Numerics of Partial Differential Equations: Stationary Problems

Lecture Notes

Michael Feischl and Dirk Praetorius

November 11, 2020

Chapter 1

Introduction

1.1 Strong Form and Variational Form

The finite element method is a scheme for the numerical solution of partial differential equations. In this chapter, we introduce the basic concepts for elliptic problems in the frame of the Riesz theorem. To that end, we consider the most standard example, namely the Poisson equation with mixed Dirichlet-Neumann boundary conditions. We aim to solve

$$\begin{aligned} -\Delta u &= f \qquad \text{in } \Omega, \\ u &= 0 \quad \text{on } \Gamma_D, \\ \partial u/\partial n &= \phi \quad \text{on } \Gamma_N, \end{aligned} \tag{1.1}$$

which is said to be the **strong form** of the boundary value problem. Here, Ω denotes a domain in $\mathbb{R}^d$, $d = 2, 3$. The boundary $\Gamma := \partial\Omega$ is split into the Dirichlet boundary Γ_D and the Neumann boundary Γ_N, respectively. To be more precise, we assume that Γ_D and Γ_N are (relatively) open subsets of Γ with $\Gamma_D \cap \Gamma_N = \emptyset$ and $\Gamma = \overline{\Gamma}_D \cup \overline{\Gamma}_N$. The source term $f : \Omega \to \mathbb{R}$ as well as the Neumann data $\phi : \Gamma_N \to \mathbb{R}$ are given, and $u : \Omega \to \mathbb{R}$ is the unknown solution. Moreover,

$$\Delta u(x) := \sum_{j=1}^{d} \frac{\partial^2 u}{\partial x_j^2}(x) \tag{1.2}$$

denotes the Laplace operator, which is defined in the classical sense for a function $u \in C^2(\overline{\Omega})$, where $C^k(\overline{\Omega}) := \big\{ w|_{\overline{\Omega}} \,\big|\, w \in C^k(\mathbb{R}^d) \big\}$. If $u \in C^2(\overline{\Omega})$ solves (1.1), u is said to be a **strong solution** of the mixed boundary value problem.

Throughout the lecture, we shall assume that Ω is a **Lipschitz domain** in $\mathbb{R}^d$, i.e.,

- Ω is a bounded, open, and connected subset of $\mathbb{R}^d$,
- Ω is locally on one side of Γ,
- Γ can locally be parametrized by Lipschitz continuous functions.

An important consequence of this assumption is the validity of the **integration by parts formula**

$$\int_\Omega \frac{\partial u}{\partial x_j}\, v \, dx + \int_\Omega u \, \frac{\partial v}{\partial x_j} \, dx = \int_\Gamma u v n_j \, ds \quad \text{for all } u, v \in C^1(\overline{\Omega}), \tag{1.3}$$

where n_j denotes the j-th component of the outer normal vector of Ω on Γ and where ds denotes the surface measure on Γ. For a precise definition and details, we refer, e.g., to [McL].

Let $u \in C^2(\overline{\Omega})$ be a strong solution of (1.1) and $v \in C^1_D(\overline{\Omega}) := \big\{w \in C^1(\overline{\Omega}) \,\big|\, w|_{\Gamma_D} = 0\big\}$. Multiplication of $-\Delta u = f$ by v, integration over Ω, and integration by parts yield that

$$\int_\Omega f v \, dx = -\int_\Omega (\Delta u) v \, dx = -\sum_{j=1}^d \int_\Omega \frac{\partial^2 u}{\partial x_j^2} \, v \, dx = \sum_{j=1}^d \Big[\int_\Omega \frac{\partial u}{\partial x_j} \, \frac{\partial v}{\partial x_j} \, dx - \int_\Gamma \frac{\partial u}{\partial x_j} \, v n_j \, ds \Big].$$

With $x \cdot y = \sum_{j=1}^d x_j y_j$ the usual scalar product in $\mathbb{R}^d$, we obtain the **first Green formula**

$$\int_\Omega f v \, dx = \int_\Omega \nabla u \cdot \nabla v \, dx - \int_\Gamma \frac{\partial u}{\partial n} \, v \, ds, \tag{1.4}$$

where we have used $\nabla u \cdot n = \partial u / \partial n$. Together with $v|_{\Gamma_D} = 0$ and $\Gamma_N = \Gamma \backslash \overline{\Gamma}_D$, we may plug-in the Neumann data to see that

$$\int_\Omega f v \, dx = \int_\Omega \nabla u \cdot \nabla v \, dx - \int_{\Gamma_N} \frac{\partial u}{\partial n} \, v \, ds = \int_\Omega \nabla u \cdot \nabla v \, dx - \int_{\Gamma_N} \phi v \, ds.$$

Altogether we thus have proven the following proposition:

Proposition 1.1. *Let $u \in C^2(\overline{\Omega})$ solve the strong form (1.1). Then, it holds that*

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx + \int_{\Gamma_N} \phi v \, ds \quad \text{for all } v \in C^1_D(\overline{\Omega}), \tag{1.5}$$

*which is the **variational form** of the boundary value problem (1.1).* ■

This proposition gives a necessary condition for a function u to solve the strong form (1.1). We stress that any strong solution belongs to $C^1_D(\overline{\Omega})$ and that the variational form (1.5) can be understood for $u \in C^1_D(\overline{\Omega})$. This leads to a symmetric variational formulation: Find $u \in C^1_D(\overline{\Omega})$ such that (1.5) holds.

Exercise 1. Prove the following well-known integral formulae:

- For $f \in C^1(\Omega)^d$, let $\operatorname{div} f := \sum_{j=1}^d \frac{\partial f_j}{\partial x_j}$ denote the divergence operators. Then, there holds the **Gauss divergence theorem**

$$\int_\Omega \operatorname{div} f \, dx = \int_\Gamma f \cdot n \, ds \quad \text{for all } f \in C^1(\overline{\Omega})^d. \tag{1.6}$$

- Besides the first Green formula, there holds the **second Green formula**

$$\int_\Omega (-\Delta u) v \, dx + \int_\Gamma \frac{\partial u}{\partial n} \, v \, ds = \int_\Omega u(-\Delta v) \, dx + \int_\Gamma u \, \frac{\partial v}{\partial n} \, ds \quad \text{for all } u, v \in C^2(\overline{\Omega}). \tag{1.7}$$

Both are easily obtained from the integration by parts formula. □

1.2 Solvability of Variational Form

To look for solutions of the weak form (1.5), we will employ the following Riesz theorem.

Theorem 1.2 (Riesz). *For a Hilbert space H (over $\mathbb{R}$), the mapping*

$$I_H : H \to H^*, \quad I_H(u) := (u\,;\,\cdot)_H \tag{1.8}$$

is linear, isometric, and bijective, i.e., for any $F \in H^*$ there is a unique $u \in H$ such that

$$(u\,;\,v)_H = F(v) \quad \text{for all } v \in H. \tag{1.9}$$

Moreover, it holds that $\|u\|_H = \|F\|_{H^*}$. ■

The proof of this theorem can be found in each textbook of functional analysis.

First, we observe that the left-hand side

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

of the variational form (1.5) defines a scalar product on $C^1_D(\overline{\Omega})$, provided the Dirichlet boundary Γ_D is nontrivial: Clearly, $(u\,;\,v)$ is a symmetric bilinear form on $C^1_D(\overline{\Omega})$. It thus only remains to prove definiteness. Note that $0 = (u\,;\,u) = \|\nabla u\|^2_{L^2(\Omega)}$ implies $\nabla u = 0$, whence u is constant in Ω. Together with $u|_{\Gamma_D} = 0$, this proves $u = 0$. Moreover, the right-hand side

$$F(v) := \int_\Omega f v \, dx + \int_{\Gamma_N} \phi v \, ds$$

defines a linear functional on $C^1_D(\overline{\Omega})$ which is continuous with respect to the induced norm $\|v\| := (v\,;\,v)^{1/2}$. We prove this claim only in the special situation $\Gamma = \Gamma_D$ and postpone the abstract proof to a subsequent section.

Lemma 1.3 (Friedrichs' inequality). *Suppose that $\Omega = [a,b] \times [c,d] \subset \mathbb{R}^2$ and $\Gamma_D = \partial\Omega$. Then, it holds that $\|v\|_{L^2(\Omega)} \le \operatorname{diam}(\Omega)\, \|\nabla v\|_{L^2(\Omega)}$ for all $v \in C^1_D(\overline{\Omega})$.*

Proof. For $x = (x_1, x_2) \in \Omega$, it holds that $v(x_1, c) = 0$. Therefore, the fundamental theorem of calculus yields that

$$v(x) = \int_c^{x_2} \partial_2 v(x_1, t) \, dt.$$

The Hölder inequality yields that

$$|v(x)| \le |d - c|^{1/2} \Big(\int_c^{x_2} |\partial_2 v(x_1, t)|^2 \, dt \Big)^{1/2}.$$

Integration over Ω gives

$$\begin{aligned}\|v\|^2_{L^2(\Omega)} = \int_\Omega |v(x)|^2\,dx &\le |d-c| \int_\Omega \int_c^{x_2} |\partial_2 v(x_1,t)|^2\,dt\,dx \\ &= |d-c| \int_c^d \int_a^b \int_c^{x_2} |\partial_2 v(x_1,t)|^2\,dt\,dx_1\,dx_2 \\ &\le |d-c| \int_c^d \|\partial_2 v\|^2_{L^2(\Omega)}\,dx_2 \\ &= |d-c|^2 \|\partial_2 v\|^2_{L^2(\Omega)}.\end{aligned}$$

This results in $\|v\|_{L^2(\Omega)} \le |d-c|\,\|\partial_2 v\|_{L^2(\Omega)} \le \operatorname{diam}(\Omega)\,\|\nabla v\|_{L^2(\Omega)}$. ■

According to the Hölder and the Friedrichs inequality, we obtain that

$$|F(v)| \le \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \le \operatorname{diam}(\Omega)\|f\|_{L^2(\Omega)}\|\nabla v\|_{L^2(\Omega)} = \operatorname{diam}(\Omega)\|f\|_{L^2(\Omega)}\|v\|.$$

Therefore, the linear functional F is continuous with respect to $\|\cdot\| := \|\nabla(\cdot)\|_{L^2(\Omega)}$ with operator norm $\|F\|_* \le \operatorname{diam}(\Omega)\|f\|_{L^2(\Omega)}$. If $C^1_D(\overline{\Omega})$ associated with the norm $\|\cdot\|$ *were* a Hilbert space, the Riesz theorem *would* therefore imply the unique solvability of the variational form (1.5). However, $C^1_D(\overline{\Omega})$ is *not* complete and therefore the Riesz theorem does *not* apply.

The remedy is to consider the (unique) completion of $C^1_D(\overline{\Omega})$ with respect to $\|\cdot\|$. This leads to a so-called **Sobolev space** $H^1_D(\Omega)$, which is —by definition— complete and hence a Hilbert space. Density arguments then lead to an extended variational form: Find $u \in H^1_D(\Omega)$ such that

$$\int_\Omega \nabla u \cdot \nabla v\,dx = \int_\Omega f v\,dx + \int_{\Gamma_N} \phi v\,ds \quad \text{for all } v \in H^1_D(\Omega), \tag{1.10}$$

which is the **weak form** of the boundary value problem (1.1). Now, the Riesz theorem applies and proves the unique existence of a **weak solution** $u \in H^1_D(\Omega)$ of (1.10). Later on, we are going to show that

- each strong solution $u \in C^2(\overline{\Omega})$ of (1.1) belongs to $H^1_D(\Omega)$ and is also the unique weak solution of (1.10).
- provided the weak solution $u \in H^1_D(\Omega)$ is smooth, i.e., $u \in C^2(\overline{\Omega})$, the weak solution also solves the strong form (1.1).

In this sense, the strong form (1.1) and the weak form (1.10) are equivalent.

1.3 Finite Element Method

The finite element method for (1.10) essentially consists of replacing the (infinite dimensional) Sobolev space $H^1_D(\Omega)$ by a finite dimensional subspace $X_h \subset H^1_D(\Omega)$: Find $u_h \in X_h$ such that

$$\int_\Omega \nabla u_h \cdot \nabla v_h\,dx = \int_\Omega f v_h\,dx + \int_{\Gamma_N} \phi v_h\,ds \quad \text{for all } v_h \in X_h. \tag{1.11}$$

This problem is equivalent to the solution of a system of linear equations $\mathbf{Ax} = \mathbf{b}$, where the system matrix $\mathbf{A}$ is symmetric and positive definite. Of course, the question of convergence depends on the choice of X_h. Thus, there remain some topics for mathematical discussions later on.

The finite element method is a special **Galerkin scheme**. In this section, we collect the most simple properties of Galerkin schemes. Throughout, H is a (real) Hilbert space, and $\langle\!\langle \cdot\ ;\ \cdot \rangle\!\rangle$ is an equivalent scalar product on H, i.e., there are constants $\alpha, \beta > 0$ such that

$$\alpha \|v\|_H \leq |\!|\!| v |\!|\!| \leq \beta \|v\|_H \quad \text{for all } v \in H, \tag{1.12}$$

where $|\!|\!| v |\!|\!| := \langle\!\langle v\ ;\ v \rangle\!\rangle^{1/2}$ denotes the induced norm. We stress that $\langle\!\langle \cdot\ ;\ \cdot \rangle\!\rangle$ and $|\!|\!| \cdot |\!|\!|$ are often called **energy scalar product** and **energy norm**, respectively (see also Exercise 5).

Remark. In the following, we state all results with respect to the norm $\| \cdot \|_H$, which involves the constants $\alpha, \beta > 0$. Analogously, one may state the results with respect to the energy norm $|\!|\!| \cdot |\!|\!| = \| \cdot \|_H$, which corresponds to $\alpha = \beta = 1$. □

For given $F \in H^*$, the Riesz theorem proves the existence and uniqueness of a solution $u \in H$ of

$$\langle\!\langle u\ ;\ v \rangle\!\rangle = F(v) \quad \text{for all } v \in H, \tag{1.13}$$

for what we use the short-hand notation

$$\langle\!\langle u\ ;\ \cdot \rangle\!\rangle = F \in H^* \tag{1.14}$$

to implicitly indicate that this equation holds (pointwise) for all $v \in H$. Now, the Galerkin method simply consists in replacing the continuous space H by some finite dimensional subspace: Let X_h be a finite-dimensional (and hence closed) subspace of H. Since the Riesz theorem applies to the Hilbert space X_h as well, there is a unique **Galerkin solution** $u_h := \mathbb{G}_h u \in X_h$ such that

$$\langle\!\langle \mathbb{G}_h u\ ;\ \cdot \rangle\!\rangle = F \in X_h^*. \tag{1.15}$$

For $u \in H$ and the corresponding functional $\langle\!\langle u\ ;\ \cdot \rangle\!\rangle \in H^*$, this defines the **Galerkin projection**

$$\mathbb{G}_h : H \to X_h \quad \text{where } \mathbb{G}_h u \in X_h \text{ solves } \quad \langle\!\langle \mathbb{G}_h u\ ;\ \cdot \rangle\!\rangle = \langle\!\langle u\ ;\ \cdot \rangle\!\rangle \in X_h^*. \tag{1.16}$$

Note that $\mathbb{G}_h u \in X_h$ is characterized by the **Galerkin orthogonality**

$$\langle\!\langle u - \mathbb{G}_h u\ ;\ v_h \rangle\!\rangle = 0 \quad \text{for all } v_h \in X_h. \tag{1.17}$$

Before we proceed with the theoretical analysis of Galerkin schemes, we treat an implementational issue. The following theorem is the fundamental observation: Usually, only the scalar product $\langle\!\langle \cdot\ ;\ \cdot \rangle\!\rangle$ and the right-hand side $F \in H^*$ are known, while the exact solution $u \in H$ of (1.13) is unknown. Then, the Galerkin solution $\mathbb{G}_h u \in X_h$ can be computed by solving a linear system of equations — without knowledge of u.

Theorem 1.4. *Let $\{\phi_1, \ldots, \phi_N\}$ be a basis of X_h. We define the Galerkin matrix $A \in \mathbb{R}^{N \times N}$ and the vector $b \in \mathbb{R}^N$ by*

$$A_{jk} := \langle\!\langle \phi_k\ ;\ \phi_j \rangle\!\rangle \quad \textit{and} \quad b_j := F(\phi_j). \tag{1.18}$$

Then, A is symmetric and positive definite and, in particular, a regular matrix. Moreover, there holds $\mathbb{G}_h u = \sum_{j=1}^N x_j \phi_j$, where the vector $x \in \mathbb{R}^N$ solves $Ax = b$.

Proof. **1. step.** Symmetry of A clearly follows from the symmetry of $\langle\!\langle \cdot\,;\,\cdot \rangle\!\rangle$.

2. step. For any $x \in \mathbb{R}^N$ and $v_h := \sum_{j=1}^N x_j \phi_j$, it holds that

$$|\!|\!| v_h |\!|\!|^2 = \langle\!\langle v_h\,;\,v_h \rangle\!\rangle = \sum_{j,k=1}^N x_j x_k \langle\!\langle \phi_j\,;\,\phi_k \rangle\!\rangle = x \cdot Ax.$$

This proves $Ax \cdot x > 0$ for all $x \neq 0$. By definition, A is positive definite and hence regular.

3. step. Determine Galerkin solution: Let $x \in \mathbb{R}^n$ be the unique solution of the linear Galerkin system $Ax = b$. We use the basis representation $\mathbb{G}_h u = \sum_{j=1}^N y_j \phi_j$ of the Galerkin solution with some coefficient vector $y \in \mathbb{R}^n$. By use of the linearity of $\langle\!\langle \cdot\,;\,\cdot \rangle\!\rangle$, equation (1.15) becomes

$$b_k = F(\phi_k) = \langle\!\langle \mathbb{G}_h u\,;\,\phi_k \rangle\!\rangle = \sum_{j=1}^N y_j \langle\!\langle \phi_j\,;\,\phi_k \rangle\!\rangle = (Ay)_k \quad \text{for all } k = 1, \dots, N.$$

Therefore, the coefficient vector $y \in \mathbb{R}^N$ satisfies $Ay = b$. This proves $x = y$, i.e., we obtain $\mathbb{G}_h u$ by solving $Ax = b$. ■

Remark. We just remark that Theorem 1.4 can be applied for *any* orthogonal-type projection, e.g., the L^2-orthogonal projection onto a discrete space. □

We now proceed with the abstract analysis of Galerkin schemes. The following two lemmata provide elementary properties of the Galerkin projection. The first lemma proves stability of the method with respect to changes of the right-hand side F.

Lemma 1.5. *The Galerkin projection $\mathbb{G}_h$ is a linear and continuous projection onto X_h with*

$$\|\mathbb{G}_h u\|_H \leq \frac{\beta}{\alpha} \|u\|_H \quad \text{for all } u \in H, \tag{1.19}$$

where $\alpha, \beta > 0$ are the norm equivalence constants from (1.12)*. Moreover, $\mathbb{G}_h$ is the orthogonal projection onto X_h with respect to the energy scalar product $\langle\!\langle \cdot\,;\,\cdot \rangle\!\rangle$.*

Proof. For $u_h \in X_h$, the Galerkin orthogonality (1.17) implies $\mathbb{G}_h u_h = u_h$. Therefore $\mathbb{G}_h$ is a projection onto X_h. Also the linearity of $\mathbb{G}_h$ follows from the Galerkin orthogonality (1.17). To see the continuity of $\mathbb{G}_h$, it remains to estimate the operator norm: For $u \in H$ holds

$$|\!|\!| \mathbb{G}_h u |\!|\!|^2 = \langle\!\langle \mathbb{G}_h u\,;\,\mathbb{G}_h u \rangle\!\rangle = \langle\!\langle u\,;\,\mathbb{G}_h u \rangle\!\rangle \leq |\!|\!| u |\!|\!| |\!|\!| \mathbb{G}_h u |\!|\!|,$$

whence $|\!|\!| \mathbb{G}_h u |\!|\!| \leq |\!|\!| u |\!|\!|$ and

$$\alpha \|\mathbb{G}_h u\|_H \leq |\!|\!| \mathbb{G}_h u |\!|\!| \leq |\!|\!| u |\!|\!| \leq \beta \|u\|_H,$$

where we have used the norm equivalence (1.12) on H as well as the Cauchy inequality for the scalar product $\langle\!\langle \cdot\,;\,\cdot \rangle\!\rangle$. This proves that $\|\mathbb{G}_h u\|_H \leq (\alpha/\beta)\|u\|_H$ and thus continuity of $\mathbb{G}_h$. Finally,

we remark that the *unique* orthogonal projection with respect to $\langle\!\langle \cdot\,;\,\cdot \rangle\!\rangle$, is characterized by the orthogonality relation (1.17). ■

The following Céa lemma states that the **Galerkin error** $\|u - \mathbb{G}_h u\|_H$ is quasi-optimal, i.e., it behaves like the best approximation error up to multiplicative constants, which depend only on the continuous setting but not on X_h.

Lemma 1.6 (Céa). *The Galerkin error is quasi-optimal, i.e.,*

$$\|u - \mathbb{G}_h u\|_H \leq \frac{\beta}{\alpha} \min_{v_h \in X_h} \|u - v_h\|_H \quad \text{for all } u \in H, \tag{1.20}$$

where $\alpha, \beta > 0$ are the norm equivalence constants from (1.12). With respect to the energy norm, it holds that

$$|||u - \mathbb{G}_h u||| = \min_{v_h \in X_h} |||u - v_h||| \quad \text{for all } u \in H, \tag{1.21}$$

i.e., the Galerkin solution $\mathbb{G}_h u$ is the best approximation of u with respect to the energy norm.

Proof. For arbitrary $v_h \in X_h$, the Galerkin orthogonality (1.17) proves that

$$|||u - \mathbb{G}_h u|||^2 = \langle\!\langle u - \mathbb{G}_h u\,;\, u - v_h \rangle\!\rangle \leq |||u - \mathbb{G}_h u||| \, |||u - v_h|||,$$

which yields (1.21) with an infimum on the right-hand side. Of course, the minimum in (1.21) is attained for $v_h = \mathbb{G}_h u$. With the same arguments as in the proof of the last lemma, we even see that

$$\alpha \|u - \mathbb{G}_h u\|_H \leq |||u - \mathbb{G}_h u||| \leq |||u - v_h||| \leq \beta \|u - v_h\|_H,$$

which implies (1.20) with an infimum on the right-hand side. This minimum is attained for $v_h = \Pi_h u$ with $\Pi_h : X \to X_h$ being the orthogonal projection onto X_h with respect to $\|\cdot\|_H$. ■

Exercise 2. Let X be a normed vector space over $\mathbb{R}$ and $X_h \subseteq X$ be a finite dimensional subspace of X. Then, for any $x \in X$, there exists some (not necessarily unique) $x_h \in X_h$ such that

$$\|x - x_h\|_X = \min_{v_h \in X_h} \|x - v_h\|_X,$$

i.e., best approximation errors on finite dimensional spaces as in (1.20) are always attained. Prove that the set of minimizers is convex, closed and bounded (and hence even compact). □

A major advantage of Galerkin methods is that one can prove convergence for any exact solution $u \in H$ if one knows that smooth functions can be approximated well. In the following, think of the subscript $h > 0$ as a mesh-size parameter with corresponding finite dimensional spaces X_h:

Proposition 1.7. *For all $h > 0$, let X_h be a finite-dimensional subspace of H. We assume that there is a dense subspace D of H with approximation property, namely*

$$\lim_{h\to 0} \min_{v_h \in X_h} \|v - v_h\|_H = 0 \quad \text{for all } v \in D. \tag{1.22}$$

Then, for any $u \in H$, it holds that

$$\lim_{h\to 0} \|u - \mathbb{G}_h u\|_H = 0, \tag{1.23}$$

i.e., the sequence of Galerkin solutions converges to the exact solution u.

Proof. For $v \in D$, the quasi-optimality (1.20) yields that

$$\|u - \mathbb{G}_h u\|_H \le \frac{\beta}{\alpha} \min_{v_h \in X_h} \|u - v_h\|_H \le \frac{\beta}{\alpha} \big(\|u - v\|_H + \min_{v_h \in X_h} \|v - v_h\|_H \big).$$

We have to show that

$$\exists C > 0 \, \forall \varepsilon > 0 \, \exists h_0 > 0 \, \forall h \in (0, h_0) \quad \|u - \mathbb{G}_h u\|_H \le C\, \varepsilon.$$

For $\varepsilon > 0$, let $v \in D$ with $\|u - v\|_H \le \varepsilon$. Choose $h_0 > 0$ according to the approximation assumption (1.22) so that $\min_{v_h \in X_h} \|v - v_h\|_H \le \varepsilon$ for all $h \in (0, h_0)$. We thus finally obtain $\|u - \mathbb{G}_h u\|_H \le 2\beta\varepsilon/\alpha$, which concludes the proof. ■

Although the result of the preceding lemma seems to be very attractive, we stress, however, that the convergence of a Galerkin scheme can be arbitrarily slow. We argue in the abstract setting: If H is a separable Hilbert space, e.g., H is a Sobolev space, there is a countable orthonormal basis $\{\phi_j \,|\, j \in \mathbb{N}\}$. Any $u \in H$ can be written as $u = \sum_{j=1}^{\infty} x_j \phi_j$ with coefficients $(x_n) \in \ell_2$. If we define $X_j := \operatorname{span}\{\phi_1, \dots, \phi_j\}$, it holds that

$$\min_{v_h \in X_h} \|u - v_h\|_H^2 = \sum_{j=k+1}^{\infty} x_j^2.$$

Finally, the decay of the right-hand side can be very slow. One may think of, e.g., $x_j^2 = j^{-(1+\varepsilon)}$ for any $\varepsilon > 0$, so that the series converges but is — in the beginning — almost the divergent harmonic series.

The following exercise shows that the approximation property (1.22) in particular implies that the Hilbert space H has to be separable.

Exercise 3. Suppose that X is a normed space with finite dimensional subspaces $X_\ell \subseteq X_{\ell+1} \subseteq X$ for all $\ell \in \mathbb{N}$. Suppose that $\mathcal{D} \subseteq X$ is a dense subspace such that, for all $x \in X$,

$$\lim_{\ell\to\infty} \min_{x_\ell \in X_\ell} \|x - x_\ell\|_X = 0. \tag{1.24}$$

Then, X is separable, i.e., there is a countable and dense subset $M \subseteq X$. □

Exercise 4. Let $X = \ell_\infty$ and $X_\ell := \big\{(x_n) \in \ell_\infty \,\big|\, x_j = 0 \text{ for all } j \geq \ell\big\}$. Prove that (1.24) fails to hold for any dense subspace $\mathcal{D}$. Note that this also follows if one proves that ℓ_∞ is not separable. □

Remark. All foregoing results of this section hold (in a slightly modified form) in case that $\langle\!\langle \cdot\,;\,\cdot \rangle\!\rangle$ only is a continuous and elliptic bilinear form on the Hilbert space H, i.e., in all proofs, one can avoid to use the symmetry of $\langle\!\langle \cdot\,;\,\cdot \rangle\!\rangle$. □

The following exercise explains why $|\!|\!| \cdot |\!|\!|$ is called energy norm. In many situations, the function $J(\cdot)$ has the interpretation of a physical energy.

Exercise 5. Let $\langle\!\langle \cdot\,;\,\cdot \rangle\!\rangle$ be a scalar product on the Hilbert space H such that the norm $|\!|\!| \cdot |\!|\!|$ is equivalent to $\|\cdot\|_H$. Let $F \in H^*$ and $u \in H$. Then, the following assertions are equivalent:

- $\langle\!\langle u\,;\,\cdot \rangle\!\rangle = F \in H^*$;
- $J(u) = \min\limits_{v \in H} J(v)$, where $J(v) := \frac{1}{2}\langle\!\langle v\,;\,v \rangle\!\rangle - F(v)$.

In particular, the variational formulation is equivalent to energy minimization, and this result also covers the discrete setting. Derive a formula for the energy error $J(\mathbb{G}_h u) - J(u)$, where $\mathbb{G}_h : H \to X_h$ denotes the Galerkin projection. □

Finally, we comment on an extension of the concept of Galerkin schemes to some nonlinear problems. We note that this framework does, in particular, cover the frame of the Lax–Milgram lemma.

