LTCC Course

Variational and Computational Methods for PDEs

Lecture Notes [*]

Sergey E. Mikhailov

December 5, 2010

1 Introduction to the finite element method

1.1 Weak formulation of the homogeneous Dirichlet problem for the Poisson equation

Our first model problem is the **Poisson equation** with homogeneous **Dirichlet boundary condition** on a sufficiently smooth, simply connected bounded domain $\Omega \subset \mathbb{R}^2$ (although all stated results will be valid for Lipschitz domains as well):

$$\begin{aligned} -\Delta u &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \Gamma. \end{aligned} \tag{1.1}$$

Here, f is a given function and Δ is the **Laplace operator** or **Laplacian** defined by

$$\Delta u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}$$

where $x = (x_1, x_2) \in \mathbb{R}^2$ are Cartesian coordinates. Even though the Poisson equation looks very special it is an important model case representing several problems from physics and engineering, e.g. electrostatics, stationary heat transfer and other diffusion problems. Variations of the techniques we will study apply to a wide class of second order so-called **elliptic** problems.

It is known that there are cases where no classical (i.e. twice continuously differentiable) solution of (1.1) exists. In order to deal with a uniquely solvable problem one therefore derives a weak formulation.

It is convenient to write the Laplace operator in the following form:

$$\Delta u = \operatorname{div} \nabla u$$

[*] Based on the Lecture Notes for 2007-2008 by Norbert Heuer

where ∇u is the **gradient** of $u(x)$ defined by

$$\nabla u = (\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}) \quad \text{in } \Omega$$

and div is the **divergence** operator defined for a vector-valued function $A = \{A_1(x), A_2(x)\}$ by

$$\operatorname{div} A = \frac{\partial A_1}{\partial x_1} + \frac{\partial A_2}{\partial x_2} \quad \text{in } \Omega.$$

We will also need the **normal derivative** of a function w defined by

$$\partial_n w := \frac{\partial w}{\partial n} := n \cdot \nabla w = \frac{\partial w}{\partial x_1} n_1 + \frac{\partial w}{\partial x_2} n_2 \quad \text{on } \Gamma.$$

Here, $n(x) = \{n_1(x), n_2(x)\}$ denotes the outward unit normal vector to Γ.

Recall the following integration-by-parts formula.

Lemma 1.1 (Gauss formula) *For sufficiently smooth functions v and $w = (w_1, w_2)$ there holds*

$$\int_\Omega \nabla v \cdot w \, dx = \int_\Gamma v \, n \cdot w \, ds - \int_\Omega v \operatorname{div} w \, dx. \tag{1.2}$$

The first integral on the right-hand side denotes integration with respect to the arc length s along Γ.

Remark 1.1 *Remember that, for a differentiable curve Γ with parameter representation $\gamma = (\gamma_1, \gamma_2) : (0, R) \to \Gamma \subset \mathbb{R}^2$, integration along Γ with respect to the arc length is defined by*

$$\int_\Gamma f \, ds = \int_0^R f(\gamma(t)) \left| \frac{d\gamma}{dt}(t) \right| dt = \int_0^R f(\gamma(t)) \sqrt{\left(\frac{d\gamma_1(t)}{dt}\right)^2 + \left(\frac{d\gamma_2(t)}{dt}\right)^2} \, dt$$

An analogous relation holds for a continuous, piecewise differentiable curve.

If we put $w = \nabla u$ in the Gauss formula, we arrive at the following statement.

Lemma 1.2 (First Green identity) *For sufficiently smooth functions v and u there holds*

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Gamma (n \cdot \nabla u) v \, ds - \int_\Omega v \, \Delta u \, dx. \tag{1.3}$$

If u satisfies the Poisson equation, then using the first Green identity we find that there holds

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Gamma (\partial_n u) v \, ds + \int_\Omega f v \, dx.$$

Let us select a space $\mathcal{H}$ as

$$\mathcal{H} := H_0^1(\Omega) := \{v \in L_2(\Omega);\ \nabla v \in (L_2(\Omega))^2,\ v = 0 \text{ on } \Gamma\},$$

(We will discuss later in our course how to understand v on Γ for some functions v discontinuous in $\overline{\Omega}$.)

This leads to the following formulation of the Dirichlet problem (1.1):

$$\text{Find } u \in \mathcal{H} = H_0^1(\Omega): \quad a(u,v) = \langle f, v\rangle \quad \forall v \in \mathcal{H} \tag{1.4}$$

with

$$a(u,v) := \int_\Omega \nabla u \cdot \nabla v \, dx \quad \text{and} \quad \langle f, v\rangle = (f,v)_{L_2(\Omega)} := \int_\Omega f v \, dx. \tag{1.5}$$

Problem (1.4) is called the **variational** or **weak formulation** of (1.1). In this particular case there is an equivalent **minimisation problem**:

$$\text{Find } u \in \mathcal{H} = H_0^1(\Omega): \quad F(u) \le F(v) \quad \forall v \in \mathcal{H} \quad \text{where} \quad F(v) := \frac{1}{2} a(v,v) - \langle f, v\rangle. \tag{1.6}$$

Notations and definitions

For the discussion and analysis of (1.4) we need to introduce some definitions and derivatives used for the space $\mathcal{H}$.

Let $\mathcal{H}$ be a linear space. A mapping $L: \mathcal{H} \to \mathbb{R}$ is called a **linear form** (or **linear functional**) if

$$L(\beta v + \theta w) = \beta L(v) + \theta L(w) \quad \forall v, w \in \mathcal{H}, \quad \forall \beta, \theta \in \mathbb{R}.$$

A mapping $a(\cdot,\cdot)$ is a **bilinear form** (or **bilinear functional**) on $\mathcal{H} \times \mathcal{H}$ if $a: \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ and if it is linear in both arguments:

$$\begin{aligned} a(u, \beta v + \theta w) &= \beta a(u,v) + \theta a(u,w), \\ a(\beta u + \theta v, w) &= \beta a(u,w) + \theta a(v,w) \end{aligned}$$

for all $u, v, w \in \mathcal{H}$ and all $\beta, \theta \in \mathbb{R}$. The bilinear form a is called **symmetric** if

$$a(v,w) = a(w,v) \quad \forall v, w \in \mathcal{H}.$$

A symmetric bilinear form on $\mathcal{H} \times \mathcal{H}$ is a **scalar** or **inner product** on $\mathcal{H}$ if it is positive definite:

$$a(v,v) > 0 \quad \forall v \in \mathcal{H},\ v \ne 0.$$

Every inner product $(\cdot,\cdot)$ on $\mathcal{H} \times \mathcal{H}$ defines a norm $\|\cdot\|$ on $\mathcal{H}$ as $\|u\| = \sqrt{(u,u)}$, and there holds the **Cauchy-Schwarz inequality**

$$|(v,w)| \le \|v\| \, \|w\| \quad \forall v, w \in \mathcal{H}. \tag{1.7}$$

Also, remember that a **complete** normed space with inner product is called a **Hilbert space**.

Now we introduce a weak form of derivatives. Let $I \subset \mathbb{R}$ be an (open) interval. The space $C_0^\infty(I)$ is the set of functions that have continuous derivatives of any order in I and for each function ϕ from this space there exists a segment (closed interval) $\bar{I}_\phi \subset I$ such that ϕ equals zero outside $\bar{I}_\phi$ (i.e., ϕ has a compact support in I).

Definition 1.1 *An element $v \in L_2(I)$ (we call it function) is* **weakly differentiable** *if there exists $g \in L_2(I)$ such that*

$$\int_I v\phi' \, dx = -\int_I g\phi \, dx \qquad \forall \phi \in C_0^\infty(I).$$

Here, the derivative ϕ' is the classical one. When such a function g exists then one defines $v' := g$ is a weak derivative of v.

Note that the weak derivative coincides with the classical derivative for a differentiable function. This follows from the integration-by-parts formula. The extension of this definition to higher orders is by induction and to higher dimensions by replacing the above integration-by-parts formula by the Gauss formula (cf. Lemma 1.1).

Summary. The boundary value problem (1.1) has the weak formulation (1.4) where $a(\cdot, \cdot)$ is a symmetric bilinear form on $H_0^1(\Omega) \times H_0^1(\Omega)$ (one can prove that it is also positive definite) and where $(f, \cdot)$ is a linear form on $H_0^1(\Omega)$. The spaces $L_2(\Omega)$ and

$$H^1(\Omega) := \{v \in L_2(\Omega);\ \nabla v \in (L_2(\Omega))^2\}$$

(derivatives are defined in the weak sense) are Hilbert spaces with inner products and norms

$$\begin{aligned}(v,w)_{L_2(\Omega)} &:= \int_\Omega vw \, dx, & \|v\|_{L_2(\Omega)} &:= \left(\textstyle\int_\Omega v^2 \, dx\right)^{1/2},\\ (v,w)_{H^1(\Omega)} &:= \textstyle\int_\Omega \left(vw + \nabla v \cdot \nabla w\right) dx, & \|v\|_{H^1(\Omega)} &:= \left(\textstyle\int_\Omega \left(v^2 + |\nabla v|^2\right) dx\right)^{1/2}.\end{aligned}$$

Moreover, $H_0^1(\Omega)$ provided with the $H^1(\Omega)$-norm is a closed subspace of $H^1(\Omega)$.

The spaces $H^1(\Omega)$, $H_0^1(\Omega)$ and $H^0(\Omega) = L_2(\Omega)$ are **Sobolev spaces**.

Theorem 1.1 *Any solution of* (1.1) *solves* (1.4)*, and the problems* (1.4) *and* (1.6) *are equivalent in $\mathcal{H} = H_0^1(\Omega)$. Any sufficiently regular solution of* (1.4) *solves* (1.1).

Proof. We have already seen that any solution of (1.1) solves (1.4). Now we show that (1.4) and (1.6) are equivalent. Let u solve (1.4) and let v be an arbitrary element of $\mathcal{H}$. Let $w = v - u$, then $v = u + w$ with $w \in \mathcal{H}$. We obtain

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

since $a(u,w) - \langle f, w\rangle = 0$ by (1.4), and $a(w,w) \geq 0$. Therefore, u solves (1.6).

Now, if u is a solution of (1.6) then for any $v \in \mathcal{H}$ and any real number ϵ there holds

$$F(u) \leq F(u + \epsilon v),$$

since $u + \epsilon v \in \mathcal{H}$. Therefore, the differentiable function g defined by

$$g(\epsilon) := F(u + \epsilon v) = \frac{1}{2} a(u,u) + \epsilon a(u,v) + \frac{\epsilon^2}{2} a(v,v) - \langle f, u \rangle - \epsilon \langle f, v \rangle$$

has a minimum at $\epsilon = 0$ and, thus, $g'(0) = 0$. This yields

$$g'(0) = a(u,v) - \langle f, v \rangle = 0 \qquad \forall v \in \mathcal{H},$$

i.e. u solves (1.4).

Now, to show that a sufficiently smooth solution of (1.4) is also a solution to (1.1) we need that Δu exists and is continuous. Then, considering the property of u that it satisfies

$$-\int_\Omega \nabla u \cdot \nabla v \, dx + \int_\Omega f v \, dx = 0 \quad \forall v \in \mathcal{H}$$

and integrating by parts (using the first Green identity), we obtain

$$\int_\Omega v \Delta u \, dx - \int_\Gamma v \, n \cdot \nabla u \, ds + \int_\Omega f v \, dx = \int_\Omega (\Delta u + f) v \, dx = 0 \quad \forall v \in \mathcal{H}.$$

By the continuity of $\Delta u + f$ this requires that

$$\Delta u + f = 0 \quad \text{pointwise on } \Omega.$$

Since u is continuous, $u \in \mathcal{H} = H_0^1(\Omega)$ in particular means that the homogeneous Dirichlet boundary condition is satisfied for u. This proves that u solves (1.1). □

Exercise 1.1 *Derive the variational formulation and corresponding minimisation problem of the boundary value problem*

$$\begin{aligned} u^{(\mathrm{iv})}(x) &= f(x) \qquad \textit{for } \ 0 < x < 1, \\ u(0) &= u(1) = u'(0) = u'(1) = 0. \end{aligned}$$

Here, $u^{(\mathrm{iv})}$ denotes the fourth order derivative of u.

1.2 The finite element method for the Poisson equation

The finite element method (FEM) for the solution of (1.1) consists in solving (1.4) or (1.6) within a finite-dimensional subspace $\mathcal{H}_h$ of $\mathcal{H}$. This so-called **finite element** or **ansatz space** is usually constructed by piecewise polynomial functions. The idea is that basis functions of $\mathcal{H}_h$ have small support. Here we consider the simplest case of continuous piecewise linear functions.

