2

The Finite Element Method for the Poisson Equation

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

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

2.1 Variational Formulation for the Model Problem

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

For the present we define as the basic space V,

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

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

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

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

Then the following properties hold:

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

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

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

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

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

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

Equation (2.10) is also called a variational equation.

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

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

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

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

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

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

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

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

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

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

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

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

where

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

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

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

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

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

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

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

It follows analogously to (2.12):

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

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

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

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

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

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

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

and the minimization problem as

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

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

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

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

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

By subtraction, it follows that

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

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

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

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

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

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

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

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

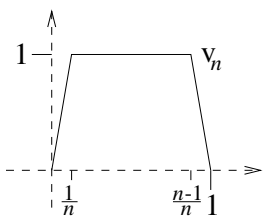

Figure 2.1. The function v_n.

Then

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

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

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

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

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

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

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

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

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

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

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

Then

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

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

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

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

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

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

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

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

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

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

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

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

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

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

of the first partial derivatives.

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

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

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

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

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

Therefore, we make the following definition:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

with the norm

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

induced by this scalar product.

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

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

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

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

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

Exercises

2.1

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

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

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

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

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

2.2 The Finite Element Method with Linear Elements

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

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

Variational equation:

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

Minimization problem:

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

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

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

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

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

This approach is called the *Ritz method*.

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

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

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

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

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

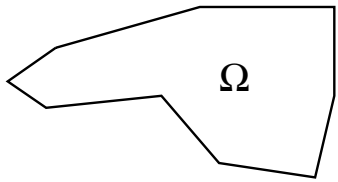

Figure 2.2. Domain with a polygonal boundary.

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

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

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

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

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

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

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

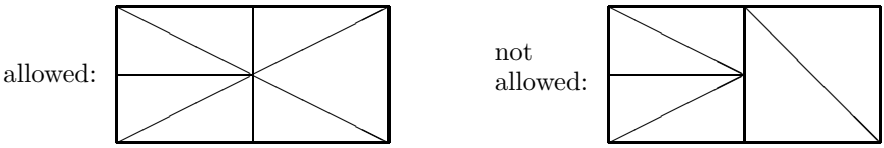

Figure 2.3. Triangulations.

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

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

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

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

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

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

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

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

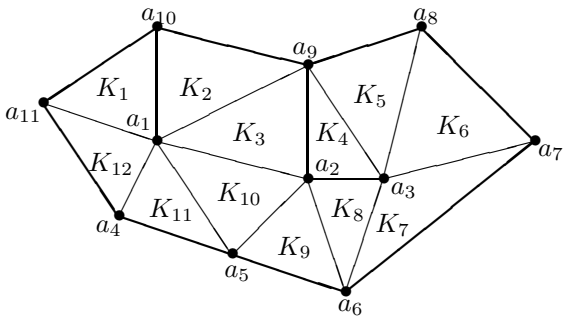

Figure 2.4. A conforming triangulation with $N = 12$, $M = 11$, $M_1 = 3$, $M_2 = 8$.

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

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

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

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

We have

$$V_h \subset V\,.$$

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

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

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

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

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

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

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

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

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

$$u \in V_h\,.$$

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

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

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

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

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

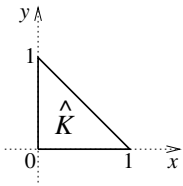

Figure 2.5. Reference element $\hat{K}$.

with the *shape functions*

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

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

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

and thus

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Since

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

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

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

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

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

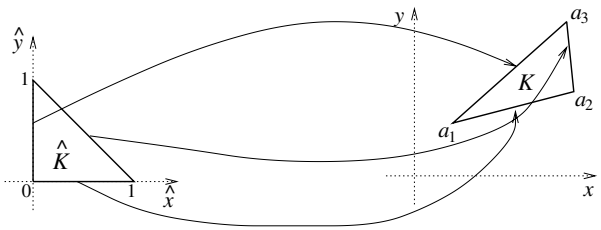

Figure 2.6. Affine-linear transformation.

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

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

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

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

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

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

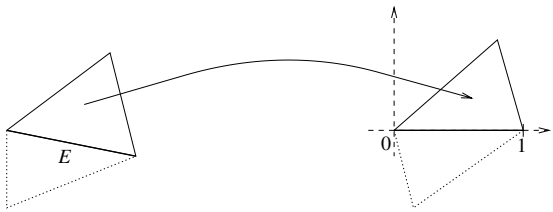

Figure 2.7. Affine-linear transformation of E on the reference element $[0, 1]$.

Therefore, the definition of u by means of

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

is unique, and this function satisfies $u \in C(\bar{\Omega})$ and $u \in X_h$.

(3) Finally, we will show that $u = 0$ on $\partial\Omega$ for u defined by (2.31) if $u_i = 0\,(i = M_1 + 1, \ldots, M)$ for the boundary nodes.
The boundary $\partial\Omega$ consists of edges of elements $K \in \mathcal{T}_h$. Let E be such an edge; i.e., E has the vertices a_{i_1}, a_{i_2} with $i_j \in \{M_1 + 1, \ldots, M\}$. The given boundary values yield $u(a_{i_j}) = 0$ for $j = 1, 2$. By means of an affine transformation analogously to the above one we obtain that $u|_E$ is a polynomial of first degree in one variable and that $u|_E$ vanishes at two points. So $u|_E = 0$, and the assertion follows. □

The following statement is an important consequence of the unique solvability of the interpolation problem in X_h irrespective of its particular definition: The interpolation conditions

$$\varphi_i(a_j) = \delta_{ij}\,, \quad j = 1, \ldots, M\,, \tag{2.32}$$

uniquely determine functions $\varphi_i \in X_h$ for $i = 1, \ldots, M$. For any $u \in X_h$, we have

$$u(x) = \sum_{i=1}^{M} u(a_i)\varphi_i(x) \quad \text{for} \quad x \in \Omega\,, \tag{2.33}$$

because both the left-hand side and the right-hand side functions belong to X_h and are equal to $u(a_i)$ at $x = a_i$.

The representation $u = \sum_{i=1}^{M} \alpha_i\varphi_i$ is unique, too, for otherwise, a function $w \in X_h$, $w \neq 0$, such that $w(a_i) = 0$ for all $i = 1, \ldots, M$ would exist. Thus $\{\varphi_1, \ldots, \varphi_M\}$ is a basis of X_h, especially $\dim X_h = M$. This basis is called a *nodal basis* because of (2.33). For the particular case of a piecewise linear ansatz space on triangles, the basis functions are called

pyramidal functions because of their shape. If the set of indices is restricted to $\{1,\dots,M_1\}$; i.e., we omit the basis functions corresponding to the boundary nodes, then a basis of V_h will be obtained and $\dim V_h = M_1$.

Summary: The function values $u(a_i)$ at the nodes $a_1,\dots,a_M$ are the *degrees of freedom* of $u \in X_h$, and the values at the interior points $a_1,\dots,a_{M_1}$ are the *degrees of freedom* of $u \in V_h$.

The following consideration is valid for an arbitrary ansatz space V_h with a basis $\{\varphi_1,\dots,\varphi_M\}$. The Galerkin method (2.23) reads as follows: Find $u_h = \sum_{i=1}^M \xi_i\varphi_i \in V_h$ such that $a(u_h,v) = b(v)$ for all $v \in V_h$. Since $v = \sum_{i=1}^M \eta_i\varphi_i$ for $\eta_i \in \mathbb{R}$, this is equivalent to

$$\begin{aligned} a(u_h,\varphi_i) &= b(\varphi_i) \quad \text{for all } i=1,\dots,M \iff \\ a\left(\sum_{j=1}^M \xi_j\varphi_j,\varphi_i\right) &= b(\varphi_i) \quad \text{for all } i=1,\dots,M \iff \\ \sum_{j=1}^M a(\varphi_j,\varphi_i)\,\xi_j &= b(\varphi_i) \quad \text{for all } i=1,\dots,M \iff \\ A_h\boldsymbol{\xi} &= \boldsymbol{q}_h \end{aligned} \tag{2.34}$$

with $A_h = (a(\varphi_j,\varphi_i))_{ij} \in \mathbb{R}^{M,M}$, $\boldsymbol{\xi} = (\xi_1,\dots,\xi_M)^T$ and $\boldsymbol{q}_h = (b(\varphi_i))_i$. Therefore, the Galerkin method is equivalent to the system of equations (2.34).

The considerations for deriving (2.34) show that, in the case of equivalence of the Galerkin method with the Ritz method, the system of equations (2.34) is equivalent to the minimization problem

$$F_h(\boldsymbol{\xi}) = \min\left\{F_h(\boldsymbol{\eta}) \;\middle|\; \boldsymbol{\eta} \in \mathbb{R}^M\right\}, \tag{2.35}$$

where

$$F_h(\boldsymbol{\eta}) = \frac{1}{2}\boldsymbol{\eta}^T A_h\boldsymbol{\eta} - \boldsymbol{q}_h^T\boldsymbol{\eta}\,.$$

Because of the symmetry and positive definiteness, the equivalence of (2.34) and (2.35) can be easily proven, and it forms the basis for the CG methods that will be discussed in Section 5.2.

Usually, A_h is called *stiffness matrix*, and $\boldsymbol{q}_h$ is called the *load vector*. These names originated from mechanics. For our model problem, we have

$$\begin{aligned} (A_h)_{ij} &= a(\varphi_j,\varphi_i) = \int_\Omega \nabla\varphi_j\cdot\nabla\varphi_i\,dx\,, \\ (\mathbf{q}_h)_i &= b(\varphi_i) = \int_\Omega f\varphi_i\,dx\,. \end{aligned}$$

By applying the finite element method, we thus have to perform the following steps:

(1) Determination of $A_h, \boldsymbol{q}_h$. This step is called *assembling*.

(2) Solution of $A_h \boldsymbol{\xi} = \boldsymbol{q}_h$.

If the basis functions φ_i have the property $\varphi_i(a_j) = \delta_{ij}$, then the solution of system (2.34) satisfies the relation $\xi_i = u_h(a_i)$, i.e., we obtain the vector of the nodal values of the finite element approximation.