Exercise 6 (Main Theorem on Strongly Monotone Operators (Zarantonello '60)). Let H be a Hilbert space and $A : H \to H^*$ be a Lipschitz continuous and strongly monotone operator, i.e.,

$$\|Au - Av\|_{H^*} \leq L\|u - v\|_H \quad \text{and} \quad \langle Au - Av\,;\,u - v\rangle_{H^* \times H} \geq M\|u - v\|_H^2 \quad \text{for all } u, v \in H$$

with constants $L, M > 0$ that only depend on A. Then, A is bijective. **Hint:** Injectivity of A follows from the monotonicity of A. To prove surjectivity, we apply a fixed point argument: Let $I_H : H \to H^*$, $I_H(u) := (u\,;\,\cdot)_H$ denote the Riesz mapping. For given $F \in H^*$ and a certain choice of $C > 0$, the mapping $\Phi(u) := u - CI_H^{-1}(Au - F)$ is a contraction on H. Therefore, the Banach contraction theorem applies and provides a unique $u \in H$ with $u = \Phi(u)$. □

Exercise 7 (Lemma of Lax–Milgram). Use Exercise 6 to derive the Lemma of Lax–Milgram: Let H be a Hilbert space and $a(\cdot,\cdot)$ be a continuous and elliptic bilinear form on H, i.e.,

$$a(u, v) \leq L\,\|u\|_H\|v\|_H \quad \text{and} \quad a(u, u) \geq M\,\|u\|_H^2 \quad \text{for all } u, v \in H,$$

where the constants $L, M > 0$ depend only on $a(\cdot,\cdot)$. Then, given a right-hand side $F \in H^*$,

there is a unique $u \in H$ with $a(u, \cdot) = F \in H^*$. □

Chapter 2

Sobolev Spaces and Poisson Problem

2.1 Sobolev Spaces on Domains

This section briefly recalls the definition of Sobolev spaces $H^m(\Omega)$, for integer order $m \in \mathbb{N}_0$, on domains $\Omega \subseteq \mathbb{R}^d$. While this section requires Ω only to be open and connected, the following sections will implicitly assume that Ω is a bounded Lipschitz domain.

Definition. A function $u \in L^1_{\ell oc}(\Omega) := \big\{ w : \Omega \to \mathbb{R} \text{ measurable} \,\big|\, \forall K \subset \Omega \text{ compact } w \in L^1(K) \big\}$ has a **weak partial derivative** $\partial_j u \in L^1_{\ell oc}(\Omega)$, if the pair $(u, \partial_j u)$ satisfies the integration by parts formula with smooth test functions that vanish on the boundary, i.e., it holds that

$$\int_\Omega u(\partial_j v)\,dx = -\int_\Omega (\partial_j u) v\,dx \quad \text{for all } v \in \mathcal{D}(\Omega) := C_c^\infty(\Omega). \tag{2.1}$$

Note that $\partial_j u$ is (so far) only a symbol, whereas $\partial_j v := \partial v / \partial x_j$ is the classical j-th derivative of $v \in \mathcal{D}(\Omega)$. We say that $u \in L^1_{\ell oc}(\Omega)$ is **weakly differentiable with weak gradient** $\nabla u \in L^1_{\ell oc}(\Omega)$, if all weak derivatives $\partial_j u$, for $j = 1, \dots, d$, exist. □

The following main theorem of calculus, which will not be proven in this lecture, we infer that the weak derivative is unique, if it exists. Moreover, the weak derivative and the classical derivative coincide, if the classical derivative exists.

Theorem 2.1 (Fundamental Theorem of Calculus of Variations). *Let $f \in L^1_{\ell oc}(\Omega)$ satisfy $\int_\Omega f v\,dx = 0$ for all $v \in \mathcal{D}(\Omega)$. Then, it holds that $f = 0$ almost everywhere in Ω.* ■

Remark. Note that $C(\Omega) \subset L^1_{\ell oc}(\Omega)$. For $f \in C(\Omega)$, the fundamental theorem of calculus of variations can be proven by elementary calculus: Note that for any $x \in \mathbb{R}^d$ and any radius $\varepsilon > 0$, there is a function $\psi \in \mathcal{D}(\mathbb{R}^d)$ such that $\big\{ y \in \mathbb{R}^d \,\big|\, \psi(y) > 0 \big\} = U(x, \varepsilon) := \big\{ y \in \mathbb{R}^d \,\big|\, |x - y| < \varepsilon \big\}$; see the following Exercise 8. Provided $f \in C(\Omega)$ with $f(x) \neq 0$ for some $x \in \Omega$, we may assume $f(x) > 0$. By continuity, there is a small radius $\varepsilon > 0$ such that $U(x, \varepsilon) \subset \Omega$ and that $f(y) > 0$ for all $y \in U(x, \varepsilon)$. With the associated function $\psi \in \mathcal{D}(\Omega)$, we thus see that $\int_\Omega f \psi\,dx > 0$. Note that this argument provides the (logically equivalent) contraposition of the fundamental theorem of calculus of variations in the case of a continuous function f. □

Exercise 8. (i) Show that the following definition provides $\phi \in C^\infty(\mathbb{R})$ with $\operatorname{supp}(\phi) = [-1,1]$:

$$\phi(t) := \begin{cases} \exp\big(-1/(1-t^2)\big), & \text{for } |t|<1, \\ 0 & \text{else.} \end{cases}$$

(ii) For $\varepsilon > 0$ and $x \in \mathbb{R}^d$, define the function $\psi_{x,\varepsilon}(y) := \phi(|x-y|^2/\varepsilon)$. Show that $\psi_{x,\varepsilon} \in C^\infty(\mathbb{R}^d)$ with $\operatorname{supp}(\psi_{x,\varepsilon}) = \big\{y \in \mathbb{R}^d \,\big|\, |x-y| \le \varepsilon\big\}$ and $\psi_{x,\varepsilon}(y) > 0$ for all $y \in \big\{y \in \mathbb{R}^d \,\big|\, |x-y| < \varepsilon\big\}$. □

Corollary 2.2. (i) *The weak derivative $\partial_j u$ is unique, if it exists: If $\partial_j u, \widetilde{\partial_j u} \in L^1_{loc}(\Omega)$ satisfy (2.1), it holds that $\partial_j u = \widetilde{\partial_j u}$ almost everywhere in Ω.*
(ii) *A function $u \in C^1(\Omega)$ is weakly differentiable, and the weak derivative coincides with the classical derivative.*

Proof. (i) It holds that $\int_\Omega (\partial_j u - \widetilde{\partial_j u}) v \, dx = 0$ for all $v \in \mathcal{D}(\Omega)$ and thus $\partial_j u - \widetilde{\partial_j u} = 0$ almost everywhere in Ω. (ii) follows from (i) and the integration by parts formula. ■

A deeper result is the following, which is somehow, nevertheless, quite natural and expected.

Theorem 2.3. *If $u \in L^1_{loc}(\Omega)$ is weakly differentiable with $\nabla u = 0$, then the function u is constant, i.e., there is a constant $c \in \mathbb{R}$ such that $u = c$ almost everywhere in Ω.* ■

Definition. For $m = 0$, we define $H^0(\Omega) := L^2(\Omega)$ as the classical Lebesgue space of square integrable functions. For $m = 1$, the **Sobolev space** $H^1(\Omega)$ is defined by

$$H^1(\Omega) := \big\{u \in L^2(\Omega) \,\big|\, u \text{ weakly differentiable, } \nabla u \in L^2(\Omega)\big\} \tag{2.2}$$

and associated with the graph norm

$$\|u\|_{H^1(\Omega)} := \big(\|u\|^2_{L^2(\Omega)} + \|\nabla u\|^2_{L^2(\Omega)}\big)^{1/2}. \tag{2.3}$$

Higher-order Sobolev spaces of integer order $m \in \mathbb{N}$ may be defined inductively by

$$H^m(\Omega) := \big\{u \in L^2(\Omega) \,\big|\, u \text{ weakly differentiable, } \nabla u \in H^{m-1}(\Omega)\big\}, \tag{2.4}$$

with associated norm

$$\|u\|_{H^m(\Omega)} := \big(\|u\|^2_{L^2(\Omega)} + \|\nabla u\|^2_{H^{m-1}(\Omega)}\big)^{1/2}. \tag{2.5}$$

Remark. Clearly, $C^1(\overline{\Omega}) \subseteq H^1(\Omega)$ and we note below that $C^1(\overline{\Omega})$ is even dense in $H^1(\overline{\Omega})$. □

Theorem 2.4. *For all $m \in \mathbb{N}_0$, the Sobolev space $H^m(\Omega)$ is a Hilbert space.*

Proof. The proof uses the (hopefully) well-known fact that $H^0(\Omega) = L^2(\Omega)$ is a Hilbert space. We shall proceed by induction on m. However, we explicitely consider the case $m = 1$ first: Obviously, the H^1-norm is induced by the scalar product

$$(u\,;\,v)_{H^1(\Omega)} := (u\,;\,v)_{L^2(\Omega)} + (\nabla u\,;\,\nabla v)_{L^2(\Omega)} \quad \text{for all } u, v \in H^1(\Omega),$$

i.e., $\|u\|^2_{H^1(\Omega)} = (u\,;\,u)_{H^1(\Omega)}$. Therefore, it only remains to prove the completeness of $H^1(\Omega)$. Let (u_n) be a Cauchy sequence in $H^1(\Omega)$. Note that, by definition of the H^1-norm, (u_n) as well as (∇u_n) are Cauchy sequences in $L^2(\Omega)$. Since $L^2(\Omega)$ is complete, there are unique $u \in L^2(\Omega)$ and $g \in L^2(\Omega)^d$ such that

$$\lim_{n\to\infty} \|u - u_n\|_{L^2(\Omega)} = 0 = \lim_{n\to\infty} \|g - \nabla u_n\|_{L^2(\Omega)}.$$

By definition of $H^1(\Omega)$, it thus only remains to prove that u is weakly differentiable with gradient $\nabla u = g$. Let $v \in \mathcal{D}(\Omega)$ be an arbitrary test function. From the weak differentiability of each u_n and L^2-convergence, we obtain that

$$(u\,;\,\partial_j v)_{L^2(\Omega)} = \lim_{n\to\infty} (u_n\,;\,\partial_j v)_{L^2(\Omega)} = -\lim_{n\to\infty} (\partial_j u_n\,;\,v)_{L^2(\Omega)} = -(g_j\,;\,v)_{L^2(\Omega)}.$$

Therefore, g_j is the j-th weak derivative of u and consequently $g = \nabla u$. This concludes the case $m = 1$. The induction step for $H^m(\Omega)$ is left to the reader, but obviously follows from the same arguments, where we replace $g \in L^2(\Omega)^d$ by $g \in H^{m-1}(\Omega)^d$. ■

2.2 Main Theorems on Sobolev Spaces

From now on, it will be important and thus assumed that $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain. By definition of the Sobolev spaces $H^m(\Omega)$, there holds $H^m(\Omega) \subset H^{m-1}(\Omega)$ with $\|u\|_{H^{m-1}(\Omega)} \leq \|u\|_{H^m(\Omega)}$. In other words, the identity operator $id : H^m(\Omega) \to H^{m-1}(\Omega)$ is well-defined and continuous. The following Rellich theorem states that it is also compact. This is a pretty strong result. The impact of which will become clear in our proofs of the Poincaré inequality and the Friedrichs inequality.

Theorem 2.5 (Rellich Compactness Theorem). *For any integer order $m \in \mathbb{N}$, the embedding $H^m(\Omega) \subseteq H^{m-1}(\Omega)$ is compact.* ■

We recall that an operator $A \in L(X;Y)$ between normed spaces X and Y is compact, if and only if each bounded set $S \subseteq X$ is mapped to a pre-compact set $A(S) \subseteq Y$, i.e., $\overline{A(S)} \subseteq Y$ is compact.

Before the statement and the proof of the Poincaré inequality, we need a further technical lemma. The result is rather standard in the analysis of variational problems.

Lemma 2.6. *A continuous and convex functional $f : X \to \mathbb{R}$ on a normed space X is weakly lower semicontinuous, i.e., for each weakly convergent sequence (x_n) in X with $x_n \rightharpoonup x \in X$,*

it holds that

$$f(x) \leq \liminf_{n\in\mathbb{N}} f(x_n). \tag{2.6}$$

Proof. **1. step.** We prove that the epigraph $G := \{(x,\alpha) \in X \times \mathbb{R} \,|\, f(x) \leq \alpha\}$ is convex: For $(x,\alpha),(y,\beta) \in G$ and $0 \leq \theta \leq 1$, the convexity of f proves that

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y) \leq \theta\alpha + (1-\theta)\beta,$$

whence $\theta(x,\alpha) + (1-\theta)(y,\beta) \in G$, i.e., $G \subseteq X \times \mathbb{R}$ is convex.

2. step. We use the continuity of f to prove that G is also closed: Let (x_n,α_n) be a convergent sequence in G, i.e., it holds that $x_n \to x \in X$ and $\alpha_n \to \alpha \in \mathbb{R}$. We prove that $(x,\alpha) \in G$, which follows from

$$f(x) = \lim_{n\to\infty} f(x_n) \leq \lim_{n\to\infty} \alpha_n = \alpha.$$

3. step. The following step in the proof is known as *Mazur's lemma*: We prove that the closed and convex set G is also weakly closed in $X \times \mathbb{R} =: Y$, i.e., closed with respect to the weak topology on Y. We argue by contradiction and assume that G is not weakly closed. Then, there is an element $y \in \overline{G}^{\sigma} \backslash G$, where $\overline{G}^{\sigma}$ denotes the weak closure of G. According to the Hahn-Banach separation theorem, there is a functional $\phi \in Y^*$ and a scalar $\lambda \in \mathbb{R}$ such that $\phi(y) < \lambda \leq \inf \phi(G)$. Therefore $U := \phi^{-1}(-\infty,\lambda)$ is weakly open with $y \in U$ and $U \cap G = \emptyset$. This contradicts topologically that y is in the weak closure of G. Hence, $G = \overline{G}^{\sigma}$ is weakly closed, and we may proceed with the proof of (2.6).

4. step. We show the weak lower semicontinuity of f: Suppose that $x_n \rightharpoonup x \in X$. For $\alpha := \liminf_n f(x_n) = \infty$, (2.6) is trivial. We thus may assume $\alpha < \infty$. Let $\beta > \alpha$ and define $\alpha_n := \max\{\beta, f(x_n)\} \to \beta$. Clearly, $(x_n,\alpha_n) \in G$. Moreover, this sequence is weakly convergent $(x_n,\alpha_n) \rightharpoonup (x,\beta)$. We deduce $(x,\beta) \in G$. Thus, $f(x) \leq \beta$ for all $\beta > \alpha$ and therefore finally $f(x) \leq \alpha = \lim_{n\to\infty} f(x_n)$. ∎

A first consequence of the preceding abstract results is that one can easily construct equivalent norms on the Sobolev space $H^1(\Omega)$.

Proposition 2.7. *Let $|\cdot|_{H^1}$ be a continuous seminorm on $H^1(\Omega)$ which is definite on the constant functions, i.e., $|c|_{H^1} = 0$ implies $c = 0$ for all $c \in \mathbb{R}$. Then, there are constants $C_1, C_2 > 0$ such that*

$$|v|_{H^1} \leq C_1 \|v\|_{H^1(\Omega)} \quad \text{as well as} \quad C_2^{-1} \|v\|_{L^2(\Omega)} \leq |\!|\!| v |\!|\!| := \|\nabla v\|_{L^2(\Omega)} + |v|_{H^1} \text{ for all } v \in H^1(\Omega).$$

In particular, $|\!|\!|\cdot|\!|\!|$ defines an equivalent norm on $H^1(\Omega)$, i.e.,

$$(1+C_1)^{-1} |\!|\!| v |\!|\!| \leq \|v\|_{H^1(\Omega)} \leq (1+C_2) |\!|\!| v |\!|\!| \quad \text{for all } v \in H^1(\Omega).$$

Proof. **1. step.** Existence of C_1: By definition of continuity, there exists an open neighborhood $O \subseteq H^1(\Omega)$ of 0 such that $|v|_{H^1} \leq 1$ for all $v \in O$. Without loss of generality, we may choose a

radius $r > 0$ sufficiently small such that $\overline{B_r(0)} \subset O \subset H^1(\Omega)$ for the closed ball with radius r and center zero. This implies

$$|v|_{H^1} = \frac{1}{r}\|v\|_{H^1(\Omega)}|r\frac{v}{\|v\|_{H^1(\Omega)}}|_{H^1} \leq \frac{1}{r}\|v\|_{H^1(\Omega)}.$$

This proves existence of $C_1 := 1/r$.

2. step. Existence of C_2: We assume that there is no constant $C_2 > 0$ such that $\|v\|_{L^2(\Omega)} \leq C_2 \,|||v|||$ for all $v \in H^1(\Omega)$. Therefore, there exists a sequence (v_n) in $H^1(\Omega)$ such that

$$\frac{1}{n}\,\|v_n\|_{L^2(\Omega)} > |||v_n||| = \|\nabla v_n\|_{L^2(\Omega)} + |v_n|_{H^1}$$

The definition of $w_n := v_n/\|v_n\|_{L^2(\Omega)}$ leads to to a sequence (w_n) in $H^1(\Omega)$ such that

$$\|w_n\|_{L^2(\Omega)} = 1, \quad \|\nabla w_n\|_{L^2(\Omega)} \leq 1/n, \quad |w_n|_{H^1} \leq 1/n.$$

Therefore, (w_n) is a bounded sequence in the Hilbert space $H^1(\Omega)$. A Hilbert space is reflexive. By virtue of the Banach-Alaoglou theorem, each bounded sequence thus has a weakly convergent subsequence. Therefore, we may assume that $w_n \rightharpoonup w \in H^1(\Omega)$. An application of Lemma 2.6 proves that

$$\|\nabla w\|_{L^2(\Omega)} \leq \liminf_{n\to\infty} \|\nabla w_n\|_{L^2(\Omega)} = 0,$$

whence the weak limit w is constant. Another application of Lemma 2.6 proves that

$$|w|_{H^1} \leq \liminf_{n\to\infty} |w_n|_{H^1} = 0$$

since a seminorm is always convex. Therefore, $w = 0$. On the other hand, the Rellich theorem states the strong convergence $w_n \to w \in L^2(\Omega)$ and thus $\|w\|_{L^2(\Omega)} = \lim_{n\to\infty} \|w_n\|_{L^2(\Omega)} = 1$. This contradiction concludes the existence of C_2. In particular, we hence observe $\|v\|_{H^1(\Omega)} \leq \|v\|_{L^2(\Omega)} + \|\nabla v\|_{L^2(\Omega)} \leq (C_2 + 1)\,|||v|||$. ■

Corollary 2.8 (Poincaré Inequality). *It holds that*

$$\|v\|_{L^2(\Omega)} \leq \widetilde{C}_P \left(\|\nabla v\|_{L^2(\Omega)} + \Big| \int_\Omega v\,dx \Big| \right) \quad \text{for all } v \in H^1(\Omega), \tag{2.7}$$

where the constant $\widetilde{C}_P > 0$ depends only on Ω. Moreover, $|||v||| := \|\nabla v\|_{L^2(\Omega)} + \big| \int_\Omega v\,dx \big|$ defines even an equivalent norm on $H^1(\Omega)$.

Proof. According to Proposition 2.7, it only remains to show that

$$|v|_{H^1} := \Big| \int_\Omega v\,dx \Big| \quad \text{for } v \in H^1(\Omega)$$

defines a continuous seminorm on $H^1(\Omega)$ which is definite on the constant functions. The equality $|c|_{H^1} = |\Omega||c|$ for $c \in \mathbb{R}$ verifies the definiteness. Lipschitz continuity follows from

$$\big| |v|_{H^1} - |w|_{H^1} \big| \leq \Big| \int_\Omega v - w\,dx \Big| \leq |\Omega|^{1/2}\|v - w\|_{L^2(\Omega)} \leq |\Omega|^{1/2}\|v - w\|_{H^1(\Omega)}$$

and from the boundedness of Ω. ■

Corollary 2.9 (Poincaré Inequality). *There is a constant $C_P > 0$, which depends only on the shape of Ω but not on its diameter, such that*

$$\|v\|_{L^2(\Omega)} \leq C_P \operatorname{diam}(\Omega) \|\nabla v\|_{L^2(\Omega)} \quad \text{for all } v \in H^1_*(\Omega) := \big\{ w \in H^1(\Omega) \,\big|\, \textstyle\int_\Omega w \, dx = 0 \big\}, \tag{2.8}$$

where $\operatorname{diam}(\Omega) := \sup\big\{ |x-y| \,\big|\, x, y \in \Omega \big\}$ *denotes the diameter of* Ω.

Proof. The proof is a so-called **scaling argument**: We define $\lambda := \operatorname{diam}(\Omega)$ and $\widetilde{\Omega} := \lambda^{-1}\Omega$. Note that the scaled domain $\widetilde{\Omega}$ satisfies $\operatorname{diam}(\widetilde{\Omega}) = 1$ and depends only on the shape of Ω. We consider the affine bijection $\Phi : \Omega \to \widetilde{\Omega}$, $\Phi(x) := \lambda^{-1} x$. Recall the transformation theorem, which holds for arbitrary diffeomorphisms $\Phi : \Omega \to \widetilde{\Omega}$ and states that

$$\int_{\widetilde{\Omega}} \widetilde{f} \, dy = \int_\Omega \widetilde{f}(\Phi(x)) \, |\det D\Phi(x)| \, dx \quad \text{for all } \widetilde{f} \in L^1(\widetilde{\Omega}).$$

Note that $\det D\Phi(x) = \lambda^{-d}$ since $D\Phi = \lambda^{-1}\boldsymbol{I}$ in our case. For $v \in H^1(\Omega)$, we define $\widetilde{v} := v \circ \Phi^{-1} \in H^1(\widetilde{\Omega})$. Then,

$$\|\widetilde{v}\|^2_{L^2(\widetilde{\Omega})} = \int_{\widetilde{\Omega}} |\widetilde{v}|^2 \, dy = \lambda^{-d} \int_\Omega |v|^2 \, dx = \lambda^{-d} \|v\|^2_{L^2(\Omega)}.$$

According to the chain rule, it holds that $\nabla \widetilde{v} = \lambda \, (\nabla v) \circ \Phi^{-1}$ and consequently that

$$\|\nabla \widetilde{v}\|^2_{L^2(\widetilde{\Omega})} = \lambda^{2-d} \|\nabla v\|^2_{L^2(\Omega)}.$$

With $\widetilde{C}_P > 0$ the Poincaré constant from (2.7) for $\widetilde{\Omega}$, we thus infer

$$\|v\|^2_{L^2(\Omega)} = \lambda^d \, \|\widetilde{v}\|^2_{L^2(\widetilde{\Omega})} \leq \lambda^d \, \widetilde{C}_P^2 \, \|\nabla \widetilde{v}\|^2_{L^2(\widetilde{\Omega})} = \lambda^2 \, \widetilde{C}_P^2 \, \|\nabla v\|^2_{L^2(\Omega)}.$$

Note that $\widetilde{C}_P$ depends only on $\widetilde{\Omega}$ und thus only on the shape of Ω. This concludes the proof. ■

Remark. We stress that $Iv := \int_\Omega v \, dx$ defines a linear and continuous functional on $H^1(\Omega)$. In particular, $H^1_*(\Omega) = \ker(I)$ is a closed subspace of $H^1(\Omega)$ and hence a Hilbert space. According to the Poincaré inequality, it holds that $\|\nabla v\|_{L^2(\Omega)} \leq \|v\|_{H^1(\Omega)} \leq (1 + \widetilde{C}_P^2)^{1/2} \|\nabla v\|_{L^2(\Omega)}$ for all $v \in H^1_*(\Omega)$. In particular, $\|\nabla v\|_{L^2(\Omega)}$ defines an equivalent Hilbert norm on $H^1_*(\Omega)$ with associated scalar product $(\nabla u \,;\, \nabla v)_{L^2(\Omega)}$. □

Theorem 2.10 (Meyers-Serrin). *For each integer order $m \in \mathbb{N}$, $C^\infty(\overline{\Omega})$ and, in particular, $C^\infty(\Omega) \cap H^m(\Omega)$ are dense subspaces of $H^m(\Omega)$.* ■

Theorem 2.11 (Trace Operator). *There is a unique operator $\gamma \in L\big(H^1(\Omega); L^2(\Gamma)\big)$ such that $\gamma v = v|_\Gamma$ for all $v \in C^1(\overline{\Omega})$, i.e., γ extends the classical trace defined as restriction $v|_\Gamma$ on the boundary for smooth functions v.* ■

As a first corollary to Theorem 2.11, we can prove that the integration by parts formula also holds for Sobolev functions $u, v \in H^1(\Omega)$.

Corollary 2.12 (Integration by Parts). *For all $u, v \in H^1(\Omega)$, it holds that*

$$\int_\Omega u \, \frac{\partial v}{\partial x_j} \, dx + \int_\Omega \frac{\partial u}{\partial x_j} \, v \, dx = \int_\Gamma \gamma u \, \gamma v \, n_j \, ds. \tag{2.9}$$

Proof. The formula (2.9) holds for $u, v \in C^1(\overline{\Omega})$. All three terms define continuous bilinear forms on $H^1(\Omega) \times H^1(\Omega)$. Therefore (2.9) follows, for arbitrary $u, v \in H^1(\Omega)$ from the density of $C^1(\overline{\Omega})$ in $H^1(\Omega)$: Given $u, v \in H^1(\Omega)$, there are sequences (u_n) and (v_n) in $C^1(\overline{\Omega})$ which converge to u resp. v in $H^1(\Omega)$. Therefore, if $a(\cdot,\cdot) : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ is continuous, then it holds that $\lim_{n\to\infty} a(u_n, v_n) = a(u, v)$. This concludes the proof. ■

The analytical treatment of the Dirichlet problem makes use of the so-called Friedrichs inequality, whereas the analytical treatment of the Neumann problem uses the previously proven Poincaré inequality.

Corollary 2.13 (Friedrichs Inequality). *Assume that the Dirichlet boundary $\Gamma_D \subseteq \Gamma$ has positive surface measure $|\Gamma_D| > 0$. Then, it holds that*

$$\|v\|_{L^2(\Omega)} \leq \widetilde{C}_F \left(\|\nabla v\|_{L^2(\Omega)} + \|\gamma v\|_{L^2(\Gamma_D)} \right) \quad \text{for all } v \in H^1(\Omega) \tag{2.10}$$

with a constant $\widetilde{C}_F > 0$, which depends only on Ω and Γ_D. Moreover, the right-hand side $|||v||| := \|\nabla v\|_{L^2(\Omega)} + \|\gamma v\|_{L^2(\Gamma_D)}$ even defines an equivalent norm on $H^1(\Omega)$.

Proof. We again apply Proposition 2.7. It only remains to show that

$$|v|_{H^1} := \|\gamma v\|_{L^2(\Gamma_D)} \quad \text{for } v \in H^1(\Omega)$$

defines a continuous seminorm on $H^1(\Omega)$ which is definite on the constant functions. The definiteness is again easily obtained from $|c|_{H^1} = |\Gamma_D|^{1/2}|c|$ for $c \in \mathbb{R}$. Lipschitz continuity follows from

$$\big| |v|_{H^1} - |w|_{H^1} \big| \leq \|\gamma v - \gamma w\|_{L^2(\Gamma_D)} = \|\gamma(v - w)\|_{L^2(\Gamma_D)} \leq C \, \|v - w\|_{H^1(\Omega)}$$

according to the continuity of the trace operator $\gamma \in L(H^1(\Omega); L^2(\Gamma))$. ■

Definition. We define $H^1_0(\Omega) := \overline{\mathcal{D}(\Omega)}^{\|\cdot\|_{H^1}}$ and $H^1_D(\Omega) := \overline{C^1_D(\overline{\Omega})}^{\|\cdot\|_{H^1}}$, where the subscript D indicates the Dirichlet boundary Γ_D. By definition, $H^1_0(\Omega)$ as well as $H^1_D(\Omega)$ are closed subspaces of $H^1(\Omega)$ and thus Hilbert spaces. In particular, it holds that $H^1_0(\Omega) \subseteq H^1_D(\Omega)$. □

The same scaling argument as for the Poincaré inequality proves the following variant of the Friedrichs inequality, where we note that continuity of the trace operator γ proves that $\gamma v = 0$, for $v \in H^1_0(\Omega)$, as well as $(\gamma v)|_{\Gamma_D} = 0$, for $v \in H^1_D(\Omega)$.