Let us assume, for simplicity, that Ω is a polygonal domain. To define the finite element space we consider a **triangulation** $\mathcal{T}_h = \{T_j : \ j = 1, \ldots, m\}$ of Ω into elements (triangles in the two-dimensional case) T_j, i.e.

$$\bar{\Omega} = \bigcup_{T \in \mathcal{T}_h} T = T_1 \cup T_2 \cup \ldots \cup T_m.$$

Here we assume that any two triangles are disjoint or intersect at a single vertex or at an entire edge. The triangulation $\mathcal{T}_h$ is also called a **mesh** on Ω. With any such mesh we associate the **mesh size** or **mesh width** defined by

$$h = \max_{T \in \mathcal{T}_h} \operatorname{diam}(T) \quad \text{where} \quad \operatorname{diam}(T) := \text{ diameter of } T = \text{ longest side of } T.$$

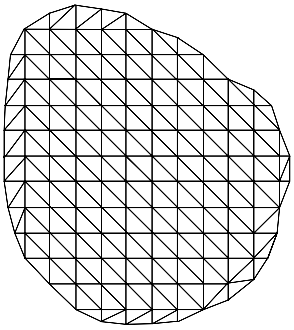

Figure 1.1: Triangulation example.

Our finite element space then is

$$\mathcal{H}_h := \{v : \ v \text{ is continuous on } \Omega, \ v|_T \text{ is linear for } T \in \mathcal{T}_h, \ v = 0 \text{ on } \Gamma\}\,.$$

The **finite element method** for (1.1) reads:

$$\text{find } u_h \in \mathcal{H}_h \text{ such that} \qquad F(u_h) \le F(v) \qquad \forall v \in \mathcal{H}_h \tag{1.8}$$

in the form of a minimisation problem, or

$$\text{find } u_h \in \mathcal{H}_h \text{ such that} \qquad a(u_h, v) = \langle f, v \rangle \qquad \forall v \in \mathcal{H}_h \tag{1.9}$$

in discrete variational form. Of course, as in the proof of Theorem 1.1 one sees that (1.8) and (1.9) are equivalent. Historically, (1.8) is called the **Ritz method** and (1.9) the **Galerkin method**.

To calculate u_h (theoretically, manually or on a computer) one transforms the discrete minimisation or variational problem (i.e. (1.8) or (1.9)) into a system of linear algebraic equations.

One can identify any element of $\mathcal{H}_h$ by its values at the **nodes** N_j $(j = 1, \ldots, M)$ of the mesh (the set of vertices of the triangles). In particular, the dimension of $\mathcal{H}_h$ is the number M of interior nodes of the mesh $\mathcal{T}_h$ (the values on boundary nodes, the ones on Γ, are fixed by definition of $\mathcal{H}_h$, when the problems with the Dirichlet condition on the boundary are considered). It is immediate that the hat-shaped functions $\varphi_j(x)$ from $\mathcal{H}_h$, that are defined by

$$\varphi_j(N_i) = \delta_{ij} \equiv \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}, \quad i, j = 1, \ldots, M$$

form a basis of $\mathcal{H}_h$ (see Figure 1.2); they are called **basis functions**.

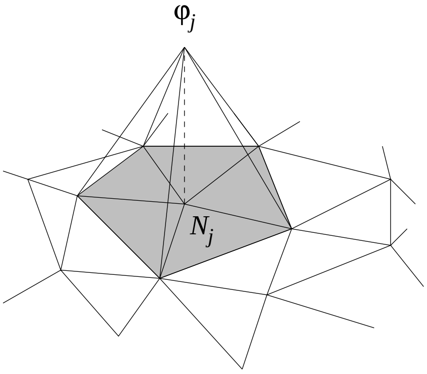

Figure 1.2: Piecewise linear basis function φ_j.

The support of φ_j consists of all elements that have N_j as a vertex. Note that this number of elements depends on the mesh construction and can be different for different nodes. One can represent any $v \in \mathcal{H}_h$ as a linear combination of the basis functions,

$$v(x) = \sum_{j=1}^{M} \eta_j \varphi_j(x) \quad \text{where} \quad \eta_j = v(N_j).$$

In particular, the finite element approximation u_h has the unique representation

$$u_h(x) = \sum_{i=1}^{M} \xi_j \varphi_j(x), \qquad \xi_j = u_h(x_j) \tag{1.10}$$

and it is enough to determine $\xi = (\xi_1, \ldots, \xi_M) \in \mathbb{R}^M$ in order to determine u_h.

The following lemma immediately follows from discrete variational formulation (1.9) if one employs the basis functions φ_i as the test functions v there.

Lemma 1.3 *The solution u_h of* (1.9) *is given by* (1.10) *where ξ is the solution of the linear system*

$$A\xi = b \tag{1.11}$$

where $A = (a_{ij})$ is the $M \times M$ **stiffness matrix** *with elements*

$$a_{ij} = a(\varphi_i, \varphi_j) = \int_\Omega \nabla\varphi_i \cdot \nabla\varphi_j \, dx, \quad i, j = 1, \ldots, M,$$

and $b = (b_i) \in \mathbb{R}^M$ *is the* **load vector** *with*

$$b_i = \langle f, \varphi_i \rangle = \int_\Omega f \varphi_i \, dx, \quad i = 1, \dots, M.$$

for problem (1.1).

1.2.1 Properties and assembly of the stiffness matrix

The stiffness matrix A of (1.11) is symmetric and positive definite:

$$\eta \cdot A\eta > 0 \qquad \forall \eta \in \mathbb{R}^M, \eta \neq 0.$$

This follows from the symmetry and positive definiteness of the bilinear form $a(\cdot, \cdot)$.

The symmetry and positive definiteness of A are important properties when solving the linear system (1.11). For moderate dimensions M it can be solved by the **Cholesky method**, and large systems can be solved iteratively by the **conjugate gradient method** (CG-method). Both methods are the most efficient ones in their class (of direct and iterative methods, respectively) and require symmetric, positive definite matrices.

Another property of A is that it has only a few non-zero elements, it is a **sparse** matrix. Indeed, whenever two basis functions φ_i, φ_j are associated with nodes of different triangles then the measure of the intersection of the supports of φ_i and φ_j is zero so that $a_{ij} = a(\varphi_i, \varphi_j) = 0$. For large numbers of unknowns M the number of non-zero elements of A grows only linearly in M (whereas there are M^2 entries of A in total). This fact, and the special structure of A, can be used to solve the linear system efficiently by only storing $O(M)$ numbers. (Here, $O(M)$ denotes a number that grows at most linearly in M when $M \to \infty$.)

To **assemble** the stiffness matrix one uses an element-oriented strategy. Using the decomposition $\bar{\Omega} = \cup_{T \in \mathcal{T}_h} T$ we find for any $i, j \in \{1, \dots, M\}$ that

$$a(\varphi_i, \varphi_j) = \int_\Omega \nabla\varphi_i \cdot \nabla\varphi_j \, dx = \sum_{T \in \mathcal{T}_h} \int_T \nabla\varphi_i \cdot \nabla\varphi_j \, dx =: \sum_{T \in \mathcal{T}_h} a_T(\varphi_i, \varphi_j). \tag{1.12}$$

There holds $a_T(\varphi_i, \varphi_j) = 0$ unless both nodes N_i and N_j are vertices of the triangle T. Therefore, to calculate $a_T(\varphi_i, \varphi_j)$, one only needs to consider the numbers $i, j \in \{1, \dots, M\}$ which coincide with one of the (global) numbers m_1, m_2, m_3 of the three vertices N_{m_1}, N_{m_2}, N_{m_3} of T. We then call the 3×3-matrix

$$A_T := \begin{pmatrix} a_T(\varphi_{m_1}, \varphi_{m_1}) & a_T(\varphi_{m_1}, \varphi_{m_2}) & a_T(\varphi_{m_1}, \varphi_{m_3}) \\ & a_T(\varphi_{m_2}, \varphi_{m_2}) & a_T(\varphi_{m_2}, \varphi_{m_3}) \\ \text{sym} & & a_T(\varphi_{m_3}, \varphi_{m_3}) \end{pmatrix} \tag{1.13}$$

the **element** or **local stiffness matrix** for T. In order to calculate the stiffness matrix A one calculates all the element stiffness matrices A_T and then forms A by using (1.12). This process is called the **assembly** of A. A is sometimes called **global stiffness matrix** to distinguish it from the local stiffness matrices. An analogous procedure is used to construct the load vector b.

To calculate A_T one obviously needs only the restrictions of the basis functions φ_{m_1}, φ_{m_2}, φ_{m_3} onto T. Let us denote these restrictions by

$$\psi_{m_1} := \varphi_{m_1}|_T, \quad \psi_{m_2} := \varphi_{m_2}|_T, \quad \psi_{m_3} := \varphi_{m_3}|_T.$$

Each of these three functions is linear (on T) and has the value 1 at exactly one vertex and vanishes at the other two vertices. Any linear function w on T can be represented by

$$w(x) = w(N_{m_1})\psi_{m_1}(x) + w(N_{m_2})\psi_{m_2}(x) + w(N_{m_3})\psi_{m_3}(x).$$

The functions ψ_{m_1}, ψ_{m_2}, ψ_{m_3} are called **local** (or **element**) **basis functions** on T.

Exercise 1.2 *Consider the triangle $\tilde{T}$ with vertices $\tilde{N}_1 = (0,0)$, $\tilde{N}_2 = (h,0)$ and $\tilde{N}_3 = (0,h)$. Define the local (linear) basis functions associated with the vertices and show that the local stiffness matrix for $\tilde{T}$ is given by*

$$\tilde{A} = (\tilde{a}_{ij})_{i,j=1}^3 = \begin{pmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}.$$

Also, convince yourself that a translation or rotation of $\tilde{T}$ does not alter this matrix.

Example 1.1 *Let us consider a square Ω with side length 1 and let $\mathcal{T}_h = \{K_j\}_{j=1}^{32}$ be a uniform triangulation of Ω with $h = 1/4$, see Figure 1.3.*

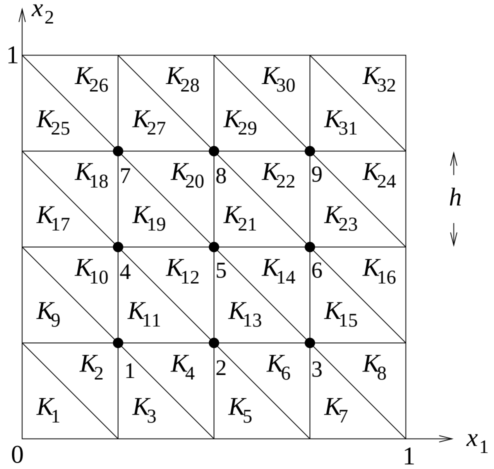

Figure 1.3: Uniform triangulation with $h = 1/4$ for Example 1.1.

(Here, for simplicity, h denotes the smallest side length of the triangles which is proportional to their diameter since they are shape-regular.) The nodes N_i appear as numbers $i = 1, \ldots, 9$ and

the elements are K_i, $i = 1, \dots, 32$. We use the local stiffness matrix $\tilde{A} = (\tilde{a}_{ij})$ from Exercise 1.2 and formula (1.12) to assemble the global stiffness matrix. For instance, noting that the supports of φ_4, $\varphi_1\varphi_4$ and $\varphi_2\varphi_4$ are $\cup_{i\in\{10,11,12,19,18,17\}}K_i$, $K_{10}\cup K_{11}$ and $K_{11}\cup K_{12}$, respectively, we obtain

$$\begin{aligned} a_{4,4} = \sum_{i\in\{10,11,12,19,18,17\}} a_{K_i}(\varphi_4,\varphi_4) &= \tilde{a}_{1,1} + \tilde{a}_{3,3} + \tilde{a}_{2,2} + \tilde{a}_{1,1} + \tilde{a}_{3,3} + \tilde{a}_{2,2} \\ &= 1 + 1/2 + 1/2 + 1 + 1/2 + 1/2 = 4, \end{aligned}$$

$$\begin{aligned} a_{1,4} &= \sum_{i\in\{10,11\}} a_{K_i}(\varphi_1,\varphi_4) = \tilde{a}_{3,1} + \tilde{a}_{1,3} = -1/2 - 1/2 = -1, \\ a_{2,4} &= \sum_{i\in\{11,12\}} a_{K_i}(\varphi_2,\varphi_4) = \tilde{a}_{2,3} + \tilde{a}_{3,2} = 0 + 0 = 0. \end{aligned}$$

Proceeding in this way we obtain the global stiffness matrix

$$A = \begin{pmatrix} 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 4 & 0 & 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & -1 & 4 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 4 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 \end{pmatrix}.$$

Exercise 1.3 *Consider the situation described in Example 1.1 but with $h = 1/3$ instead of $h = 1/4$. For right-hand side function $f(x) = 1$ ($x \in \Omega$) assemble the linear system (1.11), determine the solution u_h of (1.9) and calculate $u_h(1/2, 1/2)$.*

1.3 Galerkin orthogonality

Any variational formulation

$$u \in \mathcal{H}: \quad a(u,v) = L(v) \quad \forall v \in \mathcal{H} \tag{1.14}$$

with corresponding finite element scheme

$$u_h \in \mathcal{H}_h \subset \mathcal{H}: \quad a(u_h,v) = L(v) \quad \forall v \in \mathcal{H}_h$$

translates into the **Galerkin orthogonality**

$$a(u - u_h, v) = 0 \quad \forall v \in \mathcal{H}_h, \tag{1.15}$$

which is the usual orthogonality for the inner product $(w, v)_{\mathcal{H}} := a(w, v)$ if a is symmetric and positive definite bilinear form in $\mathcal{H}$.

