\}.$$

□

Remark 5.34. Finite elements on d-rectangles. For simplicity of presentation, the examples below consider d-rectangles. In this case, the finite elements are just tensor products of one-dimensional finite elements. In particular, the basis functions can be written as products of one-dimensional basis functions. □

Example 5.35. Q_0 *: piecewise constant finite element.* Similarly to the P_0 space, the space Q_0 consists of piecewise constant, discontinuous functions. The functional is the value of the function in the barycenter of the mesh cell K and it holds $\dim Q_0(K) = 1$. □

Example 5.36. Q_1 *: conforming piecewise d-linear finite element.* This finite element space is a subspace of $C(\overline{\Omega})$. The functionals are the values of the function in the vertices of the mesh cell, see Figure 5.10. Hence, it is $\dim Q_1(K) = 2^d$.

The one-dimensional local basis functions, which will be used for the tensor product, are given by

$$\hat{\phi}_1(\hat{x}) = \frac{1}{2}(1 - \hat{x}), \quad \hat{\phi}_2(\hat{x}) = \frac{1}{2}(1 + \hat{x}).$$

With these functions, e.g., the basis functions in two dimensions are computed by

$$\hat{\phi}_1(\hat{x})\hat{\phi}_1(\hat{y}),\ \hat{\phi}_1(\hat{x})\hat{\phi}_2(\hat{y}),\ \hat{\phi}_2(\hat{x})\hat{\phi}_1(\hat{y}),\ \hat{\phi}_2(\hat{x})\hat{\phi}_2(\hat{y}).$$

The continuity of the functions of the finite element space Q_1 is proved in the same way as for simplicial finite elements. It is used that the restriction of a function from $Q_k(K)$ to a face E is a function from the space $Q_k(E)$, $k \geq 1$. □

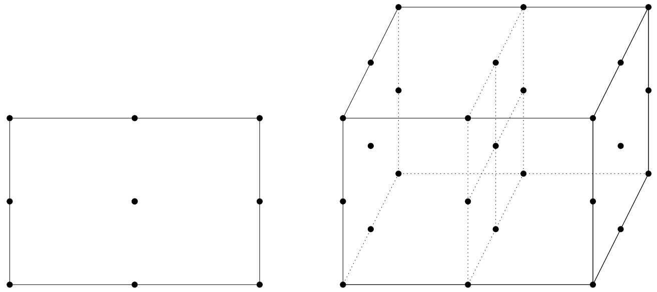

Fig. 5.11 The finite element $Q_2(K)$.

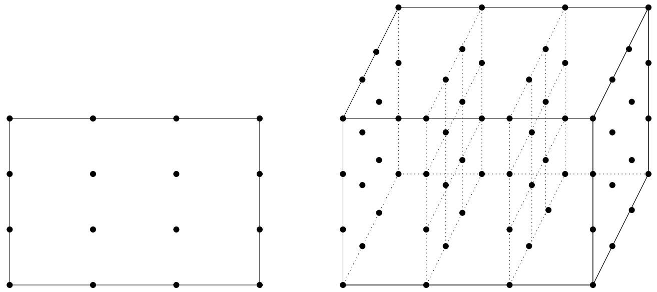

Fig. 5.12 The finite element $Q_3(K)$.

Example 5.37. Q_2 : conforming piecewise d-quadratic finite element. It holds that $Q_2 \subset C(\overline{\Omega})$. The functionals in one dimension are the values of the function at both ends of the interval and in the center of the interval, see Figure 5.11. In d dimensions, they are the corresponding values of the tensor product of the intervals. It follows that $\dim Q_2(K) = 3^d$.

The one-dimensional basis function on the reference interval are defined by

$$\hat{\phi}_1(\hat{x}) = -\frac{1}{2}\hat{x}(1-\hat{x}), \quad \hat{\phi}_2(\hat{x}) = (1-\hat{x})(1+\hat{x}), \quad \hat{\phi}_3(\hat{x}) = \frac{1}{2}(1+\hat{x})\hat{x}.$$

The basis function $\prod_{i=1}^{d} \hat{\phi}_2(\hat{x}_i)$ is called cell bubble function. □

Example 5.38. Q_3 : conforming piecewise d-cubic finite element. This finite element space is a subspace of $C(\overline{\Omega})$. The functionals on the reference interval are given by the values at the end of the interval and the values at the points $\hat{x} = -1/3$, $\hat{x} = 1/3$. In multiple dimensions, it is the corresponding tensor product, see Figure 5.12. The dimension of the local space is $\dim Q_3(K) = 4^d$.

The one-dimensional basis functions in the reference interval are given by

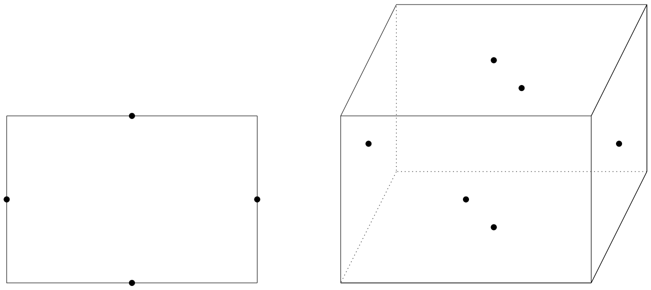

Fig. 5.13 The finite element $Q_1^{\text{rot}}(K)$.

$$\hat{\phi}_1(\hat{x}) = -\frac{1}{16}(3\hat{x}+1)(3\hat{x}-1)(\hat{x}-1), \quad \hat{\phi}_2(\hat{x}) = \frac{9}{16}(\hat{x}+1)(3\hat{x}-1)(\hat{x}-1),$$
$$\hat{\phi}_3(\hat{x}) = -\frac{9}{16}(\hat{x}+1)(3\hat{x}+1)(\hat{x}-1), \quad \hat{\phi}_4(\hat{x}) = \frac{1}{16}(3\hat{x}+1)(3\hat{x}-1)(\hat{x}+1).$$

□

Example 5.39. Q_1^{rot} : rotated non-conforming element of lowest order, Rannacher–Turek element, Rannacher & Turek (1992): This finite element space is a generalization of the P_1^{nc} finite element to quadrilateral and hexahedral mesh cells. It consists of discontinuous functions that are continuous at the barycenter of the faces. The dimension of the local finite element space is $\dim Q_1^{\text{rot}}(K) = 2d$. The space on the reference mesh cell is defined by

$$Q_1^{\text{rot}}\left(\hat{K}\right) = \left\{\hat{p} \ : \ \hat{p} \in \operatorname{span}\{1, \hat{x}, \hat{y}, \hat{x}^2 - \hat{y}^2\}\right\} \qquad \text{for } d = 2,$$
$$Q_1^{\text{rot}}\left(\hat{K}\right) = \left\{\hat{p} \ : \ \hat{p} \in \operatorname{span}\{1, \hat{x}, \hat{y}, \hat{z}, \hat{x}^2 - \hat{y}^2, \hat{y}^2 - \hat{z}^2\}\right\} \text{ for } d = 3.$$

Note that the transformed space

$$Q_1^{\text{rot}}(K) = \{p = \hat{p} \circ F_K^{-1}, \hat{p} \in Q_1^{\text{rot}}(\hat{K})\}$$

contains polynomials of the form $ax^2 - by^2$, where a, b depend on F_K.

For $d = 2$, the local basis on the reference cell is given by

$$\hat{\phi}_1(\hat{x}, \hat{y}) = -\frac{3}{8}(\hat{x}^2 - \hat{y}^2) - \frac{1}{2}\hat{y} + \frac{1}{4}, \quad \hat{\phi}_2(\hat{x}, \hat{y}) = \frac{3}{8}(\hat{x}^2 - \hat{y}^2) + \frac{1}{2}\hat{x} + \frac{1}{4},$$
$$\hat{\phi}_3(\hat{x}, \hat{y}) = -\frac{3}{8}(\hat{x}^2 - \hat{y}^2) + \frac{1}{2}\hat{y} + \frac{1}{4}, \quad \hat{\phi}_4(\hat{x}, \hat{y}) = \frac{3}{8}(\hat{x}^2 - \hat{y}^2) - \frac{1}{2}\hat{x} + \frac{1}{4}. \tag{5.6}$$

Analogously to the Crouzeix–Raviart finite element, the functionals can be defined as point values of the functions in the barycenters of the faces, see Figure 5.13, or as integral mean values of the functions at the faces.

Consequently, the finite element spaces are defined in the same way as (5.4) or (5.5), with $P_1^{\mathrm{nc}}(K)$ replaced by $Q_1^{\mathrm{rot}}(K)$.

In the code ParMooN Wilbrandt *et al.* (2017), the mean value oriented Q_1^{rot} finite element space is implemented for two dimensions and the point value oriented Q_1^{rot} finite element space for three dimensions. For $d = 3$, the integrals on the faces of mesh cells, whose equality is required in the mean value oriented Q_1^{rot} finite element space, involve a weighting function which depends on the particular mesh cell K. The computation of these weighting functions for all mesh cells is an additional computational overhead. For this reason, it was suggested in (Schieweck, 1997, p. 21) to use for $d = 3$ the simpler point value oriented form of the Q_1^{rot} finite element. □

Remark 5.40. Parametric mappings. The image of an affine mapping of the reference mesh cell $\hat{K} = [-1,1]^d$, $d \in \{2,3\}$, is a parallelepiped. If one wants to consider finite elements on general d-quadrilaterals, then the class of admissible reference maps has to be enlarged.

The simplest non-affine parametric finite element on quadrilaterals in two dimensions uses bilinear mappings. Let $\hat{K} = [-1,1]^2$ and let

$$F_K(\hat{\boldsymbol{x}}) = \begin{pmatrix} F_K^1(\hat{\boldsymbol{x}}) \\ F_K^2(\hat{\boldsymbol{x}}) \end{pmatrix} = \begin{pmatrix} a_{11} + a_{12}\hat{x} + a_{13}\hat{y} + a_{14}\hat{x}\hat{y} \\ a_{21} + a_{22}\hat{x} + a_{23}\hat{y} + a_{24}\hat{x}\hat{y} \end{pmatrix}, \ F_K^i \in Q_1, \ i = 1, 2,$$

be a bilinear mapping from $\hat{K}$ on the class of admissible quadrilaterals. A quadrilateral K is called admissible if

- the length of all edges of K is larger than zero,
- the interior angles of K are smaller than π, i.e., K is convex.

This class contains, e.g., trapezoids and rhombi. □

Remark 5.41. Parametric finite element functions. The functions of the local space $P(K)$ on the mesh cell K are defined by $p = \hat{p} \circ F_K^{-1}$. These functions are in general rational functions. However, using d-linear mappings, then the restriction of F_K on an edge of $\hat{K}$ is an affine map. For instance, in the case of the Q_1 finite element, the functions on K are linear functions on each edge of K. It follows that the functions of the corresponding finite element space are continuous, compare Example 5.26. □

5.4 Transform of Integrals

Remark 5.42. Motivation. The transformation of integrals from the reference mesh cell to mesh cells of the grid and vice versa is used as well for the analysis as for the implementation of finite element methods. This section provides an overview of the most important formulae for transformations.