Corollary 2.14 (Friedrichs Inequality). *It holds that*

$$\|v\|_{L^2(\Omega)} \leq C_F \operatorname{diam}(\Omega) \, \|\nabla v\|_{L^2(\Omega)} \quad \text{for all } v \in H^1_D(\Omega) \tag{2.11}$$

with a constant $C_F > 0$ that depends only on the shape of Ω and Γ_D. ■

We finally note the relation between $H^1_D(\Gamma)$ and the trace operator, cf. the Theorem of Meyers-Serrin.

Theorem 2.15. *There holds $H^1_0(\Omega) = \ker(\gamma)$ with $\gamma \in L(H^1(\Omega); L^2(\Gamma))$ the trace operator. Moreover, $H^1_D(\Omega) = \big\{v \in H^1(\Omega) \,\big|\, (\gamma v)|_{\Gamma_D} = 0\big\}$.* ■

Exercise 9. Usually, one defines the range of the trace operator as $H^{1/2}(\Gamma) := \mathrm{range}(\gamma) \subseteq L^2(\Gamma)$. This space is associated with the norm $\|v\|_{H^{1/2}(\Gamma)} := \inf\big\{\|\widehat{v}\|_{H^1(\Omega)} \,\big|\, \widehat{v} \in H^1(\Omega) \text{ with } \gamma\widehat{v} = v\big\}$. Prove that $H^{1/2}(\Gamma)$ associated with this norm is a Hilbert space with continuous inclusion $H^{1/2}(\Gamma) \subseteq L^2(\Gamma)$. **Hint:** Recall the definition and the standard results on quotient spaces and the associated quotient norm! □

For $X = H^1(\Omega)$ and $Y = L^2(\Omega)$, the following exercise shows that the L^2-scalar products $(f\,;\cdot)_{L^2(\Omega)}$ for $f \in L^2(\Omega)$ give (up to density) all linear and continuous functionals on $H^1(\Omega)$, i.e., the embedding $L^2(\Omega) \to H^1(\Omega)^*, f \mapsto (f\,;\cdot)_{L^2(\Omega)}$ is well-defined, linear, continuous, and injective with dense image.

Exercise 10. Let X and Y be Hilbert spaces with continuous embedding $X \subseteq Y$. Show that the mapping $I : Y^* \to X^*, Iy^* := y^*|_X$ is well-defined, linear, and continuous. Prove that $I(Y^*) \subseteq X^*$ is a dense subspace. Moreover, if $X \subseteq Y$ is dense with respect to $\|\cdot\|_Y$, then the embedding I is even injective. □

2.3 Weak Form of Laplace Problem

2.3.1 Dirichlet Problem

In this section, we generalize the variational form derived in the introductory section to our Hilbert space setting. We start with the homogeneous Dirichlet problem

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{2.12}$$

Recall that this formulation is called the **strong form** of the boundary value problem. The following proposition provides the — in some sense — equivalent and always uniquely solvable weak form of the boundary value problem.

Proposition 2.16. (i) *Provided that $u \in C^2(\overline{\Omega})$ solves (2.12) for a given source term $f \in C(\overline{\Omega})$, it holds that $u \in H^1_0(\Omega)$ as well as*

$$(\nabla u\,;\nabla v)_{L^2(\Omega)} = (f\,;v)_{L^2(\Omega)} \quad \textit{for all } v \in H^1_0(\Omega). \tag{2.13}$$

(ii) *Given $f \in L^2(\Omega)$, the* ***weak form*** (2.13) *has a unique solution $u \in H^1_0(\Omega)$. It holds that*

$$\|u\|_{H^1(\Omega)} \le C \sup_{v \in H^1_0(\Omega)\setminus\{0\}} \frac{(f\,;\,v)_{L^2(\Omega)}}{\|v\|_{H^1(\Omega)}} \le C\,\|f\|_{L^2(\Omega)}, \tag{2.14}$$

where the constant $C > 0$ depends only on Ω.

(iii) *Provided that $f \in C(\overline{\Omega})$ and that the weak solution $u \in H^1_0(\Omega)$ of* (2.13) *additionally satisfies $u \in C^2(\overline{\Omega})$, then u even solves the strong form* (2.12).

Proof. (i) We have already seen before that a strong solution $u \in C^2(\overline{\Omega})$ solves the variational form (2.13) for test functions $v \in C^1_0(\overline{\Omega}) := \big\{w \in C^1(\overline{\Omega}) \,\big|\, w|_\Gamma = 0\big\}$ replacing $H^1_0(\Omega)$; see Proposition 1.1. If we keep u fixed, the left-hand side as well as the right-hand side of (2.13) define continuous and linear functionals on $H^1(\Omega)$. Note that the closure of $C^1_0(\overline{\Omega})$ with respect to the H^1-norm leads to the Hilbert space $H^1_0(\Omega)$. Therefore, standard density arguments prove (2.13).

(ii) According to the Friedrichs inequality, it holds that

$$\|\nabla v\|^2_{L^2(\Omega)} \le \|v\|^2_{H^1(\Omega)} \le (1 + \widetilde{C}_F^2)\,\|\nabla v\|^2_{L^2(\Omega)} \quad \text{for all } v \in H^1_0(\Omega).$$

Therefore, the left-hand side of (2.13) defines an equivalent scalar product on $H^1_0(\Omega)$. The Riesz theorem thus provides a unique weak solution $u \in H^1_0(\Omega)$ of (2.13). Plugging-in $u = v \in H^1_0(\Omega)$, the weak form yields that

$$(1 + \widetilde{C}_F^2)^{-1}\|u\|^2_{H^1(\Omega)} \le \|\nabla u\|^2_{L^2(\Omega)} = (f\,;\,u)_{L^2(\Omega)} \le \sup_{v \in H^1_0(\Omega)\setminus\{0\}} \frac{(f\,;\,v)_{L^2(\Omega)}}{\|v\|_{H^1(\Omega)}}\,\|u\|_{H^1(\Omega)}$$

which results in the first estimate of (2.14). The second estimate follows from the Cauchy inequality

$$(f\,;\,v)_{L^2(\Omega)} \le \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \le \|f\|_{L^2(\Omega)}\|v\|_{H^1(\Omega)}.$$

(iii) Since the weak solution u is smooth, we may use integration by parts to see that

$$(\nabla u\,;\,\nabla v)_{L^2(\Omega)} = (-\Delta u\,;\,v)_{L^2(\Omega)} \quad \text{for all } v \in H^1_0(\Omega).$$

The difference with the weak form (2.13) thus yields that

$$0 = (f + \Delta u\,;\,v)_{L^2(\Omega)} \quad \text{for all } v \in H^1_0(\Omega).$$

Note that $F := f + \Delta u \in C(\overline{\Omega})$. With $\mathcal{D}(\Omega) \subseteq H^1_0(\Omega)$, Theorem 2.1 proves $F = 0$; see also the remark right after Theorem 2.1. Consequently, it holds that $-\Delta u = f$ in Ω. The Dirichlet boundary conditions (in the strong form) follow from $0 = \gamma u = u|_\Gamma$. Altogether, u solves (2.12) ■

2.3.2 Mixed Boundary Value Problem

Second, we consider the mixed boundary value problem

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \Gamma_D, \\ \partial u/\partial n &= \phi \quad \text{on } \Gamma_N, \end{aligned} \tag{2.15}$$

with $\Gamma = \overline{\Gamma}_D \cup \overline{\Gamma}_N$, $\Gamma_D \cap \Gamma_N = \emptyset$, and $|\Gamma_D| > 0$. The limit case $|\Gamma_D| = 0$ corresponds to the Neumann problem which is treated in Section 2.3.3. Recall the trace norm $\|\cdot\|_{H^{1/2}(\Gamma)}$ from Exercise 9. Then, the main proposition reads as follows:

Proposition 2.17. (i) *Suppose that Γ_N is smooth, i.e., the outer normal vector depends continuously on $x \in \Gamma_N$. Provided that $u \in C^2(\overline{\Omega})$ solves the* ***strong form*** (2.15) *for a given source term $f \in C(\overline{\Omega})$ and Neumann data $\phi \in C(\overline{\Gamma}_N)$, it holds that $u \in H^1_D(\Omega)$ as well as*

$$(\nabla u \,;\, \nabla v)_{L^2(\Omega)} = (f \,;\, v)_{L^2(\Omega)} + (\phi \,;\, \gamma v)_{L^2(\Gamma_N)} \quad \text{for all } v \in H^1_D(\Omega). \tag{2.16}$$

(ii) *Given $f \in L^2(\Omega)$ and $\phi \in L^2(\Gamma_N)$, the* ***weak form*** (2.16) *has a unique solution $u \in H^1_D(\Omega)$. It holds that*

$$\begin{aligned} \|u\|_{H^1(\Omega)} &\le C_1 \left(\sup_{v \in H^1_D(\Omega)\setminus\{0\}} \frac{(f \,;\, v)_{L^2(\Omega)}}{\|v\|_{H^1(\Omega)}} + \sup_{w \in H^{1/2}(\Gamma)\setminus\{0\}} \frac{(\phi \,;\, w)_{L^2(\Gamma_N)}}{\|w\|_{H^{1/2}(\Gamma)}} \right) \\ &\le C_2 \left(\|f\|_{L^2(\Omega)} + \|\phi\|_{L^2(\Gamma_N)} \right) \end{aligned} \tag{2.17}$$

where the constants $C_1, C_2 > 0$ depend only on Ω and Γ_D.

(iii) *Provided that $f \in C(\overline{\Omega})$ and $\phi \in C(\overline{\Gamma}_N)$ and that the weak solution $u \in H^1_D(\Omega)$ of (2.16) additionally satisfies $u \in C^2(\overline{\Omega})$, then u even solves the strong form (2.15).*

Proof is done in the exercises. ■

2.3.3 Neumann Problem

Finally, we consider the Neumann problem

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ \partial u/\partial n &= \phi \quad \text{on } \Gamma. \end{aligned} \tag{2.18}$$

Note that the solution u of (2.18) cannot be unique: If $u \in C^2(\overline{\Omega})$ solves the **strong form** (2.18), also $u + c$ solves (2.18), for all $c \in \mathbb{R}$. To fix the additive constant, we seek a solution which additionally satisfies, e.g., that

$$\int_\Omega u \, dx = 0. \tag{2.19}$$

Moreover, the Gauss divergence theorem shows

$$-\int_\Omega f \, dx = \int_\Omega \Delta u \, dx = \int_\Omega \operatorname{div}(\nabla u) \, dx = \int_\Gamma \frac{\partial u}{\partial n} \, ds = \int_\Gamma \phi \, ds.$$

Therefore, the data f and ϕ have to satisfy the compatibility condition

$$\int_\Omega f \, dx + \int_\Gamma \phi \, ds = 0 \tag{2.20}$$

to allow for the existence of (strong) solutions. Recall the trace norm $\|\cdot\|_{H^{1/2}(\Gamma)}$ from Exercise 9.

Proposition 2.18. (i) *Suppose that Γ is smooth, i.e., the outer normal vector depends continuously on $x \in \Gamma$. Provided that $u \in C^2(\overline{\Omega})$ solves (2.18) for a given source term $f \in C(\overline{\Omega})$ and Neumann data $\phi \in C(\Gamma)$, it holds that $u \in H^1(\Omega)$ and*

$$(\nabla u \,;\, \nabla v)_{L^2(\Omega)} = (f \,;\, v)_{L^2(\Omega)} + (\phi \,;\, \gamma v)_{L^2(\Gamma)} \quad \text{for all } v \in H^1(\Omega). \tag{2.21}$$

(ii) *Given $f \in L^2(\Omega)$ and $\phi \in L^2(\Gamma)$, the variational formulation*

$$(\nabla u \,;\, \nabla v)_{L^2(\Omega)} = (f \,;\, v)_{L^2(\Omega)} + (\phi \,;\, \gamma v)_{L^2(\Gamma)} \quad \text{for all } v \in H^1_*(\Omega) \tag{2.22}$$

has a unique solution $u \in H^1_*(\Omega) := \big\{v \in H^1(\Omega) \,\big|\, \int_\Omega v\,dx = 0\big\}$.

(iii) *Provided that the data $f \in L^2(\Omega)$ and $\phi \in L^2(\Gamma)$ satisfy (2.20), the unique solution $u \in H^1_*(\Omega)$ of (2.22) even solves the **weak form** (2.21). Moreover, it holds that*

$$\begin{aligned} \|u\|_{H^1(\Omega)} &\le C_1 \left(\sup_{v \in H^1(\Omega)\setminus\{0\}} \frac{(f \,;\, v)_{L^2(\Omega)}}{\|v\|_{H^1(\Omega)}} + \sup_{w \in H^{1/2}(\Gamma)\setminus\{0\}} \frac{(\phi \,;\, w)_{L^2(\Gamma)}}{\|w\|_{H^{1/2}(\Gamma)}} \right) \\ &\le C_2 \big(\|f\|_{L^2(\Omega)} + \|\phi\|_{L^2(\Gamma)} \big) \end{aligned} \tag{2.23}$$

where the constants $C_1, C_2 > 0$ depend only on Ω.

(iv) *Provided that $f \in C(\overline{\Omega})$ and $\phi \in C(\Gamma)$ satisfy (2.20) and that the weak solution $u \in H^1_*(\Omega)$ of (2.21) resp. (2.22) additionally satisfies $u \in C^2(\overline{\Omega})$, then u even solves the strong form (2.18).*

Proof. (i) The variational form (2.21) holds for test functions $v \in C^1(\overline{\Omega})$ according to integration by parts. For fixed u, the left-hand as well as the right-hand side define continuous linear functionals on $H^1(\Omega)$. Thus, (2.21) follows for $v \in H^1(\Omega)$ by density arguments. (ii) According to the Poincaré inequality, it holds that

$$\|\nabla v\|^2_{L^2(\Omega)} \le \|v\|^2_{H^1(\Omega)} \le (1 + \widetilde{C}_P^2) \|\nabla v\|^2_{L^2(\Omega)} \quad \text{for all } v \in H^1_*(\Omega).$$

Therefore, the left-hand side of (2.22) defines an equivalent scalar product on $H^1_*(\Omega)$. Note that $H^1_*(\Omega)$ is a closed subspace of $H^1(\Omega)$ and hence a Hilbert space. Therefore, (2.22) follows from the Riesz theorem. (iii) For a function $v \in H^1(\Omega)$, we define $\widetilde{v} := v - v_\Omega \in H^1_*(\Omega)$, where $v_\Omega \in \mathbb{R}$ denotes the integral mean $v_\Omega := (1/|\Omega|) \int_\Omega v\,dx \in \mathbb{R}$. Note that (2.20) implies that

$$(f \,;\, v_\Omega)_{L^2(\Omega)} + (\phi \,;\, v_\Omega)_{L^2(\Gamma)} = 0.$$

Thus, (2.22) proves that

$$(\nabla u \,;\, \nabla v)_{L^2(\Omega)} = (\nabla u \,;\, \nabla \widetilde{v})_{L^2(\Omega)} = (f \,;\, \widetilde{v})_{L^2(\Omega)} + (\phi \,;\, \gamma \widetilde{v})_{L^2(\Gamma)} = (f \,;\, v)_{L^2(\Omega)} + (\phi \,;\, \gamma v)_{L^2(\Gamma)},$$

i.e., u even solves (2.21). Plugging-in $u = v$, we see that

$$\|\nabla u\|^2_{L^2(\Omega)} \le \sup_{v \in H^1(\Omega)\setminus\{0\}} \frac{(f \,;\, v)_{L^2(\Omega)}}{\|v\|_{H^1(\Omega)}} \|u\|_{H^1(\Omega)} + \sup_{w \in H^{1/2}(\Gamma)\setminus\{0\}} \frac{(\phi \,;\, w)_{L^2(\Gamma)}}{\|w\|_{H^{1/2}(\Gamma)}} \|\gamma u\|_{H^{1/2}(\Gamma)},$$

where we have used that $H^{1/2}(\Gamma) = \mathrm{range}(\gamma)$. Note that the $H^{1/2}$-norm is defined in such a way that $\gamma \in L(H^1(\Omega); H^{1/2}(\Gamma))$ with $\|\gamma u\|_{H^{1/2}(\Gamma)} \leq \|u\|_{H^1(\Omega)}$. Therefore,

$$\|\nabla u\|^2_{L^2(\Omega)} \leq \|u\|_{H^1(\Omega)} \left(\sup_{v \in H^1(\Omega)\setminus\{0\}} \frac{(f\,;\,v)_{L^2(\Omega)}}{\|v\|_{H^1(\Omega)}} + \sup_{w \in H^{1/2}(\Gamma)\setminus\{0\}} \frac{(\phi\,;\,w)_{L^2(\Gamma)}}{\|w\|_{H^{1/2}(\Gamma)}} \right).$$

Together with $(1+\widetilde{C}_P^2)^{-1}\|u\|^2_{H^1(\Omega)} \leq \|\nabla u\|^2_{L^2(\Omega)}$, this proves the first estimate in (2.23). As above, the first supremum may be estimated by $\|f\|_{L^2(\Omega)}$. With the continuous embedding $H^{1/2}(\Gamma) \subset L^2(\Gamma)$, the numerator of the second supremum can be dominated by

$$(\phi\,;\,w)_{L^2(\Gamma)} \leq \|\phi\|_{L^2(\Gamma)}\|w\|_{L^2(\Gamma)} \leq \widetilde{C}\,\|\phi\|_{L^2(\Gamma)}\|w\|_{H^{1/2}(\Gamma)}.$$

This provides the upper bound $\widetilde{C}\,\|\phi\|_{L^2(\Gamma)}$ for the second supremum. (iv) As above, we may use integration by parts to see that

$$(f + \Delta u\,;\,v)_{L^2(\Omega)} + (\phi - \partial u/\partial n\,;\,\gamma v)_{L^2(\Gamma)} = 0 \quad \text{for all } v \in H^1(\Omega).$$

From this, we first conclude $f = -\Delta u$ by use of Theorem 2.1 for test functions $v \in \mathcal{D}(\Omega) \subset H^1_0(\Omega) \subset H^1(\Omega)$. To prove $\phi = \partial u/\partial n$, one proceeds analogously to the remark right after Theorem 2.1. ■

Chapter 3

A Priori Analysis

3.1 P1-Finite Element Method in 2D

A set $T \subset \mathbb{R}^2$ is called a **non-degenerate triangle** provided that there are nodes $x_T, y_T, z_T \in \mathbb{R}^2$ with $T = \operatorname{conv}\{x_T, y_T, z_T\}$ and provided that $|T| > 0$, i.e., T has positive measure. We note that T is in particular bounded and closed, whence compact. We denote by

$$\mathcal{K}_T := \{x_T, y_T, z_T\} \tag{3.1}$$

the **set of nodes** of T and by

$$\mathcal{E}_T := \big\{ \operatorname{conv}\{x_T, y_T\}, \operatorname{conv}\{y_T, z_T\}, \operatorname{conv}\{z_T, x_T\} \big\} \tag{3.2}$$

the **set of edges** of T. The **diameter** of T is denoted by

$$h_T := \operatorname{diam}(T) := \max\big\{ |x - y| \,\big|\, x, y \in T \big\}. \tag{3.3}$$

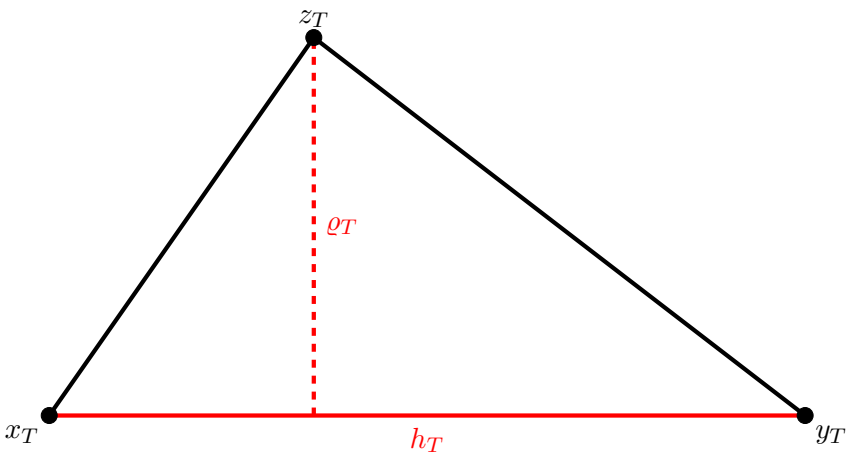


FIGURE 3.1. *The diameter h_T of the triangle T is the length of the longest edge (possibly non unique). The quantity ϱ_T denotes the corresponding height.*

Moreover, we define the **edge length**

$$h_E := \operatorname{diam}(E) := \max\big\{|x-y| \,\big|\, x,y \in E\big\} \tag{3.4}$$

for all edges $E \in \mathcal{E}_T$. Clearly, the diameter h_T of a triangle is the length of the longest edge (possibly non unique), i.e., there is some $E \in \mathcal{E}_T$ with $h_T = h_E$. The **height** over the longest edge E of T is denoted by ϱ_T, cf. Figure 3.1. Recall that the measure of the triangle reads

$$|T| = \frac{h_T \rho_T}{2}. \tag{3.5}$$

The most important example is the **reference triangle**

$$T_{\mathrm{ref}} := \operatorname{conv}\{(0,0),(1,0),(0,1)\} \tag{3.6}$$

which has measure $|T_{\mathrm{ref}}| = 1/2$.

Exercise 11. Give a formal proof that the diameter of a triangle T is the length of one longest edge, i.e., $h_T = \max_{E\in\mathcal{E}_T} h_E$. *Hint:* Use that the convex hull $\operatorname{conv}(M) := \bigcap\big\{\widehat{M} \subseteq \mathbb{R}^d \,\big|\, \widehat{M} \text{ is convex with } M \subseteq \widehat{M}\big\}$ of a set $M \subseteq \mathbb{R}^d$ is also characterized by $\operatorname{conv}(M) = \big\{\sum_{j=1}^N \lambda_j x_j \,\big|\, N \in \mathbb{N}, x_j \in M, \lambda_j \geq 0 \text{ with } \sum_{j=1}^N \lambda_j = 1\big\}$. The proof then directly applies to general simplices in $\mathbb{R}^d$, i.e., $T = \operatorname{conv}\{x_0,\dots,x_d\} \subset \mathbb{R}^d$. □

Definition. A set $\mathcal{T}$ is a **triangulation** of Ω (consisting of triangles) if and only if

- $\mathcal{T}$ is a finite set of non-degenerate triangles,
- the closure of Ω is covered by $\mathcal{T}$, i.e., $\overline{\Omega} = \bigcup \mathcal{T}$,
- for all $T, T' \in \mathcal{T}$ with $T \neq T'$, it holds that $|T \cap T'| = 0$, i.e., the overlap is a set of measure zero.

By $\mathcal{K} := \bigcup\big\{x \in \mathcal{K}_T \,\big|\, T \in \mathcal{T}\big\}$, we then denote the **set of nodes** of the triangulation $\mathcal{T}$ and by $\mathcal{E} := \bigcup\big\{E \in \mathcal{E}_T \,\big|\, T \in \mathcal{T}\big\}$ the **set of edges** of the triangulation $\mathcal{T}$. A triangulation of Ω is called **conforming** or **regular (in the sense of Ciarlet)** provided that the intersection of two elements $T, T' \in \mathcal{T}$ with $T \neq T'$ is

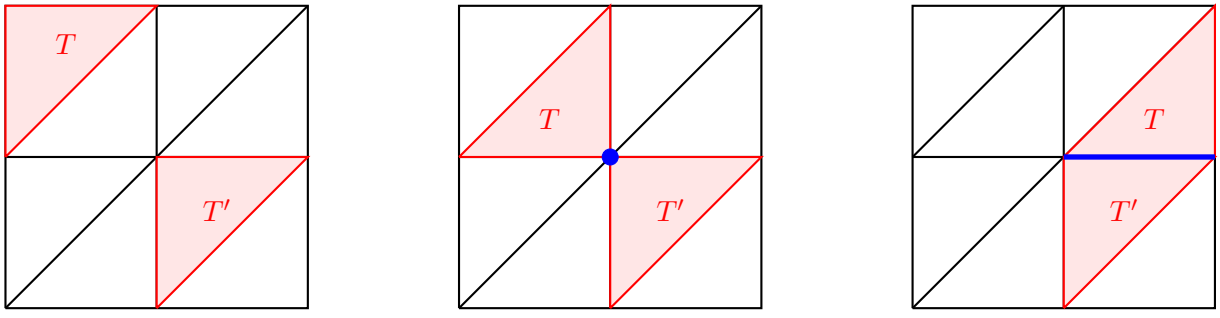


FIGURE 3.2. *For a regular triangulation $\mathcal{T}$, the intersection of two elements $T \neq T'$ is either empty, a joint node, or a joint edge.*

- either empty,
- or a joint node, i.e., $T \cap T' = \{z\} = \mathcal{K}_T \cap \mathcal{K}_{T'}$,
- or a joint edge, i.e., $E := T \cap T' \in \mathcal{E}_T \cap \mathcal{E}_{T'}$,

cf. Figure 3.2. According to this regularity assumption, an edge $E \in \mathcal{E}$ with surface measure $|E \cap \Gamma| > 0$ automatically satisfies $E \subseteq \Gamma$, i.e., an edge E is either a boundary edge or an interior edge. Additionally, we always assume that a regular triangulation resolves the boundary conditions: If $\Gamma = \partial\Omega$ is partitioned into Dirichlet and Neumann boundary Γ_D and Γ_N, respectively, each boundary edge $E \in \mathcal{E}$ with $E \subseteq \Gamma$ satisfies

- either $E \subseteq \overline{\Gamma}_D$
- or $E \subseteq \overline{\Gamma}_N$.

With this assumption, we define the (disjoint) sets of boundary edges

$$\mathcal{E}_D := \big\{E \in \mathcal{E} \,\big|\, E \subseteq \overline{\Gamma}_D\big\} \quad \text{and} \quad \mathcal{E}_N := \big\{E \in \mathcal{E} \,\big|\, E \subseteq \overline{\Gamma}_N\big\} \tag{3.7}$$

as well as the set of all interior edges

$$\mathcal{E}_\Omega := \mathcal{E} \backslash (\mathcal{E}_D \cup \mathcal{E}_N). \tag{3.8}$$

We finally note that, for each $E \in \mathcal{E}_\Omega$, there are two elements $T, T' \in \mathcal{T}$ with $E = T \cap T'$.

Exercise 12. Let $\mathcal{T}$ be a regular triangulation of Ω and $v : \Omega \to \mathbb{R}$ such that $v|_T \in C^1(T)$ for all $T \in \mathcal{T}$. Prove that $v \in H^1(\Omega)$ if and only if $v \in C(\Omega)$. □

The following proposition essentially follows from the regularity of the triangulation $\mathcal{T}$.