Let us consider the homogeneous Dirichlet problem for the Klein-Gordon equation (the Helmholtz equation with imaginary coefficient):

$$\begin{aligned} -\Delta u + u &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \Gamma. \end{aligned} \tag{1.16}$$

The corresponding variational formulation is

$$u \in H_0^1(\Omega): \quad \langle \nabla u, \nabla v \rangle + \langle u, v \rangle = \langle f, v \rangle \quad \forall v \in H_0^1(\Omega), \tag{1.17}$$

and can also be written in the general form (1.14) with

$$\mathcal{H} = H_0^1(\Omega), \quad a(u, v) = \langle \nabla u, \nabla v \rangle + \langle u, v \rangle \quad \text{and} \quad L(v) = \langle f, v \rangle.$$

We note that in fact $a(u, v) = (u, v)_{H^1(\Omega)} = (u, v)_{H_0^1(\Omega)}$, i.e., in this case $a(u, v)$ is the standard inner product in $H^1(\Omega)$ and thus in $H_0^1(\Omega)$. Then the variational formulation renders like

$$u \in H_0^1(\Omega): \quad (u, v)_{H^1(\Omega)} = L(v) \quad \forall v \in H_0^1(\Omega).$$

Introducing a finite element space $\mathcal{H}_h \subset H_0^1(\Omega)$ one has a corresponding finite element scheme and the Galerkin orthogonality (1.15) becomes

$$(u - u_h, v)_{H^1(\Omega)} = 0 \quad \forall v \in \mathcal{H}_h. \tag{1.18}$$

The relation (1.18) means that the finite element error $u - u_h$ is orthogonal to the finite element subspace $\mathcal{H}_h$ of $\mathcal{H}$. In particular, u_h is the **projection** with respect to the inner product $(\cdot, \cdot)_{H^1(\Omega)}$ of u onto $\mathcal{H}_h$. Figure 1.4 gives a geometric description of this fact for the case $\mathcal{H} = \mathbb{R}^2$ with Euclidean inner product and a one-dimensional subspace $\mathcal{H}_h \subset \mathcal{H}$.

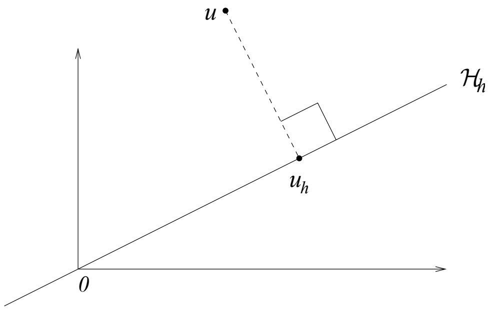

Figure 1.4: Projection of u onto $\mathcal{H}_h$.

This property proves the following best approximation property:

$$\|u - u_h\|_{H^1(\Omega)} \leq \|u - v\|_{H^1(\Omega)} \quad \forall v \in \mathcal{H}_h. \tag{1.19}$$

Note that with the finite element method for (1.16) we calculate the projection of the exact solution u onto $\mathcal{H}_h$ without actually knowing the exact solution. This only requires the solution of a sparse linear system $A\xi = b$ with symmetric, positive definite matrix A.

Exercise 1.4 *Show that* (1.18) *and* (1.19) *are equivalent.*

1.4 Natural and essential boundary conditions

So far we have considered only Dirichlet boundary conditions where the sought solution is prescribed on the boundary of the domain. There is another important type of boundary conditions where the normal derivative is prescribed. Such boundary condition is called **the Neumann boundary condition**, and a boundary value problem with the Neumann boundary condition is called **the Neumann problem**. Consider, for example, the following Neumann BVP,

$$\begin{aligned} -\Delta u + u &= f \qquad \text{in } \Omega, \\ \partial_n u &= g \qquad \text{on } \Gamma. \end{aligned} \tag{1.20}$$

Here, f and g are given functions and $\partial_n u$ denotes, as introduced before, the outward normal derivative of u on Γ.

The variational formulation of (1.20) is:

$$\text{Find } u \in H^1(\Omega): \quad a(u,v) = \langle f, v\rangle + \langle g, v\rangle_\Gamma \quad \forall v \in H^1(\Omega), \tag{1.21}$$

where

$$a(u,v) := \langle \nabla u, \nabla v\rangle + \langle u, v\rangle \quad \text{and} \quad \langle g, v\rangle_\Gamma := (g,v)_{L_2(\Gamma)} := \int_\Gamma gv\,ds.$$

Correspondingly, the minimisation problem is:

$$\text{Find } u \in H^1(\Omega): \quad F(u) \le F(v) \quad \forall v \in H^1(\Omega) \tag{1.22}$$

where

$$F(v) := \frac{1}{2}a(v,v) - \langle f, v\rangle - \langle g, v\rangle_\Gamma.$$

Theorem 1.2 *Any solution u of* (1.20) *solves* (1.21). *If u is a sufficiently regular solution of* (1.21) *then it solves* (1.20). *Moreover, problems* (1.21) *and* (1.22) *are equivalent.*

Proof. The equivalence of (1.21) and (1.22) is analogous to the situation in Theorem 1.1. Now assume that u solves (1.20). We multiply the differential equation in (1.20) by a test function $v \in H^1(\Omega)$ and integrate over Ω. Using that $\partial_n u = g$ on Γ, the first Green formula (Lemma 1.2) gives

$$\begin{aligned} \langle f, v\rangle &= \int_\Omega (-\Delta u + u)\, v\,dx = -\int_\Gamma \partial_n u v\,ds + \int_\Omega \nabla u \cdot \nabla v\,dx + \int_\Omega uv\,dx \\ &= -\langle g, v\rangle_\Gamma + \langle \nabla u, \nabla v\rangle + \langle u, v\rangle = a(u,v) - \langle g, v\rangle_\Gamma. \end{aligned}$$

This is (1.21). Now let u be a sufficiently smooth function that solves (1.21). Using again Green's first formula we obtain

$$\langle f, v\rangle + \langle g, v\rangle_\Gamma = a(u,v) = \int_\Gamma \partial_n u v \, ds + \int_\Omega (-\Delta u + u)\, v \, dx,$$

that is

$$\int_\Omega (-\Delta u + u - f)\, v \, dx + \int_\Gamma (\partial_n u - g)\, v \, ds = 0 \quad \forall v \in H^1(\Omega). \tag{1.23}$$

In particular, (1.23) holds for any $v \in H^1(\Omega)$ with $v = 0$ on Γ, that is

$$\int_\Omega (-\Delta u + u - f)\, v \, dx = 0 \quad \forall v \in H_0^1(\Omega).$$

For sufficiently smooth u this is only possible if

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

Taking this relation (it is the wanted differential equation) into equation (1.23) gives

$$\int_\Gamma (\partial_n u - g)\, v \, ds = 0 \quad \forall v \in H^1(\Omega).$$

By varying the test function $v \in H^1(\Omega)$ appropriately it can be seen that this requires

$$\partial_n u - g = 0 \quad \text{on } \Gamma.$$

This finishes the proof of the theorem. □

Remark 1.2 *Note that the Neumann boundary condition appears in the variational formulation (via the linear form on the right-hand side) and is not incorporated in the space $\mathcal{H} = H^1(\Omega)$. It is therefore called* **natural boundary condition**. *In contrast, a Dirichlet boundary condition of the type $u = 0$ on Γ enters the variational formulation by choosing $\mathcal{H}$ appropriately to reflect this condition, $\mathcal{H} = H_0^1(\Omega) \subset H^1(\Omega)$ in this case. Therefore, Dirichlet boundary conditions are also called* **essential boundary conditions**. *This difference in incorporating boundary conditions is inherited by the finite element schemes.*

To define a finite element scheme for the approximate solution of (1.21) we choose a finite-dimensional subspace $\mathcal{H}_h \subset H^1(\Omega)$. To this end we consider as before a mesh $\mathcal{T}_h$ consisting of triangles T. The simplest choice of $\mathcal{H}_h$ is

$$\mathcal{H}_h := \{v : \ v \text{ is continuous on } \Omega, \ v|_T \text{ is linear } \forall T \in \mathcal{T}_h\}.$$

Note that we do not ask $v \in \mathcal{H}_h$ to vanish on Γ. All the nodes of $\mathcal{T}_h$ including the ones on Γ are now participating in the finite element formulation. The finite element method then is:

$$\text{Find } u_h \in \mathcal{H}_h : \quad a(u_h, v) = \langle f, v\rangle + \langle g, v\rangle_\Gamma \quad \forall v \in \mathcal{H}_h. \tag{1.24}$$

Note that u_h in general does not satisfy the Neumann boundary condition. One can rather show that $\partial_n u_h \to g$ $(h \to 0)$ in an appropriate norm.

Exercise 1.5 *Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with Lipschitz continuous boundary Γ and let Γ be decomposed into two non-empty curves Γ_1 and Γ_2: $\Gamma = \bar{\Gamma}_1 \cup \bar{\Gamma}_2$ and $\Gamma_1 \cap \Gamma_2 = \emptyset$. For sufficiently smooth functions f and g give a variational formulation of the* **mixed boundary value problem**

$$\begin{aligned} -\Delta u + u &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \Gamma_1, \\ \partial_n u &= g && \text{on } \Gamma_2. \end{aligned}$$

2 Unique solvability of variational formulations

In this section we deal with existence and uniqueness of a solution to the abstract variational problem

$$u \in \mathcal{H}: \ a(u,v) = L(v) \quad \forall v \in \mathcal{H}. \tag{2.1}$$

Here, $\mathcal{H}$ is a Hilbert space with inner product $(\cdot,\cdot)$ and norm $\|\cdot\|$, $a(\cdot,\cdot)$ is a bilinear form and $L: \mathcal{H} \to \mathbb{R}$ is a linear form. We will need some properties of a and L.

Definition 2.1

1. *The linear form $L: \mathcal{H} \to \mathbb{R}$ is called linear functional. It is* **continuous** *or* **bounded** *if*
$$\exists C > 0: \ |L(v)| \le C\|v\| \quad \forall v \in \mathcal{H}.$$
2. *The space consisting of all linear bounded functionals $\mathcal{H} \to \mathbb{R}$ is called* **dual space of $\mathcal{H}$** *and is denoted by $\mathcal{H}'$ or $\mathcal{L}(\mathcal{H}, \mathbb{R})$. The norm of any element $L \in \mathcal{H}'$ is defined by*
$$\|L\|_{\mathcal{H}'} := \sup_{v \in \mathcal{H}\setminus\{0\}} \frac{|L(v)|}{\|v\|}.$$
3. *The bilinear form $a: \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is called* **continuous** *or* **bounded** *if*
$$\exists C_a > 0: \ |a(v,w)| \le C_a\|v\|\,\|w\| \quad \forall v, w \in \mathcal{H}.$$
4. *The bilinear form $a: \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is called* **$\mathcal{H}$-elliptic** *(or just* **elliptic** *if the space is clear) if*
$$\exists \alpha > 0: \ a(v,v) \ge \alpha\|v\|^2 \quad \forall v \in \mathcal{H}.$$

Instead of $\mathcal{H}$-elliptic, the terms **coercive on $\mathcal{H}$** *or* **positive definite on $\mathcal{H}$** *are sometimes used.*

We now have all the properties of bilinear and linear forms to formulate the main result of this section (Theorem 2.1 below). However, for its proof we need two more classical results from functional analysis.

