Numerics of Partial Differential Equations: Stationary Problems

Lecture Notes

Michael Feischl and Dirk Praetorius

December 19, 2024

Contents

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

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

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 $\|u\|_H = \|F\|_{H^*}$. ■

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

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

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 $C_D^1(\overline{\Omega})$ 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, $C_D^1(\overline{\Omega})$ is *not* complete and therefore the Riesz theorem does *not* apply.

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

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

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** $u \in H_D^1(\Omega)$ 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 $H_D^1(\Omega)$ and is also the unique weak solution of (1.10).
- provided the weak solution $u \in H_D^1(\Omega)$ 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 $H_D^1(\Omega)$ by a finite dimensional subspace $X_h \subset H_D^1(\Omega)$: 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 (1.23) 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 $a(u, \cdot) = F \in H^*$. □

Exercise 8. Define the Galerkin method in the context of monotone operators: Under the assumptions of Exercise 6, we aim to approximate the solution $u \in H$ of $Au = F \in H^*$. How does the Galerkin method look like in this setting? Prove that the Galerkin operator $\mathbb{G}_h : H \to X_h$ onto a finite dimensional subspace $X_h \subset H$ is a well-defined (in general nonlinear) and Lipschitz-continuous projection, i.e., $\mathbb{G}_h^2 = \mathbb{G}_h$ with

$$\|\mathbb{G}_h u - \mathbb{G}_v v\|_H \leq C \|u - v\|_H \quad \text{for all } u, v \in H.$$

Céa lemma

$$\|u - \mathbb{G}_h u\|_H \leq C \min_{v_h \in X_h} \|u - v_h\|_H \quad \text{for all } u \in H.$$

Show that the constants $C > 0$ depend only on A. □

Exercise 9. We stick with the setting of monotone operators from Exercise 7 and 8: How can one compute the Galerkin approximation $u_h = \mathbb{G}_h u \in X_h$ of a solution $u \in H$ of $Au = F \in H^*$? For $N = \dim X_h$, provide a (nonlinear) system of equations in $\mathbb{R}^N$ which characterizes the unique solution $u_h = \mathbb{G}_h u \in X_h$. What happens if the operator A is linear? □

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 $j = 1, \ldots, d$, exist. □

From the main theorem of calculus, 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 10. 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 10. (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 $\partial_j u, \widetilde{\partial_j u} \in L^1_{\ell oc}(\Omega)$ 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 $u \in L^1_{\ell oc}(\Omega)$ 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.

Lemma 2.6. *Suppose that $A \in L(X;Y)$ is a compact operator between a Banach space X and a normed space Y and that (x_n) is a weakly convergent sequence, i.e., $x_n \rightharpoonup x \in X$. Then, the image (Ax_n) is strongly convergent to Ax in Y, i.e., $Ax_n \to Ax \in Y$.*

Proof. Using the adjoint operator $A^* \in L(Y^*;X^*)$, one sees that $Ax_n \rightharpoonup Ax \in Y$. Assume that (Ax_n) does not strongly converge to Ax. Then, there is a subsequence (Ax_{n_k}) with

$\inf_{k\in\mathbb{N}} \|Ax_{n_k} - Ax\|_Y \geq \varepsilon$ for some $\varepsilon > 0$. Recall that weakly convergent sequence are always bounded. Compactness thus provides a further subsequence $(Ax_{n_{k\ell}})$ of (Ax_{n_k}) with $Ax_{n_{k\ell}} \to y \in Y$. In particular, $Ax_{n_{k\ell}} \rightharpoonup y \in Y$ and therefore $y = Ax$. This contradicts the choice of the subsequence (Ax_{n_k}). ■

Exercise 11. Let X be a reflexive Banach space and Y be a normed space. Suppose that $A \in L(X,Y)$ is completely continuous, i.e., for all (x_n) in X, weak convergence $x_n \rightharpoonup x$ in X implies strong convergence $Ax_n \to Ax$ in Y. Prove that A is compact, i.e., for X being reflexive, the operator A is compact if and only if it is completely continuous. □

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 2.7. *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 2.8. *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 2.7 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 2.7 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 2.9 (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 2.8, 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 2.10 (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 2.11 (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 2.12 (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 2.12, we can prove that the integration by parts formula also holds for Sobolev functions $u, v \in H^1(\Omega)$.

Corollary 2.13 (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 $\lim\limits_{n\to\infty} a(u_n, v_n) = a(u, v)$. 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 2.14 (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 2.8. 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 2.15 (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 2.16. *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 12. Usually, one defines the range of the trace operator as $H^{1/2}(\Gamma) := \operatorname{range}(\gamma) \subseteq L^2(\Gamma)$. 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 13. 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 2.17. (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*

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

(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 12. Then, the main proposition reads as follows:

Proposition 2.18. (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 12.

Proposition 2.19. (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*

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

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

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

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

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

where we have used that $H^{1/2}(\Gamma) = \operatorname{range}(\gamma)$. 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,

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

Together with $(1 + \widetilde{C}_P^2)^{-1}\|u\|_{H^1(\Omega)}^2 \leq \|\nabla u\|_{L^2(\Omega)}^2$, 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

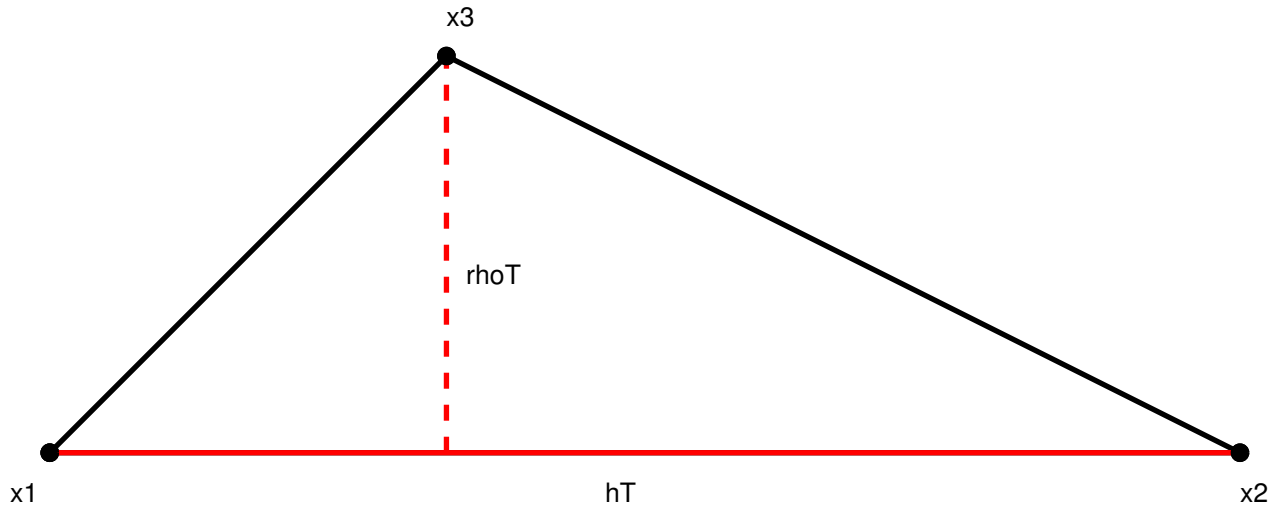


FIGURE 3.1. *The diameter h_T of the triangle T is the length of the longest edge (possibly non unique). The quantity ϱ_T denotes the corresponding height.*

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 $T = \text{conv}\{x_T, y_T, z_T\}$ 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

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

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

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

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 14. 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., $T = \operatorname{conv}\{x_0,\dots,x_d\} \subset \mathbb{R}^d$. □

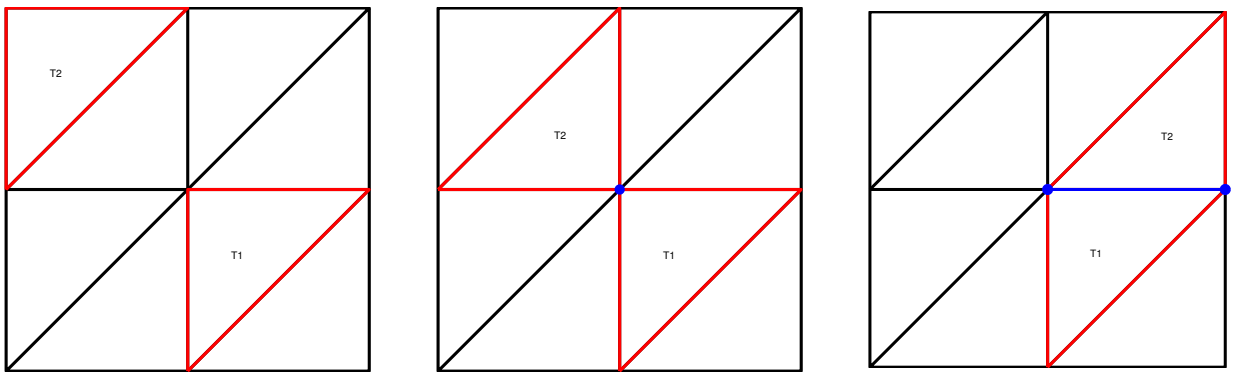


FIGURE 3.2. *For a regular triangulation $\mathcal{T}$, the intersection of two elements $T \neq T'$ is either empty, a joint node, or a joint edge.*

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 $\mathcal{K} := \bigcup \{x \in \mathcal{K}_T \,|\, T \in \mathcal{T}\}$, we then denote the **set of nodes** of the triangulation $\mathcal{T}$ and by $\mathcal{E} := \bigcup \{E \in \mathcal{E}_T \,|\, T \in \mathcal{T}\}$ 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

- 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

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

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 15. 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*

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

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

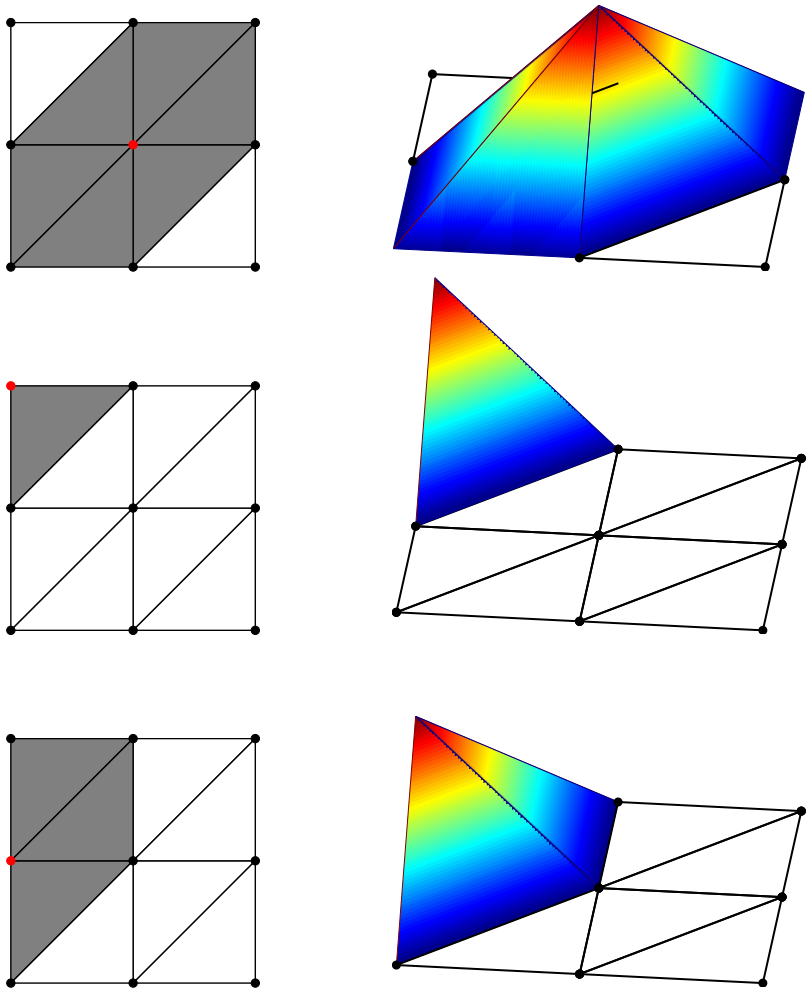


FIGURE 3.3. *Examples of P1 hat functions ζ_z: The left figures show the mesh as well as the support* $\operatorname{supp}(\zeta_z)$ *in grey, where the corresponding node $z \in \mathcal{K}$ is indicated in red. The right figures show the plots of the hat functions. Triangles $T \in \mathcal{T}$ with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

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 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 15. ■

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

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

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

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

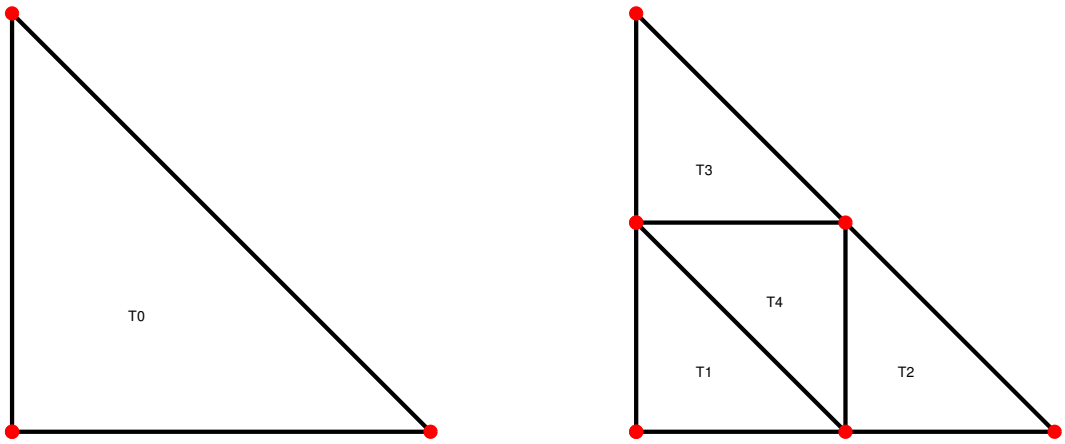


FIGURE 3.4. *Red-refinement refines the element $T \in \mathcal{T}^{(\text{old})}$ into 4 similar elements $T_1, \dots, T_4 \in \mathcal{T}^{(\text{new})}$. The new nodes $\mathcal{K}^{(\text{new})} \backslash \mathcal{K}^{(\text{old})}$ are just the edge midpoints for all edges $E \in \mathcal{E}^{(\text{old})}$. In particular, regularity of $\mathcal{T}^{(\text{old})}$ implies regularity of $\mathcal{T}^{(\text{new})}$.*

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 $\mathcal{T}$. 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 $\mathcal{T}^{(\mathrm{old})}$, we obtain a new triangulation $\mathcal{T}^{(\mathrm{new})}$ as follows: Each element $T \in \mathcal{T}^{(\mathrm{old})}$ is split into 4 similar triangles $T_1, \dots, T_4 \in \mathcal{T}^{(\mathrm{new})}$, cf. Figure 3.4. Therefore, each node $z \in \mathcal{K}^{(\mathrm{new})}$ either belongs to $\mathcal{K}^{(\mathrm{old})}$ or is the midpoint of an edge $E \in \mathcal{E}^{(\mathrm{old})}$. We stress some simple observations:

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

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

Exercise 16. 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 17. 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 $B(x, r_T) := \{y \in \mathbb{R}^2 \,|\, |x-y| \leq r_T\}$ 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

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

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

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

Before the proof of Theorem 3.5, 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*

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

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

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

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,

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

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

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

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 $C^\infty_*(\overline{\Omega}) := \{v \in C^\infty(\overline{\Omega}) \,\big|\, \int_\Omega v\,dx = 0\}$ 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 18. 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

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

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 2.7, 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 19, 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 19. 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 standard chain rule for $\Phi=(\Phi_1,\dots,\Phi_d)^T\colon\mathbb{R}^d\to\mathbb{R}^d$:

$$\partial_j(u\circ\Phi)(x) = \sum_{k=1}^{d}(\partial_k u(\Phi(x))\partial_j\Phi_k = \sum_{k=1}^{d}(\partial_k u(\Phi(x))B_{kj}. \tag{3.31}$$

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

$$\begin{aligned}
\partial_{j_1}\cdots\partial_{j_{m+1}}(u\circ\Phi)(x) &\overset{(3.30)}{=} \partial_{j_1}\Bigg(\sum_{k_2=1}^{d}\cdots\sum_{k_{m+1}=1}^{d}(\partial_{k_2}\cdots\partial_{k_{m+1}}u)(\Phi(x))\prod_{\ell=2}^{m+1}B_{k_\ell j_\ell}\Bigg)\\
&= \sum_{k_2=1}^{d}\cdots\sum_{k_{m+1}=1}^{d}\partial_{j_1}\big((\partial_{k_2}\cdots\partial_{k_{m+1}}u)(\Phi(x))\big)\prod_{\ell=2}^{m+1}B_{k_\ell j_\ell}\\
&\overset{(3.31)}{=} \sum_{k_2=1}^{d}\cdots\sum_{k_{m+1}=1}^{d}\sum_{k_1=1}^{d}(\partial_{k_1}\partial_{k_2}\cdots\partial_{k_{m+1}}u)(\Phi(x))B_{k_1 j_1}\prod_{\ell=2}^{m+1}B_{k_\ell j_\ell}\\
&= \sum_{k_1=1}^{d}\cdots\sum_{k_{m+1}=1}^{d}(\partial_{k_1}\partial_{k_2}\cdots\partial_{k_{m+1}}u)(\Phi(x))\prod_{\ell=1}^{m+1}B_{k_\ell j_\ell},
\end{aligned}$$

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

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

where the last equality follows from the general fact $\prod_{\ell=1}^{m}\sum_{k=1}^{d}a_{k\ell} = \sum_{k_1=1}^{d}\cdots\sum_{k_m=1}^{d}\prod_{\ell=1}^{m}a_{k_\ell\ell}$ (which can be proved by another simple induction argument).