Let $\hat{K} \subset \mathbb{R}^d$ be the reference mesh cell, K be an arbitrary mesh cell, and $F_K \ : \ \hat{K} \to K$ with $\boldsymbol{x} = F_K(\hat{\boldsymbol{x}})$ be the reference map. It is assumed that the reference map is a continuous differentiable one-to-one map. The inverse map is denoted by $F_K^{-1} \ : \ K \to \hat{K}$. For the integral transforms, the derivatives (Jacobians) of F_K and F_K^{-1} are needed

$$DF_K(\hat{\boldsymbol{x}})_{ij} = \frac{\partial x_i}{\partial \hat{x}_j}, \quad DF_K^{-1}(\boldsymbol{x})_{ij} = \frac{\partial \hat{x}_i}{\partial x_j}, \quad i,j = 1,\ldots,d.$$

□

Remark 5.43. Integral with a function without derivatives. This integral transforms with the standard rule of integral transforms

$$\int_K v(\boldsymbol{x}) \ d\boldsymbol{x} = \int_{\hat{K}} \hat{v}(\hat{\boldsymbol{x}}) \left|\det DF_K(\hat{\boldsymbol{x}})\right| \ d\hat{\boldsymbol{x}}, \tag{5.7}$$

where $\hat{v}(\hat{\boldsymbol{x}}) = v(F_K(\hat{\boldsymbol{x}}))$. □

Remark 5.44. Transform of derivatives. Using the chain rule, one obtains

$$\begin{aligned}
\frac{\partial v}{\partial x_i}(\boldsymbol{x}) &= \sum_{j=1}^d \frac{\partial \hat{v}}{\partial \hat{x}_j}(\hat{\boldsymbol{x}}) \frac{\partial \hat{x}_j}{\partial x_i} = \nabla_{\hat{\boldsymbol{x}}} \hat{v}(\hat{\boldsymbol{x}}) \cdot \left(\left(DF_K^{-1}(\boldsymbol{x})\right)^T \right)_i \\
&= \nabla_{\hat{\boldsymbol{x}}} \hat{v}(\hat{\boldsymbol{x}}) \cdot \left(\left(DF_K^{-1}(F_K(\hat{\boldsymbol{x}}))\right)^T \right)_i \\
&= \left(\left(DF_K^{-1}(F_K(\hat{\boldsymbol{x}}))\right)^T \right)_i \cdot \nabla_{\hat{\boldsymbol{x}}} \hat{v}(\hat{\boldsymbol{x}}), \qquad (5.8) \\
\frac{\partial \hat{v}}{\partial \hat{x}_i}(\hat{\boldsymbol{x}}) &= \sum_{j=1}^d \frac{\partial v}{\partial x_j}(\boldsymbol{x}) \frac{\partial x_j}{\partial \hat{x}_i} = \nabla v(\boldsymbol{x}) \cdot \left(\left(DF_K(\hat{\boldsymbol{x}})\right)^T \right)_i \\
&= \nabla v(\boldsymbol{x}) \cdot \left(\left(DF_K(F_K^{-1}(\boldsymbol{x}))\right)^T \right)_i. \qquad (5.9)
\end{aligned}$$

The index i denotes the i-th row of a matrix. Derivatives on the reference mesh cell are marked with a symbol on the operator. □

Remark 5.45. Integrals with a gradients. Using the rule for transforming integrals and (5.8) gives

$$\begin{aligned}
&\int_K \boldsymbol{b}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x}) \ d\boldsymbol{x} \\
&= \int_{\hat{K}} \boldsymbol{b}\left(F_K(\hat{\boldsymbol{x}})\right) \cdot \left[\left(DF_K^{-1}\right)^T \left(F_K(\hat{\boldsymbol{x}})\right) \right] \nabla_{\hat{\boldsymbol{x}}} \hat{v}(\hat{\boldsymbol{x}}) \left|\det DF_K(\hat{\boldsymbol{x}})\right| \ d\hat{\boldsymbol{x}}. \qquad (5.10)
\end{aligned}$$

Similarly, one obtains

$$\begin{aligned}&\int_K \nabla v(\boldsymbol{x}) \cdot \nabla w(\boldsymbol{x})\ d\boldsymbol{x} \\ &= \int_{\hat{K}} \left[\left(DF_K^{-1}\right)^T (F_K(\hat{\boldsymbol{x}}))\right] \nabla_{\hat{\boldsymbol{x}}} \hat{v}(\hat{\boldsymbol{x}}) \cdot \left[\left(DF_K^{-1}\right)^T (F_K(\hat{\boldsymbol{x}}))\right] \nabla_{\hat{\boldsymbol{x}}} \hat{w}(\hat{\boldsymbol{x}}) \\ &\quad \times \left|\det DF_K(\hat{\boldsymbol{x}})\right|\ d\hat{\boldsymbol{x}}. \qquad (5.11)\end{aligned}$$

□

Example 5.46. Affine transform. The most important class of reference maps are affine transforms (5.3), where the invertible matrix B_K and the vector $\boldsymbol{b}$ are constants. It follows that

$$\hat{\boldsymbol{x}} = B_K^{-1}\left(\boldsymbol{x} - \boldsymbol{b}\right) = B_K^{-1}\boldsymbol{x} - B_K^{-1}\boldsymbol{b}.$$

In this case, there are

$$DF_K = B_K, \quad DF_K^{-1} = B_K^{-1}, \quad \det DF_K = \det\left(B_K\right).$$

One obtains for the integral transforms from (5.7), (5.10), and (5.11)

$$\int_K v(\boldsymbol{x})\ d\boldsymbol{x} = \left|\det\left(B_K\right)\right| \int_{\hat{K}} \hat{v}(\hat{\boldsymbol{x}})\ d\hat{\boldsymbol{x}}, \qquad (5.12)$$

$$\int_K \boldsymbol{b}(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x})\ d\boldsymbol{x} = \left|\det\left(B_K\right)\right| \int_{\hat{K}} \boldsymbol{b}\left(F_K(\hat{\boldsymbol{x}})\right) \cdot B_K^{-T} \nabla_{\hat{\boldsymbol{x}}} \hat{v}(\hat{\boldsymbol{x}})\ d\hat{\boldsymbol{x}}, \qquad (5.13)$$

$$\int_K \nabla v(\boldsymbol{x}) \cdot \nabla w(\boldsymbol{x})\ d\boldsymbol{x} = \left|\det\left(B_K\right)\right| \int_{\hat{K}} B_K^{-T} \nabla_{\hat{\boldsymbol{x}}} \hat{v}(\hat{\boldsymbol{x}}) \cdot B_K^{-T} \nabla_{\hat{\boldsymbol{x}}} \hat{w}(\hat{\boldsymbol{x}})\ d\hat{\boldsymbol{x}}. \qquad (5.14)$$

Setting $v(\boldsymbol{x}) = 1$ in (5.12) yields

$$\left|\det\left(B_K\right)\right| = \frac{|K|}{\left|\hat{K}\right|}. \qquad (5.15)$$

□

Chapter 6
Interpolation

Remark 6.1. Motivation. Variational forms of partial differential equations use functions in Sobolev spaces. The solution of these equations shall be approximated with the Ritz method in finite-dimensional spaces, the finite element spaces. The best possible approximation of an arbitrary function from the Sobolev space by a finite element function is a factor in the upper bound for the finite element error, e.g., see the Lemma of Cea, estimate (4.20).

This section studies the approximation quality of finite element spaces. Estimates are proved for interpolants of functions. Interpolation estimates are of course upper bounds of the best approximation error and they can serve as factors in finite element error estimates. □

6.1 Interpolation in Sobolev Spaces by Polynomials

Lemma 6.2. Unique determination of a polynomial with integral conditions. *Let Ω be a bounded domain in $\mathbb{R}^d$ with Lipschitz boundary. Let $m \in \mathbb{N}\cup\{0\}$ be given and let for all derivatives with multi-index $\boldsymbol{\alpha}$, $|\boldsymbol{\alpha}| \le m$, a value $a_{\boldsymbol{\alpha}} \in \mathbb{R}$ be prescribed. Then, there is a uniquely determined polynomial $p \in P_m(\Omega)$ such that*

$$\int_\Omega \partial_{\boldsymbol{\alpha}} p(\boldsymbol{x})\ d\boldsymbol{x} = a_{\boldsymbol{\alpha}}, \quad |\boldsymbol{\alpha}| \le m. \tag{6.1}$$

Proof. Let $p \in P_m(\Omega)$ be an arbitrary polynomial. It has the form

$$p(\boldsymbol{x}) = \sum_{|\boldsymbol{\beta}|\le m} b_{\boldsymbol{\beta}} \boldsymbol{x}^{\boldsymbol{\beta}}.$$

Inserting this representation in (6.1) leads to a linear system of equations $M\underline{b} = \underline{a}$ with

$$M = (M_{\boldsymbol{\alpha\beta}}),\ M_{\boldsymbol{\alpha\beta}} = \int_\Omega \partial_{\boldsymbol{\alpha}} \boldsymbol{x}^{\boldsymbol{\beta}}\ d\boldsymbol{x},\ \underline{b} = (b_{\boldsymbol{\beta}}),\ \underline{a} = (a_{\boldsymbol{\alpha}}),$$

for $|\boldsymbol{\alpha}|, |\boldsymbol{\beta}| \le m$. Since M is a squared matrix, the linear system of equations possesses a unique solution if and only if M is non-singular.

The proof is performed by contradiction. Assume that M is singular. Then, there exists a non-trivial solution of the homogeneous system. That means, there is a polynomial $q \in P_m(\Omega) \setminus \{0\}$ with

$$\int_\Omega \partial_{\boldsymbol{\alpha}} q(\boldsymbol{x})\ d\boldsymbol{x} = 0 \text{ for all } |\boldsymbol{\alpha}| \le m.$$

The polynomial $q(\boldsymbol{x})$ has the representation $q(\boldsymbol{x}) = \sum_{|\boldsymbol{\beta}|\le m} c_{\boldsymbol{\beta}} \boldsymbol{x}^{\boldsymbol{\beta}}$. Now, one can choose a $c_{\boldsymbol{\beta}} \ne 0$ with maximal value $|\boldsymbol{\beta}|$. Then, it is $\partial_{\boldsymbol{\beta}} q(\boldsymbol{x}) = C c_{\boldsymbol{\beta}} = \mathit{const} \ne 0$, where $C > 0$ comes from the differentiation rule for polynomials, which is a contradiction to the vanishing of the integral for $\partial_{\boldsymbol{\beta}} q(\boldsymbol{x})$. ■

Remark 6.3. To Lemma 6.2. Lemma 6.2 states that a polynomial is uniquely determined if a condition on the integral on Ω is prescribed for each derivative. □

Lemma 6.4. Poincaré-type inequality. *Denote by $D^k v(\boldsymbol{x})$, $k \in \mathbb{N} \cup \{0\}$, the total derivative of order k of a function $v(\boldsymbol{x})$, e.g., for $k = 1$ the gradient of $v(\boldsymbol{x})$. Let Ω be convex and be included into a ball of radius R. Let $l \in \mathbb{N}\cup\{0\}$ with $k \le l$ and let $p \in \mathbb{R}$ with $p \in [1, \infty)$. Assume that $v \in W^{l,p}(\Omega)$ satisfies*

$$\int_\Omega \partial_{\boldsymbol{\alpha}} v(\boldsymbol{x})\ d\boldsymbol{x} = 0 \text{ for all } |\boldsymbol{\alpha}| \le l - 1,$$

then it holds the estimate

$$\left\|D^k v\right\|_{L^p(\Omega)} \le C R^{l-k} \left\|D^l v\right\|_{L^p(\Omega)},$$

where the constant C does not depend on Ω and on $v(\boldsymbol{x})$.

Proof. There is nothing to prove if $k = l$. In addition, it suffices to prove the lemma for $k = 0$ and $l = 1$, since the general case follows by applying the result to $\partial_{\boldsymbol{\alpha}} v(\boldsymbol{x})$.