Proposition 2.1 (Banach fixed point theorem)
Let $\mathcal{B}$ be a **Banach space** *(a complete linear space not necessarily having an inner product) and let an operator $\phi : \mathcal{B} \to \mathcal{B}$ be a* **contraction***, i.e.*

$$\exists c,\ 0 \le c < 1 : \quad \|\phi(g) - \phi(w)\| \le c\|g - w\| \quad \forall g, w \in \mathcal{B}. \tag{2.2}$$

Then there exists a unique $u \in \mathcal{B}$ such that

$$\phi(u) = u.$$

Proposition 2.2 (Riesz representation theorem)
Let $\mathcal{H}$ be a Hilbert space with inner product $(\cdot, \cdot)$ and norm $\|\cdot\|$. Any element $w \in \mathcal{H}$ defines a continuous linear form $L_w \in \mathcal{H}'$ by $L_w(v) := (w, v)$. On the other hand, for any continuous linear form $L \in \mathcal{H}'$ there exists a unique element $\mathcal{R}L \in \mathcal{H}$ such that

$$L(v) = (\mathcal{R}L, v) \quad \forall v \in \mathcal{H}.$$

Moreover, $\mathcal{R} : \mathcal{H}' \to \mathcal{H}$ is a linear operator and there holds $\|\mathcal{R}L\|_{\mathcal{H}} = \|L\|_{\mathcal{H}'}$, i.e.

$$\|\mathcal{R}\|_{\mathcal{H}' \to \mathcal{H}} := \sup_{G \in \mathcal{H}' \setminus \{0\}} \frac{\|\mathcal{R}G\|_{\mathcal{H}}}{\|G\|_{\mathcal{H}'}} = 1.$$

Theorem 2.1 (Lax-Milgram lemma) *Let $\mathcal{H}$ be a Hilbert space, $a(\cdot, \cdot)$ a continuous, $\mathcal{H}$-elliptic bilinear form and L a continuous linear form on $\mathcal{H}$. Then the variational problem* (2.1) *has a unique solution $u \in \mathcal{H}$.*

Proof. By the continuity of a we obtain that for any fixed $u \in \mathcal{H}$ the mapping

$$Au : \ v \mapsto Au(v) := a(u, v) \quad \forall v \in \mathcal{H} \tag{2.3}$$

is linear and bounded:

$$|Au(v)| = |a(u, v)| \le C_a\|u\|\,\|v\| \le C\|v\| \quad \forall v \in \mathcal{H} \quad \text{with } C := C_a\|u\|.$$

This means that

$$\|Au\|_{\mathcal{H}'} = \sup_{v \in \mathcal{H} \setminus \{0\}} \frac{|Au(v)|}{\|v\|} \le C_a\|u\| \quad \forall u \in \mathcal{H},$$

i.e. the operator $A : \ \mathcal{H} \to \mathcal{H}'$ is linear and continuous:

$$\|A\|_{\mathcal{H} \to \mathcal{H}'} := \sup_{u \in \mathcal{H} \setminus \{0\}} \frac{\|Au\|_{\mathcal{H}'}}{\|u\|} \le C_a.$$

By the Riesz representation theorem there exists for any element $G \in \mathcal{H}'$ (i.e. any continuous linear form $G : \ \mathcal{H} \to \mathbb{R}$) a unique element $\mathcal{R}G \in \mathcal{H}$ such that

$$G(v) = (\mathcal{R}G, v) \quad \forall v \in \mathcal{H}.$$

Therefore, our problem (2.1) that can be re-written as

$$u \in \mathcal{H}: \; Au = L,$$

is equivalent to

$$u \in \mathcal{H}: \; \mathcal{R}Au = \mathcal{R}L.$$

We show that for sufficiently small $\rho > 0$ the mapping

$$\mathcal{C}_\rho : \begin{cases} \mathcal{H} & \rightarrow \quad \mathcal{H} \\ u & \mapsto \quad u - \rho(\mathcal{R}Au - \mathcal{R}L) \end{cases}$$

is a contraction. The Banach fixed point theorem then yields the existence of unique $u \in \mathcal{H}$ such that

$$u - \rho(\mathcal{R}Au - \mathcal{R}L) = u, \quad \text{i.e.} \quad \mathcal{R}Au = \mathcal{R}L,$$

which proves the theorem.

To establish that the mapping $\mathcal{C}_\rho$ is a contraction for small ρ, we use the ellipticity of a (with constant $\alpha > 0$), the property $\|\mathcal{R}\|_{\mathcal{H}' \to \mathcal{H}} = 1$ and the boundedness $\|A\|_{\mathcal{H} \to \mathcal{H}'} \leq C_a$. Taking in mind that $(\mathcal{R}Au, v) = Au(v) = a(u, v)$, and denoting $v = g - w$ we have

$$\begin{aligned} \|\mathcal{C}_\rho g - \mathcal{C}_\rho w\|^2 &= \|v - \rho\mathcal{R}Av\|^2 = (v - \rho\mathcal{R}Av, v - \rho\mathcal{R}Av) \\ &= \|v\|^2 - 2\rho(\mathcal{R}Av, v) + \rho^2\|\mathcal{R}Av\|^2 = \|v\|^2 - 2\rho\, a(v, v) + \rho^2\|\mathcal{R}Av\|^2 \\ &\leq \|v\|^2 - 2\rho\alpha\|v\|^2 + \rho^2\|\mathcal{R}\|^2_{\mathcal{H}' \to \mathcal{H}}\|A\|^2_{\mathcal{H} \to \mathcal{H}'}\|v\|^2 \leq (1 - 2\rho\alpha + \rho^2 C_a^2)\|v\|^2, \end{aligned}$$

i.e. $\mathcal{C}_\rho$ is a contraction for $\rho \in (0, 2\alpha/C_a^2)$. □

Remark 2.1 *The Lax-Milgram lemma implies that for any $L \in \mathcal{H}'$ there exists a unique element $u = A^{-1}L \in \mathcal{H}$ solving the equation $Au = L$, where A is defined by* (2.3)*, and thus the mapping $A: \mathcal{H} \to \mathcal{H}'$ is an isomorphism (linear and bijective). Moreover, we have the estimate*

$$\alpha\|u\|^2_{\mathcal{H}} \leq a(u, u) = Au(u) \leq \|Au\|_{\mathcal{H}'}\|u\|_{\mathcal{H}} \quad \forall u \in \mathcal{H},$$

taking there $u = A^{-1}L \in \mathcal{H}$ one finds that

$$\alpha\|A^{-1}L\|_{\mathcal{H}} \leq \|L\|_{\mathcal{H}'}$$

implying the inverse A^{-1} of A is continuous with norm

$$\|A^{-1}\|_{\mathcal{H}' \to \mathcal{H}} := \sup_{L \in \mathcal{H}' \setminus \{0\}} \frac{\|A^{-1}L\|}{\|L\|_{\mathcal{H}'}} \leq \alpha^{-1}.$$

It follows that the variational formulation (2.1) *is* **well-posed** *in the sense that there exists a unique solution which depends continuously on the data (i.e. on L):*

$$\|u\| = \|A^{-1}L\| \leq \alpha^{-1}\|L\|_{\mathcal{H}'}.$$

Exercise 2.1 *Show, by using the Lax-Milgram lemma, that* (1.17) *has a unique solution provided that* $f \in L_2(\Omega)$.

Exercise 2.2 *Under appropriate conditions on* f *and* g*, prove existence and uniqueness of the solution to the variational formulation of the mixed problem found in Exercise* 1.5.
Hint: *Use without proof that*

$$\exists C > 0: \quad \|v\|_{L_2(\Gamma)} \leq C\|v\|_{H^1(\Omega)} \quad \forall v \in H^1(\Omega).$$

3 Abstract error estimate for the finite element method

Let us recall the setting. The aim is to find an approximative solution to the continuous problem (2.1),

$$u \in \mathcal{H}: \; a(u,v) = L(v) \quad \forall v \in \mathcal{H},$$

where we use the notation from before: $\mathcal{H}$ is a Hilbert space with inner product $(\cdot,\cdot)$ and norm $\|\cdot\|$, $a(\cdot,\cdot)$ is a bilinear form and $L: \mathcal{H} \to \mathbb{R}$ is a linear form.

The discrete version is as follows. **For a given finite-dimensional subspace $\mathcal{H}_h \subset \mathcal{H}$ find $u_h \in \mathcal{H}_h$ such that**

$$a(u_h, v) = L(v) \quad \forall v \in \mathcal{H}_h. \tag{3.1}$$

Theorem 3.1 *Let* $a(\cdot,\cdot)$ *be a continuous and* $\mathcal{H}$*-elliptic bilinear form with an ellipticity constant* α *and let* $L \in \mathcal{H}'$ *(i.e.* L *is a continuous linear form on* $\mathcal{H}$*). Then the discrete variational problem* (3.1) *has a unique solution* $u_h \in \mathcal{H}_h$ *and there holds the* **stability estimate**

$$\|u_h\| \leq \alpha^{-1}\|L\|_{\mathcal{H}'}. \tag{3.2}$$

Proof. Since $\mathcal{H}_h$ is a subspace of $\mathcal{H}$ the continuity of a and L and the ellipticity of a remain true on $\mathcal{H}_h$ as forms $a: \mathcal{H}_h \times \mathcal{H}_h \to \mathbb{R}$ and $L: \mathcal{H}_h \to \mathbb{R}$. Therefore, the existence and uniqueness of u_h follow from the Lax-Milgram lemma (Theorem 2.1) and the stability estimate is a discrete version of Remark 2.1. □

The next theorem is the basis for error estimates of the finite element method.

Theorem 3.2 (Céa's lemma, quasi-optimal error estimate) *Let* a *be continuous and* $\mathcal{H}$*-elliptic bilinear form and let* $L \in \mathcal{H}'$*. Then, the solutions* $u \in \mathcal{H}$ *and* $u_h \in \mathcal{H}_h$ *of* (2.1) *and* (3.1)*, respectively, satisfy*

$$\|u - u_h\| \leq \frac{C_a}{\alpha}\|u - v\| \quad \forall v \in \mathcal{H}_h.$$

Here, α *and* C_a *are the ellipticity and continuity constants of* a*, respectively (see Section* 2*).*

Proof. If $\|u - u_h\| = 0$ then there is nothing to prove. Subtracting the equations of the continuous and discrete variational formulations, (2.1) and (3.1), yields the Galerkin orthogonality

$$a(u - u_h, w) = 0 \quad \forall w \in \mathcal{H}_h.$$

We select an arbitrary $v \in \mathcal{H}_h$ and define $w := u_h - v \in \mathcal{H}_h$ so that $v = u_h - w$. Then, using the ellipticity of a, the Galerkin orthogonality and the continuity of a, we find that there holds

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

Dividing by $\|u - u_h\| > 0$ and $\alpha > 0$ gives the result. □

Céa's lemma provides an abstract error estimate by a term that includes the unknown solution u. However, it states two important facts. First, selecting any function $v \in \mathcal{H}_h$, the norm $\|u - v\|$ is, up to a constant factor (independent of $\mathcal{H}_h$), an upper bound for the error $\|u - u_h\|$. Therefore, based on Céa's lemma more specific error estimates can be derived if certain properties of u (regularity) are known. Second, Céa's lemma states that the finite element solution u_h is almost the best approximation of u among elements of $\mathcal{H}_h$. ("Almost" refers to the factor C_a/α.) Céa's lemma is therefore also called a quasi-optimal error estimate. Note that the estimate can be formulated equivalently by

$$\|u - u_h\| \leq \frac{C_a}{\alpha} \min_{v \in \mathcal{H}_h} \|u - v\|$$

and that $\min_{v \in \mathcal{H}_h} \|u - v\|$ is the distance of u to $\mathcal{H}_h$ (in the norm of $\mathcal{H}$). Thus, the finite element method has the remarkable property of delivering the (almost) best approximation of an unknown function. Of course, this fact originates from the particular type of (elliptic) problems we are studying.

3.1 The energy norm

Let us assume that a bilinear form $a : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is **symmetric** and positive definite. This means in fact that a can be taken as another inner product on $\mathcal{H}$ inducing a norm

$$\|v\|_a := \sqrt{a(v, v)}, \quad v \in \mathcal{H}.$$

This norm is called **energy norm**. Its name has a physical motivation where $\frac{1}{2}a(v, v)$ relates to the energy of a physical system. Note that ellipticity implies positive definiteness.