Proposition 3.1. *For a regular triangulation $\mathcal{T}$ of Ω, we define the discrete space*

$$\mathcal{S}^1(\mathcal{T}) := \big\{v_h \in C(\Omega) \,\big|\, \forall T \in \mathcal{T} \quad v_h|_T \text{ affine}\big\} \tag{3.9}$$

of all $\mathcal{T}$-piecewise affine and globally continuous functions. Then, there holds the following:
(i) *$\mathcal{S}^1(\mathcal{T})$ is an N-dimensional subspace of $H^1(\Omega)$ with $N = \#\mathcal{K}$ the number of nodes.*
(ii) *For each node $z \in \mathcal{K}$, there is a unique* ***hat function***

$$\zeta_z \in \mathcal{S}^1(\mathcal{T}) \quad \text{with} \quad \zeta_z(z') = \delta_{zz'} \quad \text{for all } z' \in \mathcal{K}. \tag{3.10}$$

(iii) *The set $\mathcal{B} := \big\{\zeta_z \,\big|\, z \in \mathcal{K}\big\}$ is a basis of $\mathcal{S}^1(\mathcal{T})$, the so-called* ***nodal basis****.*

Proof. **1. step.** According to the regularity of $\mathcal{T}$, hat functions ζ_z are automatically continuous on Ω: For each element $T \in \mathcal{T}$, an affine function $v_h : T \to \mathbb{R}$ is uniquely determined by the nodal values $v_h(z)$ for $z \in \mathcal{K}_T$. Therefore, the $\mathcal{T}$-piecewise affine hat function ζ_z defined by $\zeta_z(z') = \delta_{zz'}$ is uniquely defined. We now show that $\zeta_z \in C(\Omega)$: If $T, T' \in \mathcal{T}$ are elements with $T \cap T' \neq \emptyset$, regularity of $\mathcal{T}$ implies that either $T = T'$ or $\{z'\} = T \cap T'$ is a joint point or $E = T \cap T'$ is a joint edge. In the latter case, note that the trace on E of the affine function $\zeta_z|_T$ as well as of $\zeta_z|_{T'}$ is

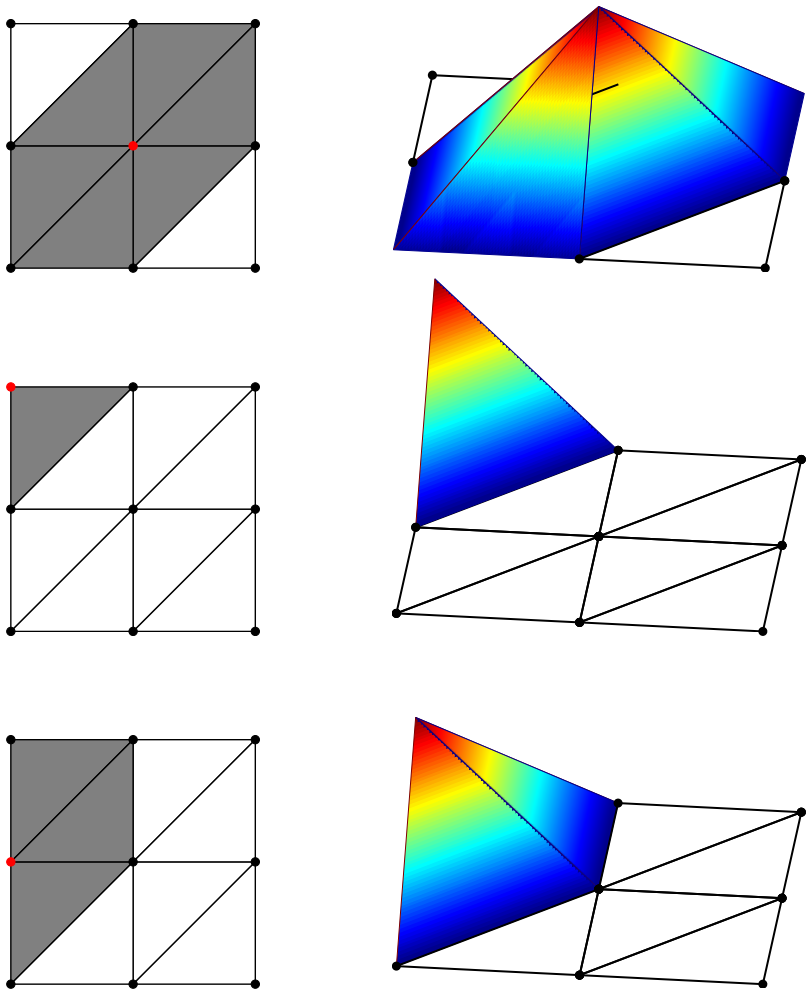


FIGURE 3.3. *Examples of P1 hat functions ζ_z: The left figures show the mesh as well as the support* $\operatorname{supp}(\zeta_z)$ *in grey, where the corresponding node $z \in \mathcal{K}$ is indicated in red. The right figures show the plots of the hat functions. Triangles $T \in \mathcal{T}$ with $\zeta_z|_T = 0$ are filled with white.*

uniquely defined on the edge E by the nodal values $\zeta_z(x_E)$ and $\zeta_z(y_E)$, where $E = \operatorname{conv}\{x_E, y_E\}$. Therefore the traces of $\zeta_z|_T$ and $\zeta_z|_{T'}$ on E coincide, i.e., ζ_z is continuous on each interior edge.

2. step. The nodal basis $\mathcal{B}$ is a basis of $\mathcal{S}^1(\mathcal{T})$ and $\dim \mathcal{S}^1(\mathcal{T}) = \#\mathcal{K}$: Clearly, the hat functions are linearly independent, $\mathcal{B} \subseteq \mathcal{S}^1(\mathcal{T})$, and $\#\mathcal{B} = \#\mathcal{K}$. Moreover, each function $v_h \in \mathcal{S}^1(\mathcal{T})$ is uniquely defined by the nodal values $v_h(z)$ for $z \in \mathcal{K}$ and can thus be written as the linear combination of the hat functions, i.e., $\mathcal{S}^1(\mathcal{T}) \subseteq \operatorname{span}(\mathcal{B})$.

3. step. The inclusion $\mathcal{S}^1(\mathcal{T}) \subset H^1(\Omega)$ follows from Exercise 12. ■

Remark. Examples for hat functions ζ_z are shown in Figure 3.3. Note that the support $\operatorname{supp}(\zeta_z)$ is always local. This leads to a sparse Galerkin matrix A, i.e., most of the entries of A are zero. □

For a given Dirichlet boundary $\Gamma_D \subseteq \Gamma$, we use the discrete space $\mathcal{S}^1_D(\mathcal{T})$ to discretize the weak form of the mixed boundary value problem. In case of $\Gamma_D = \Gamma$, we consider the space $\mathcal{S}^1_0(\mathcal{T})$.

Corollary 3.2. *Let $\mathcal{T}$ be a regular triangulation of Ω. Then, the space*

$$\mathcal{S}^1_D(\mathcal{T}) := \big\{ v_h \in \mathcal{S}^1(\mathcal{T}) \,\big|\, \forall z \in \mathcal{K} \cap \overline{\Gamma}_D \quad v_h(z) = 0 \big\} \tag{3.11}$$

is a finite dimensional subspace of $H^1_D(\Omega)$ of dimension $\#\big\{ z \in \mathcal{K} \,\big|\, z \notin \overline{\Gamma}_D \big\}$. The space

$$\mathcal{S}^1_0(\mathcal{T}) := \big\{ v_h \in \mathcal{S}^1(\mathcal{T}) \,\big|\, \forall z \in \mathcal{K} \cap \Gamma \quad v_h(z) = 0 \big\} \tag{3.12}$$

is a finite dimensional subspace of $H^1_0(\Omega)$ of dimension $\#\big\{ z \in \mathcal{K} \,\big|\, z \notin \Gamma \big\}$.

Proof. We only need to show that $v_h|_{\Gamma_D} = 0$ for $v_h \in \mathcal{S}^1_D(\mathcal{T})$. Let $x \in \Gamma_D$. According to the regularity of $\mathcal{T}$, there is an edge $E \in \mathcal{E}_D$ such that $x \in E$. Since the trace $v_h|_E$ is affine, it is uniquely determined by the nodal values $v_h(x_T) = 0 = v_h(y_T)$, where $E = \operatorname{conv}\{x_T, y_T\}$. Consequently, $v_h|_E = 0$ for all $E \in \mathcal{E}_D$ and hence $v_h \in H^1_D(\Omega)$. In particular, we obtain the claim for $\mathcal{S}^1_0(\mathcal{T})$ in case of $\Gamma_D = \Gamma$. ■

For the discretization of the Neumann problem, we are dealing with $\mathcal{S}^1_*(\mathcal{T})$.

Corollary 3.3. *For a regular triangulation $\mathcal{T}$ of Ω, the space*

$$\mathcal{S}^1_*(\mathcal{T}) := \big\{ v_h \in \mathcal{S}^1(\mathcal{T}) \,\big|\, \textstyle\int_\Omega v_h \, dx = 0 \big\} \tag{3.13}$$

is a finite dimensional subspace of $H^1_*(\Omega)$ of dimension $\#\mathcal{K} - 1$.

Proof. Clearly, it holds that $\mathcal{S}^1_*(\mathcal{T}) \subseteq H^1_*(\Omega)$. Note that $I(v_h) := \int_\Omega v_h \, dx$ is a linear functional on $\mathcal{S}^1(\mathcal{T})$ with kernel $\mathcal{S}^1_*(\mathcal{T}) = \ker(I)$. Since $\operatorname{rank}(I) = 1$, Linear Algebra yields that $\dim \mathcal{S}^1_*(\mathcal{T}) = \dim \mathcal{S}^1(\mathcal{T}) - 1$. ■

The **P1 Finite Element Method** now consists of using the Galerkin method with the discrete spaces $\mathcal{S}^1_0(\mathcal{T})$, $\mathcal{S}^1_D(\mathcal{T})$, and $\mathcal{S}^1_*(\mathcal{T})$ to approximate the weak solution of the Dirichlet problem, the mixed boundary value problem, and the Neumann problem, respectively. From now on, we shall assume that $\mathcal{T}$ is a regular triangulation of Ω. We start with the **Dirichlet problem**

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

for given data $f \in L^2(\Omega)$. The P1-FEM then reads: Find $u_h \in \mathcal{S}^1_0(\mathcal{T})$ such that

$$(\nabla u_h \,;\, \nabla v_h)_{L^2(\Omega)} = (f \,;\, v_h)_{L^2(\Omega)} \quad \text{for all } v_h \in \mathcal{S}^1_0(\mathcal{T}). \tag{3.14}$$

Second, the **mixed boundary value problem** reads

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

with $\Gamma = \overline{\Gamma}_D \cup \overline{\Gamma}_N$, $\Gamma_D \cap \Gamma_N = \emptyset$, and $|\Gamma_D| > 0$. The data satisfy $f \in L^2(\Omega)$ and $\phi \in L^2(\Gamma_N)$. The P1-FEM for the mixed BVP reads: Find $u_h \in \mathcal{S}^1_D(\mathcal{T})$ such that

$$(\nabla u_h \,;\, \nabla v_h)_{L^2(\Omega)} = (f \,;\, v_h)_{L^2(\Omega)} + (\phi \,;\, v_h)_{L^2(\Gamma_N)} \quad \text{for all } v_h \in \mathcal{S}^1_D(\mathcal{T}). \tag{3.15}$$

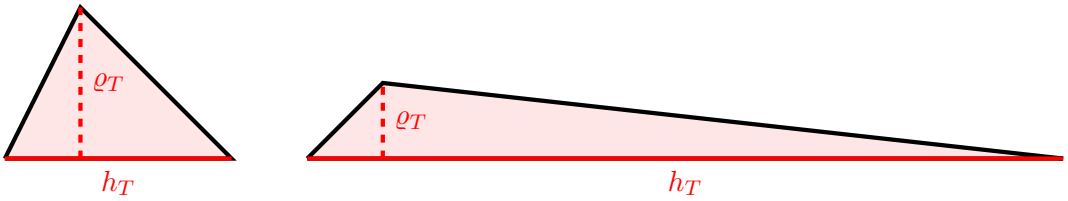


FIGURE 3.4. *two triangles with small (left) and large (right) shape regularity constant* $\sigma(T) := h_T/\varrho_T$

Finally, we consider the **Neumann problem**

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

where the data $f \in L^2(\Omega)$ and $\phi \in L^2(\Gamma)$ are assumed to satisfy $\int_\Omega f\,dx + \int_\Gamma \phi\,ds = 0$. The P1-FEM for the Neumann problem reads: Find $u_h \in \mathcal{S}^1_*(\mathcal{T})$ such that

$$(\nabla u_h \,;\, \nabla v_h)_{L^2(\Omega)} = (f \,;\, v_h)_{L^2(\Omega)} + (\phi \,;\, v_h)_{L^2(\Gamma)} \quad \text{for all } v_h \in \mathcal{S}^1_*(\mathcal{T}). \tag{3.16}$$

3.2 Approximation Theorem and Bramble-Hilbert Lemma

3.2.1 Uniform Mesh-Refinement and Shape Regularity

Let $h \in L^\infty(\Omega)$ and $\varrho \in L^\infty(\Omega)$ denote the **local mesh-width** functions which are defined by

$$h|_T := h_T = \operatorname{diam}(T) \quad \text{and} \quad \varrho|_T := \varrho_T \quad \text{for all } T \in \mathcal{T}. \tag{3.17}$$

Moreover, the quantities

$$\sigma(T) := \frac{h_T}{\varrho_T} \quad \text{and} \quad \sigma(\mathcal{T}) := \|h/\varrho\|_{L^\infty(\Omega)} = \max_{T\in\mathcal{T}} \frac{h_T}{\varrho_T} \geq 1 \tag{3.18}$$

denote the **shape regularity constant** of an element $T \in \mathcal{T}$ resp. the triangulation $\mathcal{T}$, see Fig. 3.4. Note that $|T| = h_T \varrho_T/2$ so that $2\,h_T/\varrho_T = h_T^2/|T|$. The shape regularity constant will affect all error estimates, so that mesh-refinement has to avoid a blow-up of $\sigma(\mathcal{T})$. We say that a regular mesh $\mathcal{T}$ is **γ-shape regular**, if $\sigma(\mathcal{T}) \leq \gamma < \infty$.

For this section, we stick with the so-called **uniform mesh-refinement**: Given a regular triangulation $\mathcal{T}^{(\text{old})}$, we obtain a new triangulation $\mathcal{T}^{(\text{new})}$ as follows: Each element $T \in \mathcal{T}^{(\text{old})}$ is split into 4 similar triangles $T_1, \dots, T_4 \in \mathcal{T}^{(\text{new})}$, cf. Figure 3.5. Therefore, each node $z \in \mathcal{K}^{(\text{new})}$ either belongs to $\mathcal{K}^{(\text{old})}$ or is the midpoint of an edge $E \in \mathcal{E}^{(\text{old})}$. We stress some simple observations:

- The new triangulation $\mathcal{T}^{(\text{new})}$ is also regular.
- The local mesh-width functions satisfy $h^{(\text{new})} = h^{(\text{old})}/2$ and $\varrho^{(\text{new})} = \varrho^{(\text{old})}/2$.
- In particular, the shape regularity constant satisfies that $\sigma(\mathcal{T}^{(\text{old})}) = \sigma(\mathcal{T}^{(\text{new})})$.

Further mesh-refinement strategies are discussed in the following section.

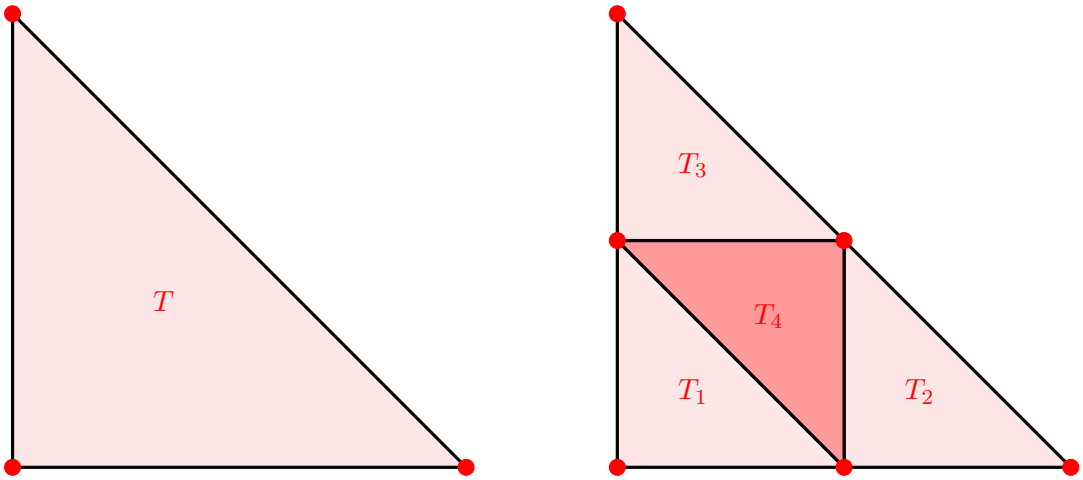


FIGURE 3.5. *Red-refinement refines the element $T \in \mathcal{T}^{(\mathrm{old})}$ into 4 similar elements $T_1, \dots, T_4 \in \mathcal{T}^{(\mathrm{new})}$. The new nodes $\mathcal{K}^{(\mathrm{new})} \backslash \mathcal{K}^{(\mathrm{old})}$ are just the edge midpoints for all edges $E \in \mathcal{E}^{(\mathrm{old})}$. In particular, regularity of $\mathcal{T}^{(\mathrm{old})}$ implies regularity of $\mathcal{T}^{(\mathrm{new})}$.*

Exercise 13. Let $T = \mathrm{conv}\{z_1, z_2, z_3\}$ be a non-degenerate triangle in $\mathbb{R}^2$. Prove that the shape regularity constant h_T/ϱ_T tends to infinity if and only if the smallest angle in T tends to zero. □

Exercise 14. Often, the shape regularity constant is defined as the maximal quotient h_T/r_T, where $r_T > 0$ denotes the maximal radius of a ball $B(x, r_T) := \{y \in \mathbb{R}^2 \,|\, |x-y| \le r_T\}$ inscribed in T, i.e., $B(x, r_T) \subseteq T$. Let $T = \mathrm{conv}\{z_1, z_2, z_3\}$ be a non-degenerate triangle in $\mathbb{R}^2$. What is the relation between ϱ_T and r_T? □

3.2.2 Statement and Interpretation of Approximation Theorem

To state our first main result in this section, we need to know that certain Sobolev functions are at least continuous.

Theorem 3.4 (Sobolev). *Let Ω be a Lipschitz domain in $\mathbb{R}^d$ and $m > d/2$. Then, there holds the continuous inclusion $H^m(\Omega) \subseteq C(\overline{\Omega})$.* ■

In particular, for $d = 2, 3$, each Sobolev function $u \in H^2(\Omega)$ is continuous so that evaluation of u at the nodes $z \in \mathcal{K}$ is well-defined. Throughout the remaining section, we assume that $\mathcal{T}$ is a regular triangulation of a bounded Lipschitz domain $\Omega \subset \mathbb{R}^2$. We stress, however, that the same results — even with the same proofs — hold for $d = 3$ as well. As in the previous section, the nodal basis function corresponding to a node $z \in \mathcal{K}$ is denoted by $\zeta_z \in \mathcal{S}^1(\mathcal{T})$.

Theorem 3.5 (Approximation Theorem). *For $u \in H^2(\Omega)$, the **nodal interpolant** reads*

$$I_h u := \sum_{z \in \mathcal{K}} u(z)\zeta_z \in \mathcal{S}^1(\mathcal{T}). \tag{3.19}$$

For all $T \in \mathcal{T}$, there hold the elementwise error estimates

$$\|u - I_h u\|_{L^2(T)} \le C \, \|h^2 D^2 u\|_{L^2(T)} \tag{3.20}$$

and

$$\|\nabla(u - I_h u)\|_{L^2(T)} \le C \, \sigma(T) \, \|hD^2 u\|_{L^2(T)}, \tag{3.21}$$

where the generic constant $C > 0$ is independent of u, $\mathcal{T}$, and Ω, but depends only on the reference triangle. In particular, this proves for all $\alpha \in \mathbb{R}$ the global error estimates

$$\|h^\alpha (u - I_h u)\|_{L^2(\Omega)} \le C \, \|h^{2+\alpha} D^2 u\|_{L^2(\Omega)} \tag{3.22}$$

and

$$\|h^\alpha \nabla(u - I_h u)\|_{L^2(\Omega)} \le C \, \sigma(\mathcal{T}) \, \|h^{1+\alpha} D^2 u\|_{L^2(\Omega)}. \tag{3.23}$$

This theorem will be shown later. First, we discuss the following immediate consequence:

Corollary 3.6. *For $u \in H^2(\Omega) \cap H^1_D(\Omega)$, it holds that $I_h u \in \mathcal{S}^1_D(\mathcal{T})$ and thus*

$$\min_{v_h \in \mathcal{S}^1_D(\mathcal{T})} \|u - v_h\|_{H^1(\Omega)} \le \|u - I_h u\|_{H^1(\Omega)} \le C \, \sigma(\mathcal{T}) \, \|hD^2 u\|_{L^2(\Omega)}. \tag{3.24}$$

For $u \in H^2(\Omega) \cap H^1_*(\Omega)$, it holds that

$$\begin{aligned} \min_{v_h \in \mathcal{S}^1_*(\mathcal{T})} \|u - v_h\|_{H^1(\Omega)} = \min_{v_h \in \mathcal{S}^1(\mathcal{T})} \|u - v_h\|_{H^1(\Omega)} &\le \|u - I_h u\|_{H^1(\Omega)} \\ &\le C \, \sigma(\mathcal{T}) \, \|hD^2 u\|_{L^2(\Omega)}. \end{aligned} \tag{3.25}$$

In either case, the constant $C > 0$ depends only on $\operatorname{diam}(\Omega)$.

Proof. Let $C_{\mathrm{apx}} > 0$ denote the constant from the approximation theorem. Then,

$$\|u - I_h u\|^2_{H^1(\Omega)} = \|u - I_h u\|^2_{L^2(\Omega)} + \|\nabla(u - I_h u)\|^2_{L^2(\Omega)} \le C^2_{\mathrm{apx}} (\operatorname{diam}(\Omega)^2 + \sigma(\mathcal{T})^2) \|hD^2 u\|^2_{L^2(\Omega)}.$$

Since $\sigma(\mathcal{T}) \ge 1$, we obtain that

$$\|u - I_h u\|_{H^1(\Omega)} \le C_{\mathrm{apx}} \sigma(\mathcal{T}) (\operatorname{diam}(\Omega)^2 + 1)^{1/2} \, \|hD^2 u\|_{L^2(\Omega)}.$$

For $u \in H^2(\Omega) \cap H^1_D(\Omega)$, it holds that $u(z) = 0$ for all $z \in \overline{\Gamma}_D$. This implies that $I_h u \in \mathcal{S}^1_D(\mathcal{T})$ and hence (3.24). Before we prove (3.25), note that $I_h u \in \mathcal{S}^1(\mathcal{T})$ does not belong to $\mathcal{S}^1_*(\mathcal{T})$ in general. However, let $\mathbb{P}_h : H^1(\Omega) \to \mathcal{S}^1(\mathcal{T})$ denote the H^1-orthogonal projection onto $\mathcal{S}^1(\mathcal{T})$. Since $1 \in \mathcal{S}^1(\mathcal{T})$, it holds that

$$0 = \int_\Omega u \, dx = (u \,;\, 1)_{H^1(\Omega)} = (\mathbb{P}_h u \,;\, 1)_{H^1(\Omega)} = \int_\Omega \mathbb{P}_h u \, dx \quad \text{for all } u \in H^1_*(\Omega).$$

Therefore, $\mathbb{P}_h u \in \mathcal{S}^1_*(\mathcal{T})$, and the best approximation property of the orthogonal projection $\mathbb{P}_h$ thus implies that

$$\|u - \mathbb{P}_h u\|_{H^1(\Omega)} = \min_{v_h \in \mathcal{S}^1(\mathcal{T})} \|u - v_h\|_{H^1(\Omega)} \leq \min_{v_h \in \mathcal{S}^1_*(\mathcal{T})} \|u - v_h\|_{H^1(\Omega)} \leq \|u - \mathbb{P}_h u\|_{H^1(\Omega)}$$

and hence equality. As before, this proves (3.25). ■

Remark. Corollary 3.6 has two important consequences: First, according to Céa's lemma, the Galerkin error is up to a constant the best approximation error. For a smooth exact solution $u \in H^2(\Omega)$, the P1-FEM thus leads (at least and in fact even) to a convergence order $\mathcal{O}(h)$. Second, $C^\infty_D(\overline{\Omega})$ is dense in $H^1_D(\Omega)$ and $C^\infty_*(\overline{\Omega}) := \big\{v \in C^\infty(\overline{\Omega}) \,\big|\, \int_\Omega v\,dx = 0\big\}$ is dense in $H^1_*(\Omega)$. Corollary 3.6 therefore implies convergence of the Galerkin scheme on a dense subspace. The abstract framework provides convergence of the P1-FEM even without any regularity assumptions on u, cf. Proposition 1.7. □

Exercise 15. Use the Poincaré inequality and the Meyers-Serrin theorem to prove that $C^\infty_*(\overline{\Omega})$ is dense in $H^1_*(\Omega)$. □

3.2.3 Bramble-Hilbert Lemma

It now remains to prove the Approximation Theorem 3.5. The proof of which needs three lemmata. The first two lemmata provide the basis for general scaling arguments. We therefore state the results even in a slightly generalized setting.

Definition. For a multiindex $\alpha \in \mathbb{N}^d_0$ and $x \in \mathbb{R}^d$, we define the **monomial** $x^\alpha := \prod_{j=1}^d x_j^{\alpha_j}$, where $|\alpha| := \sum_{j=1}^d \alpha_j$ is the **(total) degree** of α. For a Lipschitz domain $T \subseteq \mathbb{R}^d$, we define

$$\mathcal{P}^m(T) := \big\{v : T \to \mathbb{R} \,\big|\, v \text{ is linear combination of monomials of degree } \leq m\big\} \tag{3.26}$$

the space that consists of all **polynomials** of degree less than or equal to $m \in \mathbb{N}$.

Lemma 3.7 (Bramble-Hilbert). *For a Lipschitz domain $T \subset \mathbb{R}^d$ and a normed space X, let $A \in L(H^{m+1}(T); X)$ be a linear and continuous operator with $\mathcal{P}^m(T) \subseteq \ker(A)$. Besides the classical continuity estimate*

$$\|Av\|_X \leq \|A\| \|v\|_{H^{m+1}(T)} \quad \text{for all } v \in H^{m+1}(T), \tag{3.27}$$

it holds that

$$\|Av\|_X \leq C\, \|A\| \|D^{m+1} v\|_{L^2(T)} \quad \text{for all } v \in H^{m+1}(T), \tag{3.28}$$

where the constant $C > 0$ depends only on m and T.

Proof. **1. step.** Construct an equivalent norm on $H^{m+1}(T)$: Note that $\mathcal{P}^m(T)$ is a finite dimensional space. Let $\Pi : L^2(T) \to \mathcal{P}^m(T)$ denote the L^2-orthogonal projection onto $\mathcal{P}^m(T)$. We define

$$|\!|\!| v |\!|\!| := \|D^{m+1} v\|_{L^2(T)} + \|\Pi v\|_{L^2(T)} \quad \text{for } v \in H^{m+1}(T).$$

From $\|\Pi v\|_{L^2(T)} \leq \|v\|_{L^2(T)}$, we infer that

$$|||v||| \leq \|D^{m+1}v\|_{L^2(T)} + \|v\|_{L^2(T)} \leq \sqrt{2}\,\|v\|_{H^{m+1}(T)}.$$

Next, we prove the converse inequality, i.e., there exists a constant $C > 0$ such that

$$\|v\|_{H^{m+1}(T)} \leq C\,|||v||| \quad \text{for all } v \in H^{m+1}(T).$$

As above, we use the Rellich theorem and argue by contradiction: If the claim is wrong, we find $v_n \in H^{m+1}(T)$ such that $\|v_n\|_{H^{m+1}(T)} > n|||v_n|||$. We define $w_n := v_n/\|v_n\|_{H^{m+1}(T)}$. Note that

$$\|w_n\|_{H^{m+1}(T)} = 1 \quad \text{as well as} \quad |||w_n||| \leq \frac{1}{n}.$$

According to reflexivity, we may thus assume that $w_n \rightharpoonup w \in H^{m+1}(T)$. According to Lemma 2.6, convexity and continuity of $|||\cdot|||$ imply that $|||w||| = 0$. Therefore, it holds that $D^{m+1}w = 0$ as well as $\Pi w = 0$. With the help of Exercise 16, we deduce that $w \in \mathcal{P}^m(T)$ and consequently $\|w\|_{L^2(T)} = \|\Pi w\|_{L^2(T)} = 0$. According to Rellich's theorem, we have $w_n \to w = 0 \in H^m(T)$. Since $D^{m+1}w_n \to 0 \in L^2(T)$, we even conclude that $w_n \to 0 = w \in H^{m+1}(T)$. This however, contradicts $\|w_n\|_{H^{m+1}(T)} = 1$. Altogether, we have shown that $|||\cdot|||$ is an equivalent norm on $H^{m+1}(T)$.

2. step. With the norm equivalence constant $C > 0$ of step 1, it holds that

$$\|Av\|_X = \|A(v - \Pi v)\|_X \leq \|A\|\,\|v - \Pi v\|_{H^{m+1}(T)} \leq C\,\|A\|\,|||v - \Pi v||| = C\,\|A\|\,\|D^{m+1}v\|_{L^2(T)}$$

for all $v \in H^{m+1}(T)$. ■

Exercise 16. Prove that a function $v \in H^{m+1}(T)$ on a bounded Lipschitz domain $T \subset \mathbb{R}^d$ satisfies $D^{m+1}v = 0$ if and only if $v \in \mathcal{P}^m(T)$. **Hint:** You should use the case $m = 0$ without a proof, cf. Theorem 2.3. □

3.2.4 Scaling Argument and Proof of Approximation Theorem

Lemma 3.8 (Transformation Formula). *Let $T, \widehat{T} \subset \mathbb{R}^d$ be Lipschitz domains. Let $\Phi(x) := Bx + y$ with regular matrix $B \in \mathbb{R}^{d\times d}$ and vector $y \in \mathbb{R}^d$ be an affine diffeomorphism with $\Phi(\widehat{T}) = T$. For $u \in H^m(T)$, it holds that $u \circ \Phi \in H^m(\widehat{T})$ with*

$$\|D^m(u \circ \Phi)\|_{L^2(\widehat{T})} \leq |\det B|^{-1/2}\|B\|_F^m\|D^m u\|_{L^2(T)}, \tag{3.29}$$

where $\|B\|_F$ denotes the Frobenius norm of B. Moreover, for $m = 0$, there even holds equality.