Using only the properties of the bilinear form a, we obtain the following properties of A_h:

- A_h is symmetric for an arbitrary basis $\{\varphi_i\}$ because a is symmetric.
- A_h is positive definite for an arbitrary basis $\{\varphi_i\}$ because for $u = \sum_{i=1}^M \xi_i \varphi_i$,

$$\begin{aligned} \boldsymbol{\xi}^T A_h \boldsymbol{\xi} &= \sum_{i,j=1}^M \xi_j a(\varphi_j, \varphi_i) \xi_i = \sum_{j=1}^M \xi_j a\left(\varphi_j, \sum_{i=1}^M \xi_i \varphi_i\right) \\ &= a\left(\sum_{j=1}^M \xi_j \varphi_j, \sum_{i=1}^M \xi_i \varphi_i\right) = a(u,u) > 0 \end{aligned} \tag{2.36}$$

for $\boldsymbol{\xi} \neq 0$ and therefore $u \not\equiv 0$.
Here we have used only the positive definiteness of a.

Thus we have proven the following lemma.

Lemma 2.11 *The Galerkin method* (2.23) *has a unique solution if a is a symmetric, positive definite bilinear form and if b is a linear form.*

In fact, as we will see in Theorem 3.1, the symmetry of a is not necessary.

- For a special basis (i.e., for a specific finite element method), A_h is a sparse matrix, i.e., only a few entries $(A_h)_{ij}$ do not vanish. Evidently,

$$(A_h)_{ij} \neq 0 \quad \Leftrightarrow \quad \int_\Omega \nabla\varphi_j \cdot \nabla\varphi_i \, dx \neq 0 \,.$$

This can happen only if $\operatorname{supp}\varphi_i \cap \operatorname{supp}\varphi_j \neq \emptyset$, as this property is again necessary for $\operatorname{supp}\nabla\varphi_i \cap \operatorname{supp}\nabla\varphi_j \neq \emptyset$ because of

$$(\operatorname{supp}\nabla\varphi_i \cap \operatorname{supp}\nabla\varphi_j) \subset (\operatorname{supp}\varphi_i \cap \operatorname{supp}\varphi_j) \,.$$

The basis function φ_i vanishes on an element that does not contain the node a_i because of the uniqueness of the solution of the local interpolation problem. Therefore,

$$\operatorname{supp}\varphi_i = \bigcup_{\substack{K \in \mathcal{T}_h \\ a_i \in K}} K \,,$$

cf. Figure (2.8), and thus

$$(A_h)_{ij} \neq 0 \quad \Rightarrow \quad a_i, a_j \in K \text{ for some } K \in \mathcal{T}_h \,; \tag{2.37}$$

i.e., a_i, a_j are *neighbouring* nodes.

If we use the piecewise linear ansatz space on triangles and if a_i is an interior node in which L elements meet, then there exist at most L nondiagonal entries in the ith row of A_h. This number is determined only by the type of the triangulation, and it is independent of the fineness h, i.e., of the number of unknowns of the system of equations.

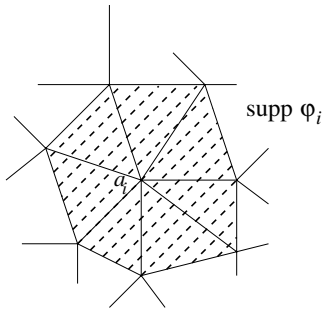

Figure 2.8. Support of the nodal basis function.

Example 2.12 We consider again the boundary value problem (2.1), (2.2) on $\Omega = (0, a) \times (0, b)$ again, i.e.

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

under the condition (1.4). The triangulation on which the method is based is created by a partition of Ω into squares with edges of length h and by a subsequent uniform division of each square into two triangles according to a fixed rule (*Friedrichs–Keller triangulation*). In order to do this, two possibilities (a) and (b) (see Figures 2.9 and 2.10) exist.

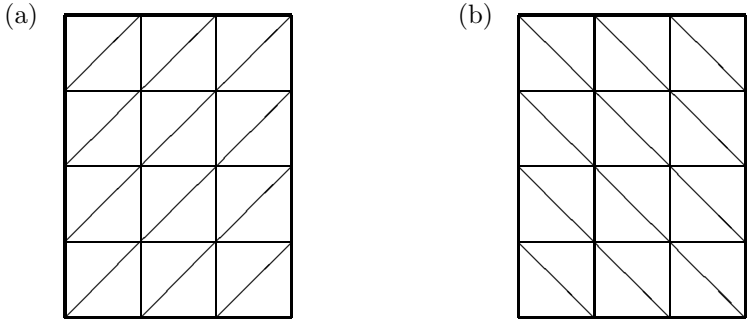

Figure 2.9. Possibilities of Friedrichs–Keller triangulation.

In both cases, a node a_Z belongs to six elements, and consequently, it has at most six neighbours:

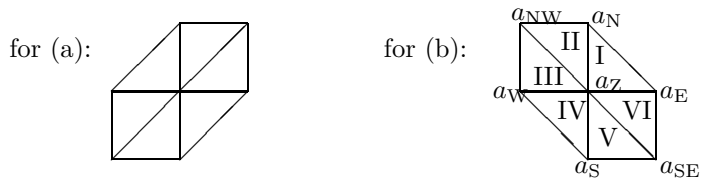

Figure 2.10. Support of the basis function.

Case (a) becomes case (b) by the transformation $x \mapsto a - x, y \mapsto y$. This transformation leaves the differential equation or the weak formulation, respectively, unchanged. Thus the Galerkin method with the ansatz space V_h according to (2.27) does not change, because $\mathcal{P}_1$ is invariant with respect to the above transformation. Therefore, the discretization matrices A_h according to (2.34) are seen to be identical by taking into account the renumbering of the nodes by the transformation.

Thus it is sufficient to consider only one case, say (b). A node which is far away from the boundary has 6 neighbouring nodes in $\{a_1, \ldots, a_{M_1}\}$, a node close to the boundary has less. The entries of the matrix in the row corresponding to a_{Z} depend on the derivatives of the basis function φ_{Z} as well as on the derivatives of the basis functions corresponding to the neighbouring nodes. The values of the partial derivatives of φ_{Z} in elements having the common vertex a_{Z} are listed in Table 2.1, where these elements are numbered according to Figure 2.10.

	I	II	III	IV	V	VI
$\partial_1 \varphi_{\mathrm{Z}}$	$-\frac{1}{h}$	0	$\frac{1}{h}$	$\frac{1}{h}$	0	$-\frac{1}{h}$
$\partial_2 \varphi_{\mathrm{Z}}$	$-\frac{1}{h}$	$-\frac{1}{h}$	0	$\frac{1}{h}$	$\frac{1}{h}$	0

Table 2.1. Derivatives of the basis functions.

Thus for the entries of the matrix in the row corresponding to a_{Z} we have

$$(A_h)_{\mathrm{Z,Z}} = a(\varphi_{\mathrm{Z}}, \varphi_{\mathrm{Z}}) = \int_{\mathrm{I}\cup\ldots\cup\mathrm{VI}} |\nabla \varphi_{\mathrm{Z}}|^2 \, dx = 2 \int_{\mathrm{I}\cup\mathrm{II}\cup\mathrm{III}} \left[(\partial_1 \varphi_{\mathrm{Z}})^2 + (\partial_2 \varphi_{\mathrm{Z}})^2 \right] dx,$$

because the integrands are equal on I and IV, on II and V, and on III and VI. Therefore

$$(A_h)_{\mathrm{Z,Z}} = 2 \int_{\mathrm{I}\cup\mathrm{III}} (\partial_1 \varphi_{\mathrm{Z}})^2 \, dx + 2 \int_{\mathrm{I}\cup\mathrm{II}} (\partial_2 \varphi_{\mathrm{Z}})^2 \, dx = 2h^{-2}h^2 + 2h^{-2}h^2 = 4\,,$$

$$\begin{aligned} (A_h)_{\mathrm{Z,N}} = a(\varphi_{\mathrm{N}}, \varphi_{\mathrm{Z}}) &= \int_{\mathrm{I}\cup\mathrm{II}} \nabla \varphi_{\mathrm{N}} \cdot \nabla \varphi_{\mathrm{Z}} \, dx \\ &= \int_{\mathrm{I}\cup\mathrm{II}} \partial_2 \varphi_{\mathrm{N}} \partial_2 \varphi_{\mathrm{Z}} \, dx = \int_{\mathrm{I}\cup\mathrm{II}} \left(-h^{-1}\right) h^{-1} dx = -1\,, \end{aligned}$$

because $\partial_1 \varphi_Z = 0$ on II and $\partial_1 \varphi_N = 0$ on I. The element I for φ_N corresponds to the element V for φ_Z; i.e., $\partial_1 \varphi_N = 0$ on I, analogously, it follows that $\partial_2 \varphi_N = h^{-1}$ on I $\cup$ II. In the same way we get

$$(A_h)_{Z,E} = (A_h)_{Z,W} = (A_h)_{Z,S} = -1$$

as well as

$$(A_h)_{Z,NW} = a(\varphi_{NW}, \varphi_Z) = \int_{II \cup III} \partial_1 \varphi_{NW}\, \partial_1 \varphi_Z + \partial_2 \varphi_{NW}\, \partial_2 \varphi_Z \, dx = 0\,.$$

The last identity is due to $\partial_1 \varphi_{NW} = 0$ on III and $\partial_2 \varphi_{NW} = 0$ on III, because the elements V and VI for φ_Z agree with the elements III and II for φ_{NW}, respectively.

Analogously, we obtain for the remaining value

$$(A_h)_{Z,SE} = 0\,,$$

such that only 5 (instead of the maximum 7) nonzero entries per row exist.

The way of assembling the stiffness matrix described above is called *node-based* assembling. However, most of the computer programs implementing the finite element method use an *element-based* assembling, which will be considered in Section 2.4.

If the nodes are numbered rowwise analogously to (1.13) and if the equations are divided by h^2, then $h^{-2} A_h$ coincides with the discretization matrix (1.14), which is known from the finite difference method. But here the right-hand side is given by

$$h^{-2} (\mathbf{q}_h)_i = h^{-2} \int_\Omega f \varphi_i \, dx = h^{-2} \int_{I \cup \ldots \cup VI} f \varphi_i \, dx$$

for $a_Z = a_i$ and thus it is not identical to $f(a_i)$, the right-hand side of the finite difference method.

However, if the *trapezoidal rule*, which is exact for $g \in \mathcal{P}_1$, is applied to approximate the right-hand side according to

$$\int_K g(x)\, dx \approx \frac{1}{3} \mathrm{vol}\,(K) \sum_{i=1}^{3} g(a_i) \tag{2.38}$$

for a triangle K with the vertices a_i, $i = 1, 2, 3$ and with the area $\mathrm{vol}\,(K)$, then

$$\int_I f \varphi_i \, dx \approx \frac{1}{3} \frac{1}{2} h^2 \left(f(a_Z) \cdot 1 + f(a_O) \cdot 0 + f(a_N) \cdot 0 \right) = \frac{1}{6} h^2 f(a_Z).$$

Analogous results are obtained for the other triangles, and thus

$$h^{-2} \int_{I \cup \ldots \cup VI} f \varphi_i \, dx \approx f(a_Z)\,.$$

In summary, we have the following result.