If a is elliptic and continuous in $\mathcal{H}$, there holds

$$\alpha\|v\|^2 \leq a(v, v) = \|v\|_a^2 = a(v, v) \leq C_a\|v\|^2 \quad \forall v \in \mathcal{H},$$

that is

$$\alpha^{1/2}\|v\| \leq \|v\|_a \leq C_a^{1/2}\|v\| \quad \forall v \in \mathcal{H}. \tag{3.3}$$

Therefore, $\|\cdot\|$ and $\|\cdot\|_a$ are equivalent norms in $\mathcal{H}$. The Galerkin orthogonality

$$a(u - u_h, v) = 0 \quad \forall v \in \mathcal{H}_h$$

then is in fact an orthogonality in the sense of the energy inner product: the finite element error $u - u_h$ is orthogonal to $\mathcal{H}_h$ with respect to the inner product $a(\cdot,\cdot)$. As we have seen in Section 1.3, the Galerkin orthogonality is equivalent to the best approximation property with respect to the norm induced by the inner product, in this case

$$\|u - u_h\|_a \leq \|u - v\|_a \quad \forall v \in \mathcal{H}_h. \tag{3.4}$$

Therefore, in the case of a symmetric, elliptic bilinear form, Céa's lemma (Theorem 3.2) can be improved to a best approximation property by switching from the norm $\|\cdot\|$ to the energy norm $\|\cdot\|_a$.

4 Sobolev spaces, trace theorem and normal derivative

Throughout, $\Omega \subset \mathbb{R}^n$ will be a bounded domain with a sufficiently smooth boundary Γ.

We use first the standard Sobolev spaces

$$H^0(\mathbb{R}^n) := L_2(\mathbb{R}^n),\ H^0(\Omega) := L_2(\Omega),\ H^k(\mathbb{R}^n),\ H^k(\Omega) \quad (k \text{ positive integer}).$$

Note that all these spaces are based on the use of weak derivatives up to order k, and the norm in $H^k(\Omega)$ is defined as

$$\|u\|_{H^t(\Omega)} := \left(\sum_{|\alpha|=0}^{t} \int_\Omega |\partial^\alpha u(x)|^2 dx \right)^{1/2} = \left(\sum_{|\alpha|=0}^{t} \|\partial^\alpha u(x)\|^2_{L_2(\Omega)} \right)^{1/2},$$

where $\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)$ is the multiindex with non-negative components and $|\alpha| := \alpha_1 + \alpha_2 + ... + \alpha_n$.

We will use the **Fourier transform** to redefine the norms in these spaces. Recall that the Fourier transform $\mathcal{F}$ is defined by

$$\hat{v}(\xi) := \mathcal{F}v(\xi) := \int_{\mathbb{R}^n} e^{-i2\pi\xi\cdot x} v(x)\, dx \qquad (\xi \in \mathbb{R}^n),$$

where a normalisation with some coefficient is also possible in the definition.

Since

$$|\hat{v}(\xi)| = |\int_{\mathbb{R}^n} e^{-i2\pi\xi\cdot x} v(x)\, dx| \leq \int_{\mathbb{R}^n} |e^{-i2\pi\xi\cdot x} v(x)|\, dx = \int_{\mathbb{R}^n} |v(x)|\, dx$$

it follows that $\hat{v}$ is well-defined whenever $v \in L_1(\mathbb{R}^n)$. The inversion formula for the Fourier transform is

$$\mathcal{F}^{-1}\hat{v}(x) := \int_{\mathbb{R}^n} e^{i2\pi\xi\cdot x} \hat{v}(\xi)\, d\xi.$$

One finds the following properties:

- If v, $\hat{v} \in L_1(\mathbb{R}^n)$ then $\mathcal{F}^{-1}\mathcal{F}v = v = \mathcal{F}\mathcal{F}^{-1}v$ at points where v is continuous.
- $\mathcal{F}$ generalises to a bounded linear mapping
$$\mathcal{F}:\ L_2(\mathbb{R}^n) \to L_2(\mathbb{R}^n)$$
and there holds
$$\langle \mathcal{F}\varphi, \mathcal{F}v\rangle = \langle \varphi, v\rangle = \langle \mathcal{F}^{-1}\varphi, \mathcal{F}^{-1}v\rangle \quad \forall \varphi, v \in L_2(\mathbb{R}^n).$$
This property is known as **Plancherel's theorem**. The symbol $\langle \cdot, \cdot\rangle$ denotes the L_2 inner product on $\mathbb{R}^n$ and will be used throughout, also for its extension by duality. When referring to the inner product in L_2 on a subset of $\mathbb{R}^n$, e.g. on Ω, we add this subset as an index, e.g. $\langle \cdot, \cdot\rangle_\Omega$.
- A conclusion from Plancherel's theorem is the relation
$$\|v\|_{L_2(\mathbb{R}^n)} = \|\hat{v}\|_{L_2(\mathbb{R}^n)} \qquad \forall v \in L_2(\mathbb{R}^n).$$
i.e., $\mathcal{F}$ is a unitary isomorphism.

Example 4.1 *Consider the one-dimensional case, i.e., $n = 1$, and let v' denote the derivative of v.*
There holds $\|v'\|_{L_2(\mathbb{R})} = \|\mathcal{F}(v')\|_{L_2(\mathbb{R})}$, and for any $v \in H^1(\mathbb{R})$ with compact support we obtain
$$\mathcal{F}(v')(\xi) = \int_{\mathbb{R}} e^{-i2\pi x\xi} v'(x)\,dx = v(x)e^{-i2\pi x\xi}\Big|_{x=-\infty}^{\infty} - \int_{\mathbb{R}} -i2\pi\xi\, e^{-i2\pi x\xi} v(x)\,dx = i2\pi\xi\,\hat{v}(\xi)$$
where we used that $v(x) = 0$ when $|x|$ is big enough since v has a compact support. Therefore,
$$\|v'\|_{L_2(\mathbb{R})} = \|i2\pi\xi\,\hat{v}(\xi)\|_{L_2(\mathbb{R})} = 2\pi\|\xi\,\hat{v}(\xi)\|_{L_2(\mathbb{R})}$$
and
$$\|v\|^2_{H^1(\mathbb{R})} = \|v\|^2_{L_2(\mathbb{R})} + \|v'\|^2_{L_2(\mathbb{R})} = \|\hat{v}\|^2_{L_2(\mathbb{R})} + 4\pi^2\|\xi\,\hat{v}\|^2_{L_2(\mathbb{R})} = \int_{\mathbb{R}} (1 + 4\pi^2\xi^2)|\hat{v}(\xi)|^2\,d\xi,$$
so that
$$\|v\|_{H^1(\mathbb{R})} \qquad \text{and} \qquad \Big(\int_{\mathbb{R}} (1 + \xi^2)|\hat{v}(\xi)|^2\,d\xi\Big)^{1/2} = \|(1 + |\xi|^2)^{1/2}\hat{v}\|_{L_2(\mathbb{R})}$$
are equivalent norms.

This example easily generalises to higher dimensions ($n > 1$). Moreover, it leads us to the definition of Sobolev spaces on $\mathbb{R}^n$ for any positive real order.

Definition 4.1 *For $s > 0$ we define*

$$H^s(\mathbb{R}^n) := \Big\{ v \in L_2(\mathbb{R}^n) : \ \|(1+|\xi|^2)^{s/2}\hat{v}\|_{L_2(\mathbb{R}^n)} < \infty \Big\}$$

with norm

$$\|v\|_{H^s(\mathbb{R}^n)} := \|(1+|\xi|^2)^{s/2}\hat{v}\|_{L_2(\mathbb{R}^n)}.$$

As in Example 4.1 one sees that, for integer s, this norm is equivalent to the usual one (based on derivatives). For non-integer s, $H^s(\mathbb{R}^n)$ is called a **fractional order Sobolev space** or **Bessel potential space**.

We are now in a position to analyse the trace operator in the half-space case. Consider the situation given in Figure 4.1. For $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ we denote $x' := (x_1, \dots, x_{n-1})$ and then $x = (x', x_n)$. For $v \in C_0^\infty(\mathbb{R}^n)$ (i.e., for v that has continuous derivatives of any order and is compactly supported in $\mathbb{R}^n$), we define its trace onto the hyperplane $\mathbb{R}^{n-1} \times \{0\}$ by

$$(\gamma v)(x') := v(x', 0), \qquad x' \in \mathbb{R}^{n-1}.$$

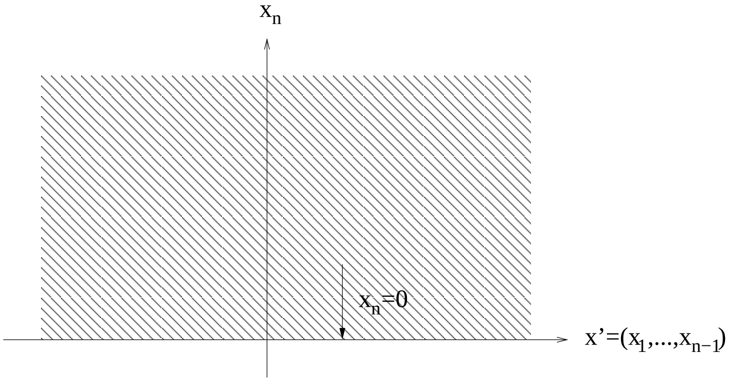

Figure 4.1: The trace in the half-space case.

The following lemma can be proved, see e.g. [McLean 2000, Lemma 3.24].

Lemma 4.1 *The space $C_0^\infty(\mathbb{R}^n)$ is dense in $H^s(\mathbb{R}^n)$ $\forall\ s \in \mathbb{R}$, i.e., for any $v \in H^s(\mathbb{R}^n)$ there exists a sequence $\{v_i\}_{i=1}^\infty \in C_0^\infty(\mathbb{R}^n)$ such that $\|v - v_i\|_{H^s(\mathbb{R}^n)} \to 0$ as $i \to \infty$.*

Theorem 4.1 (trace theorem, half-space case) *For $s > 1/2$ there exists a unique extension of γ to a bounded linear* **trace** *operator*

$$\gamma : \ H^s(\mathbb{R}^n) \to H^{s-1/2}(\mathbb{R}^{n-1}).$$

Proof. Let first $v \in C_0^\infty(\mathbb{R}^n)$. By the Fourier inversion formula we find that

$$\begin{aligned}\gamma v(x') &= \int_{\mathbb{R}^n} e^{i2\pi x\cdot\xi}\hat{v}(\xi)\,d\xi\Big|_{x_n=0} = \int_{\mathbb{R}^n} e^{i2\pi x'\cdot\xi'}\hat{v}(\xi)\,d\xi = \int_{\mathbb{R}^{n-1}}\Big(\int_{\mathbb{R}}\hat{v}(\xi',\xi_n)\,d\xi_n\Big)e^{i2\pi x'\cdot\xi'}\,d\xi' \\ &= \mathcal{F}^{-1}_{\xi'\to x'}\{V(\xi')\}, \quad \text{where} \quad V(\xi') := \int_{-\infty}^{\infty}\hat{v}(\xi',\xi_n)d\xi_n.\end{aligned}$$

Therefore,

$$\mathcal{F}(\gamma v)(\xi') = V(\xi') = \int_{\mathbb{R}}\hat{v}(\xi',\xi_n)\,d\xi_n = \int_{\mathbb{R}}(1+|\xi|^2)^{-s/2}(1+|\xi|^2)^{s/2}\hat{v}(\xi',\xi_n)\,d\xi_n$$

and an application of the Cauchy-Schwarz inequality yields

$$|\mathcal{F}(\gamma v)(\xi')|^2 \le \left[\int_{\mathbb{R}}(1+|\xi|^2)^{-s}\,d\xi_n\right]\left[\int_{\mathbb{R}}(1+|\xi|^2)^{s}|\hat{v}(\xi',\xi_n)|^2\,d\xi_n\right].$$

Now, by the substitution $\xi_n = (1+|\xi'|^2)^{1/2}t$,

$$\int_{\mathbb{R}}(1+|\xi|^2)^{-s}\,d\xi_n = \int_{\mathbb{R}}\frac{d\xi_n}{(1+|\xi'|^2+|\xi_n|^2)^s} = \frac{1}{(1+|\xi'|^2)^{s-1/2}}\int_{\mathbb{R}}\frac{dt}{(1+t^2)^s} = \frac{C_s}{(1+|\xi'|^2)^{s-1/2}},$$

where

$$C_s := \int_{\mathbb{R}}\frac{dt}{(1+t^2)^s} < \infty \quad \text{iff} \quad s > 1/2.$$