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

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

5. step. We prove the transformation formula (3.29) for general $u \in H^m(T)$: According to step 1, the linear operator $\Psi\colon L^2(T) \to L^2(\widehat{T})$, $\Psi u := u\circ\Phi$ is bounded. The Meyers-Serrin theorem shows that $C^\infty(\overline{T})$ is a dense subspace of $H^m(T)$. Note that (3.29) implies for $u \in C^\infty(\overline{T})$ the estimate $\|\Psi u\|_{H^m(\widehat{T})} \le C\,\|u\|_{H^m(T)}$, where $C > 0$ depends only on m and B. Hence, $\Psi u := u\circ\Phi$ extends uniquely to a linear and continuous mapping $\Psi : H^m(T) \to H^m(\widehat{T})$. It remains to show that this extension coincides with the composition, i.e., $\Psi u = u\circ\Phi$ on $H^m(T)$. To that end, 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 $H^m(\widehat{T})$ (and hence also in $L^2(\widehat{T})$). Moreover, according to Step 1, we have

$$
\begin{aligned}
\|\Psi u - u\circ\Phi\|_{L^2(\widehat{T})} &= \lim_{n\to\infty} \|\Psi u_n - u\circ\Phi\|_{L^2(\widehat{T})} = \lim_{n\to\infty} \|(u_n - u)\circ\Phi\|_{L^2(\widehat{T})} \\
&= |\det B| \lim_{n\to\infty} \|u_n - u\|_{L^2(T)} = 0.
\end{aligned}
$$

This shows $\Psi u = u\circ\Phi$ on $H^m(T)$.

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*

$$
\Phi_T : T_{\mathrm{ref}} \to T, \quad \Phi_T(s,t) := z_1 + B\begin{pmatrix} s \\ t \end{pmatrix}, \quad \text{where } B := \begin{pmatrix} z_2 - z_1 & z_3 - z_1 \end{pmatrix} \in \mathbb{R}^{2\times 2}. \tag{3.32}
$$

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

$$
h_T/\sqrt{2} \le \|B\|_F \le \sqrt{2}\,h_T \quad \text{as well as} \quad \varrho_T^{-1}/\sqrt{2} \le \|B^{-1}\|_F \le \sqrt{2}\,\varrho_T^{-1}. \tag{3.33}
$$

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 (3.33) 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

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

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

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

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

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

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},

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

Second, we use $v = u \circ \Phi$ 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

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

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 20. 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

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

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

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

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

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

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

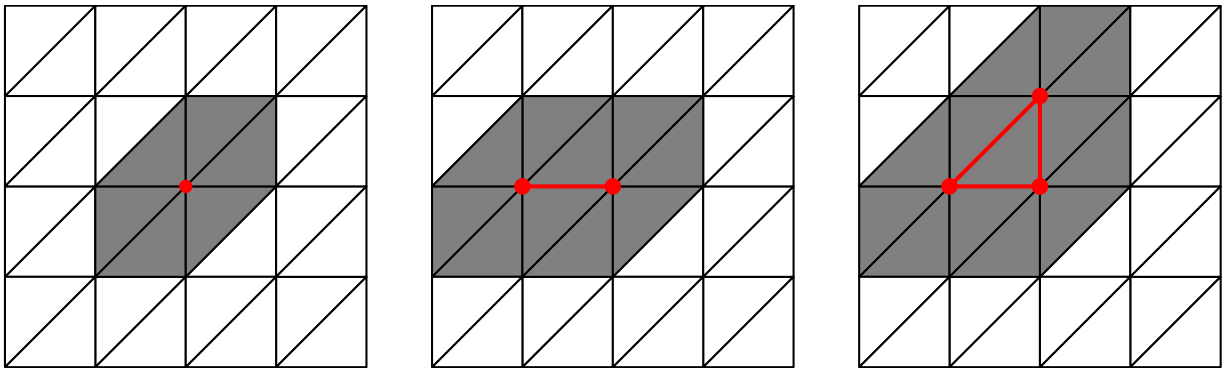
Note that the Scott-Zhang projection $J_h v$ is not defined for general L^2-functions, since $L^2(T)$ does not provide traces. However, one can define an appropriate variant as follows:

Exercise 22. 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 16. 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 $\operatorname{diam}(\Omega_T)$ of a patch is proportional to $h_T = \operatorname{diam}(T)$. This is stated in the following lemma.

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

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

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

3. step. Patch of an edge $E \in \mathcal{E}$: With $E = \operatorname{conv}\{z_1, z_2\}$ 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,

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

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 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 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. Let T_{ref} denote the reference triangle and $E_{\mathrm{ref}} = [0,1]$ be the reference edge of T_{ref}.

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

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

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 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 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) *J_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 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 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 shape regularity again, this results in

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

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

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

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 23. 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 24. Extend the approach of Exercise 22 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 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 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*

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

The constant $C > 0$ depends only on 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 25. 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 26. 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 shape regularity of $\mathcal{T}$. □

Exercise 27. Find a counter example which shows that the operator J_h from Exercise 26 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

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

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

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

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

Exercise 28. 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

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

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 29. We consider the mixed boundary value problem

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

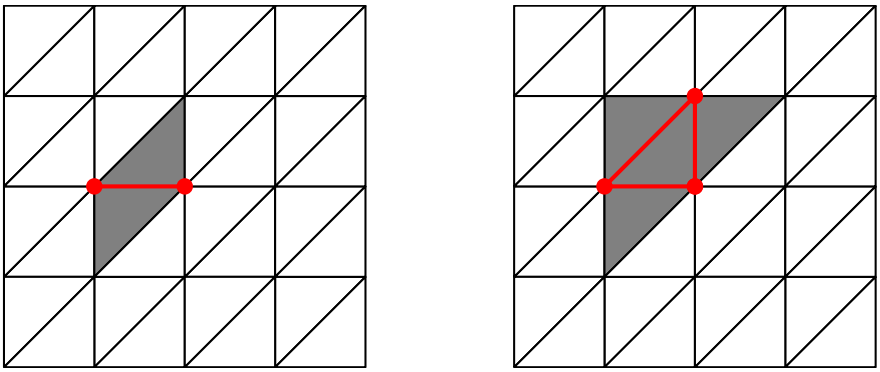
with inhomogeneous Dirichlet data $u_D \in H^{1/2}(\Gamma_D)$. Let $u \in H^1(\Omega)$ denote the weak solution and $u_h \in \mathcal{S}^1(\mathcal{T}_h)$ the P1-FEM solution for discrete Dirichlet data $u_{Dh} := \widehat{u}_{Dh}|_{\Gamma_D}$ with $\widehat{u}_{Dh} \in \mathcal{S}^1(\mathcal{T}_h)$. 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 41.*

$E = T_E^+ \cap T_E^-$. 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*

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

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*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

This finally proves

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

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

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

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

$$C_{\mathrm{ref}}^{-2}\|\phi_{\mathcal{E}} - \partial_n u_h\|_{L^2(E)}^2 \leq (\phi_{\mathcal{E}} - \partial_n u_h\,;\,v)_{L^2(E)} = (\phi_{\mathcal{E}} - \phi\,;\,v)_{L^2(E)} + (\phi - \partial_n u_h\,;\,v)_{L^2(E)}.$$

For the second term, we employ integration by parts to see

$$\begin{aligned}
(\phi - \partial_n u_h\,;\,v)_{L^2(E)} &= (\partial_n(u - u_h)\,;\,v)_{L^2(\partial T)} \\
&= (\nabla(u - u_h)\,;\,\nabla v)_{L^2(T)} - (f\,;\,v)_{L^2(T)} \\
&\leq \big(C_{\mathrm{inv}}\|\nabla(u - u_h)\|_{L^2(T)} + \|h_{\mathcal{T}} f\|_{L^2(T)}\big) h_E^{-1/2}\|\phi_{\mathcal{E}} - \partial_n u_h\|_{L^2(E)}.
\end{aligned}$$

The first term is estimated by the Cauchy inequality directly

$$(\phi_{\mathcal{E}} - \phi\,;\,v)_{L^2(E)} \leq \|h_{\mathcal{E}}^{1/2}(\phi_{\mathcal{E}} - \phi)\|_{L^2(E)}\|h_{\mathcal{E}}^{-1/2} v\|_{L^2(E)}.$$

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 30. 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 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 31. 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 $\mathcal{T}$. □

The following MATLAB code computes the vector of the element-based refinement indicators η_T from (4.40). The integral

$$h_T^2 \|f\|_{L^2(T)}^2 = h_T^2 \int_T f^2 \, dx \approx h_T^2 |T| \, f(s_T)^2 \simeq |T|^2 \, f(s_T)^2$$

is computed by 1-point quadrature associated with the center of mass s_T of T. The integral

$$\begin{aligned} h_T \|\phi - \partial_n u_h\|_{L^2(\partial T \cap \Gamma_N)}^2 &= \sum_{E \in \mathcal{E}_T \cap \mathcal{E}_N} h_T \int_E (\phi - \partial_n u_h)^2 \, ds \\ &\approx \sum_{E \in \mathcal{E}_T \cap \mathcal{E}_N} h_T h_E \big(\phi(m_E) - (\partial_n u_h)|_E\big)^2 \\ &\simeq |T| \sum_{E \in \mathcal{E}_T \cap \mathcal{E}_N} \big(\phi(m_E) - (\partial_n u_h)|_E\big)^2 \end{aligned}$$

is computed edge-wise by midpoint quadrature.

```
function etaR = computeEtaR(x,coordinates,elements,f,dirichlet,neumann,phi)

% ETAR = COMPUTEETAR(X,COORDINATES,ELEMENTS,F,DIRICHLET,NEUMANN,PHI)
% computes the element-based refinement indicators associated with
% the residual-based error estimator. ETAR is a column vector, where
% ETAR(j)^2 = |Tj| * || f ||_{L^2(Tj)}^2
%   + |Tj|^{1/2} * || jump(\partial_n u_h) ||_{L^2(\partial Tj \cap \Omega)}^2
%   + |Tj|^{1/2} * || \phi - \partial_n u_h ||_{L^2(\partial Tj\cap\Gamma_N)}^2
% The exact integrals involving F and PHI\lastmodified{11.05.2009}
 are integrated by midpoint
% quadrature.

% (c) 2007,2008 by Dirk Praetorius, last modified 08.01.2008
% dirk.praetorius@tuwien.ac.at - http://www.asc.tuwien.ac.at/~dirk

M = size(elements,1);
N = size(coordinates,1);

etaR = zeros(M,1);
int = sparse(N,N);

%*** Compute normal derivatives \partial_nT(uh) on all edges
for j = 1:M
    nodes = elements(j,:);
    B = [1 1 1 ; coordinates(nodes,:)'];
```

```
26      G = B \ [0 0 ; 1 0 ; 0 1];
27      grad = G'*x(nodes); % gradient \nabla u_h on element T_j
28      for k = 1:3
29          node1 = nodes(k);
30          node2 = nodes(mod(k,3)+1);
31          normal = [1 -1]*coordinates([node1,node2],:);
32          normal = normal*[0 1;-1 0] / norm(normal);
33          int(node1,node2) = normal*grad;
34      end
35  end
36
37  %*** Delete data in case of Dirichlet edges
38  for j = 1:size(dirichlet,1)
39      nodes = dirichlet(j,:);
40      int(nodes(1),nodes(2)) = 0;
41  end
42
43  %*** Evaluate exact Neumann data on Neumann edges
44  for j = 1:size(neumann,1)
45      nodes = neumann(j,:);
46      m = [1 1]*coordinates(nodes,:)/2;
47      int(nodes(2),nodes(1)) = -phi(m);
48  end
49
50  %*** Compute residual-based refinement indicators
51  for j = 1:M
52      nodes = elements(j,:);
53      %*** Compute volume contribution by midpoint quadrature
54      sizeT = det([1 1 1 ; coordinates(nodes,:)'])/2;
55      s = [1 1 1]*coordinates(nodes,:)/3;
56      etaR(j) = sizeT^2*f(s)^2;
57      %*** Add edge contributions
58      for k = 1:3
59          node1 = nodes(k);
60          node2 = nodes(mod(k,3)+1);
61          hE = norm([1 -1]*coordinates([node1,node2],:));
62          etaR(j) = etaR(j) + sizeT*(int(node1,node2)+int(node2,node1))^2;
63      end
64  end
65  etaR = sqrt(etaR);
```

Exercise 32. Consider the homogenous Dirichlet problems

$$\begin{aligned} -\Delta u &= 1 \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \Gamma = \partial\Omega, \end{aligned}$$

with Ω being either the square $\Omega = (-1,1)^2$ or the L-shaped domain $\Omega = (-1,1)^2 \backslash [0,1]^2$. Plot error and error estimator over the number of elements. How can one use the plot to see whether an error estimator is reliable and/or efficient? □

Exercise 33. Note that the computational time of the function `computeEtaR` grows quadratically with the number $M = \#\mathcal{T}$ of elements. This is due to the successive assembly of the sparse matrix `int`. Improve the implementation so that one observes real linear complexity. □

Exercise 34. For the computation of the residual-based refinement indicators η_T, the given MATLAB codes approximates the exact data f and ϕ for the terms

$$\|h_{\mathcal{T}} f\|_{L^2(T)} \quad \text{and} \quad \|h_{\mathcal{E}}^{1/2}(\phi - \partial_n u_h)\|_{L^2(E)} \quad \text{for } T \in \mathcal{T} \text{ resp. } E \in \mathcal{E}_N$$

by $f|_T \approx f(s_T)$ and $\phi|_E \approx \phi(m_E)$. Here, s_T and m_E denote the center of mass of T and the midpoint of E, respectively. Formally, this leads to an approximation $\widetilde{\eta}_R$ of η_R. Prove that, for $f \in H^2(\mathcal{T})$ and $\phi \in C^2(\mathcal{E}_N)$, there holds

$$|\eta_R - \widetilde{\eta}_R| \le C \left(\|h_{\mathcal{T}}^2 \nabla f\|_{H^1(\mathcal{T})} + \|h_{\mathcal{E}}^{3/2} \phi'\|_{C^1(\mathcal{E}_N)} \right)$$

with a constant $C > 0$ that only depends on $\sigma(\mathcal{T})$. Consequently, the computed estimator $\widetilde{\eta}_R$ is, in fact, reliable and efficient up to terms of higher order. □

4.4 Adaptive Mesh-Refining Algorithm

Usually, a posteriori error estimates are not only used to estimate the (unknown) error $\|\nabla(u - u_h)\|_{L^2(\Omega)}$ but even to steer the local mesh-refinement. Let

$$\eta := \Big(\sum_{T \in \mathcal{T}} \eta(T)^2 \Big)^{1/2}$$

be an a posteriori error estimator, where the quantities $\eta(T) := \eta_T$ reflect —at least heuristically— the (unknown) local error $\|\nabla(u - u_h)\|_{L^2(T)}$ for all $T \in \mathcal{T}$. We then aim to refine only the elements $T \in \mathcal{T}$, where $\eta(T)$ is large. Therefore, the quantities $\eta(T)$ are usually called **refinement indicators** (or error indicators). To state our version of an adaptive algorithm, we introduce some additional notation which will be used from now on.