Proof. **1. step.** The case $m = 0$: According to the transformation theorem and $D\Phi(x) = B$, it holds that

$$\|u\|^2_{L^2(T)} = \int_T u^2\,dy = \int_{\widehat{T}} (u \circ \Phi)^2\,|\det D\Phi|\,dx = |\det B|\,\|u \circ \Phi\|^2_{L^2(\widehat{T})}.$$

2. step. To treat the higher-order case for smooth functions $u \in C^\infty(\overline{T})$, we first prove by induction on m that for all $j_\ell \in \{1, \dots, d\}$, it holds that

$$\partial_{j_1} \cdots \partial_{j_m} (u \circ \Phi)(x) = \sum_{k_1=1}^{d} \cdots \sum_{k_m=1}^{d} \partial_{k_1} \cdots \partial_{k_m} u(\Phi(x)) \prod_{\ell=1}^{m} B_{k_\ell j_\ell}, \tag{3.30}$$

which is the special case of the Faà di Bruno formula (chain rule for partial derivatives): The case $m = 1$ follows from the chain rule $D(u \circ \Phi)(x) = Du(\Phi(x))D\Phi(x) = Du(\Phi(x))B$, where, e.g., $Du(y) = (\partial_1 u, \dots, \partial_d u)(y)$. Therefore,

$$\partial_j (u \circ \Phi)(x) = \sum_{k=1}^{d} \partial_k u(\Phi(x)) B_{kj}.$$

Assuming that (3.30) holds up to $m \in \mathbb{N}$, we now prove the equality for $m + 1$:

$$\begin{aligned}
\partial_{j_1} \cdots \partial_{j_{m+1}} (u \circ \Phi)(x) &\overset{!}{=} \partial_{j_1} \Bigg(\sum_{k_2=1}^{d} \cdots \sum_{k_{m+1}=1}^{d} \partial_{k_2} \cdots \partial_{k_{m+1}} u(\Phi(x)) \prod_{\ell=2}^{m+1} B_{k_\ell j_\ell} \Bigg) \\
&= \sum_{k_2=1}^{d} \cdots \sum_{k_{m+1}=1}^{d} \partial_{j_1} \big(\partial_{k_2} \cdots \partial_{k_{m+1}} u(\Phi(x)) \big) \prod_{\ell=2}^{m+1} B_{k_\ell j_\ell} \\
&\overset{!}{=} \sum_{k_2=1}^{d} \cdots \sum_{k_{m+1}=1}^{d} \sum_{k_1=1}^{d} \partial_{k_1} \partial_{k_2} \cdots \partial_{k_{m+1}} u(\Phi(x)) B_{k_1 j_1} \prod_{\ell=2}^{m+1} B_{k_\ell j_\ell} \\
&= \sum_{k_1=1}^{d} \cdots \sum_{k_{m+1}=1}^{d} \partial_{k_1} \partial_{k_2} \cdots \partial_{k_{m+1}} u(\Phi(x)) \prod_{\ell=1}^{m+1} B_{k_\ell j_\ell},
\end{aligned}$$

where we have used the induction hypothesis for m and the initial step $m = 1$. This verifies (3.30).

3. step. We apply the Cauchy inequality to (3.30) to see that

$$\begin{aligned}
\big| \partial_{j_1} \cdots \partial_{j_m} (u \circ \Phi)(x) \big|^2 &\le \Bigg(\sum_{k_1=1}^{d} \cdots \sum_{k_m=1}^{d} \big| \partial_{k_1} \cdots \partial_{k_m} u(\Phi(x)) \big|^2 \Bigg) \Bigg(\sum_{k_1=1}^{d} \cdots \sum_{k_m=1}^{d} \Big| \prod_{\ell=1}^{m} B_{k_\ell j_\ell} \Big|^2 \Bigg) \\
&= \Bigg(\sum_{k_1=1}^{d} \cdots \sum_{k_m=1}^{d} \big| \partial_{k_1} \cdots \partial_{k_m} u(\Phi(x)) \big|^2 \Bigg) \Bigg(\sum_{k_1=1}^{d} \cdots \sum_{k_m=1}^{d} \prod_{\ell=1}^{m} B_{k_\ell j_\ell}^2 \Bigg) \\
&\overset{!}{=} \Bigg(\sum_{k_1=1}^{d} \cdots \sum_{k_m=1}^{d} \big| \partial_{k_1} \cdots \partial_{k_m} u(\Phi(x)) \big|^2 \Bigg) \Bigg(\prod_{\ell=1}^{m} \sum_{k_\ell=1}^{d} B_{k_\ell j_\ell}^2 \Bigg),
\end{aligned}$$

where the last equality follows from another simple induction argument.

4. step. We prove the transformation formula (3.29) for $u \in C^\infty(\overline{T})$:

$$\begin{aligned}
|\det B| \|D^m(u\circ\Phi)\|^2_{L^2(\widehat{T})} &= \int_{\widehat{T}} \sum_{j_1=1}^d \cdots \sum_{j_m=1}^d |\partial_{j_1}\cdots\partial_{j_m}(u\circ\Phi)(x)|^2 |\det D\Phi(x)|\,dx \\
&\le \underbrace{\Big(\sum_{j_1=1}^d \cdots \sum_{j_m=1}^d \prod_{\ell=1}^m \sum_{k_\ell=1}^d B^2_{k_\ell j_\ell}\Big)}_{=\prod_{\ell=1}^m \sum_{j_\ell=1}^d \sum_{k_\ell=1}^d B^2_{k_\ell j_\ell}} \underbrace{\Big(\int_{\widehat{T}} \sum_{k_1=1}^d \cdots \sum_{k_m=1}^d \big|\partial_{k_1}\cdots\partial_{k_m} u(\Phi(x))\big|^2 |\det D\Phi(x)|\,dx\Big)}_{=\|D^m u\|^2_{L^2(T)}} \\
&= \|B\|_F^{2m} \|D^m u\|^2_{L^2(T)}.
\end{aligned}$$

5. step. We prove the transformation formula (3.29) for general $u \in H^m(T)$: According to the Meyers-Serrin theorem, $C^\infty(\overline{T})$ is a dense subspace of $H^m(T)$. Note that (3.29) implies for $u \in C^\infty(\overline{T})$ the estimate $\|u\circ\Phi\|_{H^m(\widehat{T})} \le C\,\|u\|_{H^m(T)}$, where $C > 0$ depends only on m and B. Hence, $\Psi u := u\circ\Phi$ extends uniquely to a linear and continuous mapping $\Psi : H^m(T) \to H^m(\widehat{T})$. For $u \in H^m(T)$, choose $(u_n) \subset C^\infty(\overline{T})$ with $u_n \to u \in H^m(T)$. By continuity of Ψ, it holds that $u_n\circ\Phi = \Psi u_n \to \Psi u$ in $H^m(\widehat{T})$. Moreover, according to step 1, it holds that $u_n\circ\Phi \to u\circ\Phi \in L^2(\widehat{T})$. This implies that $u\circ\Phi = \Psi u \in H^m(\widehat{T})$, i.e., the (unique) extension of Ψ from $C^\infty(\overline{T})$ to $H^m(T)$ is, in fact, the composition. Moreover, the left-hand side and the right-hand side of (3.29) depend continuously (with respect to $H^m(T)$) on u. This and (3.29) for $u_n \in C^\infty(\overline{T})$ prove that

$$\begin{aligned}
\|D^m(u\circ\Phi)\|_{L^2(\widehat{T})} = \lim_{n\to\infty} \|D^m(u_n\circ\Phi)\|_{L^2(\widehat{T})} &\le \lim_{n\to\infty} |\det B|^{-1/2} \|B\|_F^m \|D^m u_n\|_{L^2(T)} \\
&= |\det B|^{-1/2} \|B\|_F^m \|D^m u\|_{L^2(T)}
\end{aligned}$$

and conclude the proof. ■

Lemma 3.9. *For $\widehat{T} = T_{\mathrm{ref}}$ the reference element and $T = \mathrm{conv}\{z_1, z_2, z_3\} \subset \mathbb{R}^2$ being a non-degenerate triangle, we define*

$$\Phi_T : T_{\mathrm{ref}} \to T, \quad \Phi_T(s,t) := z_1 + B\begin{pmatrix} s \\ t \end{pmatrix}, \quad \text{where } B := \begin{pmatrix} z_2 - z_1 & z_3 - z_1 \end{pmatrix} \in \mathbb{R}^{2\times 2}. \tag{3.31}$$

Then, it holds that $|\det B| = 2\,|T|$ and

$$h_T/\sqrt{2} \le \|B\|_F \le \sqrt{2}\,h_T \quad \text{as well as} \quad \varrho_T^{-1}/\sqrt{2} \le \|B^{-1}\|_F \le \sqrt{2}\,\varrho_T^{-1}. \tag{3.32}$$

Proof. It holds that

$$\|B\|_F^2 = |z_2 - z_1|^2 + |z_3 - z_1|^2 \le 2h_T^2.$$

Moreover,

$$|z_3 - z_2| \le |z_3 - z_1| + |z_2 - z_1| \le \sqrt{2}\left(|z_3 - z_1|^2 + |z_2 - z_1|^2\right)^{1/2} \le \sqrt{2}\,\|B\|_F.$$

In particular, $h_T = \max\{|z_2 - z_1|, |z_3 - z_1|, |z_3 - z_2|\} \le \sqrt{2}\,\|B\|_F$. The transformation theorem gives

$$\frac{1}{2}\,|\det B| = |T_{\mathrm{ref}}|\,|\det B| = \int_{T_{\mathrm{ref}}} |\det D\Phi_T|\,dx = \int_T dx = |T| > 0.$$

Hence, $0 < |\det B| = 2\,|T| = h_T \varrho_T$. In particular, B^{-1} as well as ϱ_T^{-1} are well-defined. It holds that

$$B^{-1} = \frac{1}{\det B} \begin{pmatrix} b_{22} & -b_{12} \\ -b_{21} & b_{11} \end{pmatrix} \quad \text{for} \quad B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}.$$

In particular, this proves that

$$\|B^{-1}\|_F = \frac{\|B\|_F}{|\det B|} = \frac{\|B\|_F}{h_T \varrho_T},$$

and the second estimate in (3.32) follows from the first. ■

Proof of Approximation Theorem 3.5. **1. step.** Estimate on the reference element T_{ref}: Let $I_h^{\mathrm{ref}} : H^2(T_{\mathrm{ref}}) \to \mathcal{P}^1(T_{\mathrm{ref}})$ denote the nodal interpolation operator on the reference element. We consider the operator

$$A := 1 - I_h^{\mathrm{ref}} : H^2(T_{\mathrm{ref}}) \to H^k(T_{\mathrm{ref}}) \quad \text{for } k = 0, 1$$

and observe that $\mathcal{P}^1(T_{\mathrm{ref}}) \subseteq \ker(A)$. To see that A is continuous, we estimate

$$\|Av\|_{H^k(T_{\mathrm{ref}})} \leq \|v\|_{H^2(T_{\mathrm{ref}})} + \|I_h^{\mathrm{ref}} v\|_{H^k(T_{\mathrm{ref}})}.$$

Let z_1, z_2, z_3 denote the nodes of the reference element. Since all norms on the finite dimensional space $\mathcal{P}^1(T_{\mathrm{ref}})$ are equivalent, we use the Sobolev inequality to see that

$$\|I_h^{\mathrm{ref}} v\|_{H^k(T_{\mathrm{ref}})} \leq C_{\mathrm{norm}} \max_{j=1,\dots,3} |I_h^{\mathrm{ref}} v(z_j)| \leq C_{\mathrm{norm}} \|v\|_{\infty, T_{\mathrm{ref}}} \leq C_{\mathrm{norm}} C_{\mathrm{sobolev}} \|v\|_{H^2(T_{\mathrm{ref}})}.$$

Altogether, we obtain that $\|Av\|_{H^k(T_{\mathrm{ref}})} \leq (1 + C_{\mathrm{norm}} C_{\mathrm{sobolev}}) \|v\|_{H^2(T_{\mathrm{ref}})}$, whence continuity of the operator A. Consequently, the Bramble-Hilbert lemma provides a constant $C_{\mathrm{ref}} > 0$ that depends only on T_{ref} with

$$\|v - I_h^{\mathrm{ref}} v\|_{H^k(T_{\mathrm{ref}})} \leq C_{\mathrm{ref}} \|D^2 v\|_{L^2(T_{\mathrm{ref}})} \quad \text{for all } v \in H^2(T_{\mathrm{ref}}) \text{ and } k = 0, 1.$$

2. step. Scaling arguments provide the estimate on each element T: Let $\Phi = \Phi_T$ denote the affine diffeomorphism from Lemma 3.9. Note that $I_h^{\mathrm{ref}}(u \circ \Phi) = (I_h u) \circ \Phi$. Define $v := u \circ \Phi$ and observe that $(u - I_h u) \circ \Phi = (1 - I_h^{\mathrm{ref}}) v$. First, we apply the transformation formula to Φ^{-1},

$$\begin{aligned} \|D^k (u - I_h u)\|_{L^2(T)} &= \|D^k ((v - I_h^{\mathrm{ref}} v) \circ \Phi^{-1})\|_{L^2(T)} \\ &\leq |\det B^{-1}|^{-1/2} \|B^{-1}\|_F^k \|D^k (v - I_h^{\mathrm{ref}} v)\|_{L^2(T_{\mathrm{ref}})} \\ &\leq C_{\mathrm{ref}} |\det B|^{1/2} \|B^{-1}\|_F^k \|D^2 v\|_{L^2(T_{\mathrm{ref}})}. \end{aligned}$$

Second, we plug-in $v = u \circ \Phi$ and apply the transformation formula to Φ,

$$\|D^2 v\|_{L^2(T_{\mathrm{ref}})} = \|D^2 (u \circ \Phi)\|_{L^2(T_{\mathrm{ref}})} \leq |\det B|^{-1/2} \|B\|_F^2 \|D^2 u\|_{L^2(T)}.$$

The combination of the last two estimates proves that

$$\|D^k (u - I_h u)\|_{L^2(T)} \leq C_{\mathrm{ref}} \|B^{-1}\|_F^k \|B\|_F^2 \|D^2 u\|_{L^2(T)} \leq C_{\mathrm{ref}}\, 2^{(k+2)/2} h_T^2 \varrho_T^{-k} \|D^2 u\|_{L^2(T)},$$

where we have used the geometric interpretation of $\|B\|_F$ and $\|B^{-1}\|_F$. This proves that

$$\|u - I_h u\|_{L^2(T)} \leq 2C_{\mathrm{ref}} \|h^2 D^2 u\|_{L^2(T)} \quad \text{and} \quad \|\nabla(u - I_h u)\|_{L^2(T)} \leq 2^{3/2} C_{\mathrm{ref}}\, \sigma(\mathcal{T}) \|hD^2 u\|_{L^2(T)}.$$

and thus concludes the proof. ■

Remark. The proof of Theorem 3.5 shows that it is enough to assume $u \in C(\overline{\Omega}) \cap H^2(\mathcal{T})$, where $H^k(\mathcal{T}) := \big\{ u \in L^2(\Omega) \,\big|\, \forall T \in \mathcal{T} \quad u \in H^k(T) \big\}$ for $k \geq 1$. According to the Sobolev inequality, it holds that $H^2(\Omega) \subseteq C(\overline{\Omega}) \cap H^2(\mathcal{T})$. For the *broken Sobolev spaces* $H^k(\mathcal{T})$, we write $D_h^k v$ for the $\mathcal{T}$-piecewise k-th derivative of v and, in particular, $\nabla_h v = D_h^1 v$ for the $\mathcal{T}$-piecewise gradient. □

Remark. We recall the procedure of a scaling argument for proving an estimate. To that end, let $\Phi_T : T_{\mathrm{ref}} \to T$ be the affine diffeomorphism with linear part B.

- First, transfer the left-hand side from T to T_{ref}:

$$\begin{aligned} \|D^k v\|_{L^2(T)} = \|D^k (v \circ \Phi_T \circ \Phi_T^{-1})\|_{L^2(T)} &\leq |\det B^{-1}|^{-1/2} \, \|B^{-1}\|_F^k \, \|D^k (v \circ \Phi_T)\|_{L^2(T_{\mathrm{ref}})} \\ &\simeq |T| \, \varrho_T^{-k} \, \|D^k (v \circ \Phi_T)\|_{L^2(T_{\mathrm{ref}})}, \end{aligned}$$

  i.e., derivative on the left-hand side give rise to negative powers of ϱ_T.

- Second, prove an appropriate estimate on the reference element T_{ref}.

- Third, transfer the right-hand side from T_{ref} to T:

$$\|D^\ell (w \circ \Phi_T)\|_{L^2(T_{\mathrm{ref}})} \leq |\det B|^{-1/2} \, \|B\|_F^\ell \, \|D^\ell w\|_{L^2(T)} \simeq |T|^{-1/2} \, h_T^\ell \, \|D^\ell w\|_{L^2(T)},$$

  i.e., derivatives on the right-hand side give rise to positive powers of h_T.

Plugging everything together, proves the desired estimate. □

Note that the heart of the proof of the approximation theorem is the Rellich theorem and thus a compactness argument. The following exercise shows that approximation results are necessarily proved by use of compactness.

Exercise 17. Let X be a Banach space and Y be a normed space with continuous inclusion $Y \subseteq X$. For $h \to 0$, let X_h be finite dimensional subspaces of X and $I_h \in L(Y; X_h)$ be a continuous and linear operator with

$$\|u - I_h u\|_X \leq C h^\alpha \, \|u\|_Y \quad \text{for all } u \in Y,$$

where the constants $C, \alpha > 0$ are independent of u and h. Then, the continuous inclusion $Y \subseteq X$ is already compact. □

Chapter 4

A Posteriori Analysis

4.1 Introduction

We consider the model problem

$$\begin{aligned} -\Delta u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \Gamma_D, \\ \partial u/\partial n &= \phi \quad \text{on } \Gamma_N. \end{aligned} \tag{4.1}$$

Let $u \in H^1_D(\Omega)$ be the weak solution of (4.1) and $u_h \in \mathcal{S}^1_D(\mathcal{T})$ be the P1-FEM approximation of u. In the previous chapter, we aimed to control the error $\|u-u_h\|_{H^1(\Omega)}$ by a priori knowledge, e.g., regularity of the given data and the exact solution (but essentially independent of the discrete solution u_h). Since u is unknown, in general, the **a priori analysis** provides a qualitative understanding of the FEM, e.g., convergence with certain rates, but the derived bounds are non-computable in practice. In this chapter, we aim to derive *numerically computable* bounds $\eta = \eta(u_h, f, \phi, \mathcal{T})$ for the error $\|u-u_h\|_{H^1(\Omega)}$, which may depend on u_h, the triangulation $\mathcal{T}$, and the given data f and ϕ (but *not* on the exact solution u). The quantity η is referred to as **(a posteriori) error estimator**, and emphasis is laid on the fact that η can be computed algorithmically as soon as the discrete solution $u_h \in \mathcal{S}^1_D(\mathcal{T})$ has been computed. An error estimator η is called **reliable** provided that

$$\|u - u_h\|_{H^1(\Omega)} \le C_{\rm rel}\, \eta. \tag{4.2}$$

Usually, the information η provides, is used to steer a mesh-refinement that leads to a sequence $\mathcal{T}_\ell$ of regular meshes with nested spaces $\mathcal{S}^1_D(\mathcal{T}_\ell) \subseteq \mathcal{S}^1_D(\mathcal{T}_{\ell+1})$, i.e., $\mathcal{T}_{\ell+1}$ is a certain refinement of $\mathcal{T}_\ell$. If η is reliable the (numerically or algorithmically observed) decrease of η to zero implies the convergence of u_h towards u. However, it might (formally) occur that u_h tends to u, while η does not tend to zero. Therefore, an error estimator η is called **efficient** provided that

$$C_{\rm eff}\, \eta \le \|u - u_h\|_{H^1(\Omega)}. \tag{4.3}$$

For an efficient error estimator η, the convergence of u_h to u necessarily implies the convergence of η to zero. Finally, if η is reliable and efficient, we observe for η the same order of convergence as for $\|u - u_h\|_{H^1(\Omega)}$.

The aim of a posteriori error estimates is twofold:

- We want to control the accuracy $\|u - u_h\|_{H^1(\Omega)}$ of a discrete solution u_h and stop the computation if u_h is sufficiently accurate.
- The mesh-refinement should be steered automatically by the algorithm so that we are led to the highest possible accuracy with the lowest number of degrees of freedom.

Remark. Throughout, we allow the cases $\Gamma_D = \Gamma$ as well as $\Gamma_D = \emptyset$. In the latter case, (4.1) becomes the Neumann problem, for which we have to assume the compatability condition $\int_\Omega f\,dx + \int_\Gamma \phi\,ds = 0$. Then, $u \in H^1_*(\Omega)$ and, even more important, the test space $H^1_*(\Omega)$ in the weak formulation can equivalently be replaced by the entire space $H^1(\Omega)$. The same holds for the P1-FEM, where $u_h \in \mathcal{S}^1_*(\mathcal{T})$ and where the discrete test is $\mathcal{S}^1_*(\mathcal{T})$ or equivalently $\mathcal{S}^1(\mathcal{T})$. □

4.2 Scott-Zhang Projection

Since H^1-functions are in general not continuous, nodal interpolation requires additional regularity assumptions. In this section, we aim to provide some quasi-interpolation operator which is well-defined for all $u \in H^1(\Omega)$ and also has the projection property. We start with the following elementary lemma

Lemma 4.1. *For $z \in \mathcal{K}$, choose an edge $E_z \in \mathcal{E}$ with $z \in E_z$. Then, there is a unique dual function $\psi_z \in \mathcal{P}^1(E_z)$ such that*

$$\int_{E_z} \psi_z \zeta_{z'}\,ds = \delta_{zz'} \quad \text{for all } z' \in \mathcal{K}. \tag{4.4}$$

Moreover, it holds that $\|\psi_z\|_{L^\infty(E_z)} \le C\,|E_z|^{-1}$ for some generic constant $C > 0$, which is in particular independent of z and $\mathcal{T}$.

Proof. According to the Riesz theorem, there is a unique function $\widehat{\psi} \in \mathcal{P}^1[0,1]$ such that

$$\int_0^1 \widehat{\psi}\,\widehat{\phi}\,dt = \widehat{\phi}(0) \quad \text{for all } \widehat{\phi} \in \mathcal{P}^1[0,1].$$

Let $\Phi_z : [0,1] \to E_z$ be an affine parametrization of the edge E_z with $\Phi_z(0) = z$. We define

$$\psi_z := \frac{1}{|E_z|}\,\widehat{\psi} \circ \Phi_z^{-1} \in \mathcal{P}^1(E_z).$$

Clearly, $\|\psi_z\|_{L^\infty(E_z)} \le \|\widehat{\psi}\|_{L^\infty(0,1)}\,|E_z|^{-1}$. Note that $|\Phi_z'| = |E_z|$ and hence

$$\int_{E_z} \psi_z \zeta_{z'}\,ds = \int_0^1 (\psi_z \circ \Phi_z)\,(\zeta_{z'} \circ \Phi_z)\,|\Phi_z'|\,dt = \int_0^1 \widehat{\psi}(t)\,(\zeta_{z'} \circ \Phi_z)(t)\,dt = \zeta_{z'}\big(\Phi_z(0)\big) = \zeta_{z'}(z).$$

This concludes the proof. ■

Definition. For each node $z \in \mathcal{K}$ of $\mathcal{T}$, we choose an edge $E_z \in \mathcal{E}$ such that

- $E_z \subseteq \overline{\Gamma}_D$ for $z \in \overline{\Gamma}_D$,

- $E_z \subseteq \Gamma$ for $z \in \Gamma$,
- E_z arbitrary for $z \in \Omega$.

Note that the precise choice is immaterial for the following analysis. For $z \in \mathcal{K}$, let $\psi_z \in \mathcal{P}^1(E_z)$ be the corresponding dual function. Then, the **Scott-Zhang projection** is defined by

$$J_h v := \sum_{z \in \mathcal{K}} \Big(\int_{E_z} \psi_z v \, ds \Big) \zeta_z. \tag{4.5}$$

Clearly, $J_h : H^1(\Omega) \to \mathcal{S}^1(\mathcal{T})$ is well-defined and linear. Our first proposition states that J_h is in fact a projection which preserves discrete boundary data.

Proposition 4.2. *For $v \in H^1(\Omega)$ and $v_h \in \mathcal{S}^1(\mathcal{T})$, the following properties* (i)–(iii) *are true:*

(i) $J_h v_h = v_h$.

(ii) $(J_h v)|_\omega$ *depends only on the trace* $v|_\omega$ *for* $\omega \in \{\Gamma, \Gamma_D\}$.

(iii) $v|_\omega = v_h|_\omega$ *implies that* $(J_h v)|_\omega = v|_\omega$ *for* $\omega \in \{\Gamma, \Gamma_D\}$.

Proof. (i) Note that $v_h = \sum_{z' \in \mathcal{K}} v_h(z') \zeta_{z'}$. By choice of ψ_z, this shows

$$\int_{E_z} \psi_z v_h \, ds = \sum_{z' \in \mathcal{K}} v_h(z') \int_{E_z} \psi_z \zeta_{z'} \, ds = v_h(z).$$

With this, we deduce

$$J_h v_h = \sum_{z \in \mathcal{K}} \Big(\int_{E_z} \psi_z v_h \, ds \Big) \zeta_z = \sum_{z \in \mathcal{K}} v_h(z) \, \zeta_z = v_h.$$

(ii) follows from the choice of the edges E_z. (iii) We consider only $\omega = \Gamma_D$. We first note that

$$(J_h u)|_{\Gamma_D} = \sum_{z \in \mathcal{K}} \Big(\int_{E_z} \psi_z u \, ds \Big) \zeta_z|_{\Gamma_D} = \sum_{z \in \mathcal{K} \cap \overline{\Gamma}_D} \Big(\int_{E_z} \psi_z u \, ds \Big) \zeta_z|_{\Gamma_D} \quad \text{for all } u \in H^1(\Omega).$$

For $z \in \mathcal{K} \cap \overline{\Gamma}_D$, it holds that $E_z \subseteq \overline{\Gamma}_D$ and hence $\int_{E_z} \psi_z v_h \, ds = \int_{E_z} \psi_z v \, ds$. Together with the last equation and the projection property, we obtain that

$$(J_h v)|_{\Gamma_D} = (J_h v_h)|_{\Gamma_D} = v_h|_{\Gamma_D}.$$

This concludes the proof. ■

Exercise 18. Show that Lemma 4.1 holds for any dimension $d \geq 2$. □

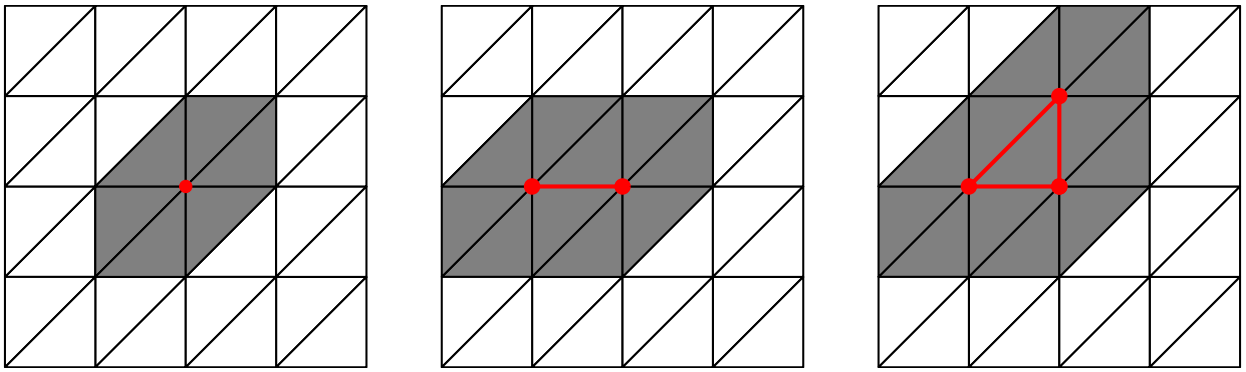
Note that the Scott-Zhang projection $J_h v$ is not defined for general L^2-functions, since $L^2(T)$ does not provide traces. However, one can define an appropriate variant as follows:

Exercise 19. Construct a linear projection $P_h : L^2(\Omega) \to \mathcal{S}^1(\mathcal{T})$ which satisfies

- $\|P_h v\|_{L^2(\Omega)} \leq C \, \|v\|_{L^2(\Omega)}$ for all $v \in L^2(\Omega)$.
- $P_h v_h = v_h$ for all $v_h \in \mathcal{S}^1(\mathcal{T})$.