Therefore, we can bound

$$(1+|\xi'|^2)^{s-1/2}|\mathcal{F}(\gamma v)(\xi')|^2 \le C_s\int_{\mathbb{R}}(1+|\xi|^2)^s|\hat{v}(\xi)|^2\,d\xi_n.$$

Then integration with respect to ξ' yields

$$\|\gamma v\|_{H^{s-1/2}(\mathbb{R}^{n-1})} \le \sqrt{C_s}\,\|v\|_{H^s(\mathbb{R}^n)}. \tag{4.1}$$

By density of $C_0^\infty(\mathbb{R}^n)$ in $H^s(\mathbb{R}^n)$ this implies the theorem for $v \in H^s(\mathbb{R}^n)$. Indeed, let a sequence $\{v_i\}_{i=1}^\infty \in C_0^\infty(\mathbb{R}^n)$ converges to $v \in H^s(\mathbb{R}^n)$ as $i \to \infty$. Then for any $\epsilon > 0$ there exists $N(\epsilon)$ such that $\|v_i - v_j\|_{H^s(\mathbb{R}^n)} < \epsilon/\sqrt{C_s}$ for any $i,j > N(\epsilon)$ implying also

$$\|\gamma v_i - \gamma v_j\|_{H^{s-1/2}(\mathbb{R}^{n-1})} \le \sqrt{C_s}\,\|v_i - v_j\|_{H^s(\mathbb{R}^n)} < \epsilon$$

This means $\{\gamma v_i\}_{i=1}^\infty$ is a fundamental sequence (Cauchy sequence) in $H^{s-1/2}(\mathbb{R}^{n-1})$ and thus has a limit v^+ in $H^{s-1/2}(\mathbb{R}^{n-1})$ that we call the trace γv of v. We have,

$$\begin{aligned}\|v^+\|_{H^{s-1/2}(\mathbb{R}^{n-1})} &\le \|v^+ - \gamma v_i\|_{H^{s-1/2}(\mathbb{R}^{n-1})} + \|\gamma v_i\|_{H^{s-1/2}(\mathbb{R}^{n-1})} \\ &\le \|v^+ - \gamma v_i\|_{H^{s-1/2}(\mathbb{R}^{n-1})} + \sqrt{C_s}\,\|v_i\|_{H^s(\mathbb{R}^n)} \\ \le \|v^+ - \gamma v_i\|_{H^{s-1/2}(\mathbb{R}^{n-1})} + \sqrt{C_s}\,\|v_i - v\|_{H^s(\mathbb{R}^n)} &+ \sqrt{C_s}\|v\|_{H^s(\mathbb{R}^n)} \to \sqrt{C_s}\|v\|_{H^s(\mathbb{R}^n)} \quad \text{as} \quad i \to \infty.\end{aligned}$$

To prove that v^+ does not depend on the sequence $\{v_i\}_{i=1}^\infty$, let us assume that two sequences, $\{v_i'\}_{i=1}^\infty$ and $\{v_i''\}_{i=1}^\infty$ converging to v produce some traces, v'^+ and v''^+, respectively. Then,

$$\|v'^+ - v''^+\|_{H^{s-1/2}(\mathbb{R}^{n-1})} \leq \|v'^+ - \gamma v_i'\|_{H^{s-1/2}(\mathbb{R}^{n-1})} + \|v''^+ - \gamma v_i''\|_{H^{s-1/2}(\mathbb{R}^{n-1})} + \|\gamma(v_i' - v_i'')\|_{H^{s-1/2}(\mathbb{R}^{n-1})}$$
$$\leq \|v'^+ - \gamma v_i'\|_{H^{s-1/2}(\mathbb{R}^{n-1})} + \|v''^+ - \gamma v_i''\|_{H^{s-1/2}(\mathbb{R}^{n-1})} + \sqrt{C_s}\,\|(v_i' - v_i'')\|_{H^s(\mathbb{R}^n)} \to 0 \quad \text{as} \quad i \to \infty.$$

□

Theorem 4.2 *If $v \in H^s(\mathbb{R}^n) \bigcap C(\mathbb{R}^n)$ for some $s > 1/2$ then the trace of the function v equals its value on the half-space boundary, i.e., $v^+(x') = v(x', 0)$ for all $x' \in \mathbb{R}^{n-1}$.*

Proof. Let first v has a compact support and a sequence $\{v_i\}_{i=1}^\infty \in C_0^\infty(\mathbb{R}^n)$ converges to $v \in H^s(\mathbb{R}^n)$ as $i \to \infty$ generating a trace v^+. One can easily check that for any function from $H^s(\mathbb{R}^n) \bigcap C(\mathbb{R}^n)$ with a compact support all the reasoning of Theorem 4.2 up to estimate (4.1) holds true, which implies,

$$\|v(\cdot, 0) - v^+\|_{H^{s-1/2}(\mathbb{R}^{n-1})} \leq \|v(\cdot, 0) - \gamma v_i\|_{H^{s-1/2}(\mathbb{R}^{n-1})} + \|v^+ - \gamma v_i\|_{H^{s-1/2}(\mathbb{R}^{n-1})}$$
$$\leq \sqrt{C_s}\,\|v - v_i\|_{H^s(\mathbb{R}^n)} + \|v^+ - \gamma v_i\|_{H^{s-1/2}(\mathbb{R}^{n-1})} \to 0 \quad \text{as} \quad i \to \infty.$$

For the function v with non-compact support one can replace v in the previous argument with μv, where $\mu \in C_0^\infty(\mathbb{R}^n)$ has an arbitrarily large but final support and $\mu = 1$ in the vicinity of an arbitrarily chosen part of the boundary. □

So far we have dealt with Sobolev spaces on $\mathbb{R}^n$. For boundary value problems on Lipschitz domains this is obviously not enough.

Definition 4.2 *Let $\Omega \subset \mathbb{R}^n$ be a Lipschitz domain. For $s \geq 0$ we introduce the following spaces:*

$$H^s(\Omega) := H^s(\mathbb{R}^n)\Big|_\Omega \quad \textit{with norm} \quad \|v\|_{H^s(\Omega)} := \inf_{V|_\Omega = v} \|V\|_{H^s(\mathbb{R}^n)},$$

$$H_0^s(\Omega) := \overline{C_0^\infty(\Omega)}^{\|\cdot\|_{H^s(\Omega)}} \quad \textit{with norm} \quad \|\cdot\|_{H^s(\Omega)},$$

and

$$\tilde{H}^s(\Omega) := \{v \in H^s(\Omega);\ v^0 \in H^s(\mathbb{R}^n)\} \quad \textit{with norm} \quad \|v\|_{\tilde{H}^s(\Omega)} := \|v^0\|_{H^s(\mathbb{R}^n)}$$

where v^0 denotes the extension of v by 0 onto $\mathbb{R}^n \setminus \bar{\Omega}$.

For $s < 0$ we define

$$H^s(\Omega) := \left(\tilde{H}^{-s}(\Omega)\right)' \quad \textit{(dual space)} \quad \textit{with operator norm}$$

and

$$\tilde{H}^s(\Omega) := \left(H^{-s}(\Omega)\right)' \quad \textit{(dual space)} \quad \textit{with operator norm.}$$

Remark 4.1 *One can show that, for $s > 0$, $\tilde{H}^s(\Omega) = H^s_0(\Omega)$ if $s \neq$ integer $+ 1/2$. In the cases $s =$ integer $+ 1/2$ the spaces are different, $\tilde{H}^s(\Omega) \subset H^s_0(\Omega)$ in general.*

Without going into the details, we mention that on a Lipschitz surface or boundary Γ all the above spaces can be defined analogously when $|s| \leq 1$. To this end one uses a partition of unity and local transformations onto subsets of $\mathbb{R}^{n-1}$. Higher order spaces require more regularity of Γ.

The trace theorem can be generalised to Lipschitz domains.

Theorem 4.3 (trace theorem, general form)
Let $\Omega \subset \mathbb{R}^n$ be a bounded Lipschitz domain with boundary Γ.
(i) *For $1/2 < s < 3/2$, γ has a unique extension to a bounded linear operator*

$$\gamma : \; H^s(\Omega) \to H^{s-1/2}(\Gamma).$$

(ii) *For $1/2 < s < 3/2$ and any $v \in H^{s-1/2}(\Gamma)$ there exists $V := \gamma_{-1}v \in H^s(\Omega)$ such that $\gamma(V) = v$ and*

$$\|\gamma_{-1}v\|_{H^s(\Omega)} \leq C_s(\Omega)\, \|v\|_{H^{s-1/2}(\Gamma)} \qquad \forall v \in H^{s-1/2}(\Gamma).$$

Remark 4.2 *Part* (ii) *of Theorem* 4.3 *means that γ has a right-inverse:*

$$v = \gamma V = \gamma\gamma_{-1}v \qquad \forall v \in H^{s-1/2}(\Gamma)$$

which is continuous, and that

$$\gamma : \; H^s(\Omega) \to H^{s-1/2}(\Gamma)$$

is surjective, i.e., $\gamma\Big(H^s(\Omega)\Big) = H^{s-1/2}(\Gamma)$. Of course, this right-inverse γ_{-1} is an extension operator (from boundary to the domain), and it is not unique.

Definition 4.3 *Let $u \in H^1(\Omega)$ and $f \in H^{-1}(\Omega)$. We say*

$$-\Delta u = f \quad in \quad \Omega$$

if

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \langle f, v\rangle_\Omega \quad \forall v \in \tilde{H}^1(\Omega),$$

where $\langle f, v\rangle_\Omega$ means the value of the functional $f \in H^{-1}(\Omega)$ applied to the function $u \in H^1(\Omega)$, in other words, it is the duality form between $H^{-1}(\Omega)$ and $\tilde{H}^1(\Omega)$.

Having the trace operator at hand we can now make an interpretation of the Dirichlet boundary condition. Studying the Dirichlet boundary value problem for the Poisson equation,

$$-\Delta u = f \quad \text{in } \Omega, \qquad u|_\Gamma = g_D$$

we understand it in the weak sense as follows:

For $f \in H^{-1}(\Omega)$ and $g_D \in H^{1/2}(\Gamma)$, find $u \in H^1(\Omega)$ such that

$$a(u,v) = L_D(v) \quad \forall v \in \tilde{H}^1(\Omega), \quad \gamma u = g_D,$$

where

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

Note that the Dirichlet condition makes sense only for $g_D \in H^{1/2}(\Gamma)$. If $g_D \notin H^{1/2}(\Gamma)$ then there does not exist a solution $u \in H^1(\Omega)$ of the given boundary value problem. This is a conclusion of the surjectivity of the trace operator.

Besides the trace operator γ, in Section 1 we were concerned about the definition of the normal derivative $\partial_n v$ of a function $v \in H^1(\Omega)$. We now deal with this operator.

First of all, if $u \in H^2(\Omega)$, then $\nabla u \in H^1(\Omega)$, which implies that ∇u has a trace $\gamma \nabla u \in H^{1/2}(\Gamma)$ and we can define the normal derivative in the classical way, $\partial_n u = \gamma \nabla u \cdot n$ on Γ. This approach does not work for $u \in H^1(\Omega)$ since then $\nabla u \in H^0(\Omega) = L_2(\Omega)$, which implies ∇u may not have a trace.

The origin for the generalised definition of the normal derivative is the first Green's identity,

$$\int_\Omega -\Delta u \, w \, dx = \int_\Omega \nabla u \cdot \nabla w \, dx - \int_\Gamma \partial_n u \, w \, ds,$$

which works for sufficiently smooth u. Using it as a hint, we can now define $\partial_n u$ for $u \in H^1(\Omega)$ such that $\Delta u \in L_2(\Omega)$ by

$$\langle \partial_n u, w \rangle_\Gamma := \int_\Omega \nabla u \cdot \nabla W \, dx + \int_\Omega \Delta u \, W \, dx$$

where $W \in H^1(\Omega)$ is any extension of $w \in H^{1/2}(\Gamma)$. The notation $\langle \Phi, \varphi \rangle_\Gamma$ is the duality form between $H^{-1/2}(\Gamma)$ and $H^{1/2}(\Gamma)$. For Φ, $\varphi \in L_2(\Gamma)$ it is simply the $L_2(\Gamma)$-inner product between Φ and φ.