Lemma 2.13 *The finite element method with linear finite elements on a triangulation according to Figure* 2.9 *and with the trapezoidal rule to approximate the right-hand side yields the same discretization as the finite difference method from* (1.7), (1.8).

We now return to the general formulation (2.21)–(2.24). The approach of the Ritz method (2.24), instead of the Galerkin method (2.23), yields an identical approximation because of the following result.

Lemma 2.14 *If a is a symmetric and positive bilinear form and b is a linear form, then the Galerkin method* (2.23) *and the Ritz method* (2.24) *have identical solutions.*

Proof: Apply Lemma 2.3 with V_h instead of V. □

Hence the finite element method is the Galerkin method (and in our problem the Ritz method, too) for an *ansatz space* V_h with the following properties:

- The coefficients have a local interpretation (here as nodal values).

The basis functions have a small support such that:

- the discretization matrix is sparse,
- the entries of the matrix can be assembled locally.

Finally, for the boundary value problem (2.1), (2.2) with the corresponding weak formulation, we consider other ansatz spaces, which to some extent do not have these properties:

(1) In Section 3.2.1, (3.28), we will show that mixed boundary conditions need not be included in the ansatz space. Then we can choose the finite dimensional polynomial space $V_h = \operatorname{span}\left\{1, x, y, xy, x^2, y^2, \ldots\right\}$ for it. But in this case, A_h is a dense matrix and ill-conditioned. Such ansatz spaces yield the *classical* Ritz–Galerkin methods.

(2) Let $V_h = \operatorname{span}\{\varphi_1, \ldots, \varphi_N\}$ and let $\varphi_i \not\equiv 0$ satisfy, for some λ_i,

$$a(\varphi_i, v) = \lambda_i \langle \varphi_i, v \rangle_0 \quad \text{for all } v \in V\,,$$

i.e., the weak formulation of the eigenvalue problem

$$\begin{aligned} -\Delta u &= \lambda u && \text{in } \Omega\,, \\ u &= 0 && \text{on } \partial\Omega\,, \end{aligned}$$

for which eigenvalues $0 < \lambda_1 \leq \lambda_2 \leq \ldots$ and corresponding eigenfunctions φ_i exist such that $\langle \varphi_i, \varphi_j \rangle_0 = \delta_{ij}$ (e.g., see [12, p. 335]). For special domains Ω, (λ_i, φ_i) can be determined explicitly, and

$$(A_h)_{ij} = a(\varphi_j, \varphi_i) = \lambda_j \langle \varphi_j, \varphi_i \rangle_0 = \lambda_j \delta_{ij}$$

is obtained. Thus A_h is a diagonal matrix, and the system of equations $A_h \boldsymbol{\xi} = \boldsymbol{q}_h$ can be solved without too great expense. But this kind of assembling is possible with acceptable costs for special cases only.

(3) The (spectral) *collocation method* consists in the requirement that the equations (2.1), (2.2) be satisfied only at certain distinct points $x_i \in \overline{\Omega}$, called *collocation points,* for a special polynomial space V_h.

The above examples describe Galerkin methods without having the typical properties of a finite element method.

2.3 Stability and Convergence of the Finite Element Method

We consider the general case of a variational equation of the form (2.21) and the Galerkin method (2.23). Here let a be a bilinear form, which is not necessarily symmetric, and let b be a linear form.

Then, if

$$e := u - u_h \, (\in V)$$

denotes the error, the important *error equation*

$$a(e, v) = 0 \quad \text{for all } v \in V_h \tag{2.39}$$

is satisfied. To obtain this equation, it is sufficient to consider equation (2.21) only for $v \in V_h \subset V$ and then to subtract from the result the Galerkin equation (2.23).

If, in addition, a is symmetric and positive definite, i.e.,

$$a(u, v) = a(v, u)\,, \quad a(u, u) \geq 0\,, \quad a(u, u) = 0 \Leftrightarrow u = 0$$

(i.e., a is a scalar product), then the error is orthogonal to the space V_h with respect to the scalar product a.

Therefore, the relation (2.39) is often called the *orthogonality of the error (to the ansatz space).* In general, the element $u_h \in V_h$ with minimal distance to $u \in V$ with respect to the induced norm $\|\cdot\|_a$ is characterized by (2.39):

Lemma 2.15 *Let $V_h \subset V$ be a subspace, let a be a scalar product on V, and let $\|u\|_a := a(u, u)^{1/2}$ be the norm induced by a. Then for $u_h \in V_h$, it follows that*

$$a(u - u_h, v) = 0 \quad \textit{for all } v \in V_h \quad \Leftrightarrow \tag{2.40}$$

$$\|u - u_h\|_a = \min \left\{ \|u - v\|_a \;\middle|\; v \in V_h \right\} . \tag{2.41}$$

Proof: For arbitrary but fixed $u \in V$, let $b(v) := a(u, v)$ for $v \in V_h$. Then b is a linear form on V_h, so (2.40) is a variational formulation on V_h.

According to Lemma 2.14 or Lemma 2.3, this variational formulation has the same solutions as

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

Furthermore, F has the same minima as the functional

$$\begin{aligned} \Big(2F(v) + a(u,u)\Big)^{1/2} &= \Big(a(v,v) - 2a(u,v) + a(u,u)\Big)^{1/2} \\ &= \Big(a(u-v,u-v)\Big)^{1/2} = \|u-v\|_a\,, \end{aligned}$$

because the additional term $a(u,u)$ is a constant. Therefore, F has the same minima as (2.41). □

If an approximation u_h of u is to be sought exclusively in V_h, then the element u_h, determined by the Galerkin method, is the optimal choice with respect to $\|\cdot\|_a$.

A general, not necessarily symmetric, bilinear form a is assumed to satisfy the following conditions, where $\|\cdot\|$ denotes a norm on V:

- a is *continuous* with respect to $\|\cdot\|$; i.e., there exists $M > 0$ such that

$$|a(u,v)| \le M\|u\|\|v\| \quad \text{for all } u, v \in V\,; \tag{2.42}$$

- a is *V-elliptic*; i.e., there exists $\alpha > 0$ such that

$$a(u,u) \ge \alpha\|u\|^2 \quad \text{for } u \in V\,. \tag{2.43}$$

If a is a scalar product, then (2.42) with $M = 1$ and (2.43) (as equality) with $\alpha = 1$ are valid for the induced norm $\|\cdot\| := \|\cdot\|_a$ due to the Cauchy–Schwarz inequality.

The V-ellipticity is an essential condition for the unique existence of a solution of the variational equation (2.21) and of the boundary value problem described by it, which will be presented in more detail in Sections 3.1 and 3.2. It also implies — without further conditions — the stability of the Galerkin approximation.

Lemma 2.16 *The Galerkin solution u_h according to* (2.23) *is* stable *in the following sense:*

$$\|u_h\| \le \frac{1}{\alpha}\|b\| \quad \textit{independently of } h\,, \tag{2.44}$$

where

$$\|b\| := \sup\left\{\frac{|b(v)|}{\|v\|} \;\middle|\; v \in V\,,\ v \neq 0\right\}\,.$$

Proof: In the case $u_h = 0$, there is nothing to prove. Otherwise, from $a(u_h, v) = b(v)$ for all $v \in V_h$, it follows that

$$\alpha \|u_h\|^2 \leq a(u_h, u_h) = b(u_h) \leq \frac{|b(u_h)|}{\|u_h\|} \|u_h\| \leq \|b\| \, \|u_h\|.$$

Dividing this relation by $\alpha \|u_h\|$, we get the assertion. □

Moreover, the approximation property (2.41) holds up to a constant:

Theorem 2.17 (Céa's lemma)
Assume (2.42), (2.43). *Then the following error estimate for the Galerkin solution holds:*

$$\|u - u_h\| \leq \frac{M}{\alpha} \min \left\{ \|u - v\| \;\middle|\; v \in V_h \right\} . \tag{2.45}$$

Proof: If $\|u - u_h\| = 0$, then there is nothing to prove. Otherwise, let $v \in V_h$ be arbitrary. Because of the error equation (2.39) and $u_h - v \in V_h$,

$$a(u - u_h, u_h - v) = 0 \, .$$

Therefore, using (2.43) we have

$$\begin{aligned} \alpha \|u - u_h\|^2 \;&\leq\; a(u - u_h, u - u_h) = a(u - u_h, u - u_h) + a(u - u_h, u_h - v) \\ &=\; a(u - u_h, u - v) \, . \end{aligned}$$

Furthermore, by means of (2.42) we obtain

$$\alpha \|u - u_h\|^2 \leq a(u - u_h, u - v) \leq M \|u - u_h\| \, \|u - v\| \text{ for arbitrary } v \in V_h \, .$$

Thus the assertion follows by division by $\alpha \|u - u_h\|$. □

Therefore also in general, in order to get an asymptotic error estimate in h, it is sufficient to estimate the *best approximation error* of V_h, i.e.,

$$\min \left\{ \|u - v\| \;\middle|\; v \in V_h \right\} .$$

However, this consideration is meaningful only in those cases where M/α is not too large. Section 3.2 shows that this condition is no longer satisfied for convection-dominated problems. Therefore, the Galerkin approach has to be modified, which will be described in Chapter 9.

We want to apply the theory developed up to now to the weak formulation of the boundary value problem (2.1), (2.2) with V according to (2.7) or (2.20) and V_h according to (2.27). According to (2.4) the bilinear form a and the linear form b read as

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

In order to guarantee that the linear form b is well-defined on V, it is sufficient to assume that the right-hand side f of the boundary value problem belongs to $L^2(\Omega)$.

Since a is a scalar product on V,

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

is an appropriate norm. Alternatively, the norm introduced in (2.19) for $V = H_0^1(\Omega)$ can be taken as

$$\|u\|_1 = \left(\int_\Omega |u(x)|^2 \, dx + \int_\Omega |\nabla u(x)|^2 \, dx \right)^{1/2} .$$

In the latter case, the question arises whether the conditions (2.42) and (2.43) are still satisfied. Indeed,

$$|a(u,v)| \le \|u\|_a \|v\|_a \le \|u\|_1 \|v\|_1 \quad \text{for all } u, v \in V .$$

The first inequality follows from the Cauchy–Schwarz inequality for the scalar product a, and the second inequality follows from the trivial estimate

$$\|u\|_a = \left(\int_\Omega |\nabla u(x)|^2 \, dx \right)^{1/2} \le \|u\|_1 \quad \text{for all } u \in V .$$

Thus a is continuous with respect to $\|\cdot\|_1$ with $M = 1$.