- the index $\ell \in \mathbb{N}_0$ denotes the step of the adaptive algorithm,
- $\mathcal{T}_\ell$ is the mesh in the ℓ-th step of the adaptive algorithm.

- $\mathcal{N}_\ell$ and $\mathcal{E}_\ell$ denote the associated sets of nodes and edges, respectively.
- $U_\ell \in \mathcal{X}_\ell := \mathcal{S}^1_D(\mathcal{T}_\ell)$ denotes the Galerkin solution in the ℓ-th step.
- $h_\ell \in \mathcal{P}^0(\mathcal{T}_\ell)$, $h_\ell|_T := \operatorname{diam}(T)$ is the local mesh-side function.

With this notation, one common strategy is the following: Let $\theta \in (0,1)$ be the parameter for the adaptive algorithm.

Algorithm 4.14 (Adaptive Mesh-Refinement). **Input:** *Initial triangulation $\mathcal{T}_0$, tolerance $\tau > 0$, adaptivity parameter $\theta \in (0,1)$, counter $\ell := 0$.*

(i) *Compute discrete solution U_ℓ.*

(ii) *Compute refinement indicators $\eta_\ell(T)$ and error estimator $\eta_\ell = \big(\sum_{T\in\mathcal{T}_\ell} \eta_\ell(T)^2\big)^{1/2}$.*

(iii) *Stop computation provided that $\eta_\ell \le \tau$*

(iv) *Choose the minimal set $\mathcal{M}_\ell \subseteq \mathcal{T}_\ell$ of marked elements such that*

$$\theta\, \eta_\ell^2 = \theta \sum_{T\in\mathcal{T}_\ell} \eta_\ell(T)^2 \le \sum_{T\in\mathcal{M}_\ell} \eta_\ell(T)^2. \tag{4.53}$$

(v) *Generate a new regular mesh $\mathcal{T}_{\ell+1}$, where at least all marked elements have been refined.*

(vi) *Update $\ell \mapsto \ell+1$ and goto* (i).

Output: *Finite sequence of discrete solutions U_ℓ and corresponding error estimators η_ℓ.*

Remark. Clearly, the stopping criterion (iii) is only meaningful if η_ℓ is reliable and if the reliability constant in $\|\nabla(u-U_\ell)\|_{L^2(\Omega)} \le C_{\mathrm{rel}}\,\eta_\ell$ is known. In practice, runtime and storage requirements are the limiting quantities for a numerical simulation. Usually, one thus uses rather a maximal runtime or a maximal storage requirement, e.g., the maximal number of elements, as a stopping criterion. Adaptivity is then used to obtain an —in some sense— optimal approximation with respect to these side constraints. □

Remark. The marking criterion (4.53) was introduced by Dörfler (1996). It will be crucial to prove convergence of U_ℓ to the exact solution $u \in H^1_D(\Omega)$ of (4.1). Note that the choice $\theta \to 0$ in (4.53) leads to highly adapted meshes, whereas $\theta \to 1$ corresponds to (almost) uniform mesh-refinement. However, for small θ, only a few elements are refined per step. This may result in too many steps in the sense that usually the assembly of the Galerkin data is the most time consuming part of the algorithm. In practice, a good compromise between sufficient mesh-adaption and as few steps in the loop as possible appears to be $\theta \approx 0.25$. □

Remark. In the beginning of the analysis of adaptive FEM, Babuška proposed the following marking criterion: An element $T \in \mathcal{T}$ is marked for refinement if and only if

$$\eta_T \ge \theta \max\big\{\eta_{T'} \,\big|\, T' \in \mathcal{T}\big\}, \tag{4.54}$$

which is called **bulk criterion** in the literature. Convergence (but *not* optimality) of this version of adaptive FEM was proven by Morin, Siebert & Veeser (2008). Very recently, Diening, Kreuzer & Stevenson (2014) proved the so-called *instance optimality* of the adaptive algorithm for some extended bulk criterion. □

Before we comment on the local mesh-refinement in step (v) of Algorithm 4.14, we give a simple Matlab realization of Algorithm 4.14. We use the number of elements $M := \#\mathcal{T}$ and stop the adaptive algorithm when $M \geq M_{\max}$.

```
function [x,M,energy,etaR] ...
    = solveLaplaceAdaptively(coordinates,elements,f,dirichlet,neumann,phi,theta,Mmax)

ell = 1;
while 1
    M(ell) = size(elements,1);

    %*** Compute discrete solution and cooresponding energy
    [x,energy(ell)] = solveLaplace(coordinates,elements,f,dirichlet,neumann,phi);

    %*** Compute refinement indicators and error estimator
    indicators = computeEtaR(x,coordinates,elements,f,dirichlet,neumann,phi);
    etaR(ell) = norm(indicators);

    %*** Stopping criterion
    if M(ell) >= Mmax
        break
    end

    %*** Use Doerfler marking to mark elements for refinement
    [indicators,idx] = sort(indicators.^2,'descend');
    sumeta = cumsum(indicators);
    m = find(theta*sumeta(end)<=sumeta,1);
    marked = idx(1:m);

    %*** Generate a new mesh by RGB-refinement
    [coordinates,elements,dirichlet,neumann] = ...
       rgbrefine(coordinates,elements,dirichlet,neumann,marked);

   %*** Update counter
   ell = ell + 1;
end
```

4.4.1 Red-Green-Blue Refinement

It now remains to discuss the mesh-refinement. Recall that all error estimates are affected by the shape regularity $\sigma(\mathcal{T}_\ell)$ in the sense that the involved constants become unbounded for $\sigma(\mathcal{T}_\ell) \xrightarrow{\ell\to\infty}$

∞. Therefore, the mesh-refinement has to take care of the interior angles of the elements $T \in \mathcal{T}_\ell$ since $\sigma(\mathcal{T}_\ell)$ tends to infinity if and only if the minimal interior angle of the triangulation tends to zero. We follow the so-called **red-green-blue strategy** (or **RGB-refinement**): This refinement strategy is based on edge-refinement. First, we thus use the following marking rule:

- If an element $T \in \mathcal{T}_\ell$ is marked for refinement, we mark all edges $E \in \mathcal{E}_T$ for refinement.

We now proceed recursively as follows:

- For each element $T \in \mathcal{T}_\ell$, we mark its *longest* edge $E \in \mathcal{E}_T$ for refinement provided that $\mathcal{E}_T$ contains a marked edge.

Each marked edge will be halved, i.e., the midpoint m_E of a marked edge belongs to the new set $\mathcal{K}_{\ell+1}$ of nodes. Finally, we have the following refinement rules, for all $T \in \mathcal{T}_\ell$:

- If no edge in $\mathcal{E}_T$ is marked for refinement, T is not refined, i.e., $T \in \mathcal{T}_{\ell+1}$.
- If all edges in $\mathcal{E}_T$ are marked, we use a **red-refinement** of T, i.e., T is refined uniformly into four similar triangles, cf. Figure 4.3.
- If one edge in $\mathcal{E}_T$ is marked (and hence the longest edge), we use a **green-refinement**, i.e., T is refined into two triangles, cf. Figure 4.4.
- If two edges in $\mathcal{E}_T$ are marked — one of which is, according to the marking rule, the longest edge of T —, we use a **blue-refinement**, i.e., T is split into three triangles, cf. Figure 4.5.

In Figure 4.6, we visualize a simple example for an RGB-refined mesh.

FIGURE 4.3. *Red-refinement: If all edges of a triangle $T \in \mathcal{T}_\ell$ are marked (left), T is refined into four similar triangles $T_1, T_2, T_3, T_4 \in \mathcal{T}_{\ell+1}$ (right).*

FIGURE 4.4. *Green-refinement: If only the longest edge of a triangle $T \in \mathcal{T}_\ell$ is marked (left), T is refined into two new triangles $T_1, T_2 \in \mathcal{T}_{\ell+1}$ (right).*

We state the following elementary but important theorem without a proof.

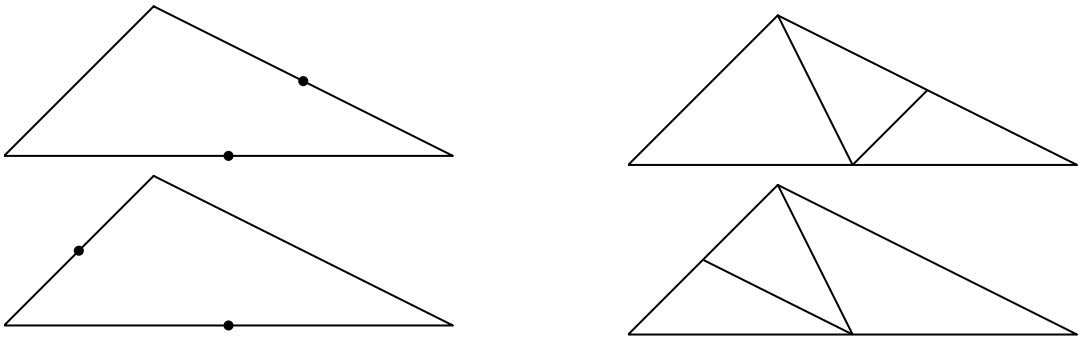


FIGURE 4.5. *Blue-refinement: If besides the longest edge of a triangle $T \in \mathcal{T}_\ell$ just one other edge is marked for refinement (left), T is refined into three new triangles $T_1, T_2, T_3 \in \mathcal{T}_{\ell+1}$ (right).*

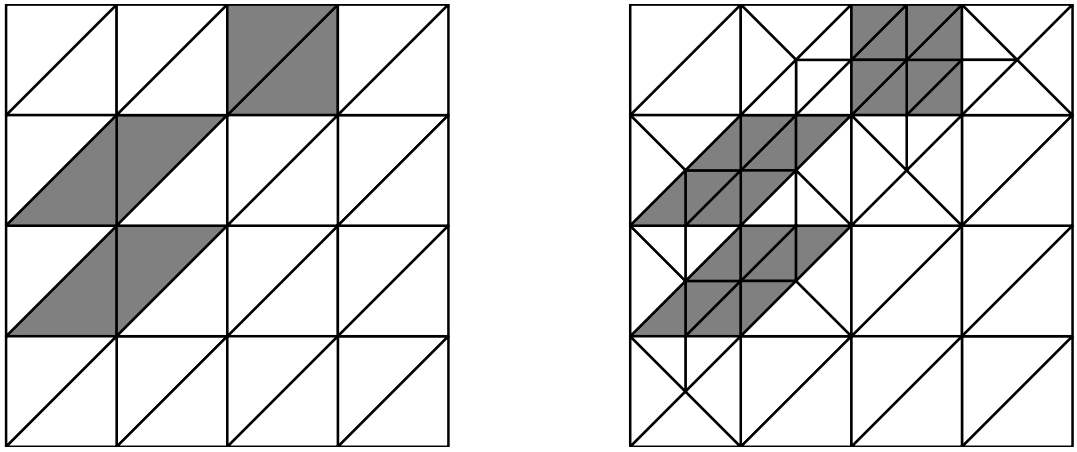


FIGURE 4.6. *The left plot shows an initial mesh $\mathcal{T}_\ell$ with marked elements coloured in grey. The right plot shows the mesh $\mathcal{T}_{\ell+1}$ obtained by RGB-refinement of the marked elements. The grey elements are obtained by uniform refinement of a marked element $T \in \mathcal{T}_0$.*

Theorem 4.15. *Let $\mathcal{T}_0$ be a regular triangulation such that $\varepsilon > 0$ is a lower bound for the smallest angle of a triangle $T \in \mathcal{T}_0$. Let $\mathcal{T}_\ell$ be a sequence of meshes, where $\mathcal{T}_\ell$ is obtained by RGB-refinement of the mesh $\mathcal{T}_{\ell-1}$ and where the set $\mathcal{M}_{\ell-1} \subseteq \mathcal{T}_{\ell-1}$ of marked elements is arbitrary. Then, $\mathcal{T}_\ell$ is regular and the smallest angle of all triangles $T \in \mathcal{T}_\ell$ is bounded from below by $\varepsilon/2$. In particular, there holds*

$$\sup_{\ell \in \mathbb{N}} \sigma(\mathcal{T}_\ell) < \infty, \tag{4.55}$$

which is an equivalent formulation for the fact that the smallest angles of the triangulations $\mathcal{T}_\ell$ do not tend to zero. ■

The following MATLAB code is an implementation of the RGB mesh-refinement strategy, which additionally takes care of the specification of the domain boundary.

```
1   function [coordinates,newelements,varargout] ...
```

```
2       = rgbrefine(coordinates,elements,varargin)
3
4   % [COORDINATES,ELEMENTS [,DIRICHLET] [,ROBIN] [,NEUMANN] ]
5   %  = RGBREFINE(COORDINATES,ELEMENTS [,DIRICHLET] [,ROBIN] [,NEUMANN], MARKED]
6   % refines the MARKED elements of a regular triangulation by a
7   % uniform refinement (red refinement). A green-blue closure leads
8   % to a new regular triangulation.
9   %
10  % vector MARKED contains the indices of all elements that will be refined
11  %
12  % Optionally, one can provide the specification of boundary conditions,
13  % e.g., Dirichlet, Robin, and/or Neumann boundaries. Then, the refined
14  % boundary conditions are returned in the same order
15
16  % (c) 2007 by Dirk Praetorius, last modified 21.11.2007
17  % dirk.praetorius@tuwien.ac.at - http://www.math.tuwien.ac.at/~dirk
18
19  M = size(elements,1);
20  N = size(coordinates,1);
21  markedelements = varargin{end};
22
23  %*** Sort elements such that the longest edge is always the first edge,
24  %*** i.e. we sort the entries in each row elements(j,:) accordingly.
25
26  for j = 1:M
27      [hmax,idx] = max(sum((coordinates(elements(j,[2,3,1]),:)- ...
28                            coordinates(elements(j,[1,2,3]),:)).^2'));
29      elements(j,:) = elements(j,[idx,mod(idx,3)+1,mod(idx+1,3)+1]);
30  end
31
32  %*** Introduce numbering of edges, stored in a sparse matrix EDGES:
33  %*** - EDGES(J,K) \neq 0 if and only if nodes J and K connected by edge,
34  %*** - EDGES(J,K) \neq EDGES(K,J) if and only if edge on boundary.
35
36  edges = sparse(N,N);
37  noedges = 0; % number of edges
38  for j = 1:M
39      for k = 1:3
40          a = [elements(j,k),elements(j,mod(k,3)+1)];
41          if edges(a(2),a(1))
42              edges(a(1),a(2)) = edges(a(2),a(1));
43          else
44              noedges = noedges+1;
45              edges(a(1),a(2)) = noedges;
46          end
```

```
47      end
48  end
49
50  %*** Transfer marking of elements to marking of edges.
51  %*** If element(j) is marked, we mark all of its edges (red-refinement).
52  %*** - MARKEDEDGES(k) \neq 0 if and only if edge K will be refined.
53
54  element2edges = zeros(M,3);
55  for j = 1:M
56      element2edges(j,:) = diag(edges(elements(j,:),elements(j,[2,3,1])))';
57  end
58  markededges = sparse(noedges,1);
59  markededges(element2edges(markedelements,:)) = ones(3*length(markedelements),1);
60
61  %*** Mark further edges according to green-blue closure:
62  %*** To ensure that the triangles do not degenerate, we always refine
63  %*** the longest edge, i.e. the first edge of an element.
64
65  edge2elements = sparse(N,N);
66  for j = 1:M
67      edge2elements(elements(j,:),elements(j,[2,3,1])) = ...
68          edge2elements(elements(j,:),elements(j,[2,3,1]))+j*eye(3,3);
69      k = j;
70      while k
71          I = element2edges(k,:);
72          if markededges(I(1))==1 | markededges(I(2:3))==[0;0]
73              k = 0;
74          else
75              markededges(I(1))=1;
76              k = edge2elements(elements(k,2),elements(k,1));
77          end
78      end
79  end
80
81  %*** For each marked edge, its midpoint becomes a new node.
82  %*** We store the number of the new nodes in MARKEDEDGES instead of 1.
83
84  idx = find(markededges);
85  markededges(idx) = N+1:N+length(idx);
86  for j = 1:nnz(markededges)
87      [a,b] = find(idx(j) == edges);
88      coordinates(markededges(idx(j)),:)=(coordinates(a(1),:)+coordinates(b(1),:))/2;
89  end
90
91  %*** Create new elements
```

```
92
93  I = reshape(edges(size(edges,1)*(elements(:,[2,3,1])-1)+elements(:,[1,2,3])),M,3);
94
95  boundaryedges = nonzeros(tril(abs(edges-edges')));
96  newelements = zeros(2*length(idx)-nnz(markededges(boundaryedges))+M,3);
97
98  counter = 0;
99  for j = 1:M
100     RefineEdge = find(markededges(I(j,:)));
101     newnodes=markededges(I(j,RefineEdge))';
102     if size(RefineEdge,1)==3 % red refinement
103         new = [ newnodes([2,3,1]);
104                 [elements(j,1) newnodes(1) newnodes(3)];
105                 [newnodes(1) elements(j,2) newnodes(2)];
106                 [newnodes(3) newnodes(2) elements(j,3)] ];
107     elseif size(RefineEdge,1)==2 % blue refinement
108         new = [ [newnodes(1), elements(j,RefineEdge(2)),newnodes(2)];
109                 [elements(j,5-RefineEdge(2)), ...
110                  elements(j,rem(5-RefineEdge(2),3)+1),newnodes(1)];
111                 [elements(j,rem(RefineEdge(2),3)+1),newnodes(1),newnodes(2)] ];
112     elseif size(RefineEdge,1)==1 % green refinement
113         new = [ [elements(j,[2,3]),newnodes];
114                 [elements(j,[3,1]),newnodes] ];
115     else % no refinement
116         new = elements(j,:);
117     end
118     newelements(counter+1:counter+size(new,1),:) = new;
119     counter = counter + size(new,1);
120 end
121
122 %*** Update boundary conditions
123
124 for j = 1:nargin-3
125     boundary = varargin{j};
126     if ~isempty(boundary)
127         counter = 0;
128         boundarynr = edges(size(edges,1)*(boundary(:,2)-1)+boundary(:,1));
129         for k = 1:size(boundary,1)
130             if markededges(boundarynr(k))
131                 boundary = [ boundary(1:k-1+counter,:);
132                              boundary(k+counter,1),markededges(boundarynr(k));
133                              markededges(boundarynr(k)),boundary(k+counter,2);
134                              boundary(k+1+counter:size(boundary,1),:) ];
135                 counter = counter + 1;
136             end
```

```
            end
        end
        varargout(j) = {boundary};
    end
```