Since Ω is assumed to be convex, the integral mean value theorem can be written in the form

$$v(\boldsymbol{x}) - v(\boldsymbol{y}) = \int_0^1 \nabla v(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \cdot (\boldsymbol{x} - \boldsymbol{y})\ dt, \quad \boldsymbol{x}, \boldsymbol{y} \in \Omega.$$

Integration with respect to $\boldsymbol{y}$ yields

$$v(\boldsymbol{x}) \int_\Omega d\boldsymbol{y} - \int_\Omega v(\boldsymbol{y})\ d\boldsymbol{y} = \int_\Omega \int_0^1 \nabla v(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \cdot (\boldsymbol{x} - \boldsymbol{y})\ dt\ d\boldsymbol{y}.$$

It follows from the assumption that the second integral on the left-hand side vanishes that

$$v(\boldsymbol{x}) = \frac{1}{|\Omega|} \int_\Omega \int_0^1 \nabla v(t\boldsymbol{x} + (1-t)\boldsymbol{y}) \cdot (\boldsymbol{x} - \boldsymbol{y})\ dt\ d\boldsymbol{y}.$$

Now, taking the absolute value on both sides, using that the absolute value of an integral is estimated from above by the integral of the absolute value, applying the Cauchy–Schwarz inequality for vectors (3.3), and the estimate $\|\boldsymbol{x} - \boldsymbol{y}\|_2 \le 2R$ yields

$$
\begin{aligned}
|v(\boldsymbol{x})| &= \frac{1}{|\Omega|}\left|\int_\Omega \int_0^1 \nabla v(t\boldsymbol{x}+(1-t)\boldsymbol{y})\cdot(\boldsymbol{x}-\boldsymbol{y})\ dt\ d\boldsymbol{y}\right| \\
&\le \frac{1}{|\Omega|}\int_\Omega \int_0^1 |\nabla v(t\boldsymbol{x}+(1-t)\boldsymbol{y})\cdot(\boldsymbol{x}-\boldsymbol{y})|\ dt\ d\boldsymbol{y} \\
&\le \frac{2R}{|\Omega|}\int_\Omega \int_0^1 \|\nabla v(t\boldsymbol{x}+(1-t)\boldsymbol{y})\|_2\ dt\ d\boldsymbol{y}. \qquad (6.2)
\end{aligned}
$$

Then, (6.2) is raised to the power p and integrated with respect to $\boldsymbol{x}$. One obtains with Hölder's inequality (3.4), with $p^{-1}+q^{-1}=1 \implies p/q-p=p(1/q-1)=-1$, that

$$
\begin{aligned}
\int_\Omega |v(\boldsymbol{x})|^p\ d\boldsymbol{x} &\le \frac{CR^p}{|\Omega|^p}\int_\Omega \left(\int_\Omega\int_0^1 \|\nabla v(t\boldsymbol{x}+(1-t)\boldsymbol{y})\|_2\ dt\ d\boldsymbol{y}\right)^p d\boldsymbol{x} \\
&\le \frac{CR^p}{|\Omega|^p}\int_\Omega \Bigg[\underbrace{\left(\int_\Omega\int_0^1 1^q\ dt\ d\boldsymbol{y}\right)^{p/q}}_{|\Omega|^{p/q}} \\
&\quad \times\left(\int_\Omega\int_0^1 \|\nabla v(t\boldsymbol{x}+(1-t)\boldsymbol{y})\|_2^p\ dt\ d\boldsymbol{y}\right)\Bigg] d\boldsymbol{x} \\
&= \frac{CR^p}{|\Omega|}\int_\Omega \left(\int_\Omega\int_0^1 \|\nabla v(t\boldsymbol{x}+(1-t)\boldsymbol{y})\|_2^p\ dt\ d\boldsymbol{y}\right) d\boldsymbol{x}.
\end{aligned}
$$

Applying the theorem of Fubini allows the commutation of the integration

$$
\int_\Omega |v(\boldsymbol{x})|^p\ d\boldsymbol{x} \le \frac{CR^p}{|\Omega|}\int_0^1\int_\Omega \left(\int_\Omega \|\nabla v(t\boldsymbol{x}+(1-t)\boldsymbol{y})\|_2^p\ d\boldsymbol{y}\right) d\boldsymbol{x}\ dt.
$$

Using the integral mean value theorem in one dimension gives that there is a $t_0\in[0,1]$ such that

$$
\int_\Omega |v(\boldsymbol{x})|^p\ d\boldsymbol{x} \le \frac{CR^p}{|\Omega|}\int_\Omega \left(\int_\Omega \|\nabla v(t_0\boldsymbol{x}+(1-t_0)\boldsymbol{y})\|_2^p\ d\boldsymbol{y}\right) d\boldsymbol{x}.
$$

The function $\|\nabla v(\boldsymbol{x})\|_2^p$ will be extended to $\mathbb{R}^d$ by zero and the extension will be also denoted by $\|\nabla v(\boldsymbol{x})\|_2^p$. Then, it is

$$
\int_\Omega |v(\boldsymbol{x})|^p\ d\boldsymbol{x} \le \frac{CR^p}{|\Omega|}\int_\Omega \left(\int_{\mathbb{R}^d} \|\nabla v(t_0\boldsymbol{x}+(1-t_0)\boldsymbol{y})\|_2^p\ d\boldsymbol{y}\right) d\boldsymbol{x}. \qquad (6.3)
$$

Let $t_0\in[0,1/2]$. Since the domain of integration is $\mathbb{R}^d$, a substitution of variables $t_0\boldsymbol{x}+(1-t_0)\boldsymbol{y}=\boldsymbol{z}$ can be applied and leads to

$$
\int_{\mathbb{R}^d} \|\nabla v(t_0\boldsymbol{x}+(1-t_0)\boldsymbol{y})\|_2^p\ d\boldsymbol{y} = \frac{1}{1-t_0}\int_{\mathbb{R}^d}\|\nabla v(\boldsymbol{z})\|_2^p\ d\boldsymbol{z} \le 2\,\|\nabla v\|_{L^p(\Omega)}^p,
$$

since $1/(1-t_0)\le 2$. Inserting this expression in (6.3) gives

$$
\int_\Omega |v(\boldsymbol{x})|^p\ d\boldsymbol{x} \le 2CR^p\,\|\nabla v\|_{L^p(\Omega)}^p.
$$

If $t_0>1/2$ then one changes the roles of $\boldsymbol{x}$ and $\boldsymbol{y}$, applies the theorem of Fubini to change the sequence of integration, and uses the same arguments. ■

Remark 6.5. On Lemma 6.4. Lemma 6.4 proves an inequality of Poincaré-type. It says that it is possible to estimate the $L^p(\Omega)$ norm of a lower derivative of a function $v(\boldsymbol{x})$ by the same norm of a higher derivative if the integral mean values of some lower derivatives vanish.

An important application of Lemma 6.4 is in the proof of the Bramble[1]–Hilbert[2] lemma. The Bramble–Hilbert lemma considers a continuous linear functional that is defined on a Sobolev space and that vanishes for all polynomials of degree less than or equal to m. It states that the value of the functional can be estimated by the Lebesgue norm of the $(m+1)$th total derivative of the functions from this Sobolev space. □

Theorem 6.6. *Bramble–Hilbert lemma.* *Let $m \in \mathbb{N} \cup \{0\}$, $p \in [1,\infty]$, and $F : W^{m+1,p}(\Omega) \to \mathbb{R}$ be a continuous linear functional, and let the conditions of Lemma 6.2 and Lemma 6.4 be satisfied. Let*

$$F(p) = 0 \quad \forall\, p \in P_m(\Omega),$$

then there is a constant $C(\Omega)$, which is independent of v and F, such that

$$|F(v)| \le C(\Omega) \left\|D^{m+1}v\right\|_{L^p(\Omega)} \quad \forall\, v \in W^{m+1,p}(\Omega).$$

Proof. Let $v \in W^{m+1,p}(\Omega)$. It follows from Lemma 6.2 that there is a polynomial from $P_m(\Omega)$ with

$$\int_\Omega \partial_{\boldsymbol{\alpha}}(v+p)(\boldsymbol{x})\; d\boldsymbol{x} = 0 \text{ for } |\boldsymbol{\alpha}| \le m.$$

Lemma 6.4 gives, with $l = m+1$ and considering each term in $\|\cdot\|_{W^{m+1,p}(\Omega)}$ individually, the estimate

$$\|v+p\|_{W^{m+1,p}(\Omega)} \le C(\Omega) \left\|D^{m+1}(v+p)\right\|_{L^p(\Omega)} = C(\Omega) \left\|D^{m+1}v\right\|_{L^p(\Omega)}.$$

From the vanishing of F for $p \in P_m(\Omega)$ and the continuity of F, it follows that

$$|F(v)| = |F(v+p)| \le C\, \|v+p\|_{W^{m+1,p}(\Omega)} \le C(\Omega) \left\|D^{m+1}v\right\|_{L^p(\Omega)}.$$

■

Remark 6.7. Strategy for estimating the interpolation error. The Bramble–Hilbert lemma, more precisely Lemma 6.4, will be used for estimating the interpolation error for finite elements. The strategy is as follows:

- Show first the estimate on the reference mesh cell $\hat{K}$.
- Transform the estimate on an arbitrary mesh cell K to the reference mesh cell $\hat{K}$.
- Apply the estimate on $\hat{K}$.
- Transform back to K.

One has to study what happens if the transforms are applied to the estimate. □

[1] James H. Bramble, born 1930

[2] Stephen R. Hilbert

Remark 6.8. Assumptions, definition of the interpolant. Let $\hat{K} \subset \mathbb{R}^d, d \in \{2,3\}$, be a reference mesh cell (compact polyhedron), $\hat{P}(\hat{K})$ a polynomial space of dimension N, and $\hat{\Phi}_1, \ldots, \hat{\Phi}_N \ : \ C^s(\hat{K}) \to \mathbb{R}$ continuous linear functionals. It will be assumed that the space $\hat{P}(\hat{K})$ is unisolvent with respect to these functionals. Then, there is a local basis $\hat{\phi}_1, \ldots, \hat{\phi}_N \in \hat{P}(\hat{K})$.

Consider $\hat{v} \in C^s(\hat{K})$, then the interpolant $I_{\hat{K}}\hat{v} \in \hat{P}(\hat{K})$ is defined by

$$I_{\hat{K}}\hat{v}(\hat{\boldsymbol{x}}) = \sum_{i=1}^{N} \hat{\Phi}_i(\hat{v})\hat{\phi}_i(\hat{\boldsymbol{x}}).$$

The operator $I_{\hat{K}}$ is a continuous and linear operator from $C^s(\hat{K})$ to $\hat{P}(\hat{K})$. From the linearity, it follows that $I_{\hat{K}}$ is the identity on $\hat{P}(\hat{K})$

$$I_{\hat{K}}\hat{p} = \hat{p} \quad \forall\, \hat{p} \in \hat{P}(\hat{K}).$$

□

Example 6.9. Interpolation operators.

- Let $\hat{K} \subset \mathbb{R}^d$ be an arbitrary reference cell, $\hat{P}(\hat{K}) = P_0(\hat{K})$, and

$$\hat{\Phi}(\hat{v}) = \frac{1}{\left|\hat{K}\right|} \int_{\hat{K}} \hat{v}(\hat{\boldsymbol{x}}) \ d\hat{\boldsymbol{x}}.$$

  The functional $\hat{\Phi}$ is bounded, and hence continuous, on $C^0(\hat{K})$ since

$$\left|\hat{\Phi}(\hat{v})\right| \le \frac{1}{\left|\hat{K}\right|} \int_{\hat{K}} \left|\hat{v}(\hat{\boldsymbol{x}})\right| \ d\hat{\boldsymbol{x}} \le \frac{\left|\hat{K}\right|}{\left|\hat{K}\right|} \max_{\hat{\boldsymbol{x}}\in\hat{K}} \left|\hat{v}(\hat{\boldsymbol{x}})\right| = \left\|\hat{v}\right\|_{C^0(\hat{K})}.$$

  For the constant function $1 \in P_0(\hat{K})$, it is $\hat{\Phi}(1) = 1 \neq 0$. Hence, $\{\hat{\phi}\} = \{1\}$ is the local basis and the space is unisolvent with respect to $\hat{\Phi}$. The operator

$$I_{\hat{K}}\hat{v}(\hat{\boldsymbol{x}}) = \hat{\Phi}(\hat{v})\hat{\phi}(\hat{\boldsymbol{x}}) = \frac{1}{\left|\hat{K}\right|} \int_{\hat{K}} \hat{v}(\hat{\boldsymbol{x}}) \ d\hat{\boldsymbol{x}}$$

  is an integral mean value operator, i.e., each continuous function on $\hat{K}$ will be approximated by a constant function whose value equals the integral mean value, see Figure 6.1
- It is possible to define $\hat{\Phi}(\hat{v}) = \hat{v}(\hat{\boldsymbol{x}}_0)$ for an arbitrary point $\hat{\boldsymbol{x}}_0 \in \hat{K}$. This functional is also linear and continuous in $C^0(\hat{K})$. The interpolation operator $I_{\hat{K}}$ defined in this way interpolates each continuous function by a constant function whose value is equal to the value of the function at $\hat{\boldsymbol{x}}_0$, see also Figure 6.1.