The V-ellipticity of a, i.e., the property

$$a(u,u) = \|u\|_a^2 \ge \alpha \|u\|_1^2 \quad \text{for some } \alpha > 0 \text{ and all } u \in V ,$$

is not valid in general for $V = H^1(\Omega)$. However, in the present situation of $V = H_0^1(\Omega)$ it is valid because of the incorporation of the boundary condition into the definition of V :

Theorem 2.18 (Poincaré) *Let $\Omega \subset \mathbb{R}^n$ be open and bounded. Then a constant $C > 0$ exists (depending on Ω) such that*

$$\|u\|_0 \le C \left(\int_\Omega |\nabla u(x)|^2 \, dx \right)^{1/2} \quad \textit{for all } u \in H_0^1(\Omega) .$$

Proof: Cf. [13]. For a special case, see Exercise 2.5. □

Thus (2.43) is satisfied, for instance with

$$\alpha = \frac{1}{1 + C^2} ,$$

(see also (3.26) below) and thus in particular

$$\alpha \|u\|_1^2 \le a(u,u) = \|u\|_a^2 \le \|u\|_1^2 \quad \text{for all } u \in V , \tag{2.46}$$

i.e., the norms $\|\cdot\|_1$ and $\|\cdot\|_a$ are equivalent on $V = H_0^1(\Omega)$ and therefore they generate the same convergence concept:

$$u_h \to u \text{ with respect to } \|\cdot\|_1 \Leftrightarrow \|u_h - u\|_1 \to 0$$

$$\Leftrightarrow \|u_h - u\|_a \to 0 \Leftrightarrow u_h \to u \text{ with respect to } \|\cdot\|_a .$$

In summary the estimate (2.45) holds for $\|\cdot\| = \|\cdot\|_1$ with the constant $1/\alpha$.

Because of the Cauchy–Schwarz inequality for the scalar product on $L^2(\Omega)$ and

$$b(v) = \int_\Omega f(x)v(x)\,dx ,$$

i.e., $|b(v)| \le \|f\|_0 \|v\|_0 \le \|f\|_0 \|v\|_1$, and thus $\|b\| \le \|f\|_0$, the stability estimate (2.44) for a right-hand side $f \in L^2(\Omega)$ takes the particular form

$$\|u_h\|_1 \le \frac{1}{\alpha}\|f\|_0 .$$

Up to now, our considerations have been independent of the special form of V_h. Now we make use of the choice of V_h according to (2.27). In order to obtain an estimate of the approximation error of V_h, it is sufficient to estimate the term $\|u - \bar{v}\|$ for some special element $\bar{v} \in V_h$. For this element $\bar{v} \in V_h$, we choose the interpolant $I_h(u)$, where

$$\begin{aligned} I_h : \left\{ u \in C(\bar{\Omega}) \;\middle|\; u = 0 \text{ on } \partial\Omega \right\} &\to V_h , \\ u &\mapsto I_h(u) \text{ with } I_h(u)(a_i) = u(a_i) . \end{aligned} \tag{2.47}$$

This interpolant exists and is unique (Lemma 2.10). Obviously,

$$\min\left\{ \|u - v\|_1 \;\middle|\; v \in V_h \right\} \le \|u - I_h(u)\|_1 \quad \text{for } u \in C(\bar{\Omega}) \text{ and } u = 0 \text{ on } \partial\Omega .$$

If the weak solution u possesses weak derivatives of second order, then for certain sufficiently fine triangulations $\mathcal{T}_h$, i.e., $0 < h \le \bar{h}$ for some $\bar{h} > 0$, an estimate of the type

$$\|u - I_h(u)\|_1 \le Ch \tag{2.48}$$

holds, where C depends on u but is independent of h (cf. (3.88)). The proof of this estimate will be explained in Section 3.4, where also sufficient conditions on the family of triangulations $(\mathcal{T}_h)_h$ will be specified.

Exercises

2.4 Let $a(u,v) := \int_0^1 x^2 u'v'dx$ for arbitrary $u, v \in H_0^1(0,1)$.

(a) Show that there is no constant $C_1 > 0$ such that the inequality

$$a(u,u) \ge C_1 \int_0^1 (u')^2\,dx \quad \text{for all } u \in H_0^1(0,1)$$

is valid.

(b) Now let $\mathcal{T}_h := \{(x_{i-1}, x_i)\}_{i=1}^N$, $N \in \mathbb{N}$, be an equidistant partition of $(0,1)$ with the parameter $h = 1/N$ and $V_h := \operatorname{span}\{\varphi_i\}_{i=1}^{N-1}$, where

$$\varphi_i(x) := \begin{cases} (x - x_{i-1})/h & \text{in } (x_{i-1}, x_i)\,, \\ (x_{i+1} - x)/h & \text{in } (x_i, x_{i+1})\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

Does there exist a constant $C_2 > 0$ with

$$a(u_h, u_h) \geq C_2 \int_0^1 (u_h')^2 \, dx \quad \text{for all } u_h \in V_h\,?$$

2.5

(a) For $\Omega := (\alpha, \beta) \times (\gamma, \delta)$ and V according to (2.7), prove the *inequality of Poincaré:* There exists a positive constant C with

$$\|u\|_0 \leq C \|u\|_a \quad \text{for all } u \in V\,.$$

Hint: Start with the relation $u(x,y) = \int_\alpha^x \partial_x u(s,y)\, ds$.

(b) For $\Omega := (\alpha, \beta)$ and $v \in C([\alpha, \beta])$ with a piecewise continuous derivative v' and $v(\gamma) = 0$ for some $\gamma \in [\alpha, \beta]$, show that

$$\|v\|_0 \leq (\beta - \alpha) \|v'\|_0\,.$$

2.6 Let $\Omega := (0,1) \times (0,1)$. Given $f \in C(\overline{\Omega})$, discretize the boundary value problem $-\Delta u = f$ in Ω, $u = 0$ on $\partial\Omega$, by means of the usual five-point difference stencil as well as by means of the finite element method with linear elements. A quadratic grid as well as the corresponding Friedrichs–Keller triangulation will be used.

Prove the following stability estimates for the matrix of the linear system of equations:

$$\text{(a) } \|A_h^{-1}\|_\infty \leq \frac{1}{8}\,, \quad \text{(b) } \|A_h^{-1}\|_2 \leq \frac{1}{16}\,, \quad \text{(c) } \|A_h^{-1}\|_0 \leq 1\,,$$

where $\|\cdot\|_\infty, \|\cdot\|_2$ denote the maximum row sum norm and the spectral norm of a matrix, respectively, and $\|A_h^{-1}\|_0 := \sup_{v_h \in V_h} \|v_h\|_0^2 / \|v_h\|_a^2$ with $\|v_h\|_a^2 := \int_\Omega |\nabla v_h|^2\, dx$.

Comment: The constant in (c) is not optimal.

2.7 Let Ω be a domain with polygonal boundary and let $\mathcal{T}_h$ be a conforming triangulation of Ω. The nodes a_i of the triangulation are enumerated from 1 to M.

Let the triangulation satisfy the following assumption: There exist constants $C_1, C_2 > 0$ such that for all triangles $K \in \mathcal{T}_h$ the relation

$$C_1 h^2 \leq \operatorname{vol}(K) \leq C_2 h^2$$

is satisfied. h denotes the maximum of the diameters of all elements of $\mathcal{T}_h$.

(a) Show the equivalence of the following norms for $u_h \in V_h$ in the space V_h of continuous, piecewise linear functions over Ω :

$$\|u_h\|_0 := \left\{ \int_\Omega |u_h|^2 \, dx \right\}^{1/2}, \quad \|u_h\|_{0,h} := h \left\{ \sum_{i=1}^{M} u_h^2(a_i) \right\}^{1/2}.$$

(b) Consider the special case $\Omega := (0,1)\times(0,1)$ with the Friedrichs–Keller triangulation as well as the subspace $V_h \cap H_0^1(\Omega)$ and find "as good as possible" constants in the corresponding equivalence estimate.

2.4 The Implementation of the Finite Element Method: Part 1

In this section we will consider some aspects of the implementation of the finite element method using linear ansatz functions on triangles for the model boundary value problem (1.1), (1.2) on a polygonally bounded domain $\Omega \subset \mathbb{R}^2$. The case of inhomogeneous Dirichlet boundary conditions will be treated also to a certain extent as far as it is possible up to now.

2.4.1 Preprocessor

The main task of the preprocessor is to determine the triangulation.

An input file might have the following format:

Let the number of variables (including also the boundary nodes for Dirichlet boundary conditions) be M. We generate the following list:

x-coordinate of node 1	y-coordinate of node 1
...	...
x-coordinate of node M	y-coordinate of node M

Let the number of (triangular) elements be N. These elements will be listed in the *element-node table*. Here, every element is characterized by the indices of the nodes corresponding to this element in a well-defined order (e.g., counterclockwise); cf. Figure 2.11.

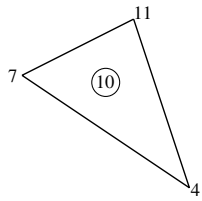

Figure 2.11. Element no. 10 with nodes nos. 4, 11, 7.

For example, the 10th row of the element-node table contains the entry

4 11 7

Usually, a triangulation is generated by a triangulation algorithm. A short overview on methods for the grid generation will be given in Section 4.1. One of the simplest versions of a grid generation algorithm has the following structure (cf. Figure 2.12):

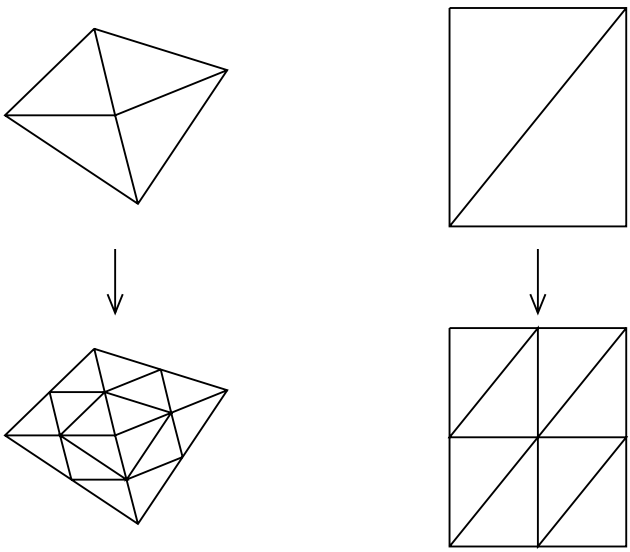

Figure 2.12. Refinement by quartering.