4.5 Convergence of Adaptive FEM

In the following, we aim to show that Algorithm 4.14 creates a sequence U_ℓ of discrete solutions which converges to the exact solution $u \in H := H^1_D(\Omega)$. The adaptive algorithm generates a sequence $\mathcal{X}_\ell = \mathcal{S}^1_D(\mathcal{T}_\ell)$ of finite dimensional nested subspaces of H, i.e., $\mathcal{X}_\ell \subsetneqq \mathcal{X}_{\ell+1}$ for all $\ell \in \mathbb{N}_0$, since $\mathcal{T}_{\ell+1}$ is some refinement of $\mathcal{T}_\ell$. We first stress that the sequence U_ℓ is always convergent to some limit $U_\infty \in H$. However, we even stress that one may in general expect that $U_\infty \neq u$.

Exercise 35. Let $\mathcal{X}_\ell$ be nested subspaces of a Hilbert space H, i.e., $\mathcal{X}_\ell \subseteq \mathcal{X}_{\ell+1}$ for all $\ell \in \mathbb{N}_0$. Let $\langle\!\langle \cdot\ ;\ \cdot \rangle\!\rangle$ be an elliptic and continuous bilinear form on H with corresponding Galerkin solutions $U_\ell \in \mathcal{X}_\ell$. Prove that the limit $U_\infty := \lim_{\ell\to\infty} U_\ell$ exists in H. **Hint:** Define $\mathcal{X}_\infty$ as the closure of $\bigcup_{\ell=0}^\infty \mathcal{X}_\ell$ in H. Let $U_\infty \in \mathcal{X}_\infty$ be the corresponding Galerkin solution, and prove that U_∞ is the limit of the sequence U_ℓ. □

Exercise 36. Let $H = H^1_D(\Omega)$ and $\mathcal{X}_\ell = \mathcal{S}^1_D(\mathcal{T}_\ell)$, where the regular initial mesh $\mathcal{T}_0$ is given and where $\mathcal{T}_\ell$ is obtained iteratively by uniform refinement of $\mathcal{T}_{\ell-1}$. Prove that $\mathcal{X}_\infty = H$ for the space $\mathcal{X}_\infty$ from Exercise 35. □

The interpretation of the last exercises is the following: For uniform mesh-refinement, there usually holds $\mathcal{X}_\infty = H$ and thus $u = U_\infty$, i.e., we have convergence of the sequence of discrete solutions U_ℓ towards the unique solution u. However, adaptive mesh-refinement may lead to $\mathcal{X}_\infty \subsetneqq H$. Consequently, the question arrises whether the adaptive algorithm guarantees $U_\infty = u$ or not. This will be discussed in the following sections.

Throughout the subsequent section, we use the following notation, which is now collected for the convenience of the reader:

- $U_\ell \in \mathcal{X}_\ell := \mathcal{S}^1_D(\mathcal{T}_\ell)$ denotes the Galerkin solution.
- For $T \in \mathcal{T}_\ell$ and some $V \in \mathcal{S}^1_D(\mathcal{T}_\ell)$, $\eta_\ell(T,V)$ denotes the associated refinement indicator, e.g.,
$$\eta_\ell(T,V)^2 = h_T^2 \|f\|^2_{L^2(T)} + h_T \|[\![\partial_n V]\!]\|^2_{L^2(\partial T\cap\Omega)} + h_T \|\phi - \partial_n V\|^2_{L^2(\partial T\cap\Gamma_N)}. \tag{4.56}$$
- For some subset $\mathcal{M} \subseteq \mathcal{T}_\ell$ and $V \in \mathcal{S}^1_D(\mathcal{T}_\ell)$, let $\eta_\ell(\mathcal{M},V) := \big(\sum_{T\in\mathcal{M}} \eta_\ell(T,V)^2\big)^{1/2}$.
- We abbreviate $\eta_\ell(\mathcal{M}) = \eta_\ell(\mathcal{M},U_\ell)$ and $\eta_\ell = \eta_\ell(\mathcal{T}_\ell)$.

Note that in case of (4.56), η_ℓ is the residual a posteriori error estimator discussed in Section 4.3. We recall some technical terms, proven above for the residual error estimator η_ℓ.

- η_ℓ is **reliable** if

$$\|u - U_\ell\|_H \le C_{\mathrm{rel}} \eta_\ell. \tag{4.57}$$

- η_ℓ is **efficient** (up to oscillation terms which depend only on $\mathcal{T}_\ell$), if

$$\eta_\ell \le C_{\mathrm{eff}} \big(\|u - U_\ell\|_H + \mathrm{osc}_\ell \big), \tag{4.58}$$

  where $\mathrm{osc}_\ell := \mathrm{osc}_\ell(\mathcal{T}_\ell)$, $\mathrm{osc}_\ell(\mathcal{M}) := \big(\sum_{T \in \mathcal{M}} \mathrm{osc}_\ell(T)^2 \big)^{1/2}$ for $\mathcal{M} \subseteq \mathcal{T}_\ell$, and

$$\mathrm{osc}_\ell(T)^2 := h_T^2 \|f - f_{\mathcal{T}}\|_{L^2(T)} + h_T \|\phi - \phi_{\mathcal{E}}\|^2_{L^2(\partial T \cap \Gamma_N)}. \tag{4.59}$$

- The set $\mathcal{M}_\ell \subseteq \mathcal{T}_\ell$ of marked elements is usually assumed to satisfy the **Dörfler marking**

$$\theta\, \eta_\ell \le \eta_\ell(\mathcal{M}_\ell) \tag{4.60}$$

  for some fixed parameter $\theta \in (0, 1)$.

Exercise 37. Prove that $\|u - U_\ell\|_H$ as well as osc_ℓ are monotonously decreasing for $\ell \to \infty$. Prove that in case of the residual-based indicators (4.56), there holds $\mathrm{osc}_\ell(T) \le \eta_\ell(T)$ for all $T \in \mathcal{T}_\ell$, i.e., the error estimator dominates the oscillation terms. □

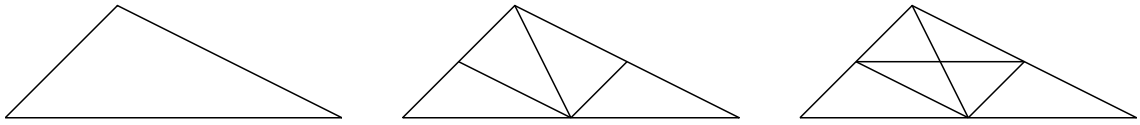


FIGURE 4.7. *Bisec(5) guarantees the inner node property: Let T be marked for refinement (left) and assume that the bottom edge is the reference edge. With five bisections, we pass the configuration of bisec(3) in the middle and end up with an inner node (right).*

The following convergence theorem is a result of CASCÓN, KREUZER, NOCHETTO & SIEBERT from 2008, where it is proven that the combined error quantity, which consists of error and error estimator, has a contraction property. We stress two important observations:

- For their analysis, CASCÓN, KREUZER, NOCHETTO, and SIEBERT re-define the **mesh width**

$$h_T := |T|^{1/2} \quad \text{for } T \in \mathcal{T}_\ell, \tag{4.61}$$

  whereas we considered $\mathrm{diam}(T)$ before. Note that, however, $|T| \le \mathrm{diam}(T)^2 \le 2\sigma(\mathcal{T})|T|$ so that both definition are equivalent for shape regular meshes, and we shall use the new definition in what follows.

- If $T \in \mathcal{T}_\ell$ is refined, each son $T' \in \mathcal{T}_{\ell+1}$ satisfies at least $|T'| \le |T|/2$, which now results in a strict reduction $h_{T'} \le h_T/\sqrt{2}$ of the local mesh-width (which fails, in general, for the usual definition $h_T = \operatorname{diam}(T)$). This observation is used in step 2 of the proof of the following theorem.

We note that the analysis holds for general symmetric problems. For non-symmetric problems, the correspoding result has been open until Feischl, Führer & Praetorius (2014).

Theorem 4.16 (Cascón, Kreuzer, Nochetto & Siebert '08). *Suppose that the set of marked elements $\mathcal{M}_\ell$ satisfies the Dörfler marking for some fixed $\theta \in (0,1)$. Then, there are constants $\kappa > 0$ and $q \in (0,1)$ which depend only on θ and uniform γ-shape regularity of $\mathcal{T}_\ell$ for all $\ell \in \mathbb{N}_0$, such that*

$$\big(\|\nabla(u-U_{\ell+1})\|_{L^2(\Omega)}^2 + \kappa\,\eta_{\ell+1}^2\big)^{1/2} \le q\,\big(\|\nabla(u-U_\ell)\|_{L^2(\Omega)}^2 + \kappa\,\eta_\ell^2\big)^{1/2} \text{ for all } \ell \in \mathbb{N}. \tag{4.62}$$

In particular, this implies convergence $\lim_{\ell\to\infty} \|\nabla(u-U_\ell)\|_{L^2(\Omega)} = 0 = \lim_{\ell\to\infty} \eta_\ell$.

Proof. **1. step.** There holds the following quasi-triangle inequality for the error estimator

$$\eta_\ell(V) \le \eta_\ell(W) + C_\Delta\,\|\nabla(V-W)\|_{L^2(\Omega)} \quad \text{for all } V, W \in \mathcal{S}^1_D(\mathcal{T}_\ell) \tag{4.63}$$

with some constant $C_\Delta > 0$ which depends only on $\sigma(\mathcal{T}_\ell)$: From the triangle inequalities in ℓ_2 and L^2, we infer

$$\begin{aligned}
\eta_\ell(V) &= \Big[\|h_\ell f\|_{L^2(\Omega)}^2 + \sum_{T\in\mathcal{T}_\ell} h_T\big(\|[\![\partial_n V]\!]\|_{L^2(\partial T\cap\Omega)}^2 + \|\phi - \partial_n V\|_{L^2(\partial T\cap\Gamma_N)}^2\big)\Big]^{1/2} \\
&\le \Big[\|h_\ell f\|_{L^2(\Omega)}^2 + \sum_{T\in\mathcal{T}_\ell} h_T\big(\|[\![\partial_n W]\!]\|_{L^2(\partial T\cap\Omega)}^2 + \|\phi - \partial_n W\|_{L^2(\partial T\cap\Gamma_N)}^2\big)\Big]^{1/2} \\
&\quad + \Big[\sum_{T\in\mathcal{T}_\ell} h_T\big(\|[\![\partial_n (V-W)]\!]\|_{L^2(\partial T\cap\Omega)}^2 + \|\partial_n (V-W)\|_{L^2(\partial T\cap\Gamma_N)}^2\big)\Big]^{1/2}.
\end{aligned}$$

For fixed $T \in \mathcal{T}_\ell$ and $E \in \mathcal{E}_T$, a scaling argument proves

$$h_T\big(\|[\![\partial_n (V-W)]\!]\|_{L^2(E\cap\Omega)}^2 + \|\partial_n (V-W)\|_{L^2(E\cap\Gamma_N)}^2\big) \lesssim \|\nabla(V-W)\|_{L^2(\omega_E)}^2,$$

where the constant depends only on $\sigma(\mathcal{T}_\ell)$. Consequently, we end up with (4.63).

2. step. There holds an estimator reduction in the sense that there is a constant $\varrho \in (0,1)$ with

$$\eta_{\ell+1}^2 \le (1+\delta)\varrho\,\eta_\ell^2 + C_\delta \|\nabla(U_{\ell+1}-U_\ell)\|_{L^2(\Omega)}^2 \quad \text{for all } \delta > 0, \tag{4.64}$$

where $C_\delta > 0$ depends only on δ and $C_\Delta > 0$. The constant ϱ depends only on θ and the reduction of the mesh-side on marked elements: Let $\Omega_* := \bigcup_{T\in\mathcal{M}_\ell} T$ denote the subdomain of Ω, where the elements are marked. Recall that $h_{T'} \le h_T/\sqrt{2}$ for all sons $T' \in \mathcal{T}_{\ell+1}$ of a marked element $T \in \mathcal{M}_\ell$. The crucial step is to observe that the error indicators

$$\eta_\ell(T,V)^2 = h_T^2\|f\|_{L^2(T)}^2 + h_T\|[\![\partial_n V]\!]\|_{L^2(\partial T\cap\Omega)}^2 + h_T\|\phi - \partial_n V\|_{L^2(\partial T\cap\Gamma_N)}^2.$$

for $h_T = |T|^{1/2}$ lead to

$$\begin{aligned}\eta_{\ell+1}(U_\ell)^2 &= \sum_{\substack{T'\in\mathcal{T}_{\ell+1}\\ T'\subseteq\Omega_*}} \eta_{\ell+1}(T',U_\ell)^2 + \sum_{\substack{T'\in\mathcal{T}_{\ell+1}\\ T'\subseteq\overline{\Omega\setminus\Omega_*}}} \eta_{\ell+1}(T',U_\ell)^2\\ &\le \frac{1}{\sqrt{2}} \sum_{\substack{T\in\mathcal{T}_\ell\\ T\subseteq\Omega_*}} \eta_\ell(T,U_\ell)^2 + \sum_{\substack{T\in\mathcal{T}_\ell\\ T\subseteq\overline{\Omega\setminus\Omega_*}}} \eta_\ell(T,U_\ell)^2\\ &= 2^{-1/2}\,\eta_\ell(\mathcal{M}_\ell)^2 + \eta_\ell(\mathcal{T}_\ell\setminus\mathcal{M}_\ell)^2\\ &= (2^{-1/2}-1)\eta_\ell(\mathcal{M}_\ell)^2 + \eta_\ell^2.\end{aligned}$$

By use of the Dörfler marking $\theta\eta_\ell^2 \le \eta_\ell(\mathcal{M}_\ell)^2$, we thus obtain

$$\eta_{\ell+1}^2(U_\ell) \le \eta_\ell^2 - (1-2^{-1/2})\eta_\ell(\mathcal{M}_\ell)^2 \le \varrho\,\eta_\ell^2 \quad \text{with} \quad \varrho := \big(1-\theta(1-2^{-1/2})\big).$$