When Δu belongs to a more general space, the analog of $\partial_n u$ cam be defined as follows.

Definition 4.4 *Let $u \in H^1(\Omega)$ and $\tilde{f} \in \tilde{H}^{-1}(\Omega)$ is such that $-\Delta u = \tilde{f}|_\Omega$ in Ω. The generalised normal derivative $\partial_n(u; \tilde{f}) \in H^{-1/2}(\Gamma)$ is defined as*

$$\langle \partial_n(u; \tilde{f}), w \rangle_\Gamma := \int_\Omega \nabla u \cdot \nabla W \, dx - \langle \tilde{f}, W \rangle_\Omega,$$

where $\langle \tilde{f}, W \rangle_\Omega$ is the duality form between $\tilde{H}^{-1}(\Omega)$ and $H^1(\Omega)$ and $W \in H^1(\Omega)$ is an extension of $w \in H^{1/2}(\Gamma)$.

The following lemma can be proved.

Lemma 4.2 *The generalised normal derivative operator does not depend on the extension W and*

$$\|\partial_n(u;\tilde{f})\|_{H^{-1/2}(\Gamma)} \leq C\left(\|u\|_{H^1(\Omega)} + \|\tilde{f}\|_{\tilde{H}^{-1}(\Omega)}\right).$$

Remark 4.3 *For a given function $u \in H^1(\Omega)$, the extension of the functional $f = \Delta u \in H^{-1}(\Omega)$ to a functional $\tilde{f} \in \tilde{H}^{-1}(\Omega)$ is generally not unique and thus the generalised normal derivative $\partial_n(u;\tilde{f})$ depends on the choice of the extension $\tilde{f}$.*

Remark 4.4 *If $-\Delta u = f \in L_2(\Omega) = H^0(\Omega)$, then its natural (canonical) extension $\tilde{f} = \begin{cases} f & \text{in } \Omega \\ 0 & \text{in } \mathbb{R}^n \backslash \Omega \end{cases}$ belongs to $\tilde{H}^0(\Omega) \subset \tilde{H}^{-1}(\Omega)$. Although other extensions of f belonging to $\tilde{H}^{-1}(\Omega)$ are also possible, the canonical extension makes the generalised normal derivative $\partial_n(u;\tilde{f})$ (with the canonical extension $\tilde{f}$) equal to the classical normal derivative $\partial_n u$ if u is smooth enough.*

The Neumann boundary value problem for the Poisson equation

$$-\Delta u = \tilde{f}|_\Omega \quad \text{in } \Omega, \qquad \partial_n(u;\tilde{f}) = g_N$$

we understand it in the weak sense as follows:

For $\tilde{f} \in \tilde{H}^{-1}(\Omega)$ and $g_N \in H^{-1/2}(\Gamma)$, find $u \in H^1(\Omega)$ such that

$$a(u,v) = L_N(v) \quad \forall v \in H^1(\Omega),$$

where

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v \, dx, \qquad L_N(v) = \langle \tilde{f}, v\rangle_\Omega + \langle \partial_n(u;\tilde{f}), \gamma v\rangle_\Gamma.$$

5 Approximation theory and finite element error analysis for elliptic problems

In this section we deal with the error analysis of the finite element method. Key steps in the error analysis are the Lax-Milgram lemma (Theorem 2.1), which proves the unique existence of u_h and its stability, and Céa's lemma (Theorem 3.2) proving

$$\|u - u_h\| \leq \frac{C_a}{\alpha}\|u - v\| \quad \forall v \in \mathcal{H}_h.$$

Here, several conditions are needed to be met, in particular the boundedness of a (with bound C_a) and its $\mathcal{H}$-ellipticity (with ellipticity constant α). Then, to bound the error in the energy norm (or the norm in $\mathcal{H}$) we only need to select an appropriate function $v \in \mathcal{H}_h$ for which we are able to further estimate $\|u - v\|$. If $\mathcal{H}_h$ consists of continuous, piecewise linear functions then a standard candidate is the piecewise linear interpolant $I_h u \in \mathcal{H}_h$ (defined below). First, in Section 5.1, we deal with approximation theory in a more general and abstract form. Then, in Section 5.2, we apply the approximation results to the finite element method.

5.1 Approximation theory

Definition 5.1 *Let X, Y be normed linear spaces, and $A \in \mathcal{L}(X,Y)$, where $\mathcal{L}(X,Y)$ denotes the space of bounded linear operators $X \to Y$. Then, A is* **compact** *if and only if the sequence $\{Ax_n\}_{n\in\mathbb{N}} \subset Y$ has a convergent subsequence for any bounded sequence $\{x_n\}_{n\in\mathbb{N}} \subset X$.*

This can be equivalently formulated as: A is compact if and only if every bounded subset of X is mapped to a relatively compact subset of Y.

Proposition 5.1 (Rellich's embedding theorem) *Let Ω be a bounded Lipschitz domain. Then for any $t > s$, the injection $i\colon H^t(\Omega) \to H^s(\Omega)$ is compact.*

Proposition 5.2 (Sobolev's embedding theorem) *Let Ω be a Lipschitz domain in $\mathbb{R}^n$. Then, the injection $i\colon H^{n/2+\varepsilon}(\Omega) \to C^0(\bar{\Omega})$ is continuous for all $\varepsilon > 0$, that is,*

$$\|u\|_\infty = \operatorname*{ess\,sup}_{x\in\Omega} |u(x)| \le C_\varepsilon \|u\|_{H^{n/2+\varepsilon}(\Omega)} \quad \textit{for all } u \in H^{n/2+\varepsilon}(\Omega).$$

The above proposition implies that for any function $u \in H^{n/2+\varepsilon}(\Omega)$ there exists a function $u_* \in C^0(\bar{\Omega})$ such that $u(x) = u_*(x)$ at almost every $x \in \bar{\Omega}$.

Remark 5.1 *A simple argument for the influence of the dimension n is the following: Using the Fourier transform, we see for $u \in C_0^\infty(\mathbb{R}^n)$ that*

$$\begin{aligned} |u(x)| &= \left| \int_{\mathbb{R}^n} \hat{u}(\xi) e^{2\pi i x\cdot\xi}\, d\xi \right| \le \int_{\mathbb{R}^n} |\hat{u}(\xi)|\, d\xi = \int_{\mathbb{R}^n} (1+|\xi|^2)^{-s/2} (1+|\xi|)^{s/2} |\hat{u}(\xi)|\, d\xi \\ &\le \left(\int_{\mathbb{R}^n} (1+|\xi|^2)^{-s}\, d\xi \right)^{1/2} \|u\|_{H^s(\mathbb{R}^n)}, \end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality. Thus, $\int_{\mathbb{R}^n} (1+|\xi|^2)^{-s}\, d\xi < \infty$ will be sufficient. As the integrand is bounded on every bounded set, we only need to study the behaviour as $|\xi| \to \infty$. Transforming to polar coordinates and choosing some $r^* > 0$,

$$\int_{\mathbb{R}^n} (1+|\xi|^2)^{-s}\, d\xi \sim \int_{r^*}^\infty r^{-2s} r^{n-1}\, dr = \int_{r^*}^\infty r^{n-2s-1}\, dr.$$

The last integral is finite if and only if $n-2s-1 < -1$. This corresponds exactly to the condition $s > \frac{n}{2}$.

For an integer $k > 0$, let us define the semi-norm in the Sobolev space $H^k(\Omega)$ as

$$|u|_{H^k(\Omega)} := \left(\sum_{|\alpha|=k} \int_\Omega |\partial^\alpha u(x)|^2 dx \right)^{1/2}.$$

Then

$$\|u\|^2_{H^k(\Omega)} = \|u\|^2_{H^{k-1}(\Omega)} + |u|^2_{H^k(\Omega)} = \sum_{l=0}^k |u|^2_{H^l(\Omega)}.$$

Lemma 5.1 *Let $\Omega \subset \mathbb{R}^2$ be a Lipschitz domain, $k > 1$ integer, $s = \frac{k(k+1)}{2}$, and $\{z_1, z_2, \ldots, z_s\} \subset \Omega$ be given points such that the interpolation operator $\mathcal{I}\colon H^k(\Omega) \to P_{k-1}$ over these points is well-defined. Here, P_{k-1} are the polynomials of degree up to $k-1$. Then, there exists $C \geq 0$ such that*

$$\|u - \mathcal{I}u\|_{H^k(\Omega)} \leq C|u|_{H^k(\Omega)} \quad \textit{for all } u \in H^k(\Omega).$$

Proof. We first prove that $\|v\|_{H^k(\Omega)}$ and $|||v||| := |v|_{H^k(\Omega)} + \sum_{i=1}^{s} |v(z_i)|$ are equivalent norms. Then it follows that

$$\begin{aligned} \|u - \mathcal{I}u\|_{H^k(\Omega)} &\leq C\,|||u - \mathcal{I}u||| = C\left(|u - \mathcal{I}u|_{H^k(\Omega)} + \sum_{i=1}^{s} |u(z_i) - (\mathcal{I}u)(z_i)|\right) \\ &= C|u - \mathcal{I}u|_{H^k(\Omega)} = C|u|_{H^k(\Omega)}, \end{aligned}$$

since the k-th derivatives of $\mathcal{I}u \in P_{k-1}$ vanish.

1. As $k > 0$ we see by Proposition 5.2 that the injection $H^k(\Omega) \to C^0(\bar{\Omega})$ is continuous. Thus, $|v(z_i)| \leq \|v\|_\infty \leq C\|v\|_{H^k(\Omega)}$, $i = 1, \ldots, s$, and

$$|||v||| = |v|_{H^k(\Omega)} + \sum_{i=1}^{s} |v(z_i)| \leq (1 + sC)\|v\|_{H^k(\Omega)} \quad \text{for all } v \in H^k(\Omega).$$

2. Let us prove that there exists a constant $C > 0$ such that $\|v\|_{H^k(\Omega)} \leq C\,|||v|||$ for any $v \in H^k(\Omega)$. Assume the contrary, i.e., that for any constant $C > 0$ there exists $v \in H^k(\Omega)$ such that $\|v\|_{H^k(\Omega)} > C\,|||v|||$ and consider the sequence $C_n = n$, $n \in \mathbb{N}$. Then, there exists a sequence $\{v_n\}_{n\in\mathbb{N}} \subset H^k(\Omega)$ such that $\|v_n\|_{H^k(\Omega)} = 1$ and $|||v_n||| \leq \frac{1}{n}$ for all $n \in \mathbb{N}$. Since $\{v_n\}$ is bounded in $H^k(\Omega)$, then by Proposition 5.1 there exists a subsequence of $\{v_n\}$ which converges in $H^{k-1}(\Omega)$. We assume without loss of generality that this subsequence is $\{v_n\}$. In particular, it follows that $\{v_n\}$ is a Cauchy sequence in $H^{k-1}(\Omega)$ and thus,

$$\begin{aligned} \|v_m - v_l\|^2_{H^k(\Omega)} &= \|v_m - v_l\|^2_{H^{k-1}(\Omega)} + |v_m - v_l|^2_{H^k(\Omega)} \\ &\leq \|v_m - v_l\|^2_{H^{k-1}(\Omega)} + (|v_m|_{H^k(\Omega)} + |v_l|_{H^k(\Omega)})^2 \to 0 \quad \text{for } m, l \to \infty \end{aligned}$$

since $|v_n|_{H^k(\Omega)} \leq |||v_n||| \to 0$ for $n \to \infty$. Therefore, $\{v_n\}$ is a Cauchy sequence in $H^k(\Omega)$ and by completeness of $H^k(\Omega)$ there exists $v^* \in H^k(\Omega)$ such that $v_n \to v^*$ in $H^k(\Omega)$ for $n \to \infty$. By the continuity of the norms it follows from $\|v_n\|_{H^k(\Omega)} = 1$ that $\|v^*\|_{H^k(\Omega)} = 1$, and from $|||v_n||| \leq \frac{1}{n}$ that $|||v^*||| = 0$ since, by the first part,

$$|||v^*||| \leq |||v^* - v_n||| + |||v_n||| \leq C\|v^* - v_n\|_{H^k(\Omega)} + |||v_n||| \to 0 \quad \text{as } n \to \infty.$$

By definition of $|||\cdot|||$ it follows that $|v^*|_{H^k(\Omega)} = 0$, that is, $v^* \in P_{k-1}$, and $|v^*(z_i)| = 0$, $i = 1, \ldots, s$. Therefore, $v^* = 0$ as a polynomial of degree $k-1$ vanishing at $\frac{k(k+1)}{2}$ distinct points (such that the interpolation by a polynomial of degree $k-1$ is well defined), which is a contradiction to $\|v^*\|_{H^k(\Omega)} = 1$.