Prescribe a coarse triangulation (according to the above format) and refine this triangulation (repeatedly) by subdividing a triangle into 4 congruent triangles by connecting the midpoints of the edges with straight lines.

If this uniform refinement is done globally, i.e., for all triangles of the coarse grid, then triangles are created that have the same interior angles as the elements of the coarse triangulation. Thus the quality of the triangulation, indicated, for example, by the ratios of the diameters of an element and of its inscribed circle (see Definition 3.28), does not change. However, if the subdivision is performed only locally, the resulting triangulation is no longer admissible, in general. Such an inadmissible triangulation can be corrected by bisection of the corresponding neighbouring (unrefined) triangles. But this implies that some of the interior angles are bisected and consequently, the quality of the triangulation becomes poorer if the bisection step is performed too frequently. The following algorithm circumvents the depicted problem. It is due to R. Bank and is implemented, for example, in the PLTMG code (see [4]).

A Possible Refinement Algorithm
Let a (uniform) triangulation $\mathcal{T}$ be given (e.g., by repeated uniform refinement of a coarse triangulation). The edges of this triangulation are called *red edges*.

(1) Subdivide the edges according to a certain local refinement criterion (introduction of new nodes) by successive bisection (cf. Figure 2.13).

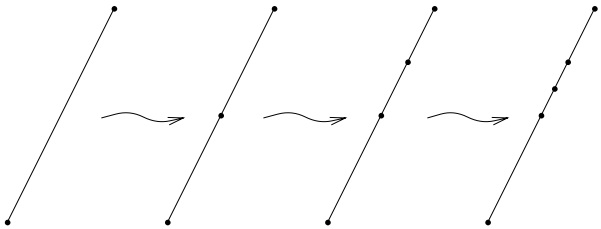

Figure 2.13. New nodes on edges.

(2) If a triangle $K \in \mathcal{T}$ has on its edges in addition to the vertices two or more nodes, then subdivide K into four congruent triangles. Iterate over step 2 (cf. Figure 2.14).

(3) Subdivide the triangles with nodes at the midpoints of the edges into 2 triangles by bisection. This step introduces the so-called *green edges*.

(4) If the refinement is to be continued, first remove the green edges.

2.4.2 Assembling

Denote by $\varphi_1, \ldots, \varphi_M$ the global basis functions. Then the stiffness matrix A_h has the following entries:

$$(A_h)_{ij} = \int_\Omega \nabla\varphi_j \cdot \nabla\varphi_i \, dx = \sum_{m=1}^{N} A_{ij}^{(m)}$$

with

$$A_{ij}^{(m)} = \int_{K_m} \nabla\varphi_j \cdot \nabla\varphi_i \, dx \,.$$

Let $a_1, \ldots, a_M$ denote the nodes of the triangulation. Because of the implication

$$A_{ij}^{(m)} \neq 0 \;\Rightarrow\; a_i, a_j \in K_m$$

(cf. (2.37)), the element K_m yields nonzero contributions for $A_{ij}^{(m)}$ only if $a_i, a_j \in K_m$ at best. Such nonzero contributions are called *element entries* of A_h. They add up to the *entries* of A_h.

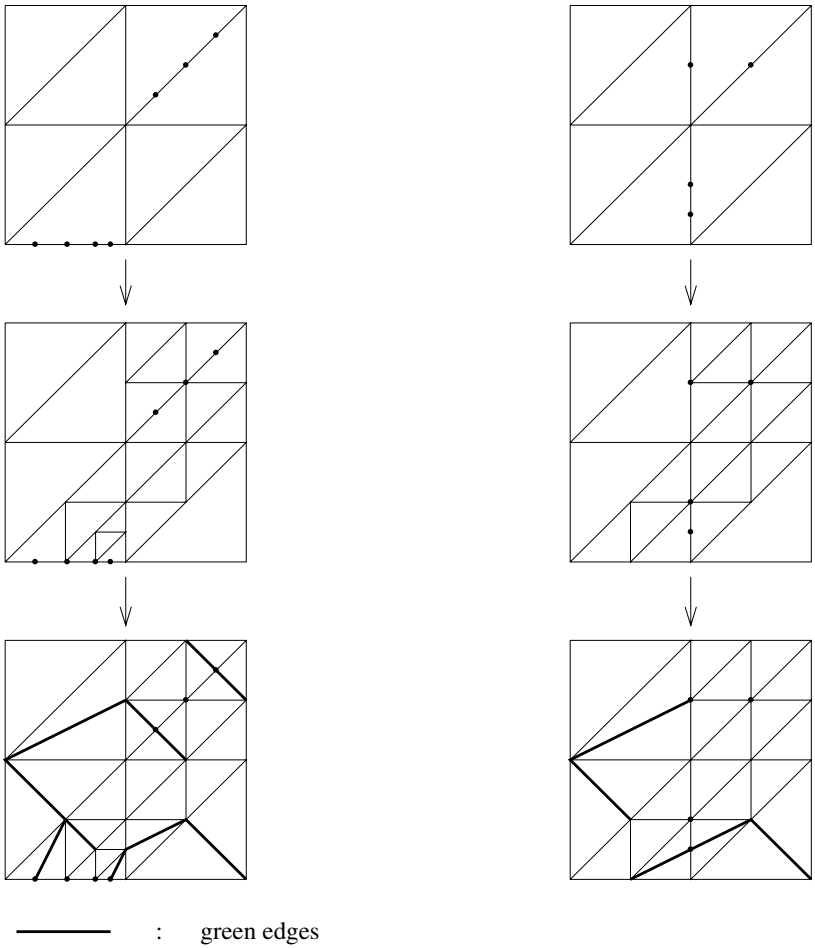

Figure 2.14. Two refinement sequences.

In Example 2.12 we explained a node-based assembling of the stiffness matrix. In contrast to this and on the basis of the above observations, in the following we will perform an *element-based assembling* of the stiffness matrix.

To assemble the entries of $A^{(m)}$, we will start from a local numbering (cf. Figure 2.15) of the nodes by assigning the local numbers 1, 2, 3 to the global node numbers r_1, r_2, r_3 (numbered counterclockwise). In contrast to the usual notation adopted in this book, here indices of vectors according to the local numbering are included in parentheses and written as superscripts.

Figure 2.15. Global (left) and local numbering.

Thus in fact, we generate

$$\left(A^{(m)}_{r_i r_j}\right)_{i,j=1,2,3} \quad \text{as} \quad \left(\tilde{A}^{(m)}_{ij}\right)_{i,j=1,2,3}.$$

To do this, we first perform a transformation of K_m onto some reference element and then we evaluate the integral on this element exactly.

Hence the entry of the *element stiffness matrix* reads as

$$\tilde{A}^{(m)}_{ij} = \int_{K_m} \nabla\varphi_{r_j} \cdot \nabla\varphi_{r_i}\, dx\,.$$

The reference element $\hat{K}$ is transformed onto the global element K_m by means of the relation $F(\hat{x}) = B\hat{x} + d$, therefore

$$D_{\hat{x}} u(F(\hat{x})) = D_x u(F(\hat{x}))\, D_{\hat{x}} F(\hat{x}) = D_x u(F(\hat{x}))\, B\,,$$

where $D_x u$ denotes the row vector $(\partial_1 u, \partial_2 u)$, i.e., the corresponding differential operator. Using the more standard notation in terms of gradients and taking into consideration the relation $B^{-T} := (B^{-1})^T$, we obtain

$$\nabla_x u\,(F(\hat{x})) = B^{-T} \nabla_{\hat{x}}\,(u\,(F(\hat{x}))) \tag{2.49}$$

and thus

$$\begin{aligned}
\tilde{A}^{(m)}_{ij} &= \int_{\hat{K}} \nabla_x \varphi_{r_j}\,(F(\hat{x})) \cdot \nabla_x \varphi_{r_i}\,(F(\hat{x}))\, |\det(DF(\hat{x}))|\; d\hat{x} \\
&= \int_{\hat{K}} B^{-T} \nabla_{\hat{x}}\left(\varphi_{r_j}\,(F(\hat{x}))\right) \cdot B^{-T} \nabla_{\hat{x}}\,(\varphi_{r_i}\,(F(\hat{x})))\, |\det(B)|\; d\hat{x} \\
&= \int_{\hat{K}} B^{-T} \nabla_{\hat{x}} \hat{\varphi}_{r_j}(\hat{x}) \cdot B^{-T} \nabla_{\hat{x}} \hat{\varphi}_{r_i}(\hat{x})\, |\det(B)|\; d\hat{x} \qquad (2.50) \\
&= \int_{\hat{K}} B^{-T} \nabla_{\hat{x}} N_j(\hat{x}) \cdot B^{-T} \nabla_{\hat{x}} N_i(\hat{x})\, |\det(B)|\; d\hat{x}\,,
\end{aligned}$$

where the transformed basis functions $\hat{\varphi}_{r_i}$, $\hat{\varphi}(\hat{x}) := \varphi(F(\hat{x}))$ coincide with the local basis functions on $\hat{K}$, i.e., with the shape functions N_i:

$$\hat{\varphi}_{r_i}(\hat{x}) = N_i(\hat{x}) \quad \text{for } \hat{x} \in \hat{K}\,.$$

The shape functions N_i have been defined in (2.29) (where (x, y) there must be replaced by $(\hat{x}_1, \hat{x}_2)$ here) for the standard reference element defined there.

Introducing the matrix $C := \left(B^{-1}\right)\left(B^{-1}\right)^T = \left(B^T B\right)^{-1}$, we can write

$$\tilde{A}_{ij}^{(m)} = \int_{\hat{K}} C\,\nabla_{\hat{x}} N_j(\hat{x}) \cdot \nabla_{\hat{x}} N_i(\hat{x})\,|\det(B)|\,d\hat{x}\,. \tag{2.51}$$

Denoting the matrix B by $B = \left(b^{(1)}, b^{(2)}\right)$, then it follows that

$$C = \begin{pmatrix} b^{(1)}\cdot b^{(1)} & b^{(1)}\cdot b^{(2)} \\ b^{(1)}\cdot b^{(2)} & b^{(2)}\cdot b^{(2)} \end{pmatrix}^{-1} = \frac{1}{\det(B)^2}\begin{pmatrix} b^{(2)}\cdot b^{(2)} & -b^{(1)}\cdot b^{(2)} \\ -b^{(1)}\cdot b^{(2)} & b^{(1)}\cdot b^{(1)} \end{pmatrix}$$

because $\det(B^T B) = \det(B)^2$. The previous considerations can be easily extended to the computation of the stiffness matrices of more general differential operators like

$$\int_\Omega K(x)\nabla\varphi_j(x)\cdot\nabla\varphi_i(x)\,dx$$

(cf. Section 3.5). For the standard reference element, which we use from now on, we have $b^{(1)} = a^{(2)} - a^{(1)}$, $b^{(2)} = a^{(3)} - a^{(1)}$. Here $a^{(i)}$, $i = 1, 2, 3$, are the locally numbered nodes of K interpreted as vectors of $\mathbb{R}^2$.