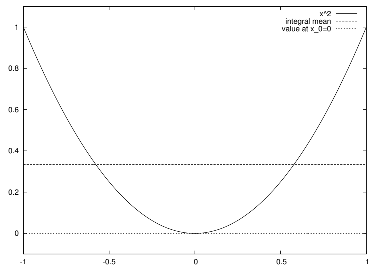

Fig. 6.1 Interpolation of x^2 in $[-1,1]$ by a P_0 function with the integral mean value and with the value of the function at $x_0 = 0$.

> Interpolation operators which are defined by using values of functions, are called Lagrangian interpolation operators.

This example demonstrates that the interpolation operator $I_{\hat{K}}$ depends on $\hat{P}(\hat{K})$ and on the functionals $\hat{\Phi}_i$. □

Theorem 6.10. Interpolation error estimate on a reference mesh cell. *Let $P_m(\hat{K}) \subset \hat{P}(\hat{K})$, let $p \in [1,\infty)$, and let $\hat{s} \in \mathbb{N} \cup \{0\}$ such that $(m+1-\hat{s})p > d \geq (m-\hat{s})p$ and $\hat{s} \geq s$, where s appears in the definition of the interpolation operator. Then there is a constant C that is independent of $\hat{v}(\hat{\boldsymbol{x}})$ such that*

$$\left\|\hat{v} - I_{\hat{K}}\hat{v}\right\|_{W^{m+1,p}(\hat{K})} \leq C \left\|D^{m+1}\hat{v}\right\|_{L^p(\hat{K})} \quad \forall\ \hat{v} \in W^{m+1,p}(\hat{K}). \tag{6.4}$$

Proof. Since $\hat{K}$ is bounded, one has the Sobolev imbedding, Theorem 3.52,

$$W^{m+1,p}(\hat{K}) = W^{(m+1-\hat{s})+\hat{s},p}(\hat{K}) \to C^{\hat{s}}(\hat{K}).$$

Because $\hat{K}$ is convex, the imbedding $C^{\hat{s}}(\hat{K}) \to C^{s}(\hat{K})$ is compact, see (Adams, 1975, Theorem 1.31), such that the interpolation operator is well defined in $W^{m+1,p}(\hat{K})$. From the identity of the interpolation operator in $P_m(\hat{K})$, the triangle inequality, the boundedness of the interpolation operator (it is a linear and continuous operator mapping $C^s(\hat{K}) \to \hat{P}(\hat{K}) \subset W^{m+1,p}(\hat{K})$), and the Sobolev imbedding, one obtains for $\hat{q} \in P_m(\hat{K})$

$$\begin{aligned}\left\|\hat{v} - I_{\hat{K}}\hat{v}\right\|_{W^{m+1,p}(\hat{K})} &= \left\|\hat{v} + \hat{q} - I_{\hat{K}}(\hat{v}+\hat{q})\right\|_{W^{m+1,p}(\hat{K})}\\ &\leq \left\|\hat{v}+\hat{q}\right\|_{W^{m+1,p}(\hat{K})} + \left\|I_{\hat{K}}(\hat{v}+\hat{q})\right\|_{W^{m+1,p}(\hat{K})}\\ &\leq \left\|\hat{v}+\hat{q}\right\|_{W^{m+1,p}(\hat{K})} + C\left\|\hat{v}+\hat{q}\right\|_{C^s(\hat{K})}\\ &\leq C\left\|\hat{v}+\hat{q}\right\|_{W^{m+1,p}(\hat{K})}.\end{aligned}$$

Now, $\hat{q}(\hat{\boldsymbol{x}})$ is chosen such that

$$\int_{\hat{K}} \partial_{\boldsymbol{\alpha}}(\hat{v}+\hat{q})\ d\hat{\boldsymbol{x}} = 0 \quad \forall\ |\boldsymbol{\alpha}| \leq m$$

holds. Hence, the assumptions of Lemma 6.4 are satisfied. It follows that

$$\|\hat{v}+\hat{q}\|_{W^{m+1,p}(\hat{K})} \le C \left\|D^{m+1}(\hat{v}+\hat{q})\right\|_{L^p(\hat{K})} = C \left\|D^{m+1}\hat{v}\right\|_{L^p(\hat{K})}.$$

■

Definition 6.11. Quasi-uniform and regular family of triangulations, (Brenner & Scott, 2008, Def. 4.4.13). Let $\{\mathcal{T}^h\}$ with $0 < h \le 1$, be a family of triangulations such that

$$\max_{K\in\mathcal{T}^h} h_K \le h \operatorname{diam}(\Omega),$$

where h_K is the diameter of $K = F_K(\hat{K})$, i.e., the largest distance of two points that are contained in K. The family is called to be quasi-uniform, if there exists a $C > 0$ such that

$$\min_{K\in\mathcal{T}^h} \rho_K \ge Ch \operatorname{diam}(\Omega) \tag{6.5}$$

for all $h \in (0,1]$, where ρ_K is the diameter of the largest ball contained in K.

The family is called to be regular, if there is exists a $C > 0$ such that for all $K \in \mathcal{T}^h$ and for all $h \in (0,1]$

$$\rho_K \ge Ch_K.$$

□

Remark 6.12. Assumptions on the reference mapping and the triangulation. For deriving the interpolation error estimate for arbitrary mesh cells K, and finally for the finite element space, one has to study the properties of the mapping from K to $\hat{K}$ and of the inverse mapping. Here, only the case of an affine family of finite elements whose mesh cells are generated by affine mappings

$$F_K\hat{\boldsymbol{x}} = B_K\hat{\boldsymbol{x}} + \boldsymbol{b},$$

will be considered, see (5.3), where B_K is a non-singular $d\times d$ matrix and $\boldsymbol{b}$ is a d vector. For the global estimate, a quasi-uniform family of triangulations will be considered. □

Lemma 6.13. Estimates of matrix norms. *For each matrix norm $\|\cdot\|$, one has the estimates*

$$\|B_K\| \le Ch_K, \quad \left\|B_K^{-1}\right\| \le Ch_K^{-1}, \tag{6.6}$$

where the constants depend on the matrix norm.

Proof. Since $\hat{K}$ is a Lipschitz domain with polyhedral boundary, it contains a ball $B(\hat{\boldsymbol{x}}_0, r)$ with $\hat{\boldsymbol{x}}_0 \in \hat{K}$ and some $r > 0$. Hence, $\hat{\boldsymbol{x}}_0 + \hat{\boldsymbol{y}} \in \hat{K}$ for all $\|\hat{\boldsymbol{y}}\|_2 = r$. It follows that the images

$$\boldsymbol{x}_0 = B_K\hat{\boldsymbol{x}}_0 + \boldsymbol{b}, \quad \boldsymbol{x} = B_K(\hat{\boldsymbol{x}}_0 + \hat{\boldsymbol{y}}) + \boldsymbol{b} = \boldsymbol{x}_0 + B_K\hat{\boldsymbol{y}}$$

are contained in K. Hence, one obtains for all $\hat{\boldsymbol{y}}$

$$\|B_K \hat{\boldsymbol{y}}\|_2 = \|\boldsymbol{x} - \boldsymbol{x}_0\|_2 \le h_K.$$

Now, it holds for the spectral norm that

$$\|B_K\|_2 = \sup_{\|\hat{\boldsymbol{z}}\|_2 = 1} \|B_K \hat{\boldsymbol{z}}\|_2 = \frac{1}{r} \sup_{\|\hat{\boldsymbol{z}}\|_2 = r} \|B_K \hat{\boldsymbol{z}}\|_2 \le \frac{h_K}{r}.$$

A bound of this form, with a possible different constant, holds also for all other matrix norms since all matrix norms are equivalent, see Remark 3.34.

The estimate for $\left\|B_K^{-1}\right\|$ proceeds in the same way with interchanging the roles of K and $\hat{K}$. ■

Theorem 6.14. Local interpolation estimate. *Let an affine family of finite elements be given by its reference cell $\hat{K}$, the functionals $\{\hat{\Phi}_i\}$, and a space of polynomials $\hat{P}(\hat{K})$. Let all assumptions of Theorem 6.10 be satisfied. Then, for all $v \in W^{m+1,p}(K)$, $p \in [1, \infty)$, there is a constant C, which is independent of v, such that*

$$\left\|D^k(v - I_K v)\right\|_{L^p(K)} \le C h_K^{m+1-k} \left\|D^{m+1} v\right\|_{L^p(K)}, \quad 0 \le k \le m+1. \quad (6.7)$$

Proof. The idea of the proof consists in transforming the left-hand side of (6.7) to the reference cell, using the interpolation estimate on the reference cell, and transforming back.

i). Denote the elements of the matrices B_K and B_K^{-1} by b_{ij} and $b_{ij}^{(-1)}$, respectively. Since $\|B_K\|_\infty = \max_{i,j} |b_{ij}|$ is also a matrix norm, it holds that

$$|b_{ij}| \le C h_K, \quad \left|b_{ij}^{(-1)}\right| \le C h_K^{-1}. \quad (6.8)$$

Using element-wise estimates for the matrix B_K (Leibniz formula for determinants), one obtains

$$|\det B_K| \le C h_K^d, \quad \left|\det B_K^{-1}\right| \le C h_K^{-d}. \quad (6.9)$$

ii). The next step consists in proving that the transformed interpolation operator is equal to the natural interpolation operator on K. The latter one is given by

$$I_K v = \sum_{i=1}^{N} \Phi_{K,i}(v) \phi_{K,i}, \quad (6.10)$$

where $\{\phi_{K,i}\}$ is the basis of the space

$$P(K) = \left\{p \,:\, K \to \mathbb{R} \,:\, p = \hat{p} \circ F_K^{-1}, \hat{p} \in \hat{P}(\hat{K})\right\},$$

which satisfies $\Phi_{K,i}(\phi_{K,j}) = \delta_{ij}$. The functionals are defined by

$$\Phi_{K,i}(v) = \hat{\Phi}_i(v \circ F_K) = \hat{\Phi}_i(\hat{v}). \quad (6.11)$$

Hence, it follows for $v = \hat{\phi}_j \circ F_K^{-1}$ from the condition on the local basis on $\hat{K}$ that

$$\Phi_{K,i}(\hat{\phi}_j \circ F_K^{-1}) = \hat{\Phi}_i(\hat{\phi}_j) = \delta_{ij},$$

i.e., the local basis on K is given by $\phi_{K,j} = \hat{\phi}_j \circ F_K^{-1}$. Using (6.11) and (6.10), one gets

$$\begin{aligned}I_{\hat K}\hat v &= \sum_{i=1}^N \hat\Phi_i(\hat v)\hat\phi_i = \sum_{i=1}^N \Phi_{K,i}(\underbrace{\hat v\circ F_K^{-1}}_{=v})\ \phi_{K,i}\circ F_K = \left(\sum_{i=1}^N \Phi_{K,i}(v)\phi_{K,i}\right)\circ F_K\\ &= I_K v\circ F_K.\end{aligned}$$

Consequently, $I_{\hat K}\hat v$ is transformed correctly.

iii). One obtains with the chain rule

$$\frac{\partial v(\boldsymbol{x})}{\partial \boldsymbol{x}_i} = \sum_{j=1}^d \frac{\partial \hat v(\hat{\boldsymbol{x}})}{\partial \hat{\boldsymbol{x}}_j} b_{ji}^{(-1)}, \quad \frac{\partial \hat v(\hat{\boldsymbol{x}})}{\partial \hat{\boldsymbol{x}}_i} = \sum_{j=1}^d \frac{\partial v(\boldsymbol{x})}{\partial \boldsymbol{x}_j} b_{ji}.$$

It follows with (6.8) that (with each derivative one obtains an additional factor of B_K or B_K^{-1}, respectively)

$$\left\|D_{\boldsymbol{x}}^k v(\boldsymbol{x})\right\|_2 \le Ch_K^{-k}\left\|D_{\hat{\boldsymbol{x}}}^k \hat v(\hat{\boldsymbol{x}})\right\|_2, \quad \left\|D_{\hat{\boldsymbol{x}}}^k \hat v(\hat{\boldsymbol{x}})\right\|_2 \le Ch_K^{k}\left\|D_{\boldsymbol{x}}^k v(\boldsymbol{x})\right\|_2.$$