Now, Young's inequality and step 1 conclude

$$\begin{aligned}\eta_{\ell+1}^2 &\le (1+\delta)\eta_{\ell+1}(U_\ell)^2 + (1+\delta^{-1})C_\Delta^2\,\|\nabla(U_{\ell+1}-U_\ell)\|_{L^2(\Omega)}^2\\ &\le (1+\delta)\varrho\,\eta_\ell^2 + (1+\delta^{-1})C_\Delta^2\,\|\nabla(U_{\ell+1}-U_\ell)\|_{L^2(\Omega)}^2.\end{aligned}$$

3. step. Proof of contraction property (4.62): Let $\kappa,\delta,\beta>0$ be constants which are fixed later. Let $\varrho\in(0,1)$ be the given constant from step 2. We recall the Galerkin orthogonality

$$\|\nabla(u-U_\ell)\|_{L^2(\Omega)}^2 = \|\nabla(u-U_{\ell+1})\|_{L^2(\Omega)}^2 + \|\nabla(U_{\ell+1}-U_\ell)\|_{L^2(\Omega)}^2$$

This and the estimator reduction imply

$$\begin{aligned}\|\nabla(u-U_{\ell+1})\|_{L^2(\Omega)}^2 + \kappa\,\eta_{\ell+1}^2 &= \|\nabla(u-U_\ell)\|_{L^2(\Omega)}^2 - \|\nabla(U_{\ell+1}-U_\ell)\|_{L^2(\Omega)}^2 + \kappa\,\eta_{\ell+1}^2\\ &\le \|\nabla(u-U_\ell)\|_{L^2(\Omega)}^2 + (\kappa C_\delta - 1)\,\|\nabla(U_{\ell+1}-U_\ell)\|_{L^2(\Omega)}^2 + \kappa(1+\delta)\varrho\,\eta_\ell^2.\end{aligned}$$

Provided that $\kappa C_\delta \le 1$, we infer

$$\begin{aligned}\|\nabla(u-U_{\ell+1})\|_{L^2(\Omega)}^2 + \kappa\,\eta_{\ell+1}^2 &\le \|\nabla(u-U_\ell)\|_{L^2(\Omega)}^2 + \kappa(1+\delta)\varrho\,\eta_\ell^2\\ &= \|\nabla(u-U_\ell)\|_{L^2(\Omega)}^2 - \kappa\beta\,\eta_\ell^2 + \kappa\big((1+\delta)\varrho+\beta\big)\,\eta_\ell^2.\end{aligned}$$

Reliability $\|\nabla(u-U_\ell)\|_{L^2(\Omega)} \le C_{\mathrm{rel}}\eta_\ell$ finally leads to

$$\begin{aligned}\|\nabla(u-U_{\ell+1})\|_{L^2(\Omega)}^2 + \kappa\,\eta_{\ell+1}^2 &\le (1-\kappa\beta C_{\mathrm{rel}}^{-2})\|\nabla(u-U_\ell)\|_{L^2(\Omega)}^2 + \kappa\big((1+\delta)\varrho+\beta\big)\,\eta_\ell^2\\ &\le \max\big\{1-\kappa\beta C_{\mathrm{rel}}^{-2}\,,\,(1+\delta)\varrho+\beta\big\}\,\big(\|\nabla(u-U_\ell)\|_{L^2(\Omega)}^2 + \kappa\,\eta_\ell^2\big).\end{aligned}$$

It remains to choose the constants κ,δ,β so that $q^2 := \max\big\{1-\kappa\beta C_{\mathrm{rel}}^{-2}\,,\,(1+\delta)\varrho+\beta\big\} \in (0,1)$:

- Choose $\delta>0$ such that $(1+\delta)\varrho<1$.
- Choose $\kappa>0$ such that $\kappa C_\delta \le 1$.
- Choose $\beta>0$ such that $(1+\delta)\kappa+\beta<1$.

This implies $q \in (0,1)$ and concludes the proof. ■

Remark. We again collect the main arguments of the preceeding proof, namely a certain quasi-triangle inequality of the estimator (4.63) and a strict reduction $\eta_{\ell+1}(\texttt{sons}(\mathcal{M}_\ell), U_\ell) \le \kappa\, \eta_\ell(\mathcal{M}_\ell, U_\ell)$ with some $\kappa \in (0,1)$ based on the strict reduction of the local mesh-width for marked elements. Besides this, only Galerkin orthogonality and Dörfler marking are used. Therefore, the proof works for a quite general class of symmetric problems and a variety of error estimators. The original work of Cascón, Kreuzer, Nochetto & Siebert (2008) considers linear second order symmetric and elliptic problems in divergence form and H^1-conforming finite element spaces with fixed polynomial degree. Finally, we stress that the proof also works for higher dimensions $d \ge 2$, where $h_T = |T|^{-1/d}$. For 2D, the usual definition $h_T := \operatorname{diam}(T)$ is sufficient if marked elements are refined, e.g., by red-refinement or bisec(3), since then all edges are bisected. □

Exercise 38. Prove the following variants of Young's inequality, for all $a, b \in \mathbb{R}$ and $\delta > 0$,
- $ab \le \frac{a^2}{2\delta} + \frac{\delta b^2}{2}$.
- $(a+b)^2 \le (1+\delta^{-1})a^2 + (1+\delta)b^2$. □

Exercise 39. Prove that the estimator reduction (4.64) with $C_\delta = (1+\delta^{-1})C_\Delta^2$ is equivalent to $\eta_{\ell+1} \le \varrho\, \eta_\ell + C_\Delta \|\nabla(U_{\ell+1} - U_\ell)\|_{L^2(\Omega)}$. □

Exercise 40. Suppose that an error estimator η_ℓ satisfies the estimator reduction (4.64) and that the discrete spaces are nested, i.e., $\mathcal{S}^1_D(\mathcal{T}_\ell) \subseteq \mathcal{S}^1_D(\mathcal{T}_{\ell+1})$ for all $\ell \in \mathbb{N}_0$. Prove that there holds $\lim_{\ell\to\infty} \eta_\ell = 0$. **Hint:** Use that there always holds convergence $U_\ell \xrightarrow{\ell\to\infty} U_\infty$ so that $\|\nabla(U_{\ell+1} - U_\ell)\|_{L^2(\Omega)} \xrightarrow{\ell\to\infty} 0$, cf. Exercise 35. □

Exercise 41. Prove that adaptive FEM based on the residual error estimator with the usual definition of $h_T := \operatorname{diam}(T)$ instead of (4.61) leads to R-linear convergence $\eta_{\ell+k} \le C\, q^k\, \eta_\ell$ for all $k, \ell \in \mathbb{N}_0$. The constants $C > 0$ and $0 < q < 1$ depend only on θ and the uniform γ-shape regularity of $\mathcal{T}_\ell$ for all $\ell \in \mathbb{N}_0$. **Hint:** Use Theorem 4.16 and consider the Dörfler marking. □

Chapter 5

A Priori Analysis II

5.1 FEM with Data Approximation

Now that we have realized that the P1-FEM is of order $\mathcal{O}(h)$, we need to show that the quadrature rules used for our MATLAB implementation are sufficiently accurate. Recall that we are approximating the right-hand side of the exact P1-FEM

$$F(v) := \int_\Omega f v \, dx + \int_{\Gamma_N} \phi v \, ds \quad \text{for } v \in H^1(\Omega) \tag{5.1}$$

by

$$F_h(v_h) := \sum_{T \in \mathcal{T}} |T| f(s_T) v_h(s_T) + \sum_{E \in \mathcal{E}_N} h_E \phi(m_E) v_h(m_E) \quad \text{for } v_h \in \mathcal{S}^1(\mathcal{T}), \tag{5.2}$$

where s_T denotes the center of mass of an element $T \in \mathcal{T}$ and where m_E denotes the midpoint of a Neumann edge $E \in \mathcal{E}_N$. Therefore, our MATLAB code realizes a perturbed P1-FEM and we need to study the convergence of this perturbed scheme.

5.1.1 First Strang Lemma

In this section, we go back to the abstract formulation of Galerkin schemes: Let H be a real Hilbert space with norm $\|\cdot\|_H$. Let $\langle\!\langle \cdot \,;\, \cdot \rangle\!\rangle$ be a bilinear form which is assumed to be elliptic and continuous, i.e., it holds that

$$\alpha \, \|v\|_H^2 \le \langle\!\langle v \,;\, v \rangle\!\rangle \quad \text{as well as} \quad \langle\!\langle v \,;\, w \rangle\!\rangle \le \beta \|v\|_H \|w\|_H \quad \text{for all } v, w \in H. \tag{5.3}$$

Let $F \in H^*$ be a given right-hand side. Then, the Lax-Milgram lemma applies and yields the existence and uniqueness of the solution $u \in H$ of

$$\langle\!\langle u \,;\, \cdot \rangle\!\rangle = F \in H^*. \tag{5.4}$$

For a discretization parameter $h > 0$, let X_h be a finite dimensional subspace of H. It is an important property of a Galerkin scheme that it is stable with respect to certain perturbations of the scalar product $\langle\!\langle \cdot \,;\, \cdot \rangle\!\rangle$ or the right-hand side F. — For the interpretation, recall that usually the right-hand side $F \in H^*$ as well as the scalar product $\langle\!\langle \cdot \,;\, \cdot \rangle\!\rangle$ involve integrals, which are computed

numerically by quadrature rules. For a fixed discrete space X_h, this leads to perturbations $F_h \in X_h^*$ and $\langle\!\langle \cdot\,;\cdot \rangle\!\rangle_h$ of F and $\langle\!\langle \cdot\,;\cdot \rangle\!\rangle$, respectively. In particular, this gives rise to additional **consistency errors**

$$\|F - F_h\|_{X_h^*} \quad \text{and} \quad \sup_{v_h \in X_h \setminus \{0\}} \frac{\|\langle\!\langle v_h\,;\cdot\rangle\!\rangle - \langle\!\langle v_h\,;\cdot\rangle\!\rangle_h\|_{X_h^*}}{\|v_h\|_H}.$$

In practice, the **best approximation error** (or **discretization error**) behaves like

$$\min_{v_h \in X_h} \|u - v_h\|_H = \mathcal{O}(h^\alpha) \quad \text{for } h \to 0,$$

where the **convergence order** $\alpha > 0$ usually depends on the regularity of the exact solution u. Then, the Céa lemma proves that

$$\|u - \mathbb{G}_h u\|_H = \mathcal{O}(h^\alpha).$$

The following result due to Strang shows that the consistency errors should be at least of the same order, i.e., one needs a sufficiently large order for the quadrature rules. Then, the perturbed Galerkin scheme

$$\langle\!\langle u_h\,;v_h\rangle\!\rangle_h = F_h(v_h) \quad \text{for all } v_h \in X_h \tag{5.5}$$

still allows for a unique solution $u_h \in X_h$. Moreover the **approximation error** still satisfies

$$\|u - u_h\|_H = \mathcal{O}(h^\alpha).$$

However, the consequence of Strang's lemma even works the other way around: You should avoid to compute integrals exactly (or with high accuracy quadrature rules) since this is usually computationally expensive and since this expense does not pay in the sense of an increased order of convergence. Finally, we note that analytic computation of integrals via antiderivatives, i.e., $\int_a^b f\,dx = F(b) - F(a)$ for the simple 1D case, necessarily leads to cancellation for small mesh-sizes. These are, however, avoided for numerical integration via Gaussian quadrature rules, since the Gaussian quadrature weights are all positive. In explicit terms, this implies that approximate computation will be numerically more accurate than analytic computation, if the quadrature rules are deliberately chosen.

Proposition 5.1 (First Strang Lemma). *Assume that $\langle\!\langle \cdot\,;\cdot \rangle\!\rangle_h$ is a bilinear form on X_h and that $F_h : X_h \to \mathbb{R}$ is linear. Then, there holds the following:*
(i) *Assume convergence of $\langle\!\langle \cdot\,;\cdot \rangle\!\rangle_h$ to $\langle\!\langle \cdot\,;\cdot \rangle\!\rangle$, i.e.,*

$$\lim_{h\to 0} E_h = 0 \quad \textit{with} \quad E_h := \sup_{v_h, w_h \in X_h \setminus \{0\}} \frac{|\langle\!\langle v_h\,;w_h\rangle\!\rangle - \langle\!\langle v_h\,;w_h\rangle\!\rangle_h|}{\|v_h\|_H \|w_h\|_H}. \tag{5.6}$$

Then, the bilinear forms are uniformly elliptic for small h, i.e.,

$$\exists \alpha_0 > 0\, \exists h_0 > 0\, \forall h \in (0,h_0)\, \forall v_h \in X_h \quad \alpha_0 \|v_h\|_H^2 \le \langle\!\langle v_h\,;v_h\rangle\!\rangle_h. \tag{5.7}$$

In particular, there exist unique $u_h \in X_h$ with $\langle\!\langle u_h \,;\, \cdot \rangle\!\rangle_h = F_h \in X_h^*$ for sufficiently small $h > 0$.
(ii) *Provided (5.7), there holds the Céa type estimate*

$$\begin{aligned} C^{-1} \, \|u - u_h\|_H &\leq \inf_{v_h \in X_h} \big(\|u - v_h\|_H + \|\langle\!\langle v_h \,;\, \cdot \rangle\!\rangle - \langle\!\langle v_h \,;\, \cdot \rangle\!\rangle_h\|_{X_h^*} \big) + \|F - F_h\|_{X_h^*} \\ &\leq (1 + E_h) \min_{v_h \in X_h} \|u - v_h\|_H + E_h \, \|u\|_H + \|F - F_h\|_{X_h^*} \end{aligned} \tag{5.8}$$

with u being the exact solution of (5.4). The constant $C > 0$ depends only on $\langle\!\langle \cdot \,;\, \cdot \rangle\!\rangle$.

Proof. Let $0 < \varepsilon < \alpha$ and $h_0 > 0$ such that

$$\forall h \in (0, h_0) \qquad \sup_{v_h \in X_h \setminus \{0\}} \frac{|\langle\!\langle v_h \,;\, v_h \rangle\!\rangle - \langle\!\langle v_h \,;\, v_h \rangle\!\rangle_h|}{\|v_h\|_H^2} \leq \varepsilon.$$

Then, $\alpha \|v_h\|_H^2 \leq \langle\!\langle v_h \,;\, v_h \rangle\!\rangle \leq \langle\!\langle v_h \,;\, v_h \rangle\!\rangle_h + |\langle\!\langle v_h \,;\, v_h \rangle\!\rangle - \langle\!\langle v_h \,;\, v_h \rangle\!\rangle_h| \leq \langle\!\langle v_h \,;\, v_h \rangle\!\rangle_h + \varepsilon \|v_h\|_H^2$, whence

$$(\alpha - \varepsilon) \, \|v_h\|_H^2 \leq \langle\!\langle v_h \,;\, v_h \rangle\!\rangle_h,$$

i.e., $\langle\!\langle \cdot \,;\, \cdot \rangle\!\rangle_h$ is an elliptic bilinear form on X_h for $h < h_0$. This concludes the proof of (i) with $\alpha_0 := \alpha - \varepsilon > 0$. To prove (ii), let $v_h \in X_h$ be arbitrary. Then,

$$\alpha_0 \, \|v_h - u_h\|_H^2 \leq \langle\!\langle v_h - u_h \,;\, v_h - u_h \rangle\!\rangle_h = \langle\!\langle v_h \,;\, v_h - u_h \rangle\!\rangle_h - F_h(v_h - u_h).$$

Together with

$$\langle\!\langle u - v_h \,;\, v_h - u_h \rangle\!\rangle = F(v_h - u_h) - \langle\!\langle v_h \,;\, v_h - u_h \rangle\!\rangle,$$

we obtain that

$$\begin{aligned} \alpha_0 \, \|v_h - u_h\|_H^2 &\leq \big[F(v_h - u_h) - F_h(v_h - u_h) \big] + \big[\langle\!\langle v_h \,;\, v_h - u_h \rangle\!\rangle_h - \langle\!\langle v_h \,;\, v_h - u_h \rangle\!\rangle \big] \\ &\quad - \langle\!\langle u - v_h \,;\, v_h - u_h \rangle\!\rangle \\ &\leq \|v_h - u_h\|_H \big[\|F - F_h\|_{X_h^*} + \|\langle\!\langle v_h \,;\, \cdot \rangle\!\rangle_h - \langle\!\langle v_h \,;\, \cdot \rangle\!\rangle\|_{X_h^*} + \beta \, \|u - v_h\|_H \big]. \end{aligned}$$

Finally, the combination with a triangle inequality yields that

$$\begin{aligned} \|u - u_h\|_H &\leq \|u - v_h\|_H + \|v_h - u_h\|_H \\ &\leq C \big[\|F_h - F\|_{X_h^*} + \|\langle\!\langle v_h \,;\, \cdot \rangle\!\rangle - \langle\!\langle v_h \,;\, \cdot \rangle\!\rangle_h\|_{X_h^*} + \|u - v_h\|_H \big] \end{aligned}$$

for any $v_h \in X_h$ with $C = 1 + \beta/\alpha_0$. This proves the first estimate in (5.8). To see the second estimate, note that

$$\|\langle\!\langle v_h \,;\, \cdot \rangle\!\rangle - \langle\!\langle v_h \,;\, \cdot \rangle\!\rangle_h\|_{X_h^*} \leq E_h \, \|v_h\|_H \leq E_h \, \|u\|_H + E_h \, \|u - v_h\|_H.$$

This concludes the proof. ∎

Under the assumptions of the Strang lemma, one can even show convergence of the perturbed Galerkin scheme.