From now on we make also use of the special form of the stiffness matrix and obtain

$$\begin{aligned} \tilde{A}_{ij}^{(m)} &= \gamma_1 \int_{\hat{K}} \partial_{\hat{x}_1} N_j\,\partial_{\hat{x}_1} N_i\,d\hat{x} \\ &\quad + \gamma_2 \int_{\hat{K}} \partial_{\hat{x}_1} N_j\,\partial_{\hat{x}_2} N_i + \partial_{\hat{x}_2} N_j\,\partial_{\hat{x}_1} N_i\,d\hat{x} \\ &\quad + \gamma_3 \int_{\hat{K}} \partial_{\hat{x}_2} N_j\,\partial_{\hat{x}_2} N_i\,d\hat{x} \end{aligned} \tag{2.52}$$

with

$$\begin{aligned} \gamma_1 &:= c_{11}|\det(B)| = \frac{1}{|\det(B)|}\left(a^{(3)} - a^{(1)}\right)\cdot\left(a^{(3)} - a^{(1)}\right), \\ \gamma_2 &:= c_{12}|\det(B)| = -\frac{1}{|\det(B)|}\left(a^{(2)} - a^{(1)}\right)\cdot\left(a^{(3)} - a^{(1)}\right), \\ \gamma_3 &:= c_{22}|\det(B)| = \frac{1}{|\det(B)|}\left(a^{(2)} - a^{(1)}\right)\cdot\left(a^{(2)} - a^{(1)}\right). \end{aligned}$$

In the implementation it is advisable to compute the values γ_i just once from the local geometrical information given in the form of the vertices $a^{(i)} = a_{r_i}$ and to store them permanently.

Thus we obtain for the local stiffness matrix

$$\tilde{A}^{(m)} = \gamma_1 S_1 + \gamma_2 S_2 + \gamma_3 S_3 \tag{2.53}$$

with

$$S_1 := \left(\int_{\hat{K}} \partial_{\hat{x}_1} N_j \partial_{\hat{x}_1} N_i\,d\hat{x}\right)_{ij},$$

$$S_2 := \left(\int_{\hat{K}} \partial_{\hat{x}_1} N_j \partial_{\hat{x}_2} N_i + \partial_{\hat{x}_2} N_j \partial_{\hat{x}_1} N_i \, d\hat{x} \right)_{ij},$$
$$S_3 := \left(\int_{\hat{K}} \partial_{\hat{x}_2} N_j \partial_{\hat{x}_2} N_i \, d\hat{x} \right)_{ij}.$$

An explicit computation of the matrices S_i is possible because the integrands are constant, and also these matrices can be stored permanently:

$$S_1 = \frac{1}{2}\begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \; S_2 = \frac{1}{2}\begin{pmatrix} 2 & -1 & -1 \\ -1 & 0 & 1 \\ -1 & 1 & 0 \end{pmatrix}, \; S_3 = \frac{1}{2}\begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix}.$$

The right-hand side $(\boldsymbol{q}_h)_i = \int_\Omega f(x)\varphi_i(x)\,dx$ can be treated in a similar manner:

$$(\boldsymbol{q}_h)_i = \sum_{m=1}^{N} \left(\boldsymbol{q}^{(m)}\right)_i$$

with

$$\left(\boldsymbol{q}^{(m)}\right)_i = \int_{K_m} f(x)\varphi_i(x)\,dx \quad (\neq 0 \;\Rightarrow\; a_i \in K_m).$$

Again, we transform the global numbering $\left(q_{r_i}^{(m)}\right)_{i=1,2,3}$ for the triangle $K_m = \operatorname{conv}\{a_{r_1}, a_{r_2}, a_{r_3}\}$ into the local numbering $\left(\tilde{q}_i^{(m)}\right)_{i=1,2,3}$. Analogously to the determination of the entries of the stiffness matrix, we have

$$\begin{aligned} \tilde{q}_i^{(m)} &= \int_{\hat{K}} f\left(F(\hat{x})\right) \varphi_{r_i}\left(F(\hat{x})\right) |\det(B)|\, d\hat{x} \\ &= \int_{\hat{K}} \hat{f}(\hat{x})\, N_i(\hat{x})\, |\det(B)|\, d\hat{x}, \end{aligned}$$

where $\hat{f}(\hat{x}) := f(F(\hat{x}))$ for $\hat{x} \in \hat{K}$.

In general, this integral cannot be evaluated exactly. Therefore, it has to be approximated by a quadrature rule.

A *quadrature rule* for $\int_{\hat{K}} g(\hat{x})\,d\hat{x}$ is of the type

$$\sum_{k=1}^{R} \omega_k \, g\left(\hat{b}^{(k)}\right)$$

with certain *weights* ω_k and *quadrature points* $\hat{b}^{(k)}$. As an example, we take the *trapezoidal rule* (cf. (2.38)), where

$$\hat{b}^{(1)} = \hat{a}_1 = (0,0), \quad \hat{b}^{(2)} = \hat{a}_2 = (1,0), \quad \hat{b}^{(3)} = \hat{a}_3 = (0,1),$$
$$\omega_k = \tfrac{1}{6}, \quad k = 1,2,3.$$

Thus for arbitrary but fixed quadrature rules, we have

$$\tilde{q}_i^{(m)} \approx \sum_{k=1}^{R} \omega_k \, \hat{f}\big(\hat{b}^{(k)}\big) \, N_i\big(\hat{b}^{(k)}\big) \, |\det(B)| \, . \tag{2.54}$$

Of course, the application of different quadrature rules on different elements is possible, too. The values $N_i\big(\hat{b}^{(k)}\big)$, $i = 1, 2, 3$, $k = 1, \ldots, R$, should be evaluated just once and should be stored. The discussion on the use of quadrature rules will be continued in Sections 3.5.2 and 3.6.

In summary, the following algorithm provides the assembling of the stiffness matrix and the right-hand side:

Loop over all elements $m = 1, \ldots, N$:

- Allocating a local numbering to the nodes based on the element-node table: $1 \mapsto r_1$, $2 \mapsto r_2$, $3 \mapsto r_3$.
- Assembling of the element stiffness matrix $\tilde{A}^{(m)}$ according to (2.51) or (2.53).
  Assembling of the right-hand side according to (2.54).
- Loop over $i, j = 1, 2, 3$:

$$\begin{aligned} (A_h)_{r_i r_j} &:= (A_h)_{r_i r_j} + \tilde{A}_{ij}^{(m)} \, , \\ (\boldsymbol{q}_h)_{r_i} &:= (\boldsymbol{q}_h)_{r_i} + \tilde{q}_i^{(m)} \, . \end{aligned}$$

For the sake of efficiency of this algorithm, it is necessary to adjust the memory structure to the particular situation; we will see how this can be done in Section 2.5.

2.4.3 Realization of Dirichlet Boundary Conditions: Part 1

Nodes where a Dirichlet boundary condition is prescribed must be labeled specially, here, for instance, by the convention $M = M_1 + M_2$, where the nodes numbered from $M_1 + 1$ to M correspond to the Dirichlet boundary nodes. In more general cases, other realizations are to be preferred.

In the first step of assembling of stiffness matrix and the load vector, the Dirichlet nodes are treated like all the other ones. After this, the Dirichlet nodes are considered separately. If such a node has the number j, the boundary condition is included by the following procedure:

Replace the jth row and the jth column (for conservation of the symmetry) of A_h by the jth unit vector and $(\boldsymbol{q}_h)_j$ by $g(a_j)$, if $u(x) = g(x)$ is prescribed for $x \in \partial\Omega$. If the jth column is replaced by the unit vector, the right-hand side $(\boldsymbol{q}_h)_i$ for $i \neq j$ must be modified to $(\boldsymbol{q}_h)_i - (A_h)_{ij} g(a_j)$. In other words, the contributions caused by the Dirichlet boundary condition are included into the right-hand side. This is exactly the elimination that led to the form (1.10), (1.11) in Chapter 1.

2.5 Solving Sparse Systems of Linear Equations by Direct Methods

Let A be an $M \times M$ matrix. Given a vector $\mathbf{q} \in \mathbb{R}^M$, we consider the system of linear equations

$$A\boldsymbol{\xi} = \mathbf{q}\,.$$

The matrices arising from the finite element discretization are *sparse*; i.e., they have a bounded number of nonzero entries per row independent of the dimension of the system of equations. For the simple example of Section 2.2, this bound is determined by the number of neighbouring nodes (see (2.37)). Methods for solving systems of equations should take advantage of the sparse structure. For iterative methods, which will be examined in Chapter 5, this is easier to reach than for direct methods. Therefore, the importance of direct methods has decreased. Nevertheless, in adapted form and for small or medium size problems, they are still the method of choice.

Elimination without Pivoting using Band Structure
In the general case, where the matrix A is assumed only to be nonsingular, there exist $M \times M$ matrices P, L, U such that

$$PA = LU\,.$$

Here P is a permutation matrix, L is a scaled lower triangular matrix, and U is an upper triangular matrix; i.e., they have the form

$$L = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ l_{ij} & & 1 \end{pmatrix}, \qquad U = \begin{pmatrix} u_{11} & & u_{ij} \\ & \ddots & \\ 0 & & u_{MM} \end{pmatrix}.$$

This decomposition corresponds to the Gaussian elimination method with pivoting. The method is very easy and has favourable properties with respect to the sparse structure, if pivoting is not necessary (i.e., $P = I$, $A = LU$). Then the matrix A is called *LU factorizable*.

Denote by A_k the leading principal submatrix of A of dimension $k \times k$, i.e.,

$$A_k := \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix},$$

and suppose that it already has been factorized as $A_k = L_k U_k$. This is obviously possible for $k = 1$: $A_1 = (a_{11}) = (1)(a_{11})$. The matrix A_{k+1} can be represented in the form of a block matrix

$$A_{k+1} = \left(\begin{array}{c|c} A_k & b \\ \hline c^T & d \end{array}\right)$$

with $b, c \in \mathbb{R}^k$, $d \in \mathbb{R}$.