One gets with (6.9)

$$\int_K \left\|D_{\boldsymbol{x}}^k v(\boldsymbol{x})\right\|_2^p\ d\boldsymbol{x} \le Ch_K^{-kp}\left|\det B_K\right|\int_{\hat K}\left\|D_{\hat{\boldsymbol{x}}}^k \hat v(\hat{\boldsymbol{x}})\right\|_2^p\ d\hat{\boldsymbol{x}} \le Ch_K^{-kp+d}\int_{\hat K}\left\|D_{\hat{\boldsymbol{x}}}^k \hat v(\hat{\boldsymbol{x}})\right\|_2^p\ d\hat{\boldsymbol{x}} \tag{6.12}$$

and

$$\int_{\hat K} \left\|D_{\hat{\boldsymbol{x}}}^k \hat v(\hat{\boldsymbol{x}})\right\|_2^p\ d\hat{\boldsymbol{x}} \le Ch_K^{kp}\left|\det B_K^{-1}\right|\int_{K}\left\|D_{\boldsymbol{x}}^k v(\boldsymbol{x})\right\|_2^p\ d\boldsymbol{x} \le Ch_K^{kp-d}\int_{K}\left\|D_{\boldsymbol{x}}^k v(\boldsymbol{x})\right\|_2^p\ d\boldsymbol{x}. \tag{6.13}$$

Using now the interpolation estimate on the reference cell (6.4) yields

$$\left\|D_{\hat{\boldsymbol{x}}}^k(\hat v - I_{\hat K}\hat v)\right\|_{L^p(\hat K)}^p \le C\left\|D_{\hat{\boldsymbol{x}}}^{m+1}\hat v\right\|_{L^p(\hat K)}^p, \quad 0\le k\le m+1. \tag{6.14}$$

It follows with (6.12), (6.14), and (6.13) that

$$\begin{aligned}\left\|D_{\boldsymbol{x}}^k(v - I_K v)\right\|_{L^p(K)}^p &\le Ch_K^{-kp+d}\left\|D_{\hat{\boldsymbol{x}}}^k(\hat v - I_{\hat K}\hat v)\right\|_{L^p(\hat K)}^p\\ &\le Ch_K^{-kp+d}\left\|D_{\hat{\boldsymbol{x}}}^{m+1}\hat v\right\|_{L^p(\hat K)}^p\\ &\le Ch_K^{(m+1-k)p}\left\|D_{\boldsymbol{x}}^{m+1} v\right\|_{L^p(K)}^p.\end{aligned}$$

Taking the p-th root proves the statement of the theorem. ■

Remark 6.15. On estimate (6.7).

- Note that the power of h_K does not depend on p and d.
- Consider a quasi-uniform triangulation and define

$$h = \max_{K\in\mathcal{T}^h}\{h_K\}.$$

Then, one obtains by summing over all mesh cells an interpolation estimate for the global finite element space

$$\begin{aligned}\left\|D^k(v-I^hv)\right\|_{L^p(\Omega)} &= \left(\sum_{K\in\mathcal{T}^h}\left\|D^k(v-I_Kv)\right\|^p_{L^p(K)}\right)^{1/p}\\ &\le \left(\sum_{K\in\mathcal{T}^h} Ch_K^{(m+1-k)p}\left\|D^{m+1}v\right\|^p_{L^p(K)}\right)^{1/p}\\ &\le Ch^{(m+1-k)}\left\|D^{m+1}v\right\|_{L^p(\Omega)}. \end{aligned} \tag{6.15}$$

□

Corollary 6.16. Finite element error estimate. *Let $u(\boldsymbol{x})$ be the solution of the model problem* (4.10) *with $u \in H^{m+1}(\Omega)$ and let $u^h(\boldsymbol{x})$ be the solution of the corresponding finite element problem. Consider a family of quasi-uniform triangulations and let the finite element spaces V^h contain polynomials of degree m. Then, the following finite element error estimate holds*

$$\left\|\nabla(u-u^h)\right\|_{L^2(\Omega)} \le Ch^m\left\|D^{m+1}u\right\|_{L^2(\Omega)} = Ch^m\left|u\right|_{H^{m+1}(\Omega)}. \tag{6.16}$$

Proof. The statement follows by combining Lemma 4.13 (for $V = H_0^1(\Omega)$) and (6.15)

$$\begin{aligned}\left\|\nabla(u-u^h)\right\|_{L^2(\Omega)} &= \inf_{v^h\in V^h}\left\|\nabla(u-v^h)\right\|_{L^2(\Omega)}\\ &\le \left\|\nabla(u-I_hu)\right\|_{L^2(\Omega)} \le Ch^m\left|u\right|_{H^{m+1}(\Omega)}.\end{aligned}$$

■

Remark 6.17. To (6.16). Note that Lemma 4.13 provides only information about the error in the norm on the left-hand side of (6.16), but not in other norms. □

6.2 Inverse Estimate

Remark 6.18. On inverse estimates. The approach for proving interpolation error estimates can be used also to prove so-called inverse estimates. With inverse estimates, a norm of a higher order derivative of a finite element function is estimated by a norm of a lower order derivative of this function. Likewise, norms in different Lebesgue spaces are estimated. One obtains as penalty a factor with negative powers of the diameter of the mesh cell. □

Theorem 6.19. Inverse estimate. *Let $0 \le k \le l$ be natural numbers and let $p, q \in [1,\infty]$. Then there is a constant C_{inv}, which depends only on $k, l, p, q, \hat{K}, \hat{P}(\hat{K})$, such that*

$$\left\|D^lv^h\right\|_{L^q(K)} \le C_{\text{inv}}h_K^{(k-l)-d(p^{-1}-q^{-1})}\left\|D^kv^h\right\|_{L^p(K)} \quad \forall\ v^h \in P(K). \tag{6.17}$$

Proof. In the first step, (6.17) is shown for $h_{\hat{K}} = 1$ and $k = 0$ on the reference mesh cell. Since all norms are equivalent in finite-dimensional spaces, one obtains

$$\left\| D^l \hat{v}^h \right\|_{L^q(\hat{K})} \le \left\| \hat{v}^h \right\|_{W^{l,q}(\hat{K})} \le C \left\| \hat{v}^h \right\|_{L^p(\hat{K})} \quad \forall\, \hat{v}^h \in \hat{P}(\hat{K}). \tag{6.18}$$

If $k > 0$, then one sets

$$\tilde{P}(\hat{K}) = \left\{ \partial_{\boldsymbol{\alpha}} \hat{v}^h \; : \; \hat{v}^h \in \hat{P}(\hat{K}), |\boldsymbol{\alpha}| = k \right\},$$

which is also a space consisting of polynomials. The application of (6.18) to $\tilde{P}(\hat{K})$ gives

$$\begin{aligned} \left\| D^l \hat{v}^h \right\|_{L^q(\hat{K})} &= \sum_{|\boldsymbol{\alpha}|=k} \left\| D^{l-k} \left(\partial_{\boldsymbol{\alpha}} \hat{v}^h \right) \right\|_{L^q(\hat{K})} \le C \sum_{|\boldsymbol{\alpha}|=k} \left\| \partial_{\boldsymbol{\alpha}} \hat{v}^h \right\|_{L^p(\hat{K})} \\ &= C \left\| D^k \hat{v}^h \right\|_{L^p(\hat{K})} . \end{aligned}$$

This estimate is transformed to an arbitrary mesh cell K analogously as for the interpolation error estimates, compare the proof of Theorem 6.14. From the estimates (6.12) and (6.13) for the transformations, one obtains

$$\begin{aligned} \left\| D^l v^h \right\|_{L^q(K)} &\le C h_K^{-l+d/q} \left\| D^l \hat{v}^h \right\|_{L^q(\hat{K})} \le C h_K^{-l+d/q} \left\| D^k \hat{v}^h \right\|_{L^p(\hat{K})} \\ &\le C_{\text{inv}} h_K^{k-l+d/q-d/p} \left\| D^k v^h \right\|_{L^p(K)} . \end{aligned}$$

■

Remark 6.20. On the proof. The crucial point in the proof is the equivalence of all norms in finite-dimensional spaces. Such a property does not hold in infinite-dimensional spaces. □

Corollary 6.21. Global inverse estimate. *Let $p = q$ and let $\left\{ \mathcal{T}^h \right\}$ be a quasi-uniform family of triangulations of Ω, then*

$$\left\| D^l v^h \right\|_{L^{p,h}(\Omega)} \le C_{\text{inv}} h^{k-l} \left\| D^k v^h \right\|_{L^{p,h}(\Omega)} , \tag{6.19}$$

where

$$\left\| \cdot \right\|_{L^{p,h}(\Omega)} = \left(\sum_{K \in \mathcal{T}^h} \left\| \cdot \right\|_{L^p(K)}^p \right)^{1/p} .$$

Remark 6.22. On $\|\cdot\|_{L^{p,h}(\Omega)}$. The cell-wise definition of the norm is important for $k \ge 2$ or $l \ge 2$ since in these cases finite element functions generally do not possess the regularity for the global norm to be well defined. It is also important for $l \ge 1$ and non-conforming finite element functions. □

Chapter 7
Finite Element Methods for Second Order Elliptic Equations

7.1 General Convergence Theorems

Remark 7.1. Motivation. In Section 5.1, non-conforming finite element methods have been introduced, i.e., methods where the finite element space V^h is not a subspace of V, which is the space in the definition of the continuous variational problem. The property $V^h \not\subset V$ is given for the Crouzeix–Raviart and the Rannacher–Turek element. Another case of non-conformity is given if the domain does not possess a polyhedral boundary and one has to apply some approximation of the boundary.

For non-conforming methods, the finite element approach is not longer a Ritz method. Hence, the convergence proof from Theorem 4.14 cannot be applied in this case. In addition, in practice, one is interested also in the order of convergence in other norms than $\|\cdot\|_V$ or one has to take into account that the values of the bilinear or linear form need to be approximated numerically. The abstract convergence theorem, which will be proved in this section, allows the numerical analysis of complex finite element methods. □

Remark 7.2. Notations, Assumptions. Let $\{h > 0\}$ be a set of mesh widths and let S^h, V^h normed spaces of functions which are defined on domains $\{\Omega^h \subset \mathbb{R}^d\}$. It will be assumed that the space S^h has a finite dimension and that S^h and V^h possess a common norm $\|\cdot\|_h$. In the application of the abstract theory, S^h will be a finite element space and V^h is defined such that the restriction and/or extension of the solution of the continuous problem to Ω^h is contained in V^h. The index h indicates that V^h might depend on h but not that V^h is finite-dimensional. Strictly speaking, the modified solution of the continuous problem does not solve the given problem any longer. Hence, it is consequent that the continuous problem does not appear explicitly in the abstract theory.

Given the bilinear forms

$$\begin{aligned} a^h &: S^h \times S^h \to \mathbb{R}, \\ \tilde{a}^h &: (S^h + V^h) \times (S^h + V^h) \to \mathbb{R}. \end{aligned}$$

Let the bilinear form a^h be regular in the sense that there is a constant $m > 0$, which is independent of h, such that for each $v^h \in S^h$ there is a $w^h \in S^h$ with $\left\|w^h\right\|_h = 1$ such that

$$m \left\|v^h\right\|_h \le a^h(v^h, w^h). \tag{7.1}$$

This condition is equivalent to the requirement that the stiffness matrix A with the entries $a_{ij} = a^h(\phi_j, \phi_i)$, where $\{\phi_i\}$ is a basis of S^h, is uniformly non-singular, i.e., its regularity is independent of h. For the second bilinear form, only its boundedness will be assumed

$$\tilde{a}^h(u, v) \le M \left\|u\right\|_h \left\|v\right\|_h \quad \forall \ u, v \in S^h + V^h. \tag{7.2}$$

Let the linear functionals $\{f^h(\cdot)\} \ : \ S^h \to \mathbb{R}$ be given. Then, the following discrete problems will be considered: Find $u^h \in S^h$ with

$$a^h(u^h, v^h) = f^h(v^h) \quad \forall \ v^h \in S^h. \tag{7.3}$$

Because the stiffness matrix is assumed to be non-singular, there is a unique solution of (7.3). □

Theorem 7.3. Abstract finite element error estimate. *Let the conditions* (7.1) *and* (7.2) *be satisfied and let u^h be the solution of* (7.3). *Then, the following error estimate holds for each $\tilde{u} \in V^h$*

$$\begin{aligned} \left\|\tilde{u} - u^h\right\|_h \le C \inf_{v^h \in S^h} &\left\{ \left\|\tilde{u} - v^h\right\|_h + \sup_{w^h \in S^h} \frac{\left|\tilde{a}^h(v^h, w^h) - a^h(v^h, w^h)\right|}{\left\|w^h\right\|_h} \right\} \\ &+ C \sup_{w^h \in S^h} \frac{\left|\tilde{a}^h(\tilde{u}, w^h) - f^h(w^h)\right|}{\left\|w^h\right\|_h} \end{aligned} \tag{7.4}$$

with $C = C(m, M)$.