The constant $C > 0$ may only depend on $\sigma(\mathcal{T})$. **Hint.** Proceed as for the standard Scott-Zhang projection. Instead of an edge E_z, associate with each node $z \in \mathcal{K}$ an arbitrary element $T_z \in \mathcal{T}$ with $z \in T_z$. □

Next, we aim to show that the Scott-Zhang projection has local stability and approximation properties. Unlike nodal interpolation, this will require appropriate patches.



Patch Ω_z of a node $z \in \mathcal{K}$ *Patch Ω_E of an edge $E \in \mathcal{E}$* *Patch Ω_T of an element $T \in \mathcal{T}$*

FIGURE 4.1. *For the a posteriori analysis, we need three types of patches $\omega \subseteq \Omega$, namely patches of nodes, edges, and elements, respectively. Note that the patch of an edge (or of an element) just is the union of the patches of its nodes.*

Definition. For the a posteriori analysis, we need certain unions of elements, called **patches**, cf. Figure 4.1: For a node $z \in \mathcal{K}$, we define

$$\widetilde{\Omega}_z := \big\{T \in \mathcal{T} \,\big|\, z \in \mathcal{K}_T\big\} \quad \text{as well as} \quad \Omega_z := \bigcup \widetilde{\Omega}_z := \big\{x \in \mathbb{R}^2 \,\big|\, \exists T \in \widetilde{\Omega}_z \quad x \in T\big\}. \tag{4.6}$$

For an edge $E \in \mathcal{E}$, we define

$$\widetilde{\Omega}_E := \big\{T \in \mathcal{T} \,\big|\, \mathcal{K}_E \cap T \neq \emptyset\big\} = \big\{T \in \widetilde{\Omega}_z \,\big|\, z \in \mathcal{K}_E\big\} \quad \text{as well as} \quad \Omega_E := \bigcup \widetilde{\Omega}_E. \tag{4.7}$$

Finally, for an element $T \in \mathcal{T}$, we define

$$\widetilde{\Omega}_T := \big\{T' \in \mathcal{T} \,\big|\, \mathcal{K}_T \cap T' \neq \emptyset\big\} = \big\{T' \in \widetilde{\Omega}_z \,\big|\, z \in \mathcal{K}_T\big\} \quad \text{as well as} \quad \Omega_T := \bigcup \widetilde{\Omega}_T. \tag{4.8}$$

The patches Ω_z, Ω_E, and Ω_T are visualized in Figure 4.1.

Lemma 4.3. *There is a constant $C > 0$ which depends only on $\sigma(\mathcal{T})$, such that*

- $\#\widetilde{\Omega}_z \leq C$ *for all* $z \in \mathcal{K}$,
- $\#\widetilde{\Omega}_E \leq C$ *for all* $E \in \mathcal{E}$,

- $\#\widetilde{\Omega}_T \leq C$ *for all* $T \in \mathcal{T}$,

i.e., the number of elements per patch is uniformly bounded. Moreover,

- $\#\big\{T' \in \mathcal{T} \,\big|\, T \in \widetilde{\Omega}_{T'}\big\} \leq C$ *for all* $T \in \mathcal{T}$,

i.e., an element $T \in \mathcal{T}$ *belongs only to finitely many patches.*

Proof. Note that $\sigma(\mathcal{T})$ provides a bound for the minimal interior angle of all elements $T \in \mathcal{T}$; see Exercise 13. Consequently, there is a maximal number $C > 0$ of elements in $\widetilde{\Omega}_z$, for all nodes $u \in \mathcal{K}$. By definition, there follows $\#\widetilde{\Omega}_E \leq 2C$ as well as $\#\widetilde{\Omega}_T \leq 3C$. ■

An essential consequence of Lemma 4.3 is that

$$\|v\|_{L^2(\Omega)} \leq \Big(\sum_{T\in\mathcal{T}} \|v\|^2_{L^2(\Omega_T)} \Big)^{1/2} \leq C_{\mathrm{patch}} \, \|v\|_{L^2(\Omega)} \quad \text{for all } v \in L^2(\Omega),$$

where $C_{\mathrm{patch}} > 0$ depends only on $\sigma(\mathcal{T})$. Another consequence of Lemma 4.3 is that the diameter $\mathrm{diam}(\Omega_T)$ of a patch is proportional to $h_T = \mathrm{diam}(T)$. This is stated in the following lemma.

Lemma 4.4. *For a regular triangulation, it holds that*

- $\mathrm{diam}(\Omega_z) \leq C\, h_T$ *for all* $z \in \mathcal{K}$ *and* $T \in \widetilde{\Omega}_z$,
- $\mathrm{diam}(\Omega_E) \leq C\, h_E \leq C\, h_T$ *for all* $E \in \mathcal{E}$ *and* $T \in \widetilde{\Omega}_E$,
- $\mathrm{diam}(\Omega_{T'}) \leq C\, h_T$ *for all* $T' \in \mathcal{T}$ *and* $T \in \widetilde{\Omega}_{T'}$.

The constant $C > 0$ *depends only on* $\sigma(\mathcal{T})$.

Proof. **1. step.** Note that $h_T \leq \sigma(\mathcal{T})\varrho_T \leq \sigma(\mathcal{T}) h_E$ for all $T \in \mathcal{T}$ and all edges $E \in \mathcal{E}_T$.

2. step. Patch of a node $z \in \mathcal{K}$: For $\widetilde{\Omega}_z = \{T_1, \dots, T_n\}$, we may choose a numbering such that T_{j-1}, T_j are neighbours, i.e., $T_{j-1} \cap T_j \in \mathcal{E}$. From step 1, we derive $h_{T_{j-1}} \leq \sigma(\mathcal{T}) h_{T_j}$, whence $h_{T'} \leq \sigma(\mathcal{T})^{n-1} h_T$ for all $T, T' \in \widetilde{\Omega}_z$. This yields that

$$\mathrm{diam}(\Omega_z) \leq 2 \max_{T'\in\widetilde{\Omega}_z} h_{T'} \leq 2\sigma(\mathcal{T})^{n-1} h_T \quad \text{for all } T \in \widetilde{\Omega}_z.$$

3. step. Patch of an edge $E \in \mathcal{E}$: With $E = \mathrm{conv}\{z_1, z_2\}$ for some $z_1, z_2 \in \mathcal{K}$, it holds that $\widetilde{\Omega}_E = \widetilde{\Omega}_{z_1} \cup \widetilde{\Omega}_{z_2}$ as well as $\widetilde{\Omega}_{z_1} \cap \widetilde{\Omega}_{z_2} \neq \emptyset$. Let $T \in \widetilde{\Omega}_E$ and $n := \max\{\#\widetilde{\Omega}_{z_1}, \#\widetilde{\Omega}_{z_2}\}$. Without loss of generality, we may assume $T \in \widetilde{\Omega}_{z_1}$. Choose $T' \in \widetilde{\Omega}_{z_1} \cap \widetilde{\Omega}_{z_2}$. Then,

$$\mathrm{diam}(\Omega_E) \leq \mathrm{diam}(\Omega_{z_1}) + \mathrm{diam}(\Omega_{z_2}) \leq 2\sigma(\mathcal{T})^{n-1}(h_T + h_{T'}) \leq 2\sigma(\mathcal{T})^{n-1}(1 + \sigma(\mathcal{T})^{n-1}) h_T.$$

4. step. Patch of an element $T \in \mathcal{T}$: Simply use the same arguments as in step 3. ■

The Scott-Zhang projection is locally H^1-stable and has a local first-order approximation property.

Proposition 4.5. *For all $T \in \mathcal{T}$, it holds that*

$$\|v - J_h v\|_{L^2(T)} + h_T \|\nabla J_h v\|_{L^2(T)} \leq C\, h_T\, \|\nabla v\|_{L^2(\Omega_T)} \quad \text{for all } v \in H^1(\Omega). \tag{4.9}$$

The constant $C > 0$ depends only on γ-shape regularity of $\mathcal{T}$.

The proof requires the following technical lemmata which are also valid in any dimension $d \geq 2$ as the proofs reveal.

Theorem 4.6 (Trace Inequality). *Let $T = \operatorname{conv}\{z_0, \dots, z_d\} \subset \mathbb{R}^d$ be a simplex in $\mathbb{R}^d$ with $|T| > 0$ and diameter $h_T := \operatorname{diam}(T)$. Let $E = \operatorname{conv}\{z_1, \dots, z_d\}$ denote one particular side of the simplex. Then, for $v \in H^1(T)$, it holds*

$$\|v\|^2_{L^2(E)} \leq \frac{|E|}{|T|} \left(\|v\|^2_{L^2(T)} + \frac{2}{d}\, h_T \|v \nabla v\|_{L^1(T)} \right) \leq \frac{|E|}{|T|} \left(1 + 2\, h_T/d\right) \|v\|^2_{H^1(T)}. \tag{4.10}$$

With the integral means $v_T := |T|^{-1} \int_T v\, dx$ and $v_E := h_E^{-1} \int_E v\, ds$, it holds that

$$\|v - v_E\|^2_{L^2(E)} \leq \|v - v_T\|^2_{L^2(E)} \leq C\, \frac{|E|\, h_T^2}{|T|}\, \|\nabla v\|^2_{L^2(T)}, \tag{4.11}$$

where $C > 0$ depends only on the reference element T_{ref} and the dimension d.

The proof of the trace inequalities (4.10)–(4.11) is done with the help of the following lemma. In particular, we shall see that both estimates are sharp. Note that, for $d = 2$, it holds that $|E|/|T| \leq 2\varrho_T^{-1}$ and $|E|\, h_T^2/|T| \leq 2\sigma(T) h_T$.

Lemma 4.7 (Trace Identity). *Let $T = \operatorname{conv}\{z_0, \dots, z_d\} \subset \mathbb{R}^d$ be a simplex in $\mathbb{R}^d$ with $|T| > 0$. Let $E = \operatorname{conv}\{z_1, \dots, z_d\}$ denote one particular side of the simplex. Then,*

$$\frac{1}{|E|} \int_E w\, ds = \frac{1}{|T|} \int_T w\, dx + \frac{1}{d|T|} \int_T (x - z_0) \cdot \nabla w\, dx \tag{4.12}$$

for all $w \in C^1(\overline{T})$.

Proof. We apply the Gauss Divergence Theorem to the function $f(x) := w(x)(x - z_0)$. With $\operatorname{div} f(x) = \nabla w(x) \cdot (x - z_0) + dw(x)$, we obtain that

$$d \int_T w\, dx + \int_T (x - z_0) \cdot \nabla w(x)\, dx = \int_T \operatorname{div} f\, dx = \int_{\partial T} f \cdot n\, ds.$$

Note that $(x - z_0) \cdot n(x) = 0$ for $x \in \partial T \backslash E$, whereas $(x - z_0) \cdot n(x) = \operatorname{dist}(z_0, H)$, where $H \subset \mathbb{R}^d$ denotes the hyperplane with $E \subseteq H$. Therefore, the boundary integral simplifies to $\int_{\partial T} f \cdot n\, ds = \operatorname{dist}(z_0, H) \int_E w\, ds$ and the latter equality reads

$$\frac{1}{|T|} \int_T w\, dx + \frac{1}{d|T|} \int_T (x - z_0) \cdot \nabla w\, dx = \frac{\operatorname{dist}(z_0, H)|E|}{d|T|}\, \frac{1}{|E|} \int_E w\, ds,$$

which holds for any $w \in C^1(\overline{T})$. The special choice $w = 1$ can be used to determine the w-independent constant $\frac{\operatorname{dist}(z_0,H)|E|}{d|T|} = 1$. This concludes the proof. ■

Remark. Note that Lemma 4.7 holds for any $w \in W^{1,1}(T) := \big\{w \in L^1(T)$ weakly differentiable $\big|\, \nabla w \in L^1(T)^d\big\}$, even with the same proof. □

Proof of Theorem 4.6. According to standard density arguments, it suffices to consider $v \in C^1(\overline{T})$. Plugging $w := v^2 \in C^1(\overline{T})$ into the trace identity (4.12), we see that

$$\frac{1}{|E|}\int_E v^2\,ds = \frac{1}{|T|}\int_T v^2\,dx + \frac{1}{d|T|}\int_T (x-z_0)\cdot(2v\nabla v)\,dx.$$

This is rewritten in the form

$$\begin{aligned}\frac{|T|}{|E|}\,\|v\|^2_{L^2(E)} = \|v\|^2_{L^2(T)} + \frac{2}{d}\int_T (x-z_0)\cdot(v\nabla v)\,dx &\le \|v\|^2_{L^2(T)} + \frac{2}{d}\,h_T\|v\nabla v\|_{L^1(T)}\\ &\le (1+2\,h_T/d)\,\|v\|^2_{H^1(T)}\end{aligned}$$

which proves (4.10). For the proof of (4.11), we simply replace v by $v - v_T$ and apply the Poincaré inequality. This leads to

$$\begin{aligned}\|v-v_T\|^2_{L^2(E)} &\le \frac{|E|}{|T|}\big(\|v-v_T\|^2_{L^2(T)} + \frac{2}{d}\,h_T\|v-v_T\|_{L^2(T)}\|\nabla v\|_{L^2(T)}\big)\\ &\le \frac{|E|}{|T|}\big(C_P^2 h_T^2\|\nabla v\|^2_{L^2(T)} + \frac{2}{d}\,C_P h_T^2\|\nabla v\|^2_{L^2(T)}\big)\\ &= (C_P^2 + 2C_P/d)\,\frac{|E|\,h_T^2}{|T|}\,\|\nabla v\|^2_{L^2(T)}.\end{aligned}$$

The remaining estimate $\|v-v_E\|_{L^2(E)} \le \|v-v_T\|_{L^2(E)}$ follows from the L^2-best approximation property of the integral mean. ■

Lemma 4.8 (Generalized Poincaré-Friedrichs inequality). *Let $v \in H^1(\Omega)$, $T \in \mathcal{T}$, $T' \in \widetilde{\Omega}_T$, and $E' \in \mathcal{E}_{T'}$. Define the integral means $v_T := (1/|T|)\int_T v\,dx$, $v_{T'} := (1/|T'|)\int_{T'} v\,dx$, and $v_{E'} := (1/|E'|)\int_{E'} v\,ds$. Then,*

$$\|v_T - v_{T'}\|_{L^2(T)} + \|v_T - v_{E'}\|_{L^2(T)} \le C\,h_T\|\nabla v\|_{L^2(\Omega_T)}. \tag{4.13}$$

In particular, this implies that

$$\|v - v_{T'}\|_{L^2(\Omega_T)} + \|v - v_{E'}\|_{L^2(\Omega_T)} \le C\,h_T\|\nabla v\|_{L^2(\Omega_T)}. \tag{4.14}$$

In either estimate, the constant $C > 0$ depends only on γ-shape regularity of $\mathcal{T}$, but is independent of Ω and the shape of Ω_T.

Proof. To ease the notation, let $v_E := (1/|E|)\int_E v\,ds$ also denote the integral mean over edges.

1. step. For any $w \in H^1(T)$, it holds that $|w_T - w_E| \le C \|\nabla w\|_{L^2(T)}$, where $C > 0$ depends only on shape regularity: With the trace inequality (4.11) on T, we see that

$$\begin{aligned} |w_T - w_E| \le |E|^{-1} \|w - w_T\|_{L^1(E)} &\le |E|^{-1/2} \|w - w_T\|_{L^2(E)} \\ &\le C \frac{h_T}{|T|^{1/2}} \|\nabla w\|_{L^2(T)} =: C \|\nabla w\|_{L^2(T)} \end{aligned}$$

2. step. For all $T, T' \in \mathcal{T}$ with $E := T \cap T' \in \mathcal{E}$, shape regularity and the triangle inequality yield that

$$\begin{aligned} \|v_T - v_{T'}\|_{L^2(T)} = |T|^{1/2} |v_T - v_{T'}| &\lesssim |T|^{1/2} |v_T - v_E| + |T'|^{1/2} |v_E - v_{T'}| \\ &= \|v_T - v_E\|_{L^2(T)} + \|v_{T'} - v_E\|_{L^2(T')}. \end{aligned}$$

With Step 1, we thus see that

$$\|v_T - v_{T'}\|_{L^2(T)} \lesssim h_T \|\nabla v\|_{L^2(T)} + h_{T'} \|\nabla v\|_{L^2(T')} \lesssim h_T \|\nabla v\|_{L^2(T \cup T')},$$

where the hidden constant now depends on $C > 0$ from step 1 and from shape regularity of $\mathcal{T}$.

3. step. If $T \cap T' \neq \emptyset$, there is a minimal $n \in \mathbb{N}$ and elements $T_0, \dots, T_n \in \mathcal{T}$ with $T_0 = T$, $T_j \cap T_{j-1} \in \mathcal{E}$ and $T_j \subseteq \Omega_T$ for all $j = 1, \dots, n$, and $T_n = T'$. Note that n is uniformly bounded in terms of the γ-shape regularity of $\mathcal{T}$. Iterating the argument from Step 2, we conclude (4.13) with $\bigcup_{j=0}^n T_j \subseteq \Omega_T$. The overall constant then depends on $C > 0$ and γ.

4. step. For each element $T'' \in \widetilde{\Omega}_T$, the Poincaré inequality and (4.13) show

$$\begin{aligned} &\|v - v_{T'}\|_{L^2(T'')} + \|v - v_{E'}\|_{L^2(T'')} \\ &\lesssim \|v - v_{T''}\|_{L^2(T'')} + \|v_T - v_{T'}\|_{L^2(T'')} + \|v_T - v_{T''}\|_{L^2(T'')} + \|v_T - v_{E'}\|_{L^2(T'')} \\ &\simeq \|v - v_{T''}\|_{L^2(T'')} + \|v_T - v_{T'}\|_{L^2(T)} + \|v_T - v_{T''}\|_{L^2(T)} + \|v_T - v_{E'}\|_{L^2(T)} \\ &\lesssim h_{T''} \|\nabla v\|_{L^2(T'')} + h_T \|\nabla v\|_{L^2(\Omega_T)} \\ &\lesssim h_T \|\nabla v\|_{L^2(\Omega_T)}. \end{aligned}$$

Summing this estimate over all $T'' \in \widetilde{\Omega}_T$, we obtain that

$$\|v - v_{T'}\|_{L^2(\Omega_T)} + \|v - v_{E'}\|_{L^2(\Omega_T)} \lesssim h_T \|\nabla v\|_{L^2(\Omega_T)},$$

where the hidden constants depends only on γ-shape regularity of $\mathcal{T}$. ∎

Proof of Proposition 4.5 (H^1-stability). For $z \in \mathcal{K}$, let $E_z \subset T_z \in \mathcal{T}$ and $h_z := \operatorname{diam}(T_z)$. Note that $T_z \subseteq \Omega_T$ for $z \in T$. The trace inequality (4.10) yields that

$$\|v\|^2_{L^2(E_z)} \lesssim h_z^{-1} \big(\|v\|^2_{L^2(T_z)} + h_z \|v\|_{L^2(T_z)} \|\nabla v\|_{L^2(T_z)} \big) \lesssim h_z^{-1} \big(\|v\|^2_{L^2(T_z)} + h_z^2 \|\nabla v\|^2_{L^2(T_z)} \big)$$

With this and Lemma 4.1, we see that

$$\begin{aligned} \Big| \int_{E_z} \psi_z v \, ds \Big| \le \|\psi_z\|_{L^\infty(E_z)} \|v\|_{L^1(E_z)} &\lesssim |E_z|^{-1/2} \|v\|_{L^2(E_z)} \\ &\lesssim |E_z|^{-1/2} h_z^{-1/2} \big(\|v\|_{L^2(T_z)} + h_z \|\nabla v\|_{L^2(T_z)} \big). \end{aligned}$$

For any hat function, an inverse estimate shows

$$\|\nabla\zeta_z\|_{L^2(T)} \lesssim h_T^{-1}\,\|\zeta_z\|_{L^2(T)} \le |T|^{1/2}h_T^{-1}.$$

Together with $|E_z|\,h_z \simeq |T_z| \simeq |T|$ and $h_z \simeq h_T$, we therefore obtain that, for all $v \in H^1(\Omega)$,

$$\|\nabla J_h v\|_{L^2(T)} \le \sum_{z\in\mathcal{K}\cap T} \Big|\int_{E_z} \psi_z v\,ds\Big| \|\nabla\zeta_z\|_{L^2(T)} \lesssim \sum_{z\in\mathcal{K}\cap T} \big(h_z^{-1}\,\|v\|_{L^2(T_z)} + \|\nabla v\|_{L^2(T_z)}\big). \tag{4.15}$$

With the integral mean $v_T := (1/|T|)\int_T v\,dx$ and the projection property $J_h v_T = v_T$, we apply the last estimate for $w := v - v_T$ and see that

$$\|\nabla J_h v\|_{L^2(T)} = \|\nabla J_h(v-v_T)\|_{L^2(T)} \lesssim \sum_{z\in\mathcal{K}\cap T} \big(h_z^{-1}\,\|v-v_T\|_{L^2(T_z)} + \|\nabla v\|_{L^2(T_z)}\big).$$

According to the Poincaré inequality and Lemma 4.8, it holds that for all $z \in \mathcal{K}\cap T$,

$$\|v-v_T\|_{L^2(T_z)} \le \|v-v_{T_z}\|_{L^2(T_z)} + \|v_{T_z}-v_T\|_{L^2(T_z)} \lesssim h_z\,\|\nabla v\|_{L^2(\Omega_T)}. \tag{4.16}$$

Combining the last two estimates, we thus conclude $\|\nabla J_h v\|_{L^2(T)} \lesssim \|\nabla v\|_{L^2(\Omega_T)}$. ■

Proof of Proposition 4.5 (approximation property). We adopt the notation from the proof of local H^1-stability. Arguing as for (4.15), we see that

$$\|J_h v\|_{L^2(T)} \le \sum_{z\in\mathcal{K}\cap T} \Big|\int_{E_z} \psi_z v\,ds\Big| \|\zeta_z\|_{L^2(T)} \lesssim \sum_{z\in\mathcal{K}\cap T} \big(\|v\|_{L^2(T_z)} + h_z\,\|\nabla v\|_{L^2(T_z)}\big). \tag{4.17}$$

With the integral mean $v_T := (1/|T|)\int_T v\,dx$ and the projection property $J_h v_T = v_T$, we apply the last estimate for $w := v - v_T$ and see that

$$\begin{aligned}
\|v - J_h v\|_{L^2(T)} &= \|(v-v_T) - J_h(v-v_T)\|_{L^2(T)} \\
&\le \|v-v_T\|_{L^2(T)} + \|J_h(v-v_T)\|_{L^2(T)} \\
&\lesssim h_T\,\|\nabla v\|_{L^2(T)} + \sum_{z\in\mathcal{K}\cap T} \big(\|v-v_T\|_{L^2(T_z)} + h_z\,\|\nabla v\|_{L^2(T_z)}\big)
\end{aligned}$$

Finally, we employ (4.16) and $h_z \simeq h_T$ to conclude $\|v - J_h v\|_{L^2(T)} \lesssim h_T\,\|\nabla v\|_{L^2(\Omega_T)}$. ■

The following theorem concludes the main properties of the Scott-Zhang projection:

Theorem 4.9. *The Scott-Zhang projection $J_h : H^1(\Omega) \to \mathcal{S}^1(\mathcal{T})$ has the following properties* (i)–(vii)*:*

(i) *J_h is linear and continuous with respect to the H^1-norm, i.e.,*

$$\|J_h v\|_{H^1(\Omega)} \le C\,(1+\operatorname{diam}(\Omega))\,\|v\|_{H^1(\Omega)} \quad \text{for all } v \in H^1(\Omega). \tag{4.18}$$

(ii) *J_h is a projection onto $\mathcal{S}^1(\mathcal{T})$, i.e.,*

$$J_h v_h = v_h \quad \text{for all } v_h \in \mathcal{S}^1(\mathcal{T}). \tag{4.19}$$

(iii) *J_h preserves discrete boundary data, i.e., for $\omega \in \{\Gamma_D, \Gamma\}$ it holds that*

$$(J_h v)|_\omega = v|_\omega \quad \text{for all } v \in H^1(\Omega) \text{ with } v|_\omega \in \mathcal{S}^1(\mathcal{T}|_\omega). \tag{4.20}$$

(iv) *J_h is locally H^1-stable, i.e.,*

$$\|\nabla J_h v\|_{L^2(T)} \leq C \, \|\nabla v\|_{L^2(\Omega_T)} \quad \text{for all } v \in H^1(\Omega) \text{ and } T \in \mathcal{T}. \tag{4.21}$$

(v) *J_h has a local first-order approximation property, i.e.,*

$$\|(1 - J_h)v\|_{L^2(T)} \leq C h_T \, \|\nabla v\|_{L^2(\Omega_T)} \quad \text{for all } v \in H^1(\Omega) \text{ and } T \in \mathcal{T}. \tag{4.22}$$

(vi) *P_h is quasi-optimal in the sense of the Céa lemma, i.e.,*

$$\|(1 - J_h)v\|_{H^1(\Omega)} \leq C(1 + \operatorname{diam}(\Omega)) \min_{v_h \in \mathcal{S}^1(\mathcal{T})} \|v - v_h\|_{H^1(\Omega)} \quad \text{for all } v \in H^1(\Omega). \tag{4.23}$$

(vii) *For all $\alpha \in \mathbb{R}$, J_h is quasi-optimal in the sense of*

$$\|h^\alpha \nabla(1 - J_h)v\|_{L^2(\Omega)} \leq C \min_{v_h \in \mathcal{S}^1(\mathcal{T})} \|h^\alpha \nabla(v - v_h)\|_{L^2(\Omega)}. \tag{4.24}$$

The constant $C > 0$ in (i)–(vii) *depends only on γ-shape regularity of $\mathcal{T}$.*

Proof. (ii)–(v) have already been shown, and (i) is a direct consequence of (vi) and the triangle inequality. (vii) Let $v_h \in \mathcal{S}^1(\mathcal{T})$. With the projection property of J_h and (iv), we see that, for all $T \in \mathcal{T}$,

$$\|\nabla(1 - J_h)v\|_{L^2(T)} = \|\nabla(1 - J_h)(v - v_h)\|_{L^2(T)} \lesssim \|\nabla(v - v_h)\|_{L^2(\Omega_T)}.$$

With γ-shape regularity and hence $h_T \simeq h_{T'}$ for all $T' \subseteq \Omega_T$, we infer

$$\|h^\alpha \nabla(1 - J_h)v\|_{L^2(T)} \lesssim \|h^\alpha \nabla(v - v_h)\|_{L^2(\Omega_T)}.$$

Using the γ-shape regularity again, this results in

$$\begin{aligned} \|h^\alpha \nabla(1 - J_h)v\|_{L^2(\Omega)}^2 = \sum_{T \in \mathcal{T}} \|h^\alpha \nabla(1 - J_h)v\|_{L^2(T)}^2 &\lesssim \sum_{T \in \mathcal{T}} \|h^\alpha \nabla(v - v_h)\|_{L^2(\Omega_T)}^2 \\ &\lesssim \|h^\alpha \nabla(v - v_h)\|_{L^2(\Omega)}^2. \end{aligned}$$

This proves (vii) with an infimum on the right-hand side. Due to finite dimension, this infimum is, in fact, attained. To prove (vi), it remains to estimate the L^2-part and use $\alpha = 0$ in (vii). With the projection property of J_h and (v), shape regularity yields that

$$\begin{aligned} \|(1 - J_h)v\|_{L^2(\Omega)}^2 = \sum_{T \in \mathcal{T}} \|(1 - J_h)(v - v_h)\|_{L^2(T)}^2 &\lesssim \sum_{T \in \mathcal{T}} h_T^2 \, \|\nabla(v - v_h)\|_{L^2(\Omega_T)}^2 \\ &\lesssim \operatorname{diam}(\Omega)^2 \, \|\nabla(v - v_h)\|_{L^2(\Omega)}^2. \end{aligned}$$

Altogether, we thus see that

$$\|(1-J_h)v\|^2_{H^1(\Omega)} \lesssim (1+\operatorname{diam}(\Omega)^2)\,\|\nabla(v-v_h)\|^2_{L^2(\Omega)}. \lesssim (1+\operatorname{diam}(\Omega))^2\,\|v-v_h\|^2_{H^1(\Omega)}.$$

This concludes the proof of (vi). ■

Remark. Theorem 4.9 holds for any dimension $d \geq 2$ and for any fixed polynomial degree $p \geq 1$. □

One drawback of the Scott-Zhang projection is that it is not positivity conserving, i.e., $v \geq 0$ does not necessarily imply that $J_h v \geq 0$.