Using the ansatz

$$L_{k+1} = \left(\begin{array}{c|c} L_k & 0 \\ \hline l^T & 1 \end{array}\right), \qquad U_{k+1} = \left(\begin{array}{c|c} U_k & u \\ \hline 0 & s \end{array}\right)$$

with unknown vectors $u, l \in \mathbb{R}^k$ and $s \in \mathbb{R}$, it follows that

$$A_{k+1} = L_{k+1}U_{k+1} \iff L_k u = b\,,\ U_k^T l = c\,,\ l^T u + s = d\,. \tag{2.55}$$

From this, we have the following result:

Let A be nonsingular. Then lower and upper triangular matrices L, U exist with $A = LU$ if and only if A_k is nonsingular for all $1 \le k \le M$. For this case, L and U are determined uniquely. (2.56)

Furthermore, from (2.55) we have the following important consequences: If the first l components of the vector b are equal to zero, then this is valid for the vector u, too:

$$\text{If } b = \begin{pmatrix} 0 \\ \beta \end{pmatrix}, \text{ then } u \text{ also has the structure } u = \begin{pmatrix} 0 \\ \varrho \end{pmatrix}.$$

Similarly,

$$c = \begin{pmatrix} 0 \\ \gamma \end{pmatrix} \text{ implies the structure } l = \begin{pmatrix} 0 \\ \lambda \end{pmatrix}.$$

For example, if the matrix A has a structure as shown in Figure 2.16, then the zeros outside of the surrounded entries are preserved after the LU factorization. Before we introduce appropriate definitions to generalize these results, we want to consider the special case of symmetric matrices.

$$A = \begin{pmatrix} * & 0 & * & 0 & 0 \\ 0 & * & * & 0 & * \\ * & * & * & * & * \\ 0 & 0 & * & * & 0 \\ 0 & * & * & 0 & * \end{pmatrix}$$

Figure 2.16. Profile of a matrix.

If A is as before nonsingular and LU factorizable, then $U = DL^T$ with a diagonal matrix $D = \operatorname{diag}(d_i)$, and therefore

$$A = LDL^T\,.$$

This is true because A has the form $A = LD\tilde{U}$, where the upper triangular matrix $\tilde{U}$ satisfies the scaling condition $\tilde{u}_{ii} = 1$ for all $i = 1, \ldots, M$. Such a factorization is unique, and thus

$$A = A^T \text{ implies } L^T = \tilde{U}\,,\ \text{therefore } A = LDL^T\,.$$

If in particular A is symmetric and positive definite, then also $d_i > 0$ is valid. Thus exactly one matrix $\tilde{L}$ of the form

$$\tilde{L} = \begin{pmatrix} l_{11} & & 0 \\ & \ddots & \\ l_{ij} & & l_{MM} \end{pmatrix} \quad \text{with } l_{ii} > 0 \quad \text{for all } i$$

exists such that

$$A = \tilde{L}\tilde{L}^T, \quad \text{the so-called } \textit{Cholesky decomposition.}$$

We have

$$\tilde{L}_{\text{Chol}} = L_{\text{Gauss}}\sqrt{D}\,, \quad \text{where} \quad \sqrt{D} := \operatorname{diag}\left(\sqrt{d_i}\right).$$

This shows that the Cholesky method for the determination of the Cholesky factor $\tilde{L}$ also preserves certain zeros of A in the same way as the Gaussian elimination without pivoting.

In what follows, we want to specify the set of zeros that is preserved by Gaussian elimination without pivoting. We will not consider a symmetric matrix; but for the sake of simplicity we will consider a matrix with a symmetric distribution of its entries.

Definition 2.19 Let $A \in \mathbb{R}^{M\times M}$ be a matrix such that $a_{ii} \neq 0$ for $i = 1, \ldots, M$ and

$$a_{ij} \neq 0 \quad \text{if and only if} \quad a_{ji} \neq 0 \quad \text{for all } i, j = 1, \ldots, M\,. \tag{2.57}$$

We define, for $i = 1, \ldots, M$,

$$f_i(A) := \min\left\{ j \;\middle|\; a_{ij} \neq 0\,,\ 1 \leq j \leq i \right\}.$$

Then

$$m_i(A) := i - f_i(A)$$

is called the *ith (left-hand side) row bandwidth* of A.

The *bandwidth* of a matrix A that satisfies (2.57) is the number

$$m(A) := \max_{1\leq i\leq M} m_i(A) = \max\left\{ i - j \;\middle|\; a_{ij} \neq 0\,,\ 1 \leq j \leq i \leq M \right\}.$$

The *band* of the matrix A is

$$B(A) := \left\{ (i,j), (j,i) \;\middle|\; i - m(A) \leq j \leq i\,,\ 1 \leq i \leq M \right\}.$$

The set

$$\operatorname{Env}(A) := \left\{ (i,j), (j,i) \;\middle|\; f_i(A) \leq j \leq i\,,\ 1 \leq i \leq M \right\}$$

is called the *hull* or *envelope* of A. The number

$$p(A) := M + 2\sum_{i=1}^{M} m_i(A)$$

is called the *profile* of A.

The profile is the number of elements of $\mathrm{Env}(A)$.

For the matrix A in Figure 2.16 we have $(m_1(A), \ldots, m_5(A)) = (0, 0, 2, 1, 3)$, $m(A) = 3$, and $p(A) = 17$.

Summarizing the above considerations, we have proved the following theorem:

Theorem 2.20 *Let A be a matrix with the symmetric structure* (2.57). *Then the Cholesky method or the Gaussian elimination without pivoting preserves the hull and in particular the bandwidth.*

The hull may contain zeros that will be replaced by (nonzero) entries during the decomposition process. Therefore, in order to keep this *fill-in* small, the profile should be as small as possible.

Furthermore, in order to exploit the matrix structure for an efficient assembling and storage, this structure (or some estimate of it) should be known in advance, before the computation of the matrix entries is started.

For example, if A is a stiffness matrix with the entries

$$a_{ij} = a(\varphi_j, \varphi_i) = \int_\Omega \nabla\varphi_j \cdot \nabla\varphi_i \, dx \,,$$

then the property

$$a_{ij} \neq 0 \quad \Rightarrow \quad a_i, a_j \text{ are neighbouring nodes}$$

can be used for the definition of an (eventually too large) symmetric matrix structure. This is also valid for the case of a nonsymmetric bilinear form and thus a nonsymmetric stiffness matrix. Also in this case, the definition of $f_i(A)$ can be replaced by

$$f_i(A) := \min \left\{ j \;\middle|\; 1 \leq j \leq i \,,\ j \text{ is a neighbouring node of } i \right\}.$$

Since the characterization (2.56) of the possibility of the Gaussian elimination without pivoting cannot be checked directly, we have to specify sufficient conditions. Examples for such conditions are the following (see [34]):

- A is symmetric and positive definite,
- A is an M-matrix.

  Sufficient conditions for this property were given in (1.32) and $(1.32)^*$. In Section 3.9, geometrical conditions for the family of triangulations $(\mathcal{T}_h)_h$ will be derived that guarantee that the finite element discretization considered here creates an M-matrix.

Data Structures

For sparse matrices, it is appropriate to store only the components within the band or the hull. A symmetric matrix $A \in \mathbb{R}^{M \times M}$ with bandwidth m can be stored in $M(m+1)$ memory positions. By means of the index

conversion $a_{ik} \rightsquigarrow b_{i,k-i+m+1}$ for $k \le i$, the matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1,m+1} & & & \\ a_{21} & a_{22} & \cdots & \vdots & \ddots & 0 & \\ \vdots & \vdots & \ddots & \vdots & \ddots & \ddots & \\ a_{m+1,1} & a_{m+1,2} & \cdots & a_{m+1,m+1} & \ddots & \ddots & \ddots \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ & 0 & \ddots & \ddots & \ddots & \ddots & \ddots \\ & & & a_{M,M-m} & \cdots & a_{M,M-1} & a_{M,M} \end{pmatrix} \in \mathbb{R}^{M\times M}$$

is mapped to the matrix

$$B = \begin{pmatrix} 0 & \cdots & \cdots & 0 & a_{11} \\ 0 & \cdots & 0 & a_{21} & a_{22} \\ \vdots & & & \vdots & \vdots \\ 0 & a_{m,1} & \cdots & \cdots & a_{m,m} \\ a_{m+1,1} & \cdots & \cdots & a_{m+1,m} & a_{m+1,m+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{M,M-m} & \cdots & \cdots & a_{M,M-1} & a_{M,M} \end{pmatrix} \in \mathbb{R}^{M\times(m+1)}.$$

The unused elements of B, i.e., $(B)_{ij}$ for $i = 1,\dots,m$, $j = 1,\dots,m+1-i$, are here filled with zeros.

For a general band matrix, the matrix $B \in \mathbb{R}^{M\times(2m+1)}$ obtained by the above conversion has the following form:

$$B = \begin{pmatrix} 0 & \cdots & 0 & a_{11} & a_{12} & \cdots & a_{1,m+1} \\ 0 & \cdots & a_{21} & a_{22} & \cdots & \cdots & a_{2,m+2} \\ \vdots & & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & a_{m,1} & \cdots & \cdots & \cdots & \cdots & a_{m,2m} \\ a_{m+1,1} & \cdots & \cdots & \cdots & \cdots & \cdots & a_{m+1,2m+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{M-m,M-2m} & \cdots & \cdots & \cdots & \cdots & \cdots & a_{M-m,M} \\ a_{M-m+1,M-2m+1} & \cdots & \cdots & \cdots & \cdots & a_{M-m+1,M} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ a_{M,M-m} & \cdots & \cdots & a_{M,M} & 0 & \cdots & 0 \end{pmatrix}.$$

Here, in the right lower part of the matrix, a further sector of unused elements arose, which is also filled with zeros.

If the storage is based on the hull, additionally a pointer field is needed, which points to the diagonal elements, for example. If the matrix is sym-

metric, again the storage of the lower triangular matrix is sufficient. For the matrix A from Figure 2.16 under the assumption that A is symmetric, the pointer field could act as shown in Figure 2.17.