Exercise 42. Assume that $\langle\!\langle \cdot\,;\,\cdot \rangle\!\rangle_h$ is a symmetric bilinear form on X_h^* and that $F_h \in X_h^*$. We assume convergence of the data in the sense that

$$\lim_{h\to 0} E_h = 0 = \lim_{h\to 0} \|F - F_h\|_{X_h^*} \quad \text{with} \quad E_h := \sup_{v_h, w_h \in X_h \setminus \{0\}} \frac{|\langle\!\langle v_h\,;\,w_h\rangle\!\rangle - \langle\!\langle v_h\,;\,w_h\rangle\!\rangle_h|}{\|v_h\|_H \|w_h\|_H}. \tag{5.9}$$

For sufficiently small $h > 0$, let $u_h \in X_h$ be the unique solutions of the perturbed Galerkin scheme (5.5). Under the approximation assumption

$$\lim_{h\to 0} \min_{v_h \in X_h} \|v - v_h\|_H = 0 \quad \text{for all } v \in D \tag{5.10}$$

for some dense subspace D of H, there holds convergence

$$\lim_{h\to 0} \|u - u_h\|_H = 0$$

with u being the exact solution of (5.4). □

5.1.2 Approximation of Volume Forces

For our Matlab implementation of P1-FEM, we compute the bilinear form $\langle\!\langle v_h\,;\,w_h\rangle\!\rangle$ analytically and perturb only the right-hand side. Let F and F_h be given by (5.1)–(5.2), respectively. According to the first Strang lemma 5.1, we only need to show that

$$\|F - F_h\|_{\mathcal{S}^1(\mathcal{T})^*} = \mathcal{O}(h)$$

to guarantee that the perturbed P1-FEM is also of order $\mathcal{O}(h)$. We consider the two contributions of the right-hand side separately.

Proposition 5.2. *Let $f \in H^2(\mathcal{T})$ and $F(v) := \int_\Omega f v\,dx$ for $v \in H^1(\Omega)$. Let $F_h(v_h) := \sum_{T\in\mathcal{T}} |T| f(s_T) v_h(s_T)$ for $v_h \in \mathcal{S}^1(\mathcal{T})$, where $s_T \in \mathbb{R}^2$ denotes the center of mass of an element $T \in \mathcal{T}$. Then, it holds that*

$$\|F - F_h\|_{\mathcal{S}^1(\mathcal{T})^*} \le C\,\|h^2 \nabla f\|_{H^1(\mathcal{T})}, \tag{5.11}$$

where the constant $C > 0$ depends only on T_{ref}, but not on Ω, $\mathcal{T}$, or f.

Proof. The proof is done elementwise. For $T \in \mathcal{T}$ and $w \in H^1(T)$, we define the integral mean $w_T := |T|^{-1} \int_T w\,dx$. According to the Poincaré inequality, it holds that $\|w - w_T\|_{L^2(T)} \le C_P h_T \|\nabla w\|_{L^2(T)}$, where the constant $C_P > 0$ is independent of T and w. Moreover, $w \mapsto w_T$ is the L^2-orthogonal projection onto $\mathcal{P}^0(T)$.

1. step. It holds that

$$\Big| \int_T f v_h\,dx - |T| f(s_T) v_h(s_T) \Big| \le C_P^2 h_T^2 \|\nabla f\|_{L^2(T)} \|\nabla v_h\|_{L^2(T)} + \|f_T - f(s_T)\|_{L^2(T)} \|v_h\|_{L^2(T)} :$$

From $\int_T v_h\,dx = |T| v_h(s_T)$, we infer that

$$\begin{aligned}\int_T f v_h\,dx - |T| f(s_T) v_h(s_T) &= (f - f(s_T)\,;\, v_h)_{L^2(T)} \\ &= (f - f(s_T)\,;\, v_h - v_{hT})_{L^2(T)} + (f - f(s_T)\,;\, v_{hT})_{L^2(T)} \\ &= (f - f_T\,;\, v_h - v_{hT})_{L^2(T)} + (f_T - f(s_T)\,;\, v_h)_{L^2(T)} \\ &\le \|f - f_T\|_{L^2(T)} \|v_h - v_{hT}\|_{L^2(T)} + \|f_T - f(s_T)\|_{L^2(T)} \|v_h\|_{L^2(T)}.\end{aligned}$$

where we have used orthogonality of $(\cdot)_T$ in the last but one step. The Poincaré inequality concludes the proof of step 1.

2. step. It holds that $\|f_T - f(s_T)\|_{L^2(T)} \le 2C_{\mathrm{ref}} h_T^2 \|D^2 f\|_{L^2(T)}$ with an independent constant $C_{\mathrm{ref}} > 0$, which is obtained from a scaling argument: Let $\Phi : T_{\mathrm{ref}} \to T$ denote an affine diffeomorphism with linear part $B \in \mathbb{R}^{2\times 2}$. Note that

$$f_T = \frac{1}{|T|}\int f\,dx = \frac{|\det B|}{|T|}\int_{T_{\mathrm{ref}}} f\circ\Phi\,dx = 2\int_{T_{\mathrm{ref}}} f\circ\Phi\,dx = \frac{1}{|T_{\mathrm{ref}}|}\int_{T_{\mathrm{ref}}} f\circ\Phi\,dx = (f\circ\Phi)_{T_{\mathrm{ref}}}.$$

Together with $f(s_T) = (f\circ\Phi)(s_{T_{\mathrm{ref}}})$, this yields that

$$\|f_T - f(s_T)\|_{L^2(T)} = |\det B^{-1}|^{-1/2}\,\|(f\circ\Phi)_{T_{\mathrm{ref}}} - (f\circ\Phi)(s_{T_{\mathrm{ref}}})\|_{L^2(T_{\mathrm{ref}})}.$$

We define $g := f\circ\Phi \in H^2(T_{\mathrm{ref}})$ and consider the operator $A : H^2(T_{\mathrm{ref}}) \to L^2(T_{\mathrm{ref}})$ defined by $Ag := g_{T_{\mathrm{ref}}} - g(s_{T_{\mathrm{ref}}})$. Then, $\mathcal{P}^1(T_{\mathrm{ref}}) \subseteq \ker A$ and continuity of A follows from the Sobolev inequality

$$\begin{aligned}\|Ag\|_{L^2(T_{\mathrm{ref}})} \le \|g_{T_{\mathrm{ref}}}\|_{L^2(T_{\mathrm{ref}})} + |T_{\mathrm{ref}}|^{1/2}|g(s_{T_{\mathrm{ref}}})| &\le \|g\|_{L^2(T_{\mathrm{ref}})} + |T_{\mathrm{ref}}|^{1/2}\|g\|_{\infty,T_{\mathrm{ref}}} \\ &\le (1 + C_{\mathrm{Sobolev}}|T_{\mathrm{ref}}|^{1/2})\|g\|_{H^2(T_{\mathrm{ref}})}\end{aligned}$$

Therefore, the Bramble-Hilbert lemma provides a constant $C_{\mathrm{ref}} > 0$ with $\|Ag\|_{L^2(T_{\mathrm{ref}})} \le C_{\mathrm{ref}}\|D^2 g\|_{L^2(T_{\mathrm{ref}})}$. We conclude the scaling argument by

$$C_{\mathrm{ref}}^{-1}\|(f\circ\Phi)_{T_{\mathrm{ref}}} - (f\circ\Phi)(s_{T_{\mathrm{ref}}})\|_{L^2(T_{\mathrm{ref}})} \le \|D^2(f\circ\Phi)\|_{L^2(T_{\mathrm{ref}})} \le |\det B|^{-1/2}\|B\|_F^2\|D^2 f\|_{L^2(T)},$$

which finally leads to

$$\|f_T - f(s_T)\|_{L^2(T)} \le 2C_{\mathrm{ref}}\,h_T^2\|D^2 f\|_{L^2(T)}.$$

3. step. It holds that $\big|\int_T f v_h\,dx - |T| f(s_T) v_h(s_T)\big| \le \max\{C_P^2, 2C_{\mathrm{ref}}\} h_T^2 \|\nabla f\|_{H^1(T)}\|v_h\|_{H^1(T)}$: The combination of step 1 and step 2 proves that

$$\begin{aligned}&\Big|\int_T f v_h\,dx - |T| f(s_T) v_h(s_T)\Big| \\ &\qquad \le \max\{C_P^2, 2C_{\mathrm{ref}}\} h_T^2 \big(\|\nabla f\|_{L^2(T)}\|\nabla v_h\|_{L^2(T)} + \|D^2 f\|_{L^2(T)}\|v_h\|_{L^2(T)}\big).\end{aligned}$$

Note that the brackets contain an $\mathbb{R}^2$-scalar product which is estimated with the help of the Cauchy inequality $ab + cd \le (a^2 + c^2)^{1/2}(b^2 + d^2)^{1/2}$. This concludes the proof of step 3.

4. step. With $C := \max\{C_P^2, 2C_{\mathrm{ref}}\}$, we finally sum over all elements $T \in \mathcal{T}$ to obtain that

$$\begin{aligned}
|F(v_h) - F_h(v_h)| &\leq \sum_{T\in\mathcal{T}} \Big| \int_T f v_h \, dx - |T| f(s_T) v_h(s_T) \Big| \\
&\leq C \sum_{T\in\mathcal{T}} \|h^2 \nabla f\|_{H^1(T)} \|v_h\|_{H^1(T)} \\
&\leq C \Big(\sum_{T\in\mathcal{T}} \|h^2 \nabla f\|_{H^1(T)}^2 \Big)^{1/2} \Big(\sum_{T\in\mathcal{T}} \|v_h\|_{H^1(T)}^2 \Big)^{1/2} \\
&= C \, \|h^2 \nabla f\|_{H^1(\mathcal{T})} \|v_h\|_{H^1(\Omega)}
\end{aligned}$$

by use of the Cauchy inequality. This concludes the proof. ∎

We stress that the proof does not work for $f \in H^1(\mathcal{T})$ since H^1-functions are in general discontinuous so that the evaluation of f at s_T is not well-defined. However, for $f \in C^1(\mathcal{T})$, everything works well.

Exercise 43. For $f \in C^1(\mathcal{T})$, define $F \in H^1(\Omega)^*$ and $F_h \in \mathcal{S}^1(\mathcal{T})^*$ as in Proposition 5.2. Then, there holds

$$\|F - F_h\|_{\mathcal{S}^1(\mathcal{T})^*} \leq C \, \|h \nabla f\|_{L^\infty(\Omega)}, \tag{5.12}$$

where the constant $C > 0$ does neither depend on Ω nor $\mathcal{T}$ or f. □

However, if the volume force only satisfies $f \in H^1(\mathcal{T})$, one can proceed as follows:

Exercise 44. For $f \in H^1(\mathcal{T})$, define $F \in H^1(\Omega)^*$ as in Proposition 5.2 and $F_h \in \mathcal{S}^1(\mathcal{T})^*$ by $F_h(v_h) := \sum_{T\in\mathcal{T}} |T| f_T v_h(s_T)$, where $f_T := |T|^{-1} \int_T f \, dx$ denotes the integral mean. Then,

$$\|F - F_h\|_{\mathcal{S}^1(\mathcal{T})^*} \leq C \, \|h^2 \nabla f\|_{L^2(\Omega)}, \tag{5.13}$$

where the constant $C > 0$ does neither depend on Ω nor $\mathcal{T}$ or f. □

5.1.3 Approximation of Neumann Data

Finally, we consider the approximation of the Neumann contribution.

Proposition 5.3. *Let $\phi \in C^2(\mathcal{E}_N) := \big\{\psi \in L^2(\Gamma_N) \,\big|\, \forall E \in \mathcal{E}_N \quad \psi|_E \in C^2(E)\big\}$ and $F(v) := \int_{\Gamma_N} \phi v \, ds$ for $v \in H^1(\Omega)$. Let $F_h(v_h) := \sum_{E\in\mathcal{E}_N} h_E \phi(m_E) v_h(m_E)$ for $v_h \in \mathcal{S}^1(\mathcal{T})$, where $m_E \in \mathbb{R}^2$ denotes the midpoint of a Neumann edge $E \in \mathcal{E}_N$. With the mesh-size function $h \in L^\infty(\Gamma_N)$, $h|_E := h_E = \operatorname{diam}(E)$, it then holds*

$$\|F - F_h\|_{\mathcal{S}^1(\mathcal{T})^*} \leq C \, \|h^{3/2} \phi'\|_{C^1(\mathcal{E}_N)} := \max_{E\in\mathcal{E}_N} \big(h_E^{3/2} \max\{\|\phi'\|_{L^\infty(E)}, \|\phi''\|_{L^\infty(E)}\} \big) \tag{5.14}$$

where the constant $C > 0$ depends only on $\sigma(\mathcal{T})$ and $|\Gamma_N|$.

Proof. We aim to follow the lines of the proof of Proposition 5.2. For a Neumann edge $E \in \mathcal{E}_N$ and $w \in L^2(E)$, let $w_E := h_E^{-1} \int_E w \, ds$ denote the integral mean.

1. step. From $\int_E v_h \, ds = h_E v_h(m_E)$, we infer that

$$\begin{aligned}\int_E \phi v_h \, ds - h_E \phi(m_E) v_h(m_E) &= (\phi - \phi(m_E) \,;\, v_h)_{L^2(E)} \\ &= (\phi - \phi(m_E) \,;\, v_h - v_{hE})_{L^2(E)} + (\phi - \phi(m_E) \,;\, v_{hE})_{L^2(E)} \\ &= (\phi - \phi_E \,;\, v_h - v_{hE})_{L^2(E)} + (\phi_E - \phi(m_E) \,;\, v_h)_{L^2(E)} \\ &\le \|\phi - \phi_E\|_{L^2(E)} \|v_h - v_{hE}\|_{L^2(E)} + \|\phi_E - \phi(m_E)\|_{L^2(E)} \|v_h\|_{L^2(E)}.\end{aligned}$$

where we have simply used orthogonality of $(\cdot)_E$. Therefore, the trace inequalities (4.10)–(4.11) yield that

$$\Big| \int_E \phi v_h \, ds - h_E \phi(m_E) v_h(m_E) \Big| \le C \Big(h_E^{1/2} \|\phi - \phi_E\|_{L^2(E)} + h_E^{-1/2} \|\phi_E - \phi(m_E)\|_{L^2(E)} \Big) \|v_h\|_{H^1(T)},$$

where $T \in \mathcal{T}$ is an arbitrary element with $E \in \mathcal{E}_T$. The constant $C > 0$ depends only on $\sigma(\mathcal{T})$ and on $|\Gamma_N|$.