Exercise 20. Suppose that $\mathcal{T}$ is a regular triangulation of $\Omega := [0,1]^2$ into 2 triangles. Find an example of a function $v \in H^1(\Omega)$ with $v \geq 0$ such that there exists some $x \in \Omega$ with $J_h v < 0$.
Hint. Compute the function $\widehat{\psi} \in \mathcal{P}^1(0,1)$ from Lemma 4.1 explicitly. □

Exercise 21. Extend the approach of Exercise 19 and construct an operator $P_h : L^2(\Omega) \to \mathcal{S}^1_D(\mathcal{T})$ with the following properties:

(i) $P_h : L^2(\Omega) \to \mathcal{S}^1_D(\mathcal{T})$ is a well-defined linear projection,

$$P_h v_h = v_h \quad \text{for all } v_h \in \mathcal{S}^1_D(\mathcal{T}).$$

(ii) P_h is locally L^2-stable, i.e., for all $T \in \mathcal{T}$, it holds that

$$\|(1-P_h)v\|_{L^2(T)} \leq C\,\|v\|_{L^2(\Omega_T)} \quad \text{for all } v \in L^2(\Omega).$$

(iii) P_h is locally H^1_D-stable, i.e., for all $T \in \mathcal{T}$, it holds that

$$\|\nabla(1-P_h)v\|_{L^2(T)} \leq C\,\|\nabla v\|_{L^2(\Omega_T)} \quad \text{for all } v \in H^1_D(\Omega).$$

(iv) P_h has a local first-order approximation property

$$\|(1-P_h)v\|_{L^2(T)} \leq C h_T\,\|\nabla v\|_{L^2(\Omega_T)} \quad \text{for all } v \in H^1_D(\Omega).$$

(v) $P_h : L^2(\Omega) \to L^2(\Omega)$ as well as $P_h : H^1_D(\Omega) \to H^1_D(\Omega)$ are bounded linear operators.

(vi) J_h is quasi-optimal in the sense of the Céa lemma, i.e.,

$$\|(1-P_h)v\|_{H^1(\Omega)} \leq C \min_{v_h \in \mathcal{S}^1_D(\mathcal{T})} \|v-v_h\|_{H^1(\Omega)} \quad \text{for all } v \in H^1_D(\Omega).$$

(vii) For all $\alpha \in \mathbb{R}$, P_h is quasi-optimal in the sense of

$$\|h^\alpha(1-P_h)v\|_{L^2(\Omega)} \leq C \min_{v_h \in \mathcal{S}^1_D(\mathcal{T})} \|h^\alpha(v-v_h)\|_{L^2(\Omega)} \quad \text{for all } v \in L^2(\Omega).$$

(viii) For all $\alpha \in \mathbb{R}$, P_h is quasi-optimal in the sense of

$$\|h^\alpha \nabla(1-P_h)v\|_{L^2(\Omega)} \le C \min_{v_h \in \mathcal{S}^1_D(\mathcal{T})} \|h^\alpha \nabla(v-v_h)\|_{L^2(\Omega)} \quad \text{for all } v \in H^1_D(\Omega).$$

The constant $C > 0$ in (i)–(viii) depends only on γ-shape regularity of $\mathcal{T}$. **Hint.** Let $\mathcal{K}_F := \mathcal{K} \backslash \Gamma_D$ denote the free nodes, where possibly $\mathcal{K}_F = \mathcal{K}$ for $\Gamma_D = \emptyset$. Then, P_h can be chosen as

$$P_h v = \sum_{z \in \mathcal{K}_F} \Big(\int_{T_z} v \psi_z \, dx \Big) \zeta_z$$

with appropriate $T_z \in \mathcal{T}$ and $\psi_z \in \mathcal{P}^1(T_z)$. □

Definition. The Scott-Zhang projection is just a special example of a Clément-type quasi-interpolation operator: We say that an operator $J_h : H^1_D(\Omega) \to \mathcal{S}^1_D(\mathcal{T})$ is a **Clément-type quasi-interpolation operator** if, for all $v \in H^1_D(\Omega)$ and all $T \in \mathcal{T}$, it holds that

- it is locally H^1-stable

$$\|\nabla(1-J_h)v\|_{L^2(T)} \le C \, \|\nabla v\|_{L^2(\Omega_T)}, \tag{4.25}$$

- and has a local first-order approximation property

$$\|(1-J_h)v\|_{L^2(T)} \le C h_T \, \|\nabla v\|_{L^2(\Omega_T)}. \tag{4.26}$$

The constant $C > 0$ may only depend on γ-shape regularity of $\mathcal{T}$ (and possibly the shapes of possible patches in $\mathcal{T}$).

For the a posteriori error analysis, we shall need the following simple consequence.

Lemma 4.10. *Suppose that $J_h : H^1_D(\Omega) \to \mathcal{S}^1_D(\mathcal{T})$ is a Clément-type operator, i.e.,* (4.25)–(4.26) *hold. Let $T \in \mathcal{T}$ and $E \in \mathcal{E}_T$. Then, it holds that*

$$\|(1-J_h)v\|_{L^2(E)} \le C h_E^{1/2} \, \|\nabla v\|_{L^2(\Omega_T)} \quad \textit{for all } v \in H^1_D(\Omega). \tag{4.27}$$

The constant $C > 0$ depends only on γ-shape regularity of $\mathcal{T}$.

Proof. We apply the trace inequality

$$\|w\|^2_{L^2(E)} \lesssim h_T^{-1} \big(\|w\|^2_{L^2(T)} + h_T \, \|w\|_{L^2(T)} \|\nabla w\|_{L^2(T)} \big)$$

for $w := (1-J_h)v \in H^1_D(\Omega)$. With the Clément properties (4.25)–(4.26), this yields that

$$\|(1-J_h)v\|^2_{L^2(E)} \lesssim h_T \, \|\nabla v\|^2_{L^2(\Omega_T)}.$$

Shape regularity and hence $h_T \simeq h_E$ concludes the proof. ■

The following example is one further *classical* example of a Clément-type operator. The analysis will be left to the reader, but requires the following simple observation:

Exercise 22. Use a scaling argument to show that

$$C^{-1} h_T \, \|\nabla v_h\|_{L^\infty(T)} \leq \|\nabla v_h\|_{L^2(T)} \leq \frac{h_T}{\sqrt{2}} \, \|\nabla v_h\|_{L^\infty(T)} \quad \text{for all } v_h \in \mathcal{P}^m(T),$$

where the constant $C > 0$ only depends on $\sigma(\mathcal{T})$ and the polynomial degree $m \in \mathbb{N}_0$. □

Exercise 23. Let $\mathcal{K}_F := \mathcal{K} \backslash \overline{\Gamma}_D$ denote the free nodes (where possibly $\mathcal{K}_F = \mathcal{K}$ if $\Gamma_D = \emptyset$). Define

$$J_h v := \sum_{z \in \mathcal{K}_F} v_z \zeta_z \quad \text{with} \quad v_z := \frac{1}{|\Omega_z|} \int_{\Omega_z} v \, dx, \tag{4.28}$$

where $\Omega_z \subseteq \overline{\Omega}$ denotes the patch of a node $z \in \mathcal{K}$. Prove that J_h satisfies the following properties:

(i) $J_h : L^2(\Omega) \to \mathcal{S}^1_D(\mathcal{T})$ is a well-defined linear operator.

(ii) J_h is locally L^2-stable, i.e., for all $T \in \mathcal{T}$, it holds that

$$\|(1 - J_h)v\|_{L^2(T)} \leq C \, \|v\|_{L^2(\Omega_T)} \quad \text{for all } v \in L^2(\Omega).$$

(iii) J_h is locally H^1_D-stable, i.e., for all $T \in \mathcal{T}$, it holds that

$$\|\nabla(1 - J_h)v\|_{L^2(T)} \leq C \, \|\nabla v\|_{L^2(\Omega_T)} \quad \text{for all } v \in H^1_D(\Omega).$$

(iv) J_h has a local first-order approximation property

$$\|(1 - J_h)v\|_{L^2(T)} \leq C h_T \, \|\nabla v\|_{L^2(\Omega_T)} \quad \text{for all } v \in H^1_D(\Omega).$$

(v) $J_h : L^2(\Omega) \to L^2(\Omega)$ as well as $J_h : H^1_D(\Omega) \to H^1_D(\Omega)$ are bounded linear operators.

(vi) J_h is positivity preserving, i.e., $J_h v \geq 0$ for all $v \in L^2(\Omega)$ with $v \geq 0$.

(vii) With $\Pi_h : L^2(\Omega) \to \mathcal{P}^0(\mathcal{T})$ the L^2-orthogonal projection onto $\mathcal{P}^0(\mathcal{T})$, it holds that $J_h \Pi_h = J_h$.

The constant $C > 0$ depends only on γ-shape regularity of $\mathcal{T}$. □

Exercise 24. Find a counter example which shows that the operator J_h from Exercise 23 is no projection, i.e., it holds that $J_h v_h \neq v_h$ for some $v_h \in \mathcal{S}^1_D(\mathcal{T})$. □

4.3 Residual-Based Error Estimator

Residual-based a posteriori error estimates follow a general strategy. Recall that the weak solution $u \in H^1_D(\Omega)$ of (4.1) solves the variational form

$$(\nabla u \,;\, \nabla v)_{L^2(\Omega)} = (f \,;\, v)_{L^2(\Omega)} + (\phi \,;\, v)_{L^2(\Gamma_N)} \quad \text{for all } v \in H^1_D(\Omega). \tag{4.29}$$

For an approximation $u_h \in \mathcal{S}^1_D(\mathcal{T})$ it is thus natural to define the **residual** $R_h \in H^1_D(\Omega)^*$ by

$$R_h(v) := (f \,;\, v)_{L^2(\Omega)} + (\phi \,;\, v)_{L^2(\Gamma_N)} - (\nabla u_h \,;\, \nabla v)_{L^2(\Omega)}, \tag{4.30}$$

i.e., $R_h = 0$ if and only if $u_h = u$. Let $|\!|\!| w |\!|\!| := \|\nabla w\|_{L^2(\Omega)}$ denote the energy norm on $H^1_D(\Omega)$ and

$$|\!|\!| \Phi |\!|\!|_* := \sup_{w \in H^1_D(\Omega)\backslash\{0\}} \frac{\Phi(w)}{|\!|\!| w |\!|\!|}$$

the induced operator norm on $H^1_D(\Omega)^*$, where we stress that both are equivalent norms on $H^1_D(\Omega)$ and its dual space, respectively. Then, the Riesz theorem and $R_h(v) = (\nabla(u - u_h) \,;\, \nabla v)_{L^2(\Omega)}$ yield

$$|\!|\!| R_h |\!|\!|_* = |\!|\!| u - u_h |\!|\!|.$$

To derive a reliable error estimator η, we thus need to prove an estimate of the type

$$R_h(v) \le \widetilde{C_{\mathrm{rel}}}\, \eta\, |\!|\!| v |\!|\!| \quad \text{for all } v \in H^1_D(\Omega). \tag{4.31}$$

To derive an efficient error estimator η, we need to show

$$R_h(v) \ge \widetilde{C_{\mathrm{eff}}}\, \eta\, |\!|\!| v |\!|\!| \quad \text{for some } v \in H^1_D(\Omega)\backslash\{0\}, \tag{4.32}$$

where this $v \in H^1_D(\Omega)$ has to be constructed appropriately.

Exercise 25. Prove that reliability (4.2) of an error estimator η is, in fact, equivalent to (4.31). Prove that efficiency (4.3) of η holds if and only if (4.32) holds. □

So far, our observations did not use that we are dealing with Galerkin schemes. We stress that the Galerkin orthogonality reads

$$R_h(v_h) = 0 \quad \text{for all } v_h \in \mathcal{S}^1_D(\mathcal{T}) \tag{4.33}$$

with respect to the residual R_h. To provide a reliable (and residual-based) error estimator η, we will use some Clément-type operator $J_h : H^1(\Omega) \to \mathcal{S}^1_D(\Omega)$ in connection with the Galerkin orthogonality (4.33).

Before introducing a first a posteriori error estimator, we introduce the following notational conventions. We define the $\mathcal{T}$-piecewise resp. $\mathcal{E}$-piecewise constant mesh-width functions

$$h_{\mathcal{T}}|_T := h_T \quad \text{and} \quad h_{\mathcal{E}}|_T := h_E$$

for elements $T \in \mathcal{T}$ and edges $E \in \mathcal{E}$, respectively. Moreover, we write

$$\|h_{\mathcal{E}}^{1/2}\psi\|_{L^2(\mathcal{E}_*)} := \Big(\sum_{E\in\mathcal{E}_*} h_E \|\psi\|_{L^2(E)}^2 \Big)^{1/2}$$

for any set $\mathcal{E}_* \subseteq \mathcal{E}$ of edges and any function ψ which belongs to $L^2(E)$ for all $E \in \mathcal{E}_*$. Recall that $\mathcal{E}_D$ and $\mathcal{E}_N$ denote the Dirichlet and Neumann edges of $\mathcal{T}$, respectively. Moreover, let $\mathcal{E}_\Omega$ denote the set of all **interior edges**, i.e., for $E \in \mathcal{E}_\Omega$, there are unique elements $T_E^+, T_E^- \in \mathcal{T}$ with $E = T_E^+ \cap T_E^-$. Finally, for $E \in \mathcal{E}_\Omega$, we define the **jump of the normal derivative** by

$$[\![\partial_n u_h]\!]_E := \frac{\partial u_h}{\partial n_E^+} + \frac{\partial u_h}{\partial n_E^-} \in \mathbb{R}, \tag{4.34}$$

where $n_E^\pm$ denote the outer normal vectors of the elements $T_E^\pm$ on the edge E. Note that $n_E^+ = -n_E^-$ so that the sum in the definition is, in fact, a difference.

Theorem 4.11. *The error estimator*

$$\eta := \big(\|h_{\mathcal{T}} f\|_{L^2(\Omega)}^2 + \|h_{\mathcal{E}}^{1/2} [\![\partial_n u_h]\!]\|_{L^2(\mathcal{E}_\Omega)}^2 + \|h_{\mathcal{E}}^{1/2}(\phi - \partial_n u_h)\|_{L^2(\mathcal{E}_N)}^2 \big)^{1/2} \tag{4.35}$$

satisfies the reliability estimate

$$\|u - u_h\|_{H^1(\Omega)} \le C\,\eta, \tag{4.36}$$

where the constant $C > 0$ depends only on γ-shape regularity of $\mathcal{T}$.

Proof. For all $w \in H_D^1(\Omega)$, elementwise integration by parts proves

$$\begin{aligned}
R_h(w) &= (f\,;w)_{L^2(\Omega)} + (\phi\,;w)_{L^2(\Gamma_N)} - \sum_{T\in\mathcal{T}} (\nabla u_h\,;\nabla w)_{L^2(T)} \\
&= (f\,;w)_{L^2(\Omega)} + \sum_{E\in\mathcal{E}_N} (\phi\,;w)_{L^2(E)} - \sum_{T\in\mathcal{T}} (\partial_n u_h\,;w)_{L^2(\partial T)} \\
&= \sum_{T\in\mathcal{T}} (f\,;w)_{L^2(T)} + \sum_{E\in\mathcal{E}_N} (\phi - \partial_n u_h\,;w)_{L^2(E)} - \sum_{E\in\mathcal{E}_\Omega} ([\![\partial_n u_h]\!]\,;w)_{L^2(E)} \\
&\le \sum_{T\in\mathcal{T}} \|f\|_{L^2(T)}\|w\|_{L^2(T)} + \sum_{E\in\mathcal{E}_N} \|\phi - \partial_n u_h\|_{L^2(E)}\|w\|_{L^2(E)} + \sum_{E\in\mathcal{E}_\Omega} \|[\![\partial_n u_h]\!]\|_{L^2(E)}\|w\|_{L^2(E)}.
\end{aligned}$$

For arbitrary $v \in H_D^1(\Omega)$, we now choose $w = v - J_h v$ and note that $R_h(v) = R_h(w)$ according to the Galerkin orthogonality. Then, we estimate the three sums separately. The approximation property of the Clément-type operator J_h and Lemma 4.3 imply

$$\begin{aligned}
\sum_{T\in\mathcal{T}} \|f\|_{L^2(T)}\|v - J_h v\|_{L^2(T)} &\lesssim \Big(\sum_{T\in\mathcal{T}} \|h_{\mathcal{T}} f\|_{L^2(T)}^2 \Big)^{1/2} \Big(\sum_{T\in\mathcal{T}} \|\nabla v\|_{L^2(\Omega_T)}^2 \Big)^{1/2} \\
&\lesssim \Big(\sum_{T\in\mathcal{T}} \|h_{\mathcal{T}} f\|_{L^2(T)}^2 \Big)^{1/2} \Big(\sum_{T\in\mathcal{T}} \|\nabla v\|_{L^2(T)}^2 \Big)^{1/2} \\
&= \|h_{\mathcal{T}} f\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)}.
\end{aligned}$$

For each edge $E \in \mathcal{E}$, we choose an arbitrary element $T_E \in \mathcal{T}$ with $E \in \mathcal{E}_{T_E}$. Let $\mathcal{E}_* \subset \mathcal{E}$ and $\psi \in L^2(E)$ for all $E \in \mathcal{E}_*$. Recall that $\|(1 - J_h)v\|_{L^2(E)} \lesssim h_E^{1/2} \|\nabla v\|_{L^2(\Omega_{T_E})}$. Therefore, the same arguments as before prove

$$\begin{aligned}\sum_{E\in\mathcal{E}_*} \|\psi\|_{L^2(E)}\|v - J_h v\|_{L^2(E)} &\lesssim \Big(\sum_{E\in\mathcal{E}_*} \|h_{\mathcal{E}}^{1/2}\psi\|_{L^2(E)}^2\Big)^{1/2}\Big(\sum_{E\in\mathcal{E}_*} \|\nabla v\|_{L^2(\Omega_{T_E})}^2\Big)^{1/2}\\ &\lesssim \|h_{\mathcal{E}}^{1/2}\psi\|_{L^2(\mathcal{E}_*)}\|\nabla v\|_{L^2(\Omega)},\end{aligned}$$

where we note that an element $T \in \mathcal{T}$ may satisfy $T = T_E$ for at most three edges. Altogether, we now see

$$\begin{aligned}R_h(v) &\lesssim \|\nabla v\|_{L^2(\Omega)} \big(\|h_{\mathcal{T}} f\|_{L^2(\Omega)} + \|h_{\mathcal{E}}^{1/2}[\![\partial_n u_h]\!]\|_{L^2(\mathcal{E}_\Omega)} + \|h_{\mathcal{E}}^{1/2}(\phi - \partial_n u_h)\|_{L^2(\mathcal{E}_N)}\big)\\ &\le \sqrt{3}\,\|\nabla v\|_{L^2(\Omega)}\,\eta.\end{aligned}$$

The hidden constant C depends only (on the Clément operator J_h and) on γ-shape regularity of $\mathcal{T}$. ■

Remark. Note that we have used $u_h \in \mathcal{S}^1(\mathcal{T})$ in the sense that the elementwise Laplacian satisfies $\Delta u_h|_T = 0$ for all $T \in \mathcal{T}$. For general $\mathcal{T}$-piecewise polynomials, the same proof applies with $\|h_{\mathcal{T}} f\|_{L^2(\Omega)}$ replaced by $\|h_{\mathcal{T}}(f + \Delta u_h)\|_{L^2(\Omega)}$. □

Exercise 26. We consider the mixed boundary value problem

$$\begin{aligned}-\Delta u &= f && \text{in } \Omega,\\ u &= u_D && \text{on } \Gamma_D,\\ \partial_n u &= \phi && \text{on } \Gamma_N.\end{aligned}$$

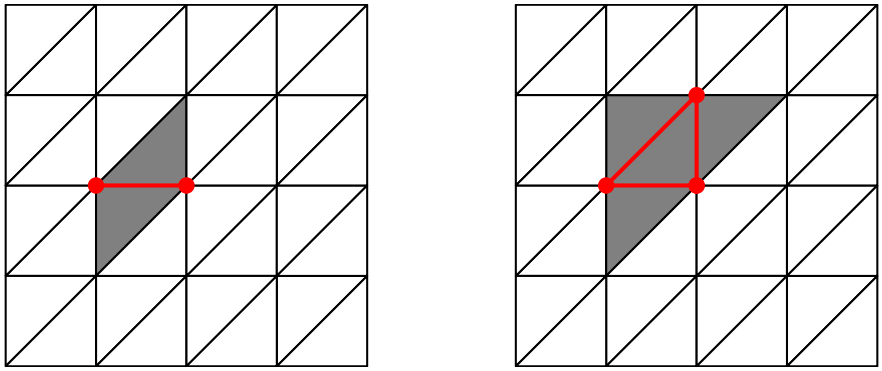
with inhomogeneous Dirichlet data $u_D \in H^{1/2}(\Gamma_D)$. Let $u \in H^1(\Omega)$ denote the weak solution and $u_h \in \mathcal{S}^1(\mathcal{T}_h)$ the P1-FEM solution for discrete Dirichlet data $u_{Dh} := \widehat{u}_{Dh}|_{\Gamma_D}$ with $\widehat{u}_{Dh} \in \mathcal{S}^1(\mathcal{T}_h)$. Use the additional problem

$$\begin{aligned}-\Delta w &= 0 && \text{in } \Omega,\\ w &= u_D - u_{Dh} && \text{on } \Gamma_D,\\ \partial_n u &= 0 && \text{on } \Gamma_N.\end{aligned}$$

with weak solution $w \in H^1(\Omega)$ to derive a reliable error estimator for $\|u - u_h\|_{H^1(\Omega)}$.

Hint. Prove that $\|w\|_{H^1(\Omega)} \simeq \|u_D - u_{Dh}\|_{H^{1/2}(\Gamma_D)}$, where the right-hand side is already an a posteriori term. Then, consider the residual $\widetilde{R}_h \in H^1_D(\Omega)^*$ corresponding to the function $(u - u_h) - w \in H^1_D(\Omega)$. □

Next, we prove the efficiency of the residual-based error estimator η from (4.35) — at least up to terms of higher order. The efficiency estimate even holds locally with refined patches ω_E and ω_T shown in Figure 4.2: For an interior edge $E \in \mathcal{E}_\Omega$, let $T_E^+, T_E^- \in \mathcal{T}$ be the unique elements with



Patch ω_E of an edge $E \in \mathcal{E}$ *Patch ω_T of an element $T \in \mathcal{T}$*

FIGURE 4.2. *To prove the efficiency estimate, it suffices to consider smaller patches $\omega_E \subseteq \Omega_E$ and $\omega_T \subseteq \Omega_T$, for edges $E \in \mathcal{E}$ and elements $T \in \mathcal{T}$, respectively. For comparison with the larger patches Ω_E and Ω_T, the reader may consider Figure 4.1 on page 40.*

$E = T_E^+ \cap T_E^-$. For a boundary edge $E \in \mathcal{E}_\Gamma$, there is a unique element $T_E \in \mathcal{T}$ with $E \in \mathcal{E}_{T_E}$. We define the refined patch of an edge $E \in \mathcal{E}$ by

$$\omega_E := \begin{cases} T_E^+ \cup T_E^- & \text{for } E \in \mathcal{E}_\Omega, \\ T_E & \text{for } E \in \mathcal{E}_\Gamma. \end{cases} \tag{4.37}$$

Moreover, we define the refined patch of an element $T \in \mathcal{T}$ by

$$\omega_T := \bigcup \{ \omega_E \,|\, E \in \mathcal{E}_T \}. \tag{4.38}$$

Note that $\omega_E \subseteq \Omega_E$ and $\omega_T \subseteq \Omega_T$, so that Lemma 4.3 and Lemma 4.4 even hold for the refined patches.

Usually, one is interested in error estimators which are localized with respect to the elements or the edges of $\mathcal{T}$, respectively. For instance, one considers the **element-based residual error estimator**

$$\eta_{\mathcal{T}} := \Big(\sum_{T \in \mathcal{T}} \eta_T^2 \Big)^{1/2}, \tag{4.39}$$

where

$$\eta_T = \big(h_T^2 \|f\|_{L^2(T)}^2 + h_T \|[\![\partial_n u_h]\!]\|_{L^2(\partial T \cap \Omega)}^2 + h_T \|\phi - \partial_n u_h\|_{L^2(\partial T \cap \Gamma_N)}^2 \big)^{1/2} \tag{4.40}$$

or the **edge-based residual error estimator**

$$\eta_{\mathcal{E}} := \Big(\sum_{E \in \mathcal{E}} \eta_E^2 \Big)^{1/2}, \tag{4.41}$$

where

$$\eta_E = \begin{cases} \big(h_E^2 \|f\|_{L^2(\omega_E)}^2 + h_E \|[\![\partial_n u_h]\!]\|_{L^2(E)}^2 \big)^{1/2} & \text{for } E \in \mathcal{E}_\Omega, \\ \big(h_E^2 \|f\|_{L^2(\omega_E)}^2 + h_E \|\phi - \partial_n u_h\|_{L^2(E)}^2 \big)^{1/2} & \text{for } E \in \mathcal{E}_N, \\ 0 & \text{for } E \in \mathcal{E}_D. \end{cases} \tag{4.42}$$

Alternatively, one could also define

$$\eta_{\mathcal{T}\cup\mathcal{E}} := \Big(\sum_{T\in\mathcal{T}} \eta_T^2 + \sum_{E\in\mathcal{E}} \eta_E^2 \Big)^{1/2}, \tag{4.43}$$

where

$$\eta_T = h_T \|f\|_{L^2(T)}, \tag{4.44a}$$

$$\eta_E = \begin{cases} h_E^{1/2} \, \|[\![\partial_n u_h]\!]\|_{L^2(E)} & \text{for } E \in \mathcal{E}_\Omega, \\ h_E^{1/2} \, \|\phi - \partial_n u_h\|_{L^2(E)} & \text{for } E \in \mathcal{E}_N, \\ 0 & \text{for } E \in \mathcal{E}_D. \end{cases} \tag{4.44b}$$

Note that $\eta_{\mathcal{T}}$ as well as $\eta_{\mathcal{E}}$ are equivalent to the error estimator η from (4.35): There holds

$$\eta = \eta_{\mathcal{T}\cup\mathcal{E}} \le \eta_{\mathcal{T}} \le \sqrt{2}\,\sigma(\mathcal{T}_h)^{1/2}\,\eta \quad \text{as well as} \quad \sigma(\mathcal{T})^{-1}\,\eta \le \eta_{\mathcal{E}} \le \sqrt{3}\,\eta,$$

since $\eta_{\mathcal{T}}$ adds the contributions of interior edges twice and $h_E \le h_T \le \sigma(\mathcal{T}_h)\,h_E$ for each edge $E \in \mathcal{E}_T$, whereas $\eta_{\mathcal{E}}$ adds the element contribution at most three times. The local quantities η_T and η_E can be used to steer an adaptive mesh-refining algorithm. They are therefore called **refinement indicators**. We are going to discuss adaptive mesh-refinement below.

Theorem 4.12 (inverse estimate). *For all polynomial degrees $m \in \mathbb{N}$ and $k, r \in \mathbb{N}$ with $k > r$, there exists a constant $C > 0$ such that*

$$\|D^k v_h\|_{L^2(T)} \le C\sigma(\mathcal{T})\,h_T^{r-k}\|D^r v_h\|_{L^2(T)} \quad \textit{for all } v_h \in \mathcal{P}^m(\mathcal{T}) \textit{ and all } T \in \mathcal{T}, \tag{4.45}$$

where $\mathcal{P}^m(\mathcal{T}) := \big\{ v_h : \Omega \to \mathbb{R} \,\big|\, \forall T \in \mathcal{T} \quad v_h|_T \in \mathcal{P}^m(T) \big\}$.