Proof. Because of (7.1), there is for each $v^h \in S^h$ a $w^h \in S^h$ with $\left\|w^h\right\|_h = 1$ and

$$m \left\|u^h - v^h\right\|_h \le a^h(u^h - v^h, w^h).$$

Using the definition of u^h from (7.3), one obtains

$$m \left\|u^h - v^h\right\|_h \le f^h(w^h) - a^h(v^h, w^h) + \tilde{a}^h(v^h, w^h) + \tilde{a}^h(\tilde{u} - v^h, w^h) - \tilde{a}^h(\tilde{u}, w^h).$$

From (7.2) and $\left\|w^h\right\|_h = 1$, it follows that

$$\tilde{a}^h(\tilde{u} - v^h, w^h) \le M \left\|\tilde{u} - v^h\right\|_h.$$

Rearranging the terms appropriately and using $\left\|w^h / \left\|w^h\right\|_h\right\|_h = 1$ gives

$$
\begin{aligned}
m \left\|u^h - v^h\right\|_h \le M \left\|\tilde{u} - v^h\right\|_h &+ \sup_{w^h \in S^h} \frac{\left|\tilde{a}^h(v^h, w^h) - a^h(v^h, w^h)\right|}{\left\|w^h\right\|_h} \\
&+ \sup_{w^h \in S^h} \frac{\left|\tilde{a}^h(\tilde{u}, w^h) - f^h(w^h)\right|}{\left\|w^h\right\|_h}.
\end{aligned}
\tag{7.5}
$$

Applying the triangle inequality

$$\left\|\tilde{u} - u^h\right\|_h \le \left\|\tilde{u} - v^h\right\|_h + \left\|u^h - v^h\right\|_h$$

and inserting the estimate (7.5) gives (7.4). ■

Remark 7.4. To Theorem 7.3.

- An important special case of this theorem is the case that the stiffness matrix is uniformly positive definite, i.e., the condition

$$m \left\|v^h\right\|_h^2 \le a^h(v^h, v^h) \quad \forall\, v^h \in S^h \tag{7.6}$$

is satisfied. Dividing (7.6) by $\left\|v^h\right\|_h$ reveals that condition (7.1) is implied by (7.6).

- If the continuous problem is also defined with the bilinear form $\tilde{a}^h(\cdot,\cdot)$, then

$$\sup_{w^h \in S^h} \frac{\left|\tilde{a}^h(v^h, w^h) - a^h(v^h, w^h)\right|}{\left\|w^h\right\|_h}$$

can be considered as consistency error of the bilinear forms and the term

$$\sup_{w^h \in S^h} \frac{\left|\tilde{a}^h(\tilde{u}, w^h) - f^h(w^h)\right|}{\left\|w^h\right\|_h}$$

as consistency error of the right-hand sides.

□

Theorem 7.5. First Strang[1] lemma *Let S^h be a conforming finite element space, i.e., $S^h \subset V$, with $\left\|\cdot\right\|_h = \left\|\cdot\right\|_V$ and let the space V^h be independent of h. Consider a continuous problem of the form*

$$\tilde{a}^h(u, v) = f(v) \quad \forall\, v \in V,$$

then the following error estimate holds

$$
\begin{aligned}
\left\|u - u^h\right\|_V \le C \inf_{v^h \in S^h} &\left\{ \left\|u - v^h\right\|_V + \sup_{w^h \in S^h} \frac{\left|\tilde{a}^h(v^h, w^h) - a^h(v^h, w^h)\right|}{\left\|w^h\right\|_V} \right\} \\
&+ C \sup_{w^h \in S^h} \frac{\left|f(w^h) - f^h(w^h)\right|}{\left\|w^h\right\|_V}.
\end{aligned}
$$

Proof. The statement of this theorem follows directly from Theorem 7.3. ■

[1] Gilbert Strang, born 1934

7.2 Finite Element Method with the Non-conforming Crouzeix–Raviart Element

Remark 7.6. The continuous problem. Let $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, be a bounded domain with polygonal Lipschitz boundary. Let

$$Lu = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega, \tag{7.7}$$

where the operator is given by

$$Lu = -\nabla \cdot (A\nabla u)$$

with

$$A(\boldsymbol{x}) = (a_{ij}(\boldsymbol{x}))_{i,j=1}^d, \quad a_{ij} \in W^{1,p}(\Omega), p > d, \tag{7.8}$$

It will be assumed that there are two positive real numbers m, M such that

$$m \left\|\boldsymbol{\xi}\right\|_2^2 \le \boldsymbol{\xi}^T A(\boldsymbol{x})\boldsymbol{\xi} \le M \left\|\boldsymbol{\xi}\right\|_2^2 \quad \forall\, \boldsymbol{\xi} \in \mathbb{R}^d, \boldsymbol{x} \in \overline{\Omega}. \tag{7.9}$$

From the Sobolev inequality, Theorem 3.52, it follows that $a_{ij} \in L^\infty(\Omega)$. With

$$a(u,v) = \int_\Omega (A(\boldsymbol{x})\nabla u(\boldsymbol{x})) \cdot \nabla v(\boldsymbol{x})\ d\boldsymbol{x}$$

and the Cauchy–Schwarz inequality, one obtains

$$|a(u,v)| \le \|A\|_{L^\infty(\Omega)} \int_\Omega |\nabla u(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x})|\ \ d\boldsymbol{x} \le C \left\|\nabla u\right\|_{L^2(\Omega)} \left\|\nabla v\right\|_{L^2(\Omega)}$$

for all $u, v \in H_0^1(\Omega)$. In addition, it follows from (7.9) that

$$m \left\|\nabla u\right\|_{L^2(\Omega)}^2 \le a(u,u) \quad \forall\, u \in H_0^1(\Omega).$$

Hence, the bilinear form is bounded and elliptic. Using the Theorem of Lax–Milgram, Theorem 4.5, it follows that there es a unique weak solution $u \in H_0^1(\Omega)$ of (7.7) with

$$a(u,v) = f(v) \quad \forall\, v \in H_0^1(\Omega).$$

□

Remark 7.7. Assumptions and the discrete problem. The non-conforming Crouzeix–Raviart finite element P_1^{nc} was introduced in Example 5.30. To simplify the presentation, it will be restricted here on the two-dimensional case. In addition, to avoid the estimate of the error coming from approximating the domain, it will be assumed that Ω is a convex domain with polygonal boundary. It can be shown that in this case the boundary is Lipschitz.

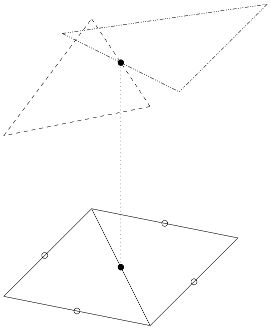

Fig. 7.1 Function from P_1^{nc}.

Let $\mathcal{T}^h$ be a regular triangulation of Ω with triangles. Let P_1^{nc} (nc – non-conforming) be denote the finite element space of piecewise linear functions which are continuous at the midpoints of the edges. This space is non-conforming if it is applied for the discretization of a second order elliptic equation since the continuous problem is given in $H_0^1(\Omega)$ and the functions of $H_0^1(\Omega)$ do not possess jumps. The functions of P_1^{nc} have generally jumps, see Figure 7.1, and they are not weakly differentiable. In addition, the space is also non-conforming with respect to the boundary condition, which is not satisfied exactly. The functions from P_1^{nc} vanish in the midpoint of the edges at the boundary. However, in the other points at the boundary, their value is generally not equal to zero.

The bilinear form

$$a(u,v) = \int_\Omega (A(\boldsymbol{x})\nabla u(\boldsymbol{x})) \cdot \nabla v(\boldsymbol{x}) \ d\boldsymbol{x}$$

will be extended to $H_0^1(\Omega) + P_1^{\mathrm{nc}}$ by

$$a^h(u,v) = \sum_{K\in\mathcal{T}^h} \int_K (A(\boldsymbol{x})\nabla u(\boldsymbol{x})) \cdot \nabla v(\boldsymbol{x}) \ d\boldsymbol{x} \quad \forall\, u,v \in H_0^1(\Omega) + P_1^{\mathrm{nc}}.$$

Then, the non-conforming finite element method is given by: Find $u^h \in P_1^{\mathrm{nc}}$ with

$$a^h(u^h,v^h) = (f,v^h) \quad \forall\, v^h \in P_1^{\mathrm{nc}}.$$

The goal of this section consists in proving the linear convergence with respect to h in the energy norm $\|\cdot\|_h = \left(a^h(\cdot,\cdot)\right)^{1/2}$. It will be assumed that the solution of the continuous problem (7.7) is smooth, i.e., that $u \in H^2(\Omega)$, that $f \in L^2(\Omega)$, and that the coefficients $a_{ij}(\boldsymbol{x})$ are weakly differentiable with bounded derivatives. □

Remark 7.8. The error equation. The first step of proving an error estimate consists in deriving an equation for the error. To this end, multiply the continuous problem (7.7) with a test function from $v^h \in P_1^{\mathrm{nc}}$, integrate the product on Ω, and apply integration by parts on each triangle. This approach gives

$$\begin{aligned}(f, v^h) &= -\sum_{K\in\mathcal{T}^h} \int_K \nabla\cdot\left(A(\boldsymbol{x})\nabla u(\boldsymbol{x})\right) v^h(\boldsymbol{x})\ d\boldsymbol{x} \\ &= \sum_{K\in\mathcal{T}^h} \int_K \left(A(\boldsymbol{x})\nabla u(\boldsymbol{x})\right)\cdot\nabla v^h(\boldsymbol{x})\ d\boldsymbol{x} \\ &\quad - \sum_{K\in\mathcal{T}^h} \int_{\partial K} \left(A(s)\nabla u(s)\right)\cdot\boldsymbol{n}_K(s) v^h(s)\ ds \\ &= a^h(u, v^h) - \sum_{K\in\mathcal{T}^h} \int_{\partial K} \left(A(s)\nabla u(s)\right)\cdot\boldsymbol{n}_K(s) v^h(s)\ ds,\end{aligned}$$

where $\boldsymbol{n}_K$ is the unit outer normal at the edges of the triangles. Subtracting the finite element equation, one obtains

$$a^h(u-u^h, v^h) = \sum_{K\in\mathcal{T}^h} \int_{\partial K} \left(A(s)\nabla u(s)\right)\cdot\boldsymbol{n}_K(s) v^h(s)\ ds \quad \forall\ v^h \in P_1^{\mathrm{nc}}. \quad (7.10)$$

□

Lemma 7.9. Estimate of the right-hand side of the error equation (7.10)**.** *Assume that* $u \in H^2(\Omega)$ *and* $a_{ij} \in W^{1,\infty}(\Omega)$, $i, j = 1, 2$, *then it is*

$$\left|\sum_{K\in\mathcal{T}^h} \int_{\partial K} A(s)\nabla u(s)\cdot\boldsymbol{n}_K(s) v^h(s)\ ds\right| \le Ch \left\|u\right\|_{H^2(\Omega)} \left\|v^h\right\|_h .$$