2. step. It holds that $\|\phi - \phi_E\|_{L^2(E)} \le h_E^{3/2} \|\phi'\|_{L^\infty(E)}$: Note that $w := \phi - \phi_E \in C^1(E)$ has necessarily a zero $\zeta \in E$. Therefore, the fundamental theorem of calculus proves that

$$|w(x)| = \Big| \int_\zeta^x w' \, ds \Big| \le h_E^{1/2} \|w'\|_{L^2(E)}.$$

Integration over E thus yields that

$$\|\phi - \phi_E\|_{L^2(E)}^2 = \|w\|_{L^2(E)}^2 = \int_E |w(x)|^2 \, ds_x \le h_E^2 \|w'\|_{L^2(E)}^2 = h_E^2 \|\phi'\|_{L^2(E)}^2 \le h_E^3 \|\phi'\|_{L^\infty(E)}^2.$$

3. step. It holds that $\|\phi_E - \phi(m_E)\|_{L^2(E)} \le (1/2) \, h_E^{5/2} \, \|\phi''\|_{L^\infty(E)}$: Let $p \in \mathcal{P}^1(E)$ be a polynomial on E (with respect to the arclength) such that $\phi(m_E) = p(m_E)$ and $\phi'(m_E) = p'(m_E)$. Then,

$$\|\phi_E - \phi(m_E)\|_{L^2(E)} = h_E^{1/2} |\phi_E - \phi(m_E)| = h_E^{-1/2} \Big| \int_E \phi \, ds - h_E p(m_E) \Big| = h_E^{-1/2} \Big| \int_E (\phi - p) \, ds \Big|$$

With $w := \phi - p$ and hence $w'' = \phi''$, this implies that

$$\|\phi_E - \phi(m_E)\|_{L^2(E)} \le h_E^{-1/2} \|w\|_{L^1(E)} \le \|w\|_{L^2(E)}.$$

Note that w as well as w' have zeros at the edge midpoint m_E. Therefore, the same arguments as in step 2 (with the zero $\zeta = m_E$ and hence integration along a segment of length $h_E/2$) prove that

$$\|w\|_{L^2(E)}^2 \le \frac{h_E^2}{2} \, \|w'\|_{L^2(E)}^2 \quad \text{as well as} \quad \|w'\|_{L^2(E)}^2 \le \frac{h_E^2}{2} \, \|w''\|_{L^2(E)}^2 = \frac{h_E^2}{2} \, \|\phi''\|_{L^2(E)}^2.$$

Altogether, we see that

$$\|\phi_E - \phi(m_E)\|_{L^2(E)}^2 \le \|w\|_{L^2(E)}^2 \le \frac{h_E^4}{4} \, \|\phi''\|_{L^2(E)}^2 \le \frac{h_E^5}{4} \, \|\phi''\|_{L^\infty(E)}^2.$$

4. step. The combination of the preceding steps proves that

$$\begin{aligned}\Big|\int_E \phi v_h\,ds - h_E\phi(m_E)v_h(m_E)\Big| &\le C\,h_E^2\big(\|\phi'\|_{L^\infty(E)} + \|\phi''\|_{L^\infty(E)}\big)\|v_h\|_{H^1(T)}\\ &\le 2C\,h_E^2\|\phi'\|_{C^1(E)}\|v_h\|_{H^1(T)}\\ &\le 2C\,h_E^{1/2}\|h^{3/2}\phi'\|_{C^1(\mathcal{E}_N)}\|v_h\|_{H^1(T)}\end{aligned}$$

by definition of $\|w\|_{C^1(E)} := \max\{\|w\|_{L^\infty(E)}, \|w'\|_{L^\infty(E)}\}$.

5. step. We obtain the final result by summing over all Neumann edges $E \in \mathcal{E}_N$: For each $E \in \mathcal{E}_N$ we choose an element $T_E \in \mathcal{T}$ with $E \in \mathcal{E}_T$. Note that the element T_E can arise at most 3 times. Therefore,

$$\begin{aligned}|F(v_h) - F_h(v_h)| &\le 2C\,\|h^{3/2}\phi'\|_{C^1(\mathcal{E}_N)}\sum_{E\in\mathcal{E}_N} h_E^{1/2}\|v_h\|_{H^1(T_E)}\\ &\le 2C\,\|h^{3/2}\phi'\|_{C^1(\mathcal{E}_N)}\Big(\sum_{E\in\mathcal{E}_N} h_E\Big)^{1/2}\Big(\sum_{E\in\mathcal{E}_N}\|v_h\|^2_{H^1(T_E)}\Big)^{1/2}\\ &\le 2\sqrt{3}\,C\,|\Gamma_N|^{1/2}\|h^{3/2}\phi'\|_{C^1(\mathcal{E}_N)}\Big(\sum_{T\in\mathcal{T}}\|v_h\|^2_{H^1(T)}\Big)^{1/2}\\ &= 2\sqrt{3}\,C\,|\Gamma_N|^{1/2}\|h^{3/2}\phi'\|_{C^1(\mathcal{E}_N)}\|v_h\|_{H^1(\Omega)}.\end{aligned}$$

This concludes the proof. ■

Exercise 45. (i) Extend the MATLAB code `solveLaplace` such that besides the coefficient vector of the Galerkin solution $u_h \in \mathcal{S}^1(\mathcal{T})$ even the energy $|||u_h|||^2 = \|\nabla u_h\|^2_{L^2(\Omega)}$ is returned. The Galerkin orthogonality yields that

$$|||u - u_h|||^2 = |||u|||^2 - |||u_h|||^2.$$

Even if the exact energy $|||u|||^2$ is unknown, it can be extrapolated by use of Aitkin's Δ^2-method to obtain a good approximation of the error $|||u - u_h|||$.
(ii) Consider the homogenous Dirichlet problems

$$\begin{aligned}-\Delta u &= 1 \quad \text{in } \Omega,\\ u &= 0 \quad \text{on } \Gamma = \partial\Omega,\end{aligned}$$

with Ω being either the square $\Omega = (-1,1)^2$ or the L-shaped domain $\Omega = (-1,1)^2\backslash[0,1]^2$. Which experimental convergence rates $|||u - u_h||| = \mathcal{O}(h^\alpha)$ are observed? Do you expect that the solutions belong to $H^2(\Omega)$? **Hint:** For a convergent sequence $(x_j)_{j\in\mathbb{N}}$, the Δ^2-sequence reads

$$y_j = x_j - \frac{(x_{j+1} - x_j)^2}{x_{j+2} - 2x_{j+1} + x_j}.$$

Under certain assumptions on $(x_j)_{j\in\mathbb{N}}$ the sequence $(y_j)_{j\in\mathbb{N}}$ then converges faster to $\lim_{j\to\infty} x_j$. □

5.2 Inhomogeneous Dirichlet Data

Under the usual assumptions of the mixed boundary value problem of Section 2.3.2, we consider the boundary value problem

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega, \\ u &= u_D && \text{on } \Gamma_D, \\ \partial_n u &= \phi && \text{on } \Gamma_N. \end{aligned} \tag{5.15}$$

The only difference to the problem treated above is the fact, that the Dirichlet data u_D might be nontrivial. A function $u \in C^2(\overline{\Omega})$ that solves (5.15) is called **strong solution** of (5.15), and the formulation (5.15) is called the **strong form** of the boundary value problem. A function $u \in H^1(\Omega)$ is **weak solution** of (5.15) provided that

$$\gamma u|_{\Gamma_D} = u_D \tag{5.16a}$$

$$(\nabla u \,;\, \nabla v)_{L^2(\Omega)} = (f \,;\, v)_{L^2(\Omega)} + (\phi \,;\, \gamma v)_{L^2(\Gamma_N)} \quad \text{for all } v \in H^1_D(\Omega). \tag{5.16b}$$

These two equations are referred to as the **weak form** of the boundary value problem (5.15). Note that the variational part (5.16b) of the weak form is the same as for the mixed boundary value problem with homogeneous Dirichlet conditions $u_D = 0$.

The following proposition shows that (5.15) and (5.16) are essentially equivalent and that the weak solution is unique. The unique solvability, however, needs certain assumptions on the Dirichlet data: If (5.16) has a solution $u \in H^1(\Omega)$, then it holds that $\gamma u|_{\Gamma_D} = u_D$, i.e., u_D can be extended from Γ_D to a function $\widehat{u}_D \in H^1(\Omega)$. With the same arguments as above, cf. Exercise 12 on page 18, one shows that

$$H^{1/2}(\Gamma_D) := \left\{ \gamma u|_{\Gamma_D} \,\middle|\, u \in H^1(\Omega) \right\} \quad \text{with norm} \quad \|v\|_{H^{1/2}(\Gamma_D)} = \inf \left\{ \|\widehat{v}\|_{H^1(\Omega)} \,\middle|\, \gamma \widehat{v}|_{\Gamma_D} = v \right\}$$

is a Hilbert space. Moreover, $H^{1/2}(\Gamma_D)$ is continuously embedded into $L^2(\Gamma_D)$, and the restriction operator $(\cdot)|_{\Gamma_D} : H^{1/2}(\Gamma) \to H^{1/2}(\Gamma_D)$ is well-defined and continuous.

Proposition 5.4. (i) *Provided that $u \in C^2(\overline{\Omega})$ solves the strong form (5.15), u solves also the weak form (5.16).*

(ii) *Provided that $f \in C(\overline{\Omega})$, $\phi \in C(\overline{\Gamma}_N)$, and $u_D \in C(\overline{\Gamma}_D)$ and that the weak solution $u \in H^1(\Omega)$ of (5.16) additionally satisfies $u \in C^2(\overline{\Omega})$, then u even solves the strong form (5.15).*

(iii) *Let $\widehat{u}_D \in H^1(\Omega)$ be an arbitrary extension of the Dirichlet data $u_D \in H^{1/2}(\Gamma)$. Given $f \in L^2(\Omega)$ and $\phi \in L^2(\Gamma_N)$, there exists a unique $u_0 \in H^1_D(\Omega)$ such that*

$$(\nabla u_0 \,;\, \nabla v)_{L^2(\Omega)} = (f \,;\, v)_{L^2(\Omega)} - (\nabla \widehat{u}_D \,;\, \nabla v)_{L^2(\Omega)} + (\phi \,;\, \gamma v)_{L^2(\Gamma_N)} \quad \textit{for all } v \in H^1_D(\Omega). \tag{5.17}$$

(iv) *Under the assumptions of* (iii)*, a function $u \in H^1(\Omega)$ with $\gamma u|_{\Gamma_D} = u_D$ solves the weak form (5.16), if and only if $u_0 := u - \widehat{u}_D \in H^1_D(\Omega)$ solves (5.17).*

(v) *Under the assumptions of* (iii)*, there exists a unique weak solution $u \in H^1(\Omega)$ of (5.16). Contrary to $u_0 \in H^1_D(\Omega)$, however, the function $u \in H^1(\Omega)$ does not depend on the special choice of $\widehat{u}_D$.*

(vi) *The weak solution* $u \in H^1(\Omega)$ *satisfies*

$$\begin{aligned}\|u\|_{H^1(\Omega)} &\leq C_1 \left(\sup_{v \in H^1_D(\Omega)\setminus\{0\}} \frac{(f\,;\,v)_{L^2(\Omega)}}{\|v\|_{H^1(\Omega)}} + \sup_{w \in H^{1/2}(\Gamma_N)\setminus\{0\}} \frac{(\phi\,;\,w)_{L^2(\Gamma_N)}}{\|w\|_{H^{1/2}(\Gamma_N)}} + \|u_D\|_{H^{1/2}(\Gamma_D)} \right) \\ &\leq C_2 \left(\|f\|_{L^2(\Omega)} + \|\phi\|_{L^2(\Gamma_N)} + \|u_D\|_{H^{1/2}(\Gamma_D)} \right)\end{aligned} \tag{5.18}$$

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

Proof. Note that the variational form (5.16b) does not consider whether u_D is zero or not. Therefore, the same proofs as for the mixed boundary value problem with homogeneous Dirichlet data apply to prove (i) and (ii). To verify (iii), simply note that the left-hand side of (5.17) defines an equivalent scalar product on the Hilbert space $H^1_D(\Omega)$. The right-hand side is linear and continuous on $H^1_D(\Omega)$. Therefore, existence and uniqueness of u_0 follows from the Riesz theorem. (iv) is obvious, and (v) thus an immediate consequence of (iii) and (iv). To prove the stability estimate, we argue as for the homogeneous Dirichlet conditions. With the Friedrichs inequality, we see that

$$\begin{aligned}C_F^{-2}\,\|u_0\|^2_{H^1(\Omega)} &\leq \|\nabla u_0\|^2_{L^2(\Omega)} \\ &= (\nabla u_0\,;\,\nabla u_0)_{L^2(\Omega)} \\ &= (f\,;\,\nabla u_0)_{L^2(\Omega)} + (\phi\,;\,\gamma u_0)_{L^2(\Gamma_N)} - (\nabla \widehat{u}_D\,;\,\nabla u_0)_{L^2(\Omega)} \\ &\leq \|u_0\|_{H^1(\Omega)} \left(\sup_{v \in H^1_D(\Omega)\setminus\{0\}} \frac{(f\,;\,v)_{L^2(\Omega)}}{\|v\|_{H^1(\Omega)}} + \sup_{w \in H^{1/2}(\Gamma_N)\setminus\{0\}} \frac{(\phi\,;\,w)_{L^2(\Gamma_N)}}{\|w\|_{H^{1/2}(\Gamma_N)}} + \|\widehat{u}_D\|_{H^1(\Omega)} \right)\end{aligned}$$

Second, the triangle inequality gives

$$\begin{aligned}\|u\|_{H^1(\Omega)} &\leq \|\widehat{u}_D\|_{H^1(\Omega)} + \|u_0\|_{H^1(\Omega)} \\ &\leq (1 + C_F^2) \left(\sup_{v \in H^1_D(\Omega)\setminus\{0\}} \frac{(f\,;\,v)_{L^2(\Omega)}}{\|v\|_{H^1(\Omega)}} + \sup_{w \in H^{1/2}(\Gamma_N)\setminus\{0\}} \frac{(\phi\,;\,w)_{L^2(\Gamma_N)}}{\|w\|_{H^{1/2}(\Gamma_N)}} + \|\widehat{u}_D\|_{H^1(\Omega)} \right).\end{aligned}$$

Taking the infimum over all $\widehat{u}_D$, we conclude the stability estimate (5.18). ■

Remark. A first idea for the numerical approximation of the weak solution $u \in H^1(\Omega)$ of (5.15) might be the following:

- Construct an extension $\widehat{u}_D \in H^1(\Omega)$ of the Dirichlet data.
- Discretize the variational form (5.17) by P1-FEM to obtain an approximation $u_{0h} \in \mathcal{S}^1_D(\mathcal{T})$ of $u_0 \in H^1_D(\Omega)$.
- Compute $u_h := u_{0h} + \widehat{u}_D$ to obtain an approximation of u.

We stress, however, that then $u_h \notin \mathcal{S}^1(\mathcal{T})$ so that a postprocessing or evaluation of u_h is nontrivial. Moreover, we have to compute the scalar product $(\nabla \widehat{u}_D\,;\,\nabla v_h)$ for discrete functions to build the load vector of the P1-FEM for u_0. This leads to additional quadrature errors. Finally and most important, it might be hard to compute $\widehat{u}_D$ unless the Dirichlet data u_D are rather simple. □

To overcome the difficulties mentioned in the previous remark, one uses the following approach in practice, which is then called P1-FEM of the weak form (5.16):

- Discretize Dirichlet data $u_D \in H^{1/2}(\Gamma_D)$ by some $u_{Dh} \in \mathcal{S}^1(\mathcal{T}|_{\Gamma_D}) := \big\{ v_h|_{\Gamma_D} \,\big|\, v_h \in \mathcal{S}^1(\mathcal{T}) \big\}$.
- Construct extension $\widehat{u}_{Dh} \in \mathcal{S}^1(\mathcal{T})$ with $\widehat{u}_{Dh}|_{\Gamma_D} = u_{Dh}$.
- With $\widehat{u}_{Dh}$ replacing $\widehat{u}_D$, compute P1-FEM approximation $u_{0h} \in \mathcal{S}^1_D(\mathcal{T})$, cf. (5.17).
- Finally, define $u_h := u_{0h} + \widehat{u}_{Dh} \in \mathcal{S}^1(\mathcal{T})$ as approximation of the weak solution $u \in H^1(\Omega)$.