Proof. The proof is done $\mathcal{T}$-elementwise and follows from a scaling argument. We start with an abstract observation.

1. step. Let X be a finite dimensional space, $\|\cdot\|_X$ be a norm on X and $|\cdot|_X$ be a seminorm on X. Then, there exists a constant $C > 0$ such that

$$|x|_X \le C\,\|x\|_X \quad \text{for all } x \in X :$$

We consider the quotient space X/Y, where $Y := \big\{ x \in X \,\big|\, |x|_X = 0 \big\}$. Note that X/Y is finite dimensional and that

$$\|x+Y\|_{X/Y} := \inf_{y\in Y} \|x+y\|_X \quad \text{as well as} \quad |x+Y|_{X/Y} := \inf_{y\in Y} |x+y|_X = |x|_X$$

are norms on the finite dimensional space X/Y. Therefore, there is a norm equivalence constant $C > 0$ such that

$$|x|_X = |x+Y|_{X/Y} \le C\,\|x+Y\|_{X/Y} \le C\,\|x\|_X \quad \text{for all } x \in X.$$

2. step. There exists a constant $C_{\mathrm{ref}} > 0$ such that

$$\|D^k w_h\|_{L^2(T_{\mathrm{ref}})} \le C_{\mathrm{ref}} \|D^r w_h\|_{L^2(T_{\mathrm{ref}})} \quad \text{for all } w_h \in \mathcal{P}^m(T_{\mathrm{ref}}).$$

This follows from the abstract framework for $X = \mathcal{P}^m(T_{\mathrm{ref}})$.

3. step. Proof of the statement: Let $\Phi : T_{\mathrm{ref}} \to T$ be an affine diffeomorphism and $B \in \mathbb{R}^{2\times 2}$ its linear part. We apply the transformation formula to Φ^{-1} to see that

$$\|D^k v_h\|_{L^2(T)} \le |\det B^{-1}|^{-1/2} \|B^{-1}\|_F^k \|D^k(v_h \circ \Phi)\|_{L^2(T_{\mathrm{ref}})}.$$

Note that the L^2-norm can be estimated by step 2 since $v_h \circ \Phi \in \mathcal{P}^m(T_{\mathrm{ref}})$. The application of the transformation formula to Φ proves that

$$\|D^r(v_h \circ \Phi)\|_{L^2(T_{\mathrm{ref}})} \le |\det B|^{-1/2} \|B\|_F^r \|D^r v_h\|_{L^2(T)}.$$

By definition of the shape regularity constant $\sigma(\mathcal{T})$, we obtain that

$$\begin{aligned}\|D^k v_h\|_{L^2(T)} &\le C_{\mathrm{ref}} \|B^{-1}\|_F^k \|B\|_F^r \|D^r v_h\|_{L^2(T)} \le \sqrt{2}\, C_{\mathrm{ref}} \varrho_T^{-k} h_T^r \|D^r v_h\|_{L^2(T)} \\ &\le \sqrt{2}\, C_{\mathrm{ref}}\, \sigma(\mathcal{T}) h_T^{r-k} \|D^r v_h\|_{L^2(T)},\end{aligned}$$

where we have used that $\|B^{-1}\|_F \le \sqrt{2}\, \varrho_T^{-1}$. This concludes the proof. ∎

Theorem 4.13. *We define $f_{\mathcal{T}} \in \mathcal{P}^0(\mathcal{T})$ by $f_{\mathcal{T}}|_T := |T|^{-1} \int_T f\,dx$ and $\phi_{\mathcal{E}} \in \mathcal{P}^0(\mathcal{E}_N)$ by $\phi_{\mathcal{E}}|_E := h_E^{-1} \int_E \phi\,ds$. For each element $T \in \mathcal{T}$, the refinement indicator η_T from (4.40) satisfies*

$$\eta_T \le C \big(\|\nabla(u-u_h)\|_{L^2(\omega_T)}^2 + \|h_{\mathcal{T}}(f-f_{\mathcal{T}})\|_{L^2(\omega_T)}^2 + \|h_{\mathcal{E}}^{1/2}(\phi-\phi_{\mathcal{E}})\|_{L^2(\partial T\cap\Gamma_N)}^2 \big)^{1/2}. \tag{4.46}$$

Moreover, the error estimator η from (4.35) is efficient in the sense that

$$\eta \le C \big(\|u-u_h\|_{H^1(\Omega)} + \|h_{\mathcal{T}}(f-f_{\mathcal{T}})\|_{L^2(\Omega)} + \|h_{\mathcal{E}}^{1/2}(\phi-\phi_{\mathcal{E}})\|_{L^2(\Gamma_N)} \big). \tag{4.47}$$

The constant $C > 0$ only depends on the shape regularity constant $\sigma(\mathcal{T})$.

Remark. For $f \in H^1(\mathcal{T})$ holds $\|h_{\mathcal{T}}(f-f_{\mathcal{T}})\|_{L^2(\Omega)} = \mathcal{O}(h^2)$. For $\phi \in C^1(\mathcal{E}_N)$ holds $\|h_{\mathcal{E}}^{1/2}(\phi - \phi_{\mathcal{E}})\|_{L^2(\Gamma_N)} = \mathcal{O}(h^{3/2})$. Even for $u \in H^2(\Omega)$, the error as well as the error estimator η only satisfy $\|u-u_h\|_{H^1(\Omega)} = \mathcal{O}(h) = \eta$. Therefore, the two terms on the right-hand side are of higher order. □

Proof of Theorem 4.13. **1. step.** Estimate (4.47) is a consequence of (4.46) since

$$\eta \le \eta_{\mathcal{T}} = \Big(\sum_{T\in\mathcal{T}} \eta_T^2 \Big)^{1/2} \le 2C \big(\|u-u_h\|_{H^1(\Omega)} + \|h_{\mathcal{T}}(f-f_{\mathcal{T}})\|_{L^2(\Omega)} + \|h_{\mathcal{E}}^{1/2}(\phi-\phi_{\mathcal{E}})\|_{L^2(\Gamma_N)} \big).$$

Here, the factor $2 = 4^{1/2}$ appears since each element $T \in \mathcal{T}$ belongs at most to four patches $\omega_{T'}$.

The proof of (4.46) is split into three steps, where we consider each of the three contributions of η_T separately.

2. step. There holds

$$\|h_{\mathcal{T}} f\|_{L^2(T)} \le C \big(\|\nabla(u-u_h)\|_{L^2(T)} + \|h_{\mathcal{T}}(f-f_{\mathcal{T}})\|_{L^2(T)} \big) : \tag{4.48}$$

For $T \in \mathcal{T}$, we define the **element bubble function**

$$b_T := \prod_{z \in \mathcal{K}_T} \zeta_z \in H^1_0(T) \cap \mathcal{P}^3(T)$$

as product of all three hat functions. It is essential to observe the following estimate

$$\|f_{\mathcal{T}} b_T\|_{L^2(T)} \le \|f_{\mathcal{T}} b_T^{1/2}\|_{L^2(T)} \le \|f_{\mathcal{T}}\|_{L^2(T)} \le C_{\mathrm{ref}} \|f_{\mathcal{T}} b_T\|_{L^2(T)}, \tag{4.49}$$

where the existence of an independent constant $C_{\mathrm{ref}} > 0$ follows from a scaling argument. We stress, however, that $\|f_{\mathcal{T}} b_T\|_{L^2(T)}$ and hence C_{ref} — since $f_{\mathcal{T}}$ is constant on T — can explicitly be computed. In the following, the main idea is to use integration by parts for $v := f_{\mathcal{T}} b_T \in H^1_0(T)$ to show

$$\begin{aligned} C_{\mathrm{ref}}^{-2} \|f_{\mathcal{T}}\|^2_{L^2(T)} \le \|f_{\mathcal{T}} b_T^{1/2}\|^2_{L^2(T)} = (f_{\mathcal{T}} \,;\, v)_{L^2(T)} &= (f_{\mathcal{T}} - f \,;\, v)_{L^2(T)} + (f - \Delta u_h \,;\, v)_{L^2(T)} \\ &= (f_{\mathcal{T}} - f \,;\, v)_{L^2(T)} + (\nabla(u - u_h) \,;\, \nabla v)_{L^2(T)}. \end{aligned}$$

Now, we estimate each of the two scalar products on the right-hand side by use of the Cauchy inequality. Together with $v = f_{\mathcal{T}} b_T \in \mathcal{P}^3(\mathcal{T})$ we observe

$$(f_{\mathcal{T}} - f \,;\, v)_{L^2(T)} \le \|f_{\mathcal{T}} - f\|_{L^2(T)} \|f_{\mathcal{T}} b_T\|_{L^2(T)} \le \|f_{\mathcal{T}} - f\|_{L^2(T)} \|f_{\mathcal{T}}\|_{L^2(T)}$$

as well as

$$\begin{aligned} (\nabla(u - u_h) \,;\, \nabla v)_{L^2(T)} &\le \|\nabla(u - u_h)\|_{L^2(T)} \|\nabla(f_T b_T)\|_{L^2(T)} \\ &\le C_{\mathrm{inv}} h_T^{-1} \|\nabla(u - u_h)\|_{L^2(T)} \|f_{\mathcal{T}} b_T\|_{L^2(T)} \\ &\le C_{\mathrm{inv}} h_T^{-1} \|\nabla(u - u_h)\|_{L^2(T)} \|f_{\mathcal{T}}\|_{L^2(T)}. \end{aligned}$$

Altogether, we see

$$h_T \|f_{\mathcal{T}}\|_{L^2(T)} \le C_{\mathrm{ref}}^2 \big(h_T \|f_{\mathcal{T}} - f\|_{L^2(T)} + C_{\mathrm{inv}} \|\nabla(u - u_h)\|_{L^2(T)} \big),$$

which finally results in

$$h_T \|f\|_{L^2(T)} \le (1 + C_{\mathrm{ref}}^2) \big(h_T \|f_{\mathcal{T}} - f\|_{L^2(T)} + C_{\mathrm{inv}} \|\nabla(u - u_h)\|_{L^2(T)} \big).$$

3. step. For an interior edge $E \in \mathcal{E}_\Omega$, there holds

$$h_E^{1/2} \|[\![\partial_n u_h]\!]\|_{L^2(E)} \le C \left(\|\nabla(u - u_h)\|_{L^2(\omega_E)} + \|h_{\mathcal{T}}(f - f_{\mathcal{T}})\|_{L^2(\omega_E)} \right) : \tag{4.50}$$

To prove this estimate, we define the **edge bubble function**

$$b_E := \prod_{z \in \mathcal{K}_E} \zeta_z \in H^1_0(\omega_E) \cap \mathcal{P}^2(\mathcal{T}).$$

The essential estimate reads

$$\|b_E\|_{L^2(E)} \le \|b_E^{1/2}\|_{L^2(E)} \le h_E^{1/2} \le C_{\mathrm{ref}} \|b_E\|_{L^2(E)}, \tag{4.51}$$

where the constant $C_{\mathrm{ref}} > 0$ is independent of E. In particular, this provides

$$C_{\mathrm{ref}}^{-2}\|[\![\partial_n u_h]\!]\|_{L^2(E)}^2 \le \|[\![\partial_n u_h]\!] b_E^{1/2}\|_{L^2(E)}^2 = ([\![\partial_n u_h]\!]\,;\,[\![\partial_n u_h]\!] b_E)_{L^2(E)}.$$

Let $T_E^+, T_E^- \in \mathcal{T}$ be the unique elements with $T_E^+ \cap T_E^- = E$ and $\omega_E = T_E^+ \cup T_E^-$. Note that $v := [\![\partial_n u_h]\!] b_E \in \mathcal{P}^2(T_E^\pm)$ satisfies $v|_{\partial T_E^\pm \setminus E} = 0$. Therefore, integration by parts on $T_E^\pm$ proves

$$\begin{aligned}
([\![\partial_n u_h]\!]\,;\,v)_{L^2(E)} &= (\partial_n u_h\,;\,v)_{L^2(\partial T_E^+)} + (\partial_n u_h\,;\,v)_{L^2(\partial T_E^-)} \\
&= (\nabla u_h\,;\,\nabla v)_{L^2(\omega_E)} \\
&= (\nabla(u_h - u)\,;\,\nabla v)_{L^2(\omega_E)} + (f\,;\,v)_{L^2(\omega_E)} \\
&\le \big(C_{\mathrm{inv}}\|\nabla(u_h - u)\|_{L^2(\omega_E)} + \|h_{\mathcal{T}} f\|_{L^2(\omega_E)}\big)\|h_{\mathcal{T}}^{-1} v\|_{L^2(\omega_E)},
\end{aligned}$$

where we have applied the Cauchy inequality and an inverse estimate for $v \in \mathcal{P}^2(\mathcal{T})$. For $T \in \{T_E^+, T_E^-\}$ holds

$$\|v\|_{L^2(T)} = |[\![\partial_n u_h]\!]_E|\|b_E\|_{L^2(T)} \le |T|^{1/2}|[\![\partial_n u_h]\!]_E| \le \frac{h_T^{1/2}}{\sqrt{2}}\|[\![\partial_n u_h]\!]\|_{L^2(E)},$$

since $|T| \le \frac{1}{2} h_T h_E$. From this, we infer

$$h_E^{1/2}\|h_{\mathcal{T}}^{-1} v\|_{L^2(\omega_E)} \le \|h_{\mathcal{T}}^{-1/2} v\|_{L^2(\omega_E)} \le \|[\![\partial_n u_h]\!]\|_{L^2(E)}.$$

This finally proves

$$h_E^{1/2}\|[\![\partial_n u_h]\!]\|_{L^2(E)}^2 \le C_{\mathrm{ref}}^2\big(C_{\mathrm{inv}}\|\nabla(u_h - u)\|_{L^2(\omega_E)} + \|h_{\mathcal{T}} f\|_{L^2(\omega_E)}\big)\,\|[\![\partial_n u_h]\!]\|_{L^2(E)}$$

and we may conclude this step by use of step 2 to dominate $\|h_{\mathcal{T}} f\|_{L^2(\omega_E)}$.

4. step. For $T \in \mathcal{T}$ and a Neumann edge $E \in \mathcal{E}_N \cap \mathcal{E}_T$, it holds

$$\begin{aligned}
&h_E^{1/2}\|\phi - \partial_n u_h\|_{L^2(E)} \\
&\qquad \le C\big(\|h_{\mathcal{E}}^{1/2}(\phi - \phi_{\mathcal{E}})\|_{L^2(E)} + \|\nabla(u - u_h)\|_{L^2(T)} + \|h_{\mathcal{T}}(f - f_{\mathcal{T}})\|_{L^2(T)}\big)\,:
\end{aligned} \tag{4.52}$$

We consider again the edge bubble function $b_E \in \mathcal{P}^2(T)$ and note that $b_E|_{\partial T \setminus E} = 0$. With $v := (\phi_{\mathcal{E}} - \partial_n u_h) b_E \in \mathcal{P}^2(T)$, we proceed as in step 3 and obtain

$$C_{\mathrm{ref}}^{-2}\|\phi_{\mathcal{E}} - \partial_n u_h\|_{L^2(E)}^2 \le (\phi_{\mathcal{E}} - \partial_n u_h\,;\,v)_{L^2(E)} = (\phi_{\mathcal{E}} - \phi\,;\,v)_{L^2(E)} + (\phi - \partial_n u_h\,;\,v)_{L^2(E)}.$$

For the second term, we employ integration by parts to see

$$\begin{aligned}
(\phi - \partial_n u_h\,;\,v)_{L^2(E)} &= (\partial_n(u - u_h)\,;\,v)_{L^2(\partial T)} \\
&= (\nabla(u - u_h)\,;\,\nabla v)_{L^2(T)} - (f\,;\,v)_{L^2(T)} \\
&\le \big(C_{\mathrm{inv}}\|\nabla(u - u_h)\|_{L^2(T)} + \|h_{\mathcal{T}} f\|_{L^2(T)}\big) h_E^{-1/2}\|\phi_{\mathcal{E}} - \partial_n u_h\|_{L^2(E)}.
\end{aligned}$$

The first term is estimated by the Cauchy inequality directly

$$(\phi_{\mathcal{E}} - \phi\,;\,v)_{L^2(E)} \le \|h_{\mathcal{E}}^{1/2}(\phi_{\mathcal{E}} - \phi)\|_{L^2(E)}\|h_{\mathcal{E}}^{-1/2} v\|_{L^2(E)}.$$

There holds

$$\|v\|_{L^2(E)} = |(\phi_{\mathcal{E}} - \partial_n u_h)|_E| \|b_E\|_{L^2(E)} \le h_E^{1/2} |(\phi_{\mathcal{E}} - \partial_n u_h)|_E| = \|\phi_{\mathcal{E}} - \partial_n u_h\|_{L^2(E)}.$$

Altogether, we thus have shown

$$\begin{aligned} &h_E^{1/2} \|\phi_{\mathcal{E}} - \partial_n u_h\|_{L^2(E)}^2 \\ &\quad \le C_{\mathrm{ref}}^2 \big(C_{\mathrm{inv}} \|\nabla(u - u_h)\|_{L^2(T)} + \|h_{\mathcal{T}} f\|_{L^2(T)} + \|h_{\mathcal{E}}^{1/2} (\phi_{\mathcal{E}} - \phi)\|_{L^2(E)} \big) \, \|\phi_{\mathcal{E}} - \partial_n u_h\|_{L^2(E)}. \end{aligned}$$

Here, $\|h_{\mathcal{T}} f\|_{L^2(T)}$ is estimated by step 2, and $\phi_{\mathcal{E}}$ on the left-hand side is replaced by ϕ with the help of the triangle inequality. ■

Exercise 27. Prove that $f_{\mathcal{T}}$ in Theorem 4.13 can be replaced by an arbitrary $\mathcal{T}$-elementwise polynomial $f_{\mathcal{T}} \in \mathcal{P}^m(\mathcal{T})$. The constant $C > 0$ in (4.46)–(4.47) then additionally depends on the polynomial degree $m \in \mathbb{N}_0$. □

Remark. With the help of a so-called extension operator that extends a polynomial $p : E \to \mathbb{R}$ to a polynomial $F_{\mathrm{ext}} p : T \to \mathbb{R}$, one can show that $\phi_{\mathcal{E}}$ in Theorem 4.13 can be replaced by an arbitrary $\mathcal{E}_N$-edgewise polynomial (with respect to the arclength). □

Actually, the volume residual contribution $\|h_{\mathcal{T}} f\|_{L^2(\Omega)} = \mathcal{O}(h)$ to η can be improved. This is done in the following exercise, where this term is replaced by some higher-order term $\mathcal{O}(h^2)$.

Exercise 28. Let $\Omega_z = \mathrm{supp}(\zeta_z)$ denote the node patch of $z \in \mathcal{K}$. For $f \in L^2(\Omega)$, let $f_z := |\Omega_z|^{-1} \int_{\Omega_z} f \, dx$ denote the corresponding integral mean. Prove the following claims:

(i) For all inner nodes $z \in \mathcal{K} \backslash \Gamma$, it holds

$$\int_{\Omega_z} f f_z \zeta_z \, dx \le C \Big(\sum_{\substack{E \in \mathcal{E}_\Omega \\ z \in E}} \|[\![\partial_n u_h]\!]\|_{L^2(E)}^2 \Big)^{1/2} \|h_{\mathcal{T}}^{-1/2} f_z\|_{L^2(\Omega_z)}.$$

(ii) For all inner nodes $z \in \mathcal{K} \backslash \Gamma$ and elements $T \in \mathcal{T}$ with $z \in T$, it holds

$$C^{-1} \|h_{\mathcal{T}} f\|_{L^2(T)}^2 \le \|h_{\mathcal{T}} (f - f_z)\|_{L^2(\Omega_z)}^2 + \sum_{\substack{E \in \mathcal{E}_\Omega \\ z \in E}} \|h_{\mathcal{E}}^{1/2} [\![\partial_n u_h]\!]\|_{L^2(E)}^2.$$

(iii) Derive the equivalence

$$\begin{aligned} C^{-1} \eta^2 \le \widetilde{\eta}^2 :=& \|h_{\mathcal{E}}^{1/2} [\![\partial_n u_h]\!]\|_{L^2(\mathcal{E}_\Omega)}^2 + \|h_{\mathcal{E}}^{1/2} (\phi - \partial_n u_h)\|_{L^2(\mathcal{E}_N)}^2 \\ &+ \sum_{z \in \mathcal{K} \backslash \Omega} \|h_{\mathcal{T}} (f - f_z)\|_{L^2(\Omega_z)}^2 \le C \, \eta^2. \end{aligned}$$

(iv) Conclude that the improved error estimator $\widetilde{\eta}$ is reliable and efficient.

(v) In what sense is the error estimator $\widetilde{\eta}$ improved when compared to η.

The constant $C > 0$ in (i)–(iii) depends only on γ-shape regularity of $\mathcal{T}$. □

Bibliography

[Bra] Dietrich Braess: *Finite elements. Theory, fast solvers, and applications in elasticity theory*, Cambridge University Press, Cambridge, 2007.

[McL] William McLean: *Strongly elliptic systems and boundary integral equations*, Cambridge University Press, Cambridge, 2000.

Appendix A

Some Facts from Functional Analysis

I this appendix we collect some results from introductory functional analysis courses which are used throughout. We stick with the case of vector spaces over $\mathbb{R}$.

A.1 Main Theorems from Functional Analysis

Theorem A.1 (Hahn-Banach Extension Theorem). *Let $p : X \to \mathbb{R}$ be a sublinear functional on a linear space X, i.e. $p(x+y) \leq p(x) + p(y)$ and $p(\lambda x) = \lambda p(x)$ for all $x, y \in X$ and $\lambda \geq 0$. If Y is a subspace of X and $f : Y \to \mathbb{R}$ is a linear functional with $f \leq p$ on Y, there is a linear extension $F : X \to \mathbb{R}$ with $F|_Y = f$ and $F \leq p$ on X.* ■

If X is a normed space and $f \in Y^*$, one may choose $p(x) = \|x\|_X \|f\|_{X^*}$ to prove the extension theorem for continuous linear functionals.

Corollary A.2. *If Y is the subspace of a normed space X and $f \in Y^*$, there is an extension $F \in X^*$ with $F|_Y = f$ and $\|F\|_{X^*} = \|f\|_{Y^*}$.* ■

One then considers the subspace $Y := \operatorname{span}\{x\}$ and $f(\lambda x) = \lambda \|x\|_X$ to derive the following corollary:

Corollary A.3. *If X is a normed space and $x \in X$, there is a linear functional $f \in X^*$ with $\|f\|_{X^*} = 1$ and $f(x) = \|x\|_X = \sup_{\|f\|_{X^*}=1} |f(x)|$.* ■

Theorem A.4 (Hahn-Banach Separation Theorem). *Let X be a normed space, and let A and B be convex, nonempty subsets of X with $A \cap B = \emptyset$.*
(i) *If A is open, there is a linear functional $f \in X^*$ and a scalar $\lambda \in \mathbb{R}$ such that $f(x) < \lambda \leq f(y)$ for all $x \in A$ and $y \in B$.*
(ii) *If A is compact and B is closed, there is a linear functional $f \in X^*$ and scalars $\lambda_1, \lambda_2 \in \mathbb{R}$ such that $f(x) \leq \lambda_1 < \lambda_2 \leq f(y)$ for all $x \in A$ and $y \in B$.* ■

If Y is a subspace of X, one can use (ii) to characterize the closure $\overline{Y}$ of Y in X. The proof only needs that each bounded linear functional $f \in Y^*$ is trivial, i.e. $f|_Y = 0$.

Corollary A.5. *Let Y be a subspace of the normed space X. Then, $x \in X$ satisfies $x \in \overline{Y}$ if and only if $f(x) = 0$ for all $f \in X^*$ with $f|_Y = 0$.*

Proof. For $x \in \overline{Y}$ and $f \in X^*$ with $f|_Y = 0$, continuity yields $f(x) = 0$. The converse implication is proven by contradiction: We assume that $x \notin \overline{Y}$ and choose $f \in X^*$ such that $f(x) < \lambda \leq f(y)$ for all $y \in Y$ and some fixed $\lambda \in \mathbb{R}$. Using that Y is a vector space, we infer that $\lambda \leq f(\pm y) = -f(\mp y) \leq -\lambda$ and thus $f(y) \in [\lambda, -\lambda]$ for all $y \in Y$. As bounded linear functionals are trivial, we obtain $f|_Y = 0$. According to our assumptions, this implies $f(x) = 0$ and thus contradicts $f(x) < \lambda \leq f(0) = 0$. ■

The following corollary is an immediate consequence of the last one.

Corollary A.6. *Let Y be a subspace of the normed space X. Then, Y is dense in X if and only if each functional $f \in X^*$ with $f|_Y = 0$ is trivial, i.e., $f = 0 \in X^*$.* ■

For an operator $T \in L(X;Y)$, one defines $(T^*y^*)(x) := y^*(Tx)$ for all $y^* \in Y^*$ and $x \in X$. From the continuity of T, we see that $T^*y^* \in X^*$, and obviously $T^* : Y^* \to X^*$ is a linear operator. From the corollary of the Hahn-Banach extension theorem, we derive for the operator norm

$$\begin{aligned}\|T^*\| = \sup_{\|y^*\|_{Y^*}=1} \|T^*y^*\|_{X^*} &= \sup_{\|y^*\|_{Y^*}=1} \sup_{\|x\|_X=1} (T^*y^*)(x) \\ &= \sup_{\|x\|_X=1} \sup_{\|y^*\|_{Y^*}=1} (y^*)(Tx) = \sup_{\|x\|_X=1} \|Tx\|_Y = \|T\|,\end{aligned}$$

i.e. there holds $T^* \in L(Y^*;X^*)$ with operator norm $\|T^*\| = \|T\|$. The operator T^* is called the **adjoint operator** of T.

Theorem A.7 (Banach Closed Range Theorem). *For an operator $T \in L(X;Y)$ between Banach spaces X and Y and $T^* \in L(Y^*;X^*)$ its adjoint, the following is pairwise equivalent:*
(i) $\operatorname{range}(T)$ *is a closed subspace of* Y.
(ii) $\operatorname{range}(T) = (\ker T^*)_\circ := \big\{y \in Y \,\big|\, \forall y^* \in \ker(T^*) \quad y^*(y) = 0\big\}$.
(iii) $\operatorname{range}(T^*)$ *is a closed subspace of* X^*.
(iv) $\operatorname{range}(T^*) = (\ker T)^\circ := \big\{x^* \in X^* \,\big|\, \forall x \in \ker(T) \quad x^*(x) = 0\big\}$. ■

A.2 Hilbert Spaces

A space X is called **Hilbert space** if it is a Banach space whose norm is induced by a scalar product.

Theorem A.8. *Let Y be the closed subspace of a Hilbert space X and $Y^\perp := \big\{x \in X \,\big|\, \forall y \in Y \quad (x\,;\,y)_X = 0\big\}$ the orthogonal complement. Then, there holds $X = Y \oplus Y^\perp$ in the sense of the linear algebra, i.e. every element $x \in X$ has a unique decomposition $x = y + y^\perp$ with some $y \in Y$ and $y^\perp \in Y^\perp$.* ■

With the orthogonal decomposition $X = Y \oplus Y^\perp$, one can define a projection $\pi_Y : X \to Y$ by $x = y + y^\perp \mapsto y$.

Corollary A.9. *Let Y be the closed subspace of a Hilbert space X. Then, there is a unique linear operator $\Pi : X \to Y$ with $\Pi|_Y = id$ and $\ker(\Pi) = Y^\perp$, which is called* **orthogonal projection** *onto Y. This projection is continuous with operator norm $\|\Pi\| = 1$ and symmetric, i.e. $(x\,;\, y)_X = (\Pi x\,;\, y)_X$ for all $x \in X$ and $y \in Y$. Moreover, the orthogonal projection is the solution operator for the best approximation problem, $\|x - \Pi x\|_X = \min_{y \in Y} \|x - y\|_X$.* ■

The dual space X^* of a Hilbert space X has a straight-forward representation, and one can somehow identify X with X^*.

Theorem A.10 (Riesz). *For a Hilbert space X, the* **Riesz mapping** *$I_X : X \to X^*$, $I_X x := (x\,;\,\cdot)_X \in X^*$, is an isometric isomorphism.* ■