Proof. Every edge of the triangulation which is in Ω appears exactly twice in the boundary integrals on ∂K. The corresponding unit normals possess opposite signs. One can choose for each edge one unit normal and then one can write the integrals in the form

$$\sum_E \int_E \left[\left|\left(A(s)\nabla u(s)\right)\cdot\boldsymbol{n}_E(s) v^h(s)\right|\right]_E\ ds = \sum_E \int_E \left(A(s)\nabla u(s)\right)\cdot\boldsymbol{n}_E(s) \left[\left|v^h\right|\right]_E(s)\ ds,$$

where the sum is taken over all edges $\{E\}$. Here, $\left[\left|v^h\right|\right]_E$ denotes the jump of v^h

$$\left[\left|v^h\right|\right]_E(s) = \begin{cases} v^h|_{K_1}(s) - v^h|_{K_2}(s) & s \in E \subset \Omega, \\ v^h(s) & s \in E \subset \partial\Omega, \end{cases}$$

where $\boldsymbol{n}_E$ is directed from K_1 to K_2 or it is the outer normal on $\partial\Omega$. For writing the integrals in this form, it was used that $\nabla u(s)$, $A(s)$, and $\boldsymbol{n}_E(s)$ are almost everywhere continuous, such that these functions can be written as factor in front of the jumps. Because of the continuity condition for the functions from P_1^{nc} and the homogeneous Dirichlet boundary condition, it is for all $v^h \in P_1^{\mathrm{nc}}$ that $\left[\left|v^h\right|\right]_E(P) = 0$ for the midpoints P of all edges. From the linearity of the functions on the edges, it follows that

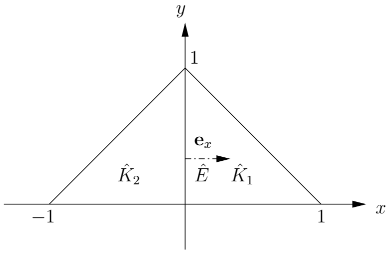

Fig. 7.2 Reference configuration.

$$\int_E [\![v^h]\!]_E(s)\,ds = 0 \quad \forall\, E. \tag{7.11}$$

Let E be an arbitrary edge in Ω which belongs to the triangles K_1 and K_2. The next goal consists in proving the estimate

$$\begin{aligned}&\left|\int_E (A(s)\nabla u(s))\cdot \boldsymbol{n}_E(s)\,[\![v^h]\!]_E(s)\,ds\right|\\ &\le Ch\,\|u\|_{H^2(K_1)}\left(\left\|\nabla v^h\right\|_{L^2(K_1)} + \left\|\nabla v^h\right\|_{L^2(K_2)}\right).\end{aligned} \tag{7.12}$$

To this end, one uses a reference configuration $\left(\hat{K}_1, \hat{K}_2, \hat{E}\right)$, where $\hat{K}_1$ is the unit triangle and $\hat{K}_2$ is the triangle that is obtained by reflecting the unit triangle at the y-axis. The common edge $\hat{E}$ is the interval $(0,1)$ on the y-axis. The unit normal on $\hat{E}$ will be chosen to be the Cartesian unit vector $\boldsymbol{e}_x$, see Figure 7.2. This choice is the other way around than in the definition of the jump, but it is just for simplicity of notation and it does not influence the estimate. The reference configuration can be transformed to (K_1, K_2, E) by a map which is continuous and on both triangles $\hat{K}_i$ affine. For this map one, can prove the same properties for the transform as proved in Chapter 6.

Using (7.11), the Cauchy–Schwarz inequality, and the trace theorem, one obtains for an arbitrary constant $\alpha \in \mathbb{R}$

$$\begin{aligned}\int_{\hat{E}} \left(\hat{A}(\hat{s})\nabla\hat{u}(\hat{s})\right)\cdot \boldsymbol{e}_x\,[\![\hat{v}^h]\!]_{\hat{E}}\,d\hat{s} &= \int_{\hat{E}} \left(\left(\hat{A}(\hat{s})\nabla\hat{u}(\hat{s})\right)\cdot \boldsymbol{e}_x - \alpha\right)[\![\hat{v}^h]\!]_{\hat{E}}\,d\hat{s}\\ &\le C\left\|\left(\hat{A}\nabla\hat{u}\right)\cdot \boldsymbol{e}_x - \alpha\right\|_{H^1(\hat{K}_1)}\left\|[\![\hat{v}^h]\!]_{\hat{E}}\right\|_{L^2(\hat{E})}\end{aligned} \tag{7.13}$$

In particular, one can choose α such that

$$\int_{\hat{E}} \left(\left(\hat{A}(\hat{s})\nabla\hat{u}(\hat{s})\right)\cdot \boldsymbol{e}_x - \alpha\right)\,d\hat{s} = 0.$$

Using first that $(a^2+b^2)^{1/2} \le a+b$ for $a, b \ge 0$, then the $L^2(\Omega)$ term in the first factor of the right-hand side of (7.13) can be bounded using the estimate from Lemma 6.4 for $k=0$ and $l=1$ and the choice of α

$$\begin{aligned}
&\left\|\left(\hat{A}\nabla\hat{u}\right)\cdot\boldsymbol{e}_x-\alpha\right\|_{H^1(\hat{K}_1)}\\
&\le\left(\left\|\left(\hat{A}\nabla\hat{u}\right)\cdot\boldsymbol{e}_x-\alpha\right\|_{L^2(\hat{K}_1)}+\left\|\nabla\left(\left(\hat{A}\nabla\hat{u}\right)\cdot\boldsymbol{e}_x-\alpha\right)\right\|_{L^2(\hat{K}_1)}\right)\\
&\le C\left\|\nabla\left(\left(\hat{A}\nabla\hat{u}\right)\cdot\boldsymbol{e}_x-\alpha\right)\right\|_{L^2(\hat{K}_1)}\\
&=C\left\|\nabla\left(\left(\hat{A}\nabla\hat{u}\right)\cdot\boldsymbol{e}_x\right)\right\|_{L^2(\hat{K}_1)}.
\end{aligned}$$

To estimate the second factor, the trace theorem and the equivalence of norms in finite-dimensional spaces are applied

$$\begin{aligned}
\left\|\left[\left|\hat{v}^h\right|\right]_{\hat{E}}\right\|_{L^2(\hat{E})}&\le C\left(\left\|\hat{v}^h\right\|_{H^1(\hat{K}_1)}+\left\|\hat{v}^h\right\|_{H^1(\hat{K}_2)}\right)\\
&\le C\left(\left\|\nabla\hat{v}^h\right\|_{L^2(\hat{K}_1)}+\left\|\nabla\hat{v}^h\right\|_{L^2(\hat{K}_2)}\right).
\end{aligned}\tag{7.14}$$

To apply the norm equivalence, one has to prove that the terms in the last line are in fact norms. Let the terms in the last line be zero, then it follows that $\hat{v}^h=c_1$ in $\hat{K}_1$ and $\hat{v}^h=c_2$ in $\hat{K}_2$. Because $\hat{v}^h$ is continuous in the midpoint of $\hat{E}$, one finds that $c_1=c_2$ and consequently that $\left[\left|\hat{v}^h\right|\right]_{\hat{E}}=0$. Hence, also the left-hand side of the estimate is zero. It follows that the right-hand side of estimate (7.14) defines a norm in the quotient space of P_1^{nc} with respect to $\left[\left|\hat{v}^h\right|\right]_{\hat{E}}=0$.

Altogether, one obtains for the reference configuration

$$\begin{aligned}
&\left|\int_{\hat{E}}\left(\hat{A}(\hat{s})\nabla u(\hat{s})\right)\cdot\boldsymbol{e}_x\left[\left|\hat{v}^h\right|\right]_{\hat{E}}\ d\hat{s}\right|\\
&\le C\left\|\nabla\left(\left(\hat{A}\nabla u\right)\cdot\boldsymbol{e}_x\right)\right\|_{L^2(\hat{K}_1)}\left(\left\|\nabla\hat{v}^h\right\|_{L^2(\hat{K}_1)}+\left\|\nabla\hat{v}^h\right\|_{L^2(\hat{K}_2)}\right).
\end{aligned}$$

This estimate has to be transformed to the triple (K_1,K_2,E). In this step, one gets for the integral on the edge the factor C (Ch for ∇ and Ch^{-1} for $d\hat{s}$). For the product of the norms on the right-hand side, one obtains the factor Ch (Ch for the first factor and C for the second factor). In addition, one uses that $A(s)$ and all first order derivatives of $A(s)$ are bounded to estimated the first term on the right-hand side (*exercise*). In summary, (7.12) is proved.

The statement of the lemma follows by summing over all edges and by applying on the right-hand side the Cauchy–Schwarz inequality. ■

Theorem 7.10. Finite element error estimate. *Let the assumptions of Lemma 7.9 be satisfied, then it holds the following error estimate*

$$\left\|u-u^h\right\|_h^2\le Ch\left\|u\right\|_{H^2(\Omega)}\left\|u-u^h\right\|_h+Ch^2\left\|u\right\|_{H^2(\Omega)}^2.$$

Proof. Applying Lemma 7.9, it follows from the error equation (7.10) that

$$\left|a^h(u-u^h,v^h)\right|\le Ch\left\|u\right\|_{H^2(\Omega)}\left\|v^h\right\|_h\quad\forall\ v^h\in P_1^{\mathrm{nc}}.$$

Let $I^h\ :\ H_0^1(\Omega)\to P_1^{\mathrm{nc}}$ be an interpolation operator with optimal interpolation order in $\left\|\cdot\right\|_h$. Then, one obtains with the Cauchy–Schwarz inequality, the triangle inequality, and the interpolation estimate

$$\begin{aligned}
\left\|u-u^h\right\|_h^2 &= a^h(u-u^h,u-u^h) = a^h(u-u^h,u-I^hu) + a^h(u-u^h,I^hu-u^h)\\
&\le \left|a^h(u-u^h,u-I^hu)\right| + Ch\left\|u\right\|_{H^2(\Omega)}\left\|I^hu-u^h\right\|_h\\
&\le \left\|u-u^h\right\|_h\left\|u-I^hu\right\|_h + Ch\left\|u\right\|_{H^2(\Omega)}\left(\left\|I^hu-u\right\|_h+\left\|u-u^h\right\|_h\right)\\
&\le Ch\left\|u-u^h\right\|_h\left\|u\right\|_{H^2(\Omega)} + Ch\left\|u\right\|_{H^2(\Omega)}\left(h\left\|u\right\|_{H^2(\Omega)}+\left\|u-u^h\right\|_h\right).
\end{aligned}$$

■

Remark 7.11. To the error estimate. If h is sufficiently small, than the second term of the error estimate is of higher order and this term can be absorbed in the constant of the first term. One obtains the asymptotic error estimate

$$\left\|u-u^h\right\|_h \le Ch\left\|u\right\|_{H^2(\Omega)},$$

i.e., first order convergence. □

7.3 $L^2(\Omega)$ Error Estimate

Remark 7.12. Motivation. A method is called quasi-optimal in a given norm, if the order of the method is the same as the optimal approximation order. Already for one dimension, one can show that at most linear convergence in $H^1(\Omega)$ can be achieved for the best approximation in P_1. This statement can be already verified with the function $v(x) = x^2$. Hence, all considered methods so far are quasi-optimal in the energy norm.

However, the best approximation error in $L^2(\Omega)$ is of one order higher than the best approximation error in $H^1(\Omega)$. A natural question is whether finite element methods converge also of higher order with respect to the error in $L^2(\Omega)$ than with respect to the error in the energy norm.

In this section, it will be shown that one can obtain for finite element methods a higher order of convergence in $L^2(\Omega)$ than in $H^1(\Omega)$. However, there are more restrictive assumptions to prove this property in comparison with the convergence proof for the energy norm. □

Remark 7.13. Model problem. Let $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, be a convex polyhedral domain with Lipschitz boundary. The model problem has the form

$$-\Delta u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega. \tag{7.15}$$

For proving an error estimate in $L^2(\Omega)$, the regularity of the solution of (7.15) plays an essential role. □

Definition 7.14. m-regular differential operator. Let L be a second order differential operator. This operator is called m-regular, $m \ge 2$, if for all $f \in H^{m-2}(\Omega)$ the solutions of $Lu = f$ in Ω, $u = 0$ on $\partial\Omega$, are in the space $H^m(\Omega)$ and the following estimate holds

$$\|u\|_{H^m(\Omega)} \le C \|f\|_{H^{m-2}(\Omega)} + C \|u\|_{H^1(\Omega)} . \tag{7.16}$$

□

Remark 7.15. On the m-regularity.