Therefore, there exists a constant C such that $\|v\|_{H^k(\Omega)} \leq C\,|||v|||$.

□

Lemma 5.1 is a special case of the following more general statement.

Theorem 5.1 (Bramble-Hilbert Lemma) *Let $\Omega \subset \mathbb{R}^2$ be a Lipschitz domain, and $k > 1$ integer. For a normed linear space Y let $L \in \mathcal{L}(H^k(\Omega), Y)$.*

If $P_{k-1} \subset \ker L$ (i.e., $Lv = 0 \ \forall v \in P_{k-1}$) then there exists a constant $C \geq 0$ such that

$$\|Lv\|_Y \leq C|v|_{H^k(\Omega)} \quad \textit{for all } v \in H^k(\Omega).$$

Proof. As L is bounded and linear there exists a constant $D \geq 0$ such that $\|Lv\|_Y \leq D\|v\|_{H^k(\Omega)}$ for all $v \in H^k(\Omega)$. Let $\mathcal{I}: \ H^k(\Omega) \to P_{k-1}$ be an interpolation operator as in Lemma 5.1. Then, $\mathcal{I}v \in P_{k-1} \subset \ker L$ for all $v \in H^k(\Omega)$ and

$$\|Lv\|_Y = \|L(v - \mathcal{I}v)\|_Y \leq D\|v - \mathcal{I}v\|_{H^k(\Omega)} \leq CD|v|_{H^k(\Omega)}$$

by Lemma 5.1. □

5.2 Finite element error estimate for elliptic problems

We deal with the case $\mathcal{H} = H^1(\Omega)$ and $\mathcal{H}_h = \{v \in \mathcal{H} : \ v|_T \in P_{k-1}(T) \ \forall T \in \mathcal{T}_h\}$ where $\mathcal{T}_h = \{T\}$ is a triangulation of a polygonal domain Ω (so that it can be discretised by triangular meshes). Here, $P_{k-1}(T)$ denotes the space of polynomials of degree $k-1$ on T. The mesh needs to satisfy certain conditions. We define

$$\begin{aligned} h_T &= \text{diameter of } T = \text{length of longest side of } T, \\ \rho_T &= \text{diameter of the largest circle in } T, \\ h &= \max_{T \in \mathcal{T}_h} h_T \end{aligned}$$

and require that $\mathcal{T}_h$ is **shape regular**, i.e., there exists $\beta > 0$ which is independent of h such that

$$\frac{\rho_T}{h_T} \geq \beta \qquad \forall T \in \mathcal{T}_h. \tag{5.1}$$

This means that the elements $T \in \mathcal{T}_h$ are not too thin, i.e. the interior angles of T are not too small (they are bounded from below by a positive constant). Since we are interested in the behaviour of the finite element error $\|u - u_h\|$ on a sequence of meshes $\{\mathcal{T}_h\}$ with decreasing mesh sizes h, the constant β in (5.1) must be independent of h.

We now apply the Bramble-Hilbert Lemma to prove a piecewise polynomial approximation result.

Theorem 5.2 *For a Lipschitz domain $\Omega \subset \mathbb{R}^2$ with polygonal boundary and a given integer $k > 1$ let $\{T : T \in \mathcal{T}_h\}$ be a shape regular triangulation of Ω with $h < 1$.*

Then, for a piecewise polynomial interpolation operator $\mathcal{I}_h$ of degree $k-1$ (piecewise with respect to $\mathcal{T}_h$) there holds

$$\left(\sum_{T \in \mathcal{T}_h} \|u - \mathcal{I}_h u\|^2_{H^m(T)} \right)^{1/2} \leq C h^{k-m} |u|_{H^k(\Omega)} \quad \text{for all } u \in H^k(\Omega) \text{ and all } 0 \leq m \leq k.$$

Here, the constant C is independent of h and u.

Proof. The idea of the proof is to transform to the reference element $\check{T}$, make a transition from H^m to H^k, and transform back. The transformations give the required powers of h since the Bramble-Hilbert Lemma gives the transition to a semi-norm on $\check{T}$.

By the assumption of shape regularity it is enough to consider the case that all T are similar to $\check{T}$. Then we can assume without loss of generality that $T = h_T \check{T} := \{(x_1, x_2) : 0 \leq x_1, x_2 \leq h_T, x_1 + x_2 \leq h_T\}$. For $v \in H^k(T)$ we define $\check{v} \in H^k(\check{T})$ by $\check{v}(\xi_1, \xi_2) := v(h_T \xi_1, h_T \xi_2)$. For a multi-index $\alpha = (\alpha_1, \alpha_2)$ of order $|\alpha| = \alpha_1 + \alpha_2$ with non-negative integers α_1, α_2 let D^α denote the partial derivative operator defined by

$$D^\alpha v(x_1, x_2) := \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2}} v(x_1, x_2).$$

We see that $D^\alpha v = h_T^{-|\alpha|} D^\alpha \check{v}$ for all multi-indices α. Thus,

$$\begin{aligned} |v|^2_{H^l(T)} &= \sum_{|\alpha|=l} \int_T (D^\alpha v)^2 \, dx = \sum_{|\alpha|=l} h_T^{-2l} h_T^2 \int_{\check{T}} (D^\alpha \check{v})^2 \, d\xi = h_T^{2-2l} |\check{v}|^2_{H^l(\check{T})}, \\ \|v\|^2_{H^m(T)} &= \sum_{l=0}^m h_T^{2-2l} |\check{v}|^2_{H^l(\check{T})} = h_T^{2-2m} \sum_{l=0}^m h_T^{2(m-l)} |\check{v}|^2_{H^l(\check{T})} \leq h_T^{2-2m} \|\check{v}\|^2_{H^m(\check{T})}. \end{aligned}$$

Since the transform $\check{\mathcal{I}}$ of the interpolation operator $\mathcal{I}_h$ is again an interpolation operator, we can transform to the reference element, apply the Bramble-Hilbert Lemma and transform back to obtain

$$\begin{aligned} \|v - \mathcal{I}_h v\|^2_{H^m(T)} &\leq h_T^{2-2m} \|\check{v} - \check{\mathcal{I}}\check{v}\|^2_{H^m(\check{T})} \leq h_T^{2-2m} \|\check{v} - \check{\mathcal{I}}\check{v}\|^2_{H^k(\check{T})} \\ &\leq C h_T^{2-2m} |\check{v}|^2_{H^k(\check{T})} = C h_T^{2-2m} h_T^{2k-2} |v|^2_{H^k(T)} = C h_T^{2(k-m)} |v|^2_{H^k(T)}. \end{aligned}$$

Summing up this yields

$$\sum_{T \in \mathcal{T}_h} \|v - \mathcal{I}_h v\|^2_{H^m(T)} \leq C \sum_{T \in \mathcal{T}_h} h_T^{2(k-m)} |v|^2_{H^k(T)} \leq C h^{2(k-m)} |v|^2_{H^k(\Omega)}.$$

□

Remark 5.2 *Note that in Theorem* 5.2 *we cannot write* $\|u-\mathcal{I}_h u\|_{H^m(\Omega)}$ *in general since* $u-\mathcal{I}_h u$ *might not be in* $H^m(\Omega)$. *The operator* $\mathcal{I}_h$ *represents only a piecewise interpolation from which, in general, no global regularity properties follow.*

Let N_i, $i = 1, \ldots, M$, be the nodes of $\mathcal{T}_h$. For a continuous function $u \in C^0(\bar{\Omega})$ we now consider the piecewise linear interpolant (again using the same operator symbol $\mathcal{I}_h$) $\mathcal{I}_h u \in \mathcal{H}_h$ by

$$\mathcal{I}_h u(N_i) = u(N_i), \qquad i = 1, \ldots, M. \tag{5.2}$$

Note that on any $T \in \mathcal{T}_h$, $\mathcal{I}_h u$ is the linear interpolant of u.

Corollary 5.1 *Let* $\Omega \subset \mathbb{R}^2$ *be a polygon with a quasi-uniform and shape-regular mesh with* $h < 1$, *and* $\mathcal{I}_h$ *be the piecewise linear interpolation operator (piecewise with respect to the triangulation) over the vertices of the mesh.*

Then,

$$\|u - \mathcal{I}_h u\|_{H^1(\Omega)} \le Ch|u|_{H^2(\Omega)} \quad \textit{for all } u \in H^2(\Omega).$$

Proof. As $\mathcal{I}_h$ interpolates over the vertices of the mesh, $\mathcal{I}_h u$ is continuous and piecewise linear, i.e. $\mathcal{I}_h u \in H^1(\Omega)$. An application of Theorem 5.2 proves

$$\|u - \mathcal{I}_h u\|^2_{H^1(\Omega)} = \sum_{T \in \mathcal{T}_h} \|u - \mathcal{I}_h u\|^2_{H^1(T)} \le Ch^2|u|^2_{H^2(\Omega)}.$$

□

Now we are in a position to present an a priori error estimate for the finite element method dealing with elliptic problems of second order. Assume that we are solving a variational problem in $\mathcal{H} = H^1_0(\Omega)$ ($\Omega \subset \mathbb{R}^2$ is a Lipschitz-continuous polygonal domain),

$$u \in \mathcal{H}: \quad a(u, v) = L(v) \quad \forall v \in \mathcal{H}, \tag{5.3}$$

where a is a continuous, $\mathcal{H}$-elliptic bilinear form, and L is a continuous linear form on $\mathcal{H}$. We then consider the finite element approximation u_h to u defined by

$$u_h \in \mathcal{H}_h: \quad a(u_h, v) = L(v) \quad \forall v \in \mathcal{H}_h. \tag{5.4}$$

Selecting any finite-dimensional subspace $\mathcal{H}_h \subset \mathcal{H}$ there holds Céa's lemma. In particular, selecting $\mathcal{H}_h$ to be the space of continuous, piecewise linear functions defined on a mesh $\mathcal{T}_h$ satisfying the shape-regularity condition (5.1) there holds (applying Céa's lemma)

$$\|u - u_h\|_{H^1(\Omega)} \le \frac{C_a}{\alpha}\|u - \mathcal{I}_h u\|_{H^1(\Omega)}. \tag{5.5}$$

Here, $\mathcal{I}_h$ is the interpolation operator defined in (5.2).

Therefore, applying Corollary 5.1, we conclude that there holds the following **a priori error estimate**.

Theorem 5.3 (a priori error estimate)
Assume that the solution u of (5.3) *satisfies $u \in H^2(\Omega)$ and that $u_h \in \mathcal{H}_h$ is the finite element approximation defined by* (5.4) *(using piecewise linear functions on a shape regular mesh). Then there exists a constant $C > 0$ which is independent of h such that*

$$\|u - u_h\|_{H^1(\Omega)} \leq C\, h\, |u|_{H^2(\Omega)}. \tag{5.6}$$

This means that u_h converges linearly in h to u in the $H^1(\Omega)$-norm.

6 Literature

- Classical FEM texts: [2, 4]
- More modern text books: [6, 3]
- FEM directed at engineers (there are many more): [8]
- Elliptic problems, Sobolev spaces: [5, 1, 7]

References

[1] R. Adams, **Sobolev Spaces**, Academic press, 2003, 305p.

[2] I. Babuška and A. K. Aziz, **Survey lectures on the mathematical foundations of the finite element method**, in The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, A. K. Aziz, ed., New York, 1972, Academic Press, pp. 5–359.

[3] S. C. Brenner and L. R. Scott, **The Mathematical Theory of Finite Element Methods**, no. 15 in Texts in Applied Mathematics, Springer-Verlag, New York, 1994.

[4] P. G. Ciarlet, **The Finite Element Method for Elliptic Problems**, North-Holland, Amsterdam, 1978.

[5] P. Grisvard, **Elliptic Problems in Nonsmooth Domains**, Pitman Publishing Inc., Boston, 1985.

[6] M. Křížek and P. Neittaanmäki, **Finite Element Approximation of Variational Problems and Applications**, no. 50 in Pitman Monographs and Surveys in Pure and Applied Mathematics, Longman Scientific & Technical, Harlow, England, 1990.

[7] W. McLean, **Strongly Elliptic Systems and Boundary Integral Equations**. Cambridge University Press, Cambridge, UK, 2000.

[8] B. Szabó and I. Babuška, **Finite Element Analysis**, Wiley, New York, 1991.