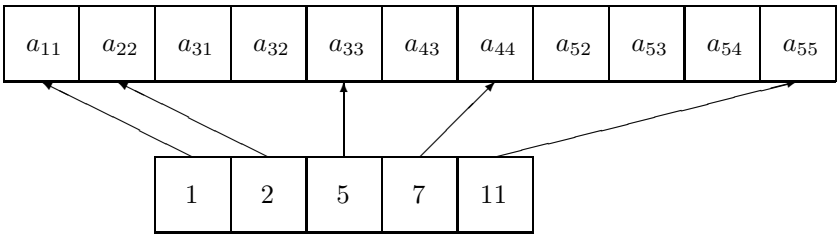

Figure 2.17. Linear storage of the hull.

Coupled Assembling and Decomposition

A formerly popular method, the so-called *frontal method*, performs simultaneously assembling and the Cholesky factorization.

We consider this method for the example of the stiffness matrix $A_h = (a_{ij}) \in \mathbb{R}^{M\times M}$ with bandwidth m (with the original numbering).

The method is based on the kth step of the Gaussian or Cholesky method (cf. Figure 2.18).

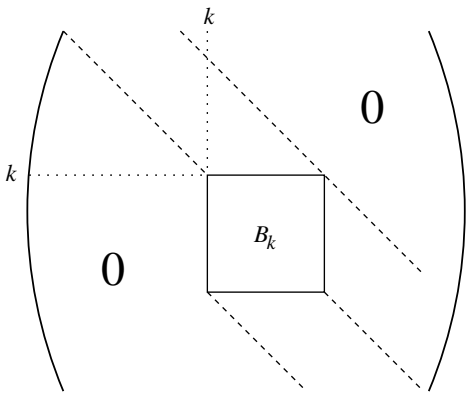

Figure 2.18. kth step of the Cholesky method.

Only the entries of B_k are to be changed, i.e., only those elements a_{ij} with $k \le i, j \le k + m$. The corresponding formula is

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)}\,, \quad i,j = k+1,\ldots,k+m\,. \tag{2.58}$$

Here, the upper indices indicate the steps of the elimination method, which we store in a_{ij}. The entries a_{ij} are generated by summation of entries of

the element stiffness matrix of those elements K that contain nodes with the indices i, j.

Furthermore, to perform the elimination step (2.58), only $a_{ik}^{(k)}, a_{kj}^{(k)}$ for $i, j = k, \ldots, k+m$ must be completely assembled; $a_{ij}^{(k)}, i, j = k+1, \ldots, k+m$, can be replaced by $\tilde{a}_{ij}^{(k)}$ if $a_{ij}^{(k+1)}$ is later defined by $a_{ij}^{(k+1)} := \tilde{a}_{ij}^{(k+1)} + a_{ij}^{(k)} - \tilde{a}_{ij}^{(k)}$. That is, for the present, a_{ij} needs to consist of only a few contributions of elements K with nodes i, j in K.

From these observations, the following algorithm is obtained. The kth step for $k = 1, \ldots, M$ reads as follows:

- Assemble all of the missing contributions of elements K that contain the node with index k.
- Compute $A^{(k+1)}$ by modification of the entries of B_k according to (2.58).
- Store the kth row of $A^{(k+1)}$, also out of the main memory.
- Define B_{k+1} (by a south-east shift).

Here the assembling is node-based and not element-based.

The advantage of this method is that A_h need not be completely assembled and stored in the main memory, but only a matrix $B_k \in \mathbb{R}^{(m+1)\times(m+1)}$. Of course, if M is not too large, there may be no advantage.

Bandwidth Reduction

The *complexity*, i.e., the number of operations, is crucial for the application of a particular method:

The Cholesky method, applied to a symmetric matrix $A \in \mathbb{R}^{M\times M}$ with bandwidth m, requires $O(m^2 M)$ operations in order to compute L.

However, the bandwidth m of the stiffness matrix depends on the numbering of the nodes. Therefore, a numbering is to be found where the number m is as small as possible.

We want to consider this again for the example of the Poisson equation on the rectangle with the discretization according to Figure 2.9. Let the interior nodes have the coordinates (ih, jh) with $i = 1, \ldots k-1$, $j = 1, \ldots, l-1$. The discretization corresponds to the finite difference method introduced beginning with (1.10); i.e., the bandwidth is equal to $k-1$ for a rowwise numbering or $l-1$ for a columnwise numbering.

For $k \ll l$ or $k \gg l$, this fact results in a large difference of the bandwidth m or of the profile (of the left triangle), which is of size $(k-1)(l-1)(m+1)$ except for a term of m^2. Therefore, the columnwise numbering is preferred for $k \gg l$; the rowwise numbering is preferred for $k \ll l$.

For a general domain Ω, a numbering algorithm based on a given triangulation $\mathcal{T}_h$ and on a basis $\{\varphi_i\}$ of V_h is necessary with the following properties:

The structure of A resulting from the numbering must be such that the band or the profile of A is as small as possible. Furthermore, the numbering algorithm should yield the numbers $m(A)$ or $f_i(A)$, $m_i(A)$ such that the matrix A can also be assembled using the element matrices $A^{(k)}$.

Given a triangulation $\mathcal{T}_h$ and a corresponding basis $\{\varphi_i \mid 1 \le i \le M\}$ of V_h, we start with the assignment of some graph G to this triangulation as follows:

The nodes of G coincide with the nodes $\{a_1, \ldots, a_M\}$ of the triangulation. The definition of its edges is:

$$(a_i, a_j) \text{ is an edge of } G \iff$$
$$\text{there exists a } K \in \mathcal{T}_h \text{ such that } \varphi_i|_K \not\equiv 0\,,\ \varphi_j|_K \not\equiv 0\,.$$

In Figure 2.19 some examples are given, where the example (2) will be introduced in Section 3.3.

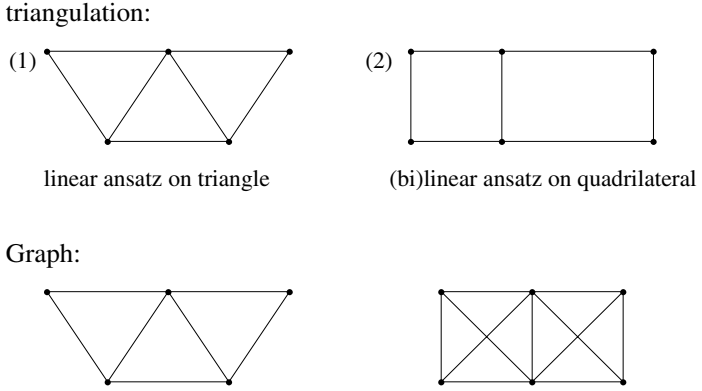

Figure 2.19. Triangulation and assigned graph.

If several degrees of freedom are assigned to some node of the triangulation $\mathcal{T}_h$, then also in G several nodes are assigned to it. This is the case, for example, if so-called Hermite elements are considered, which will be introduced in Section 3.3. The costs of administration are small if the same number of degrees of freedom is assigned to all nodes of the triangulation.

An often-used numbering algorithm is the *Cuthill–McKee method*. This algorithm operates on the graph G just defined. Two nodes a_i, a_j of G are called *neighboured* if (a_i, a_j) is an edge of G. The *degree* of a node a_i of G is defined as the number of neighbours of a_i.

The kth step of the algorithm for $k = 1, \ldots, M$ has the following form:

$k = 1$: Choose a starting node, which gets the number 1. This starting node forms the level 1.

$k > 1$: If all nodes are already numbered, the algorithm is terminated. Otherwise, the level k is formed by taking all the nodes that are not num-

bered yet and that are neighbours of a node of level $k-1$. The nodes of level k will be consecutively numbered.

Within a level, we can sort, for example, by the degree, where the node with the smallest degree is numbered first.

The *reverse Cuthill–McKee method* consists of the above method and the inversion of the numbering at the end; i.e.,

$$\text{new node number } = M + 1 - \text{old node number}\,.$$

This corresponds to a reflection of the matrix at the counterdiagonal. The bandwidth does not change by the inversion, but the profile may diminish drastically for many examples (cf. Figure 2.20).

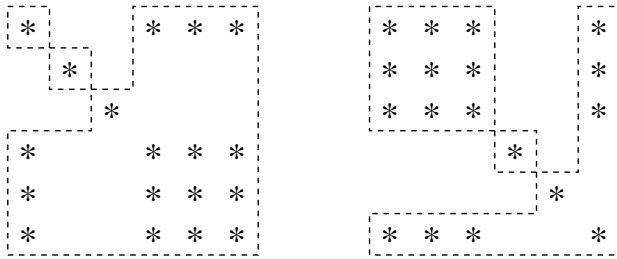

Figure 2.20. Change of the hull by reflection at the counterdiagonal.

The following estimate holds for the bandwidth m of the numbering created by the Cuthill–McKee algorithm:

$$\frac{D+i}{2} \le m \le \max_{2\le k\le \nu} (N_{k-1} + N_k - 1)\ .$$

Here D is the maximum degree of a node of G, ν is the number of levels, and N_k is the number of nodes of level k. The number i is equal to 0 if D is even, and i is equal to 1 if D is odd. The left-hand side of the above inequality is easy to understand by means of the following argument: To reach a minimal bandwidth, all nodes that are neighbours of a_i in the graph G should also be neighbours of a_i in the numbering. Then the best situation is given if the neighboured nodes would appear uniformly immediately before and after a_i. If D is odd, then one side has one node more than the other.

To verify the right-hand side, consider a node a_i that belongs to level $k-1$ as well as a node a_j that is a neighbour of a_i in the graph G and that is not yet numbered in level $k-1$. Therefore, a_j will get a number in the kth step. The largest bandwidth is obtained if a_i is the first node of the numbering of level $k-1$ and if a_j is the last node of level k. Hence exactly $(N_{k-1}-1)+(N_k-1)$ nodes lie between both of these; i.e., their distance in the numbering is $N_{k-1}+N_k-1$.

It is favourable if the number ν of levels is as large as possible and if all the numbers N_k are of the same size, if possible. Therefore, the starting node should be chosen "at one end" of the graph G if possible; if all the

starting nodes are to be checked, the expense will be $O(M\tilde{M})$, where $\tilde{M}$ is the number of edges of G. One possibility consists in choosing a node with minimum degree for the starting node. Another possibility is to let the algorithm run once and then to choose the last-numbered node as the starting node.

If a numbering is created by the (reverse) Cuthill–McKee algorithm, we can try to improve it "locally", i.e., by exchanging particular nodes.

Exercise

2.8 Show that the number of arithmetic operations for the Cholesky method for an $M \times M$ matrix with bandwidth m has order $Mm^2/2$; additionally, M square roots have to be calculated.