- The definition is formulated in a way that it can be applied also if the solution of the problem is not unique.
- For the Laplacian, the term $\|u\|_{H^1(\Omega)}$ can be estimated by $\|f\|_{L^2(\Omega)}$ such that with (7.16) one obtains (*exercise*)

$$\|u\|_{H^2(\Omega)} \le C \|f\|_{L^2(\Omega)} .$$

- Many regularity results can be found in the literature. Loosely speaking, they say that regularity is given if the data of the problem (coefficients of the operator, boundary of the domain) are sufficiently regular. For instance, an elliptic operator in divergence form ($\Delta = \nabla \cdot (A\nabla)$) is 2-regular if the coefficients are from $W^{1,p}(\Omega)$, $p \ge 1$, and if $\partial\Omega$ is a C^2 boundary. Another important result is the 2-regularity of the Laplacian on a convex domain. A comprehensive overview on regularity results can be found in Grisvard (1985).

□

Remark 7.16. Variational form and finite element formulation of the model problem. The variational form of (7.15) is: Find $u \in H_0^1(\Omega)$ with

$$(\nabla u, \nabla v) = (f, v) \quad \forall\, v \in H_0^1(\Omega).$$

The P_1 finite element space, with zero boundary conditions, will be used for the discretization. Then, the finite element problem reads as follows: Find $u^h \in P_1$ such that

$$(\nabla u^h, \nabla v^h) = (f, v^h) \quad \forall\, v^h \in P_1. \tag{7.17}$$

□

Theorem 7.17. Finite element error estimates. *Let $u(\boldsymbol{x})$ be the solution of* (7.15)*, let* (7.15) *be 2-regular, and let $u^h(\boldsymbol{x})$ be the solution of* (7.17)*. Then, the following error estimates hold*

$$\begin{aligned} \left\|\nabla(u - u^h)\right\|_{L^2(\Omega)} &\le Ch \|f\|_{L^2(\Omega)} , \\ \left\|u - u^h\right\|_{L^2(\Omega)} &\le Ch^2 \|f\|_{L^2(\Omega)} . \end{aligned}$$

Proof. With the error estimate in $H^1(\Omega)$, Corollary 6.16, and the 2-regularity, one obtains

$$\left\|\nabla(u - u^h)\right\|_{L^2(\Omega)} \le Ch \|u\|_{H^2(\Omega)} \le Ch \|f\|_{L^2(\Omega)} . \tag{7.18}$$

For proving the $L^2(\Omega)$ error estimate, let $w \in H_0^1(\Omega)$ be the unique solution of the so-called dual problem

$$(\nabla v, \nabla w) = (u - u^h, v) \quad \forall\, v \in H_0^1(\Omega).$$

For a symmetric differential operator, the dual problem has the same form like the original (primal) problem. Hence, the dual problem is also 2-regular and it holds the estimate

$$\|w\|_{H^2(\Omega)} \le C \left\|u - u^h\right\|_{L^2(\Omega)}.$$

For performing the error estimate, the Galerkin orthogonality of the error is utilized

$$(\nabla(u - u^h), \nabla v^h) = (\nabla u, \nabla v^h) - (\nabla u^h, \nabla v^h) = (f, v^h) - (f, v^h) = 0$$

for all $v^h \in P_1$. Now, the error $u - u^h$ is used as test function v in the dual problem. Let $I^h w$ be the interpolant of w in P_1. Using the Galerkin orthogonality, the interpolation estimate, and the regularity of w, one obtains

$$\begin{aligned}
\left\|u - u^h\right\|_{L^2(\Omega)}^2 &= (\nabla(u - u^h), \nabla w) = (\nabla(u - u^h), \nabla(w - I^h w)) \\
&\le \left\|\nabla(u - u^h)\right\|_{L^2(\Omega)} \left\|\nabla(w - I^h w)\right\|_{L^2(\Omega)} \\
&\le Ch \|w\|_{H^2(\Omega)} \left\|\nabla(u - u^h)\right\|_{L^2(\Omega)} \\
&\le Ch \left\|u - u^h\right\|_{L^2(\Omega)} \left\|\nabla(u - u^h)\right\|_{L^2(\Omega)}.
\end{aligned}$$

Finally, division by $\left\|u - u^h\right\|_{L^2(\Omega)}$ and the application of the already known error estimate (7.18) for $\left\|\nabla(u - u^h)\right\|_{L^2(\Omega)}$ are used for completing the proof of the theorem. ■

7.4 Outlook

Remark 7.18. Outlook to forthcoming classes. This class provided an introduction to numerical methods for solving partial differential equations and the numerical analysis of these methods. There are many further aspects that will be covered in forthcoming classes.

Further aspects for elliptic problems.

- Adaptive methods and a posteriori error estimators. It will be shown how it is possible to estimate the error of the computed solution only using known quantities and in which ways one can decide where it makes sense to refine the mesh and where not.
- Multigrid methods. Multigrid methods are for certain classes of problems optimal solvers.
- Numerical analysis of problems with other boundary conditions or taking into account quadrature rules.

Time-dependent problems. As mentioned in Remark 1.7, standard approaches for the numerical solution of time-dependent problems are based on solving stationary problems in each discrete time.

- The numerical analysis of discretizations of time-dependent problems has some new aspects, but also many tools from the analysis of steady-state problems are used.

Convection-diffusion equations. Convection-diffusion equations are of importance in many applications. Generally, the convection (first order differential operator) dominates the diffusion (second order differential operator).

- In the convection-dominated regime, the Galerkin method as presented in this class does not work. One needs new ideas for discretizations and these new discretizations create new challenges for the numerical analysis.

Problems with more than one unknown function. The fundamental equation of fluid dynamics, the Navier–Stokes equations, Section 1.3, belong to this class.

- It will turn out that the discretization of the Navier–Stokes equations requires special care in the choice of the finite element spaces. The numerical analysis becomes rather involved.

□

References

Adams, R. A. (1975) *Sobolev spaces*. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, pp. xviii+268. Pure and Applied Mathematics, Vol. 65.

Adams, R. A. & Fournier, J. J. F. (2003) *Sobolev spaces*. Pure and Applied Mathematics (Amsterdam), vol. 140, second edn. Elsevier/Academic Press, Amsterdam, pp. xiv+305.

Alt, H. (1999) *Lineare Funktionalanalysis. Eine anwendungsorientierte Einführung*, 3. edn. Springer Berlin.

Braess, D. (2001) *Finite elements*, second edn. Cambridge: Cambridge University Press, pp. xviii+352. Theory, fast solvers, and applications in solid mechanics, Translated from the 1992 German edition by Larry L. Schumaker.

Brenner, S. C. & Scott, L. R. (2008) *The mathematical theory of finite element methods*. Texts in Applied Mathematics, vol. 15, third edn. New York: Springer, pp. xviii+397.

Ciarlet, P. G. (1978) *The finite element method for elliptic problems*. Amsterdam: North-Holland Publishing Co., pp. xix+530. Studies in Mathematics and its Applications, Vol. 4.

Ciarlet, P. G. (2002) *The finite element method for elliptic problems*. Classics in Applied Mathematics, vol. 40. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), pp. xxviii+530. Reprint of the 1978 original [North-Holland, Amsterdam; MR0520174 (58 #25001)].

Crouzeix, M. & Raviart, P.-A. (1973) Conforming and nonconforming finite element methods for solving the stationary Stokes equations. I. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge*, **7**, 33–75.

Deuflhard, P. & Weiser, M. (2012) *Adaptive numerical solution of PDEs*. de Gruyter Textbook. Berlin: Walter de Gruyter & Co., pp. xii+421.

Dziuk, G. (2010) *Theorie und Numerik partieller Differentialgleichungen*. Walter de Gruyter GmbH & Co. KG, Berlin, pp. x+319.

Ern, A. & Guermond, J.-L. (2004) *Theory and practice of finite elements*. Applied Mathematical Sciences, vol. 159. New York: Springer-Verlag, pp. xiv+524.

Evans, L. C. (2010) *Partial differential equations*. Graduate Studies in Mathematics, vol. 19, second edn. Providence, RI: American Mathematical Society, pp. xxii+749.

Fefferman, C. (2000). http://www.claymath.org/millennium/Navier-Stokes_Equations/.

Galdi, G. P. (2011) *An introduction to the mathematical theory of the Navier-Stokes equations*. Springer Monographs in Mathematics, second edn. Springer, New York, pp. xiv+1018. Steady-state problems.

Goering, H., Hans-Görg, R. & Tobiska, L. (2010) *Die Finite-Elemente-Methode fr Anfänger*, fourth edn. Wiley-VCH, Berlin, pp. ix + 219.

Grisvard, P. (1985) *Elliptic problems in nonsmooth domains*. Monographs and Studies in Mathematics, vol. 24. Boston, MA: Pitman (Advanced Publishing Program), pp. xiv+410.

Grossmann, C. & Roos, H.-G. (2007) *Numerical treatment of partial differential equations*. Universitext. Berlin: Springer, pp. xii+591. Translated and revised from the 3rd (2005) German edition by Martin Stynes.

Haroske, D. D. & Triebel, H. (2008) *Distributions, Sobolev spaces, elliptic equations*. EMS Textbooks in Mathematics. European Mathematical Society (EMS), Zürich, pp. x+294.

John, V. (2016) *Finite element methods for incompressible flow problems*. Springer Series in Computational Mathematics, vol. 51. Springer, Cham, pp. xiii+812.

John, V. & Matthies, G. (2004) MooNMD—a program package based on mapped finite element methods. *Comput. Vis. Sci.*, **6**, 163–169.

Landau, L. & Lifschitz, E. (1966) *Lehrbuch der theoretischen Physik*, vol. VI, Hydrodynamik. Akademie-Verlag Berlin.

LeVeque, R. J. (2007) *Finite difference methods for ordinary and partial differential equations*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), pp. xvi+341. Steady-state and time-dependent problems.

Rannacher, R. & Turek, S. (1992) Simple nonconforming quadrilateral Stokes element. *Numer. Methods Partial Differential Equations*, **8**, 97–111.

Samarskii, A. A. (2001) *The theory of difference schemes*. Monographs and Textbooks in Pure and Applied Mathematics, vol. 240. New York: Marcel Dekker Inc., pp. xviii+761.

Samarskij, A. (1984) *Theorie der Differenzenverfahren*. Mathematik und ihre Anwendungen in Physik und Technik, vol. 40. Akademische Verlagsgesellschaft Geest & Portig K.-G., Leipzig.

Schieweck, F. (1997) *Parallele Lösung der stationären inkompressiblen Navier-Stokes Gleichungen*. Otto-von-Guericke-Universität Magdeburg, Fakultät für Mathematik. Habilitation.

Smirnow, W. I. (1967) *Lehrgang der höheren Mathematik. Teil V*. VEB Deutscher Verlag der Wissenschaften, Berlin, pp. xiii+570. Zweite, berichtigte Auflage, Übersetzung aus dem Russischen von Renate Helle und Brigitte Mai, Hochschulbücher für Mathematik, Band 6.

Šolín, P. (2006) *Partial differential equations and the finite element method*. Pure and Applied Mathematics (New York). Hoboken, NJ: Wiley-Interscience [John Wiley & Sons], pp. xviii+472.

Strang, G. & Fix, G. (2008) *An analysis of the finite element method*, second edn. Wellesley, MA: Wellesley-Cambridge Press, pp. x+402.

Wilbrandt, U., Bartsch, C., Ahmed, N., Alia, N., Anker, F., Blank, L., Caiazzo, A., Ganesan, S., Giere, S., Matthies, G., Meesala, R., Shamim, A., Venkatesan, J. & John, V. (2017) ParMooN—A modernized program package based on mapped finite elements. *Comput. Math. Appl.*, **74**, 74–88.

Wladimirow, W. S. (1972) *Gleichungen der mathematischen Physik*. VEB Deutscher Verlag der Wissenschaften, Berlin.