Note that the discrete solution $u_h \in \mathcal{S}^1(\mathcal{T})$ then belongs to the affine space $\widehat{u}_{Dh} + \mathcal{S}^1_D(\mathcal{T})$. The following result is the corresponding Céa-type lemma:

Lemma 5.5 (Céa lemma, first version). *Let $u \in H^1(\Omega)$ be the weak solution of (5.15). Let $\widehat{u}_{Dh} \in \mathcal{S}^1(\mathcal{T})$ be the approximate Dirichlet data and $u_{Dh} := \widehat{u}_{Dh}|_{\Gamma_D}$. Let $u_h \in \mathcal{S}^1(\mathcal{T})$ be the unique solution of*

$$\begin{aligned} u_h|_{\Gamma_D} &= u_{Dh} \\ (\nabla u_h \,;\, \nabla v_h)_{L^2(\Omega)} &= (f \,;\, v_h)_{L^2(\Omega)} + (\phi \,;\, v_h)_{L^2(\Gamma_N)} \quad \text{for all } v_h \in \mathcal{S}^1_D(\mathcal{T}). \end{aligned} \tag{5.19}$$

Then, u_h is quasioptimal in the sense that there exists a constant $C > 0$ such that

$$C^{-1} \, \|u - u_h\|_{H^1(\Omega)} \le \min_{v_h \in \mathcal{S}^1_D(\mathcal{T})} \|u - (v_h + \widehat{u}_{Dh})\|_{H^1(\Omega)} = \min_{\substack{w_h \in \mathcal{S}^1(\mathcal{T}) \\ w_h|_{\Gamma_D} = u_{Dh}}} \|u - w_h\|_{H^1(\Omega)}. \tag{5.20}$$

The constant $C > 0$ only depends on Ω and Γ_D.

Proof. Note that the variational formulations (5.16) and (5.19) imply the Galerkin orthogonality

$$(\nabla(u - u_h) \,;\, \nabla v_h)_{L^2(\Omega)} = 0 \quad \text{for all } v_h \in \mathcal{S}^1_D(\mathcal{T}).$$

We define $u_{0h} := u_h - \widehat{u}_{Dh} \in \mathcal{S}^1_D(\mathcal{T})$ and observe that

$$\begin{aligned} \|\nabla(u - u_h)\|^2_{L^2(\Omega)} &= (\nabla(u - u_h) \,;\, \nabla(u - [u_{0h} + \widehat{u}_{Dh}]))_{L^2(\Omega)} \\ &= (\nabla(u - u_h) \,;\, \nabla(u - [v_h + \widehat{u}_{Dh}]))_{L^2(\Omega)} \\ &\le \|\nabla(u - u_h)\|_{L^2(\Omega)} \|\nabla(u - [v_h + \widehat{u}_{Dh}])\|_{L^2(\Omega)} \end{aligned}$$

for each $v_h \in \mathcal{S}^1_D(\mathcal{T})$. Next, recall that $|\!|\!| v |\!|\!| := \|\nabla v\|_{L^2(\Omega)} + \|\gamma v\|_{L^2(\Gamma_D)}$ provides an equivalent norm on $H^1(\Omega)$, i.e., there are constants $C_1, C_2 > 0$ such that $C_1^{-1} |\!|\!| v |\!|\!| \le \|v\|_{H^1(\Omega)} \le C_2 \, |\!|\!| v |\!|\!|$ for all $v \in H^1(\Omega)$. Consequently,

$$\begin{aligned} C_2^{-1} \|u - u_h\|_{H^1(\Omega)} &\le \|\nabla(u - u_h)\|_{L^2(\Omega)} + \|\gamma(u - u_h)\|_{L^2(\Gamma_D)} \\ &= \|\nabla(u - u_h)\|_{L^2(\Omega)} + \|u_D - u_{Dh}\|_{L^2(\Gamma_D)} \\ &\le \|\nabla(u - [v_h + \widehat{u}_{Dh}])\|_{L^2(\Omega)} + \|\gamma(u - [v_h + \widehat{u}_{Dh}])\|_{L^2(\Gamma_D)} \\ &\le C_1 \|u - (v_h + \widehat{u}_{Dh})\|_{H^1(\Omega)} \end{aligned}$$

for all $v_h \in \mathcal{S}^1_D(\mathcal{T})$. This proves (5.20) with an infimum on the right-hand side. Standard arguments show that this infimum is, in fact, attained. ∎

Exercise 46. Proof that (5.19) has a unique solution $u_h \in \mathcal{S}^1(\mathcal{T})$. □

Remark. Note that Lemma 5.5 is independent of how the Dirichlet data are actually discretized, but the discretization enters the right-hand side, since it constraints the affine space for the minimum in (5.20). Later on, we shall see that appropriate discretization $u_{Dh} = J_h u_D$ by means of the Scott-Zhang projection J_h even guarantees that

$$\|u - u_h\|_{H^1(\Omega)} \leq C \min_{w_h \in \mathcal{S}^1(\mathcal{T})} \|u - w_h\|_{H^1(\Omega)},$$

where the right-hand side is independent of how u_D is actually discretized; see also Exercise 48–49 below. □

Remark. If the Dirichlet data u_D have an extension $\widehat{u}_D \in H^2(\Omega)$ with $\gamma \widehat{u}_D|_{\Gamma_D} = u_D$, then u_D is continuous. We define $\widehat{u}_{Dh} \in \mathcal{S}^1(\mathcal{T})$ nodewise by

$$\widehat{u}_{Dh}(z) = \begin{cases} u_D(z) & \text{for } z \in \overline{\Gamma}_D, \\ 0 & \text{else,} \end{cases}$$

for $z \in \mathcal{K}$. Let $u \in H^1(\Omega)$ denote the weak solution of (5.15) and $u_0 := u - \widehat{u}_D \in H^1_D(\Omega)$. We additionally define $\widetilde{u}_{Dh} \in \mathcal{S}^1_D(\mathcal{T})$ nodewise by

$$\widetilde{u}_{Dh}(z) = \begin{cases} 0 & \text{for } z \in \overline{\Gamma}_D, \\ \widehat{u}_D(z) & \text{else,} \end{cases}$$

for $z \in \mathcal{K}$. Note that the nodal interpolant of $\widehat{u}_D$ reads $I_h \widehat{u}_D = \widehat{u}_{Dh} + \widetilde{u}_{Dh}$ and that $\|\widehat{u}_D - I_h \widehat{u}_D\|_{H^1(\Omega)} = \mathcal{O}(h)$ decays with optimal order. Consequently, we may plug-in $u = \widehat{u}_D + u_0$ into Céa's lemma to observe that

$$\begin{aligned} C^{-1} \|u - u_h\|_{H^1(\Omega)} &\leq \min_{v_h \in \mathcal{S}^1_D(\mathcal{T})} \|u - (\widehat{u}_{Dh} + v_h)\|_{H^1(\Omega)} \\ &= \min_{v_h \in \mathcal{S}^1_D(\mathcal{T})} \|(\widehat{u}_D - I_h \widehat{u}_D) + (u_0 - v_h + \widetilde{u}_{Dh})\|_{H^1(\Omega)} \\ &= \min_{w_h \in \mathcal{S}^1_D(\mathcal{T})} \|(\widehat{u}_D - I_h \widehat{u}_D) + (u_0 - w_h)\|_{H^1(\Omega)} \\ &\leq \|\widehat{u}_D - I_h \widehat{u}_D\|_{H^1(\Omega)} + \min_{w_h \in \mathcal{S}^1_D(\mathcal{T})} \|u_0 - w_h\|_{H^1(\Omega)}. \end{aligned}$$

Conversely, it holds that

$$\begin{aligned} \min_{w_h \in \mathcal{S}^1_D(\mathcal{T})} \|u_0 - w_h\|_{H^1(\Omega)} &= \min_{w_h \in \mathcal{S}^1_D(\mathcal{T})} \|(u - \widehat{u}_D) - w_h\|_{H^1(\Omega)} \\ &= \min_{w_h \in \mathcal{S}^1_D(\mathcal{T})} \|u - (w_h + I_h \widehat{u}_D) - (\widehat{u}_D - I_h \widehat{u}_D)\|_{H^1(\Omega)} \\ &\leq \min_{w_h \in \mathcal{S}^1_D(\mathcal{T})} \|u - (w_h + I_h \widehat{u}_D)\|_{H^1(\Omega)} + \|\widehat{u}_D - I_h \widehat{u}_D\|_{H^1(\Omega)} \\ &\leq \|u - u_h\|_{H^1(\Omega)} + \|\widehat{u}_D - I_h \widehat{u}_D\|_{H^1(\Omega)}. \end{aligned}$$

Therefore, the proposed P1-FEM for the approximation of $u \in H^1(\Omega)$ converges with the same order as the P1-FEM for the approximation of $u_0 \in H^1_D(\Omega)$. □

The inhomogeneous Dirichlet problem allows the proof that the trace operator has a right inverse $\mathcal{L}$. This inverse is called *lifting operator*.

Exercise 47. Let $\gamma \in L(H^1(\Omega); H^{1/2}(\Gamma))$ denote the trace operator. Prove that there exists a lifting operator $\mathcal{L} \in L(H^{1/2}(\Gamma); H^1(\Omega))$ such that $\gamma \mathcal{L} v = v$ for all $v \in H^{1/2}(\Gamma)$. **Hint.** Consider an appropriate Dirichlet-Problem with inhomogeneous Dirichlet data $v \in H^{1/2}(\Gamma)$ and let $u := \mathcal{L} v \in H^1(\Omega)$ denote the unique solution. □

The assumptions of the following exercise will be satisfied for the Scott-Zhang projection.

Exercise 48 (Céa lemma, second version). Suppose that there exists a linear projection $P_h : H^1(\Omega) \to \mathcal{S}^1(\mathcal{T})$ with the following properties

(i) $\|P_h v\|_{H^1(\Omega)} \leq C_{\mathrm{stab}} \|v\|_{H^1(\Omega)}$ for all $v \in H^1(\Omega)$

(ii) $P_h v_h = v_h$ for all $v_h \in \mathcal{S}^1(\mathcal{T})$

(iii) $(P_h v)|_\omega = v|_\omega$ for all $v \in H^1(\Omega)$ with $v|_\omega \in \mathcal{S}^1(\mathcal{T}|_\omega)$ and $\omega \in \{\Gamma, \Gamma_D\}$

(iv) $(P_h v)|_\omega$ depends only on the trace $v|_\omega$ for all $v \in H^1(\Omega)$ and $\omega \in \{\Gamma, \Gamma_D\}$

Then, for $u \in H^1(\Omega)$ being the solution of (5.15) and $u_{Dh} := (P_h u)|_{\Gamma_D}$, it holds that

$$\min_{\substack{v_h \in \mathcal{S}^1(\mathcal{T}) \\ v_h|_{\Gamma_D} = u_{Dh}}} \|u - v_h\|_{H^1(\Omega)} \leq C \min_{w_h \in \mathcal{S}^1(\mathcal{T})} \|u - w_h\|_{H^1(\Omega)},$$

where $C > 0$ depends only on the stability constant C_{stab}, Ω, and Γ_D. In particular, this implies an unconstrained Céa lemma for the mixed boundary value problem with inhomogeneous Dirichlet data, i.e., under the assumptions of Lemma 5.5 and with $u_{Dh} = (P_h u)|_{\Gamma_D}$, it holds

$$\|u - u_h\|_{H^1(\Omega)} \leq C \min_{w_h \in \mathcal{S}^1(\mathcal{T})} \|u - w_h\|_{H^1(\Omega)}.$$

Hint. Let $w \in H^1(\Omega)$ be the weak solution of $\Delta w = 0$ in Ω subject to the boundary conditions $w = u - u_{Dh}$ on Γ_D and $\partial_n w = 0$ on Γ_N. Define $u_0 := u - w$. Prove that $u_0 = u_{Dh}$ on Γ_D and $\|u - u_0\|_{H^1(\Omega)} \simeq \|u - u_{Dh}\|_{H^{1/2}(\Gamma)}$. Choose $v_h := P_h u_0$. □

The existence of a Scott-Zhang-type projection is essentially equivalent to the validity of the Céa lemma.

Exercise 49. (a) Suppose that $P_h : H^1(\Omega) \to \mathcal{S}^1(\mathcal{T})$ satisfies the properties (i)–(iii) of Exercise 48 for $\omega = \Gamma$ only. Then, for all $u \in H^1(\Omega)$ and all $u_{Dh} \in \mathcal{S}^1(\mathcal{T})$, it holds that

$$\min_{\substack{v_h \in \mathcal{S}^1(\mathcal{T}) \\ v_h|_\Gamma = u_{Dh}}} \|u - v_h\|_{H^1(\Omega)} \leq C \big[\min_{w_h \in \mathcal{S}^1(\mathcal{T})} \|u - w_h\|_{H^1(\Omega)} + \|u - u_{Dh}\|_{H^{1/2}(\Gamma)} \big], \tag{5.21}$$

where $C > 0$ depends only on Ω and the stability constant C_{stab}. **Hint.** Argue along the lines of Exercise 48.
(b) Suppose that (5.21) holds true. Then, there exists a linear projection $P_h : H^1(\Omega) \to \mathcal{S}^1(\mathcal{T})$ which satisfies the properties (i)–(iii) of Exercise 48. **Hint.** For given $u \in H^1(\Omega)$, let $u_{Dh} \in \mathcal{S}^1(\mathcal{T}|_\Gamma)$ be the $H^{1/2}(\Gamma)$-best approximation of $u|_\Gamma \in H^{1/2}(\Gamma)$. With this, let

$P_h u := u_h \in \mathcal{S}^1(\mathcal{T})$ be the FEM solution of the inhomogeneous Dirichlet problem with discrete Dirichlet data u_{Dh}. □

Remark. Note that, for inhomogeneous Dirichlet data, it holds that

$$|||u - u_h|||^2 \neq |||u|||^2 - |||u_h|||^2$$

in general. Therefore, we cannot proceed as in Exercise 45 to approximate the error. Instead, in academic examples, where u is known, one has to compute

$$|||u - u_h|||^2 = \sum_{T \in \mathcal{T}} \|\nabla u - \nabla u_h\|^2_{L^2(T)}$$

by $\mathcal{T}$-piecewise numerical quadrature. □

Exercise 50. Write a MATLAB code for the P1-FEM for the mixed boundary value problem (5.15) with inhomogeneous but continuous Dirichlet data u_D. To verify the code, consider the Dirichlet problem

$$\begin{aligned} -\Delta u &= 1 \quad \text{in } \Omega = [0,1]^2, \\ u &= 1 \quad \text{on } \Gamma. \end{aligned}$$

If u_0 denotes the solution of the corresponding homogeneous problem, then it holds that $u = u_0 + 1$. □

5.3 Higher Dimensions

A set $T \subset \mathbb{R}^d$ is called **non-degenerate simplex** provided that there are nodes $z_0, \dots, z_d \in \mathbb{R}^d$ with $T = \operatorname{conv}\{z_0, \dots, z_d\}$ and provided that $|T| > 0$, i.e., T has positive measure. We note that T is in particular bounded and closed, whence compact. For $d = 2$, this definition describes non-degenerate triangles; for $d = 3$, this definition describes non-degenerate tetrahedra.

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

$$T_{\text{ref}} := \operatorname{conv}\{0, \mathbf{e}_1, \dots, \mathbf{e}_d\}, \tag{5.22}$$

where $\mathbf{e}_j$ is the j-th unit vector. There holds $|T_{\text{ref}}| = 1/d!$

The **diameter** of T is denoted by

$$h_T := \operatorname{diam}(T) := \max\big\{|x - y| \,\big|\, x, y \in T\big\}. \tag{5.23}$$

Moreover, ρ_T denotes the radius of the largest ball inscribed of T, i.e.,

$$\rho_T := \sup\big\{\rho > 0 \,\big|\, \exists x \in T \quad B(x, \rho) \subseteq T\big\}. \tag{5.24}$$

By $\mathcal{K}_T = \{z_0, \dots, z_d\}$, we denote the set of nodes of T. By $\mathcal{E}_T$, we denote the set of faces of T, i.e., $\mathcal{E}_T := \big\{\operatorname{conv}(M) \,\big|\, M \subseteq \mathcal{K}_T \text{ with } \#M = d\big\}$. Note that $E \in \mathcal{E}_T$ is a hyper-simplex of dimension $d - 1$, e.g., the faces of a tetrahedron are 2-dimensional surface triangles.

Definition. Let Ω be a bounded Lipschitz domain in $\mathbb{R}^d$, $d \geq 2$. A set $\mathcal{T}$ is a **triangulation** of Ω (consisting of simplices) if and only if

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

By $\mathcal{K} := \bigcup \{x \in \mathcal{K}_T \,|\, T \in \mathcal{T}\}$, we then denote the **set of nodes** of the triangulation $\mathcal{T}$ and by $\mathcal{E} := \bigcup \{E \in \mathcal{E}_T \,|\, T \in \mathcal{T}\}$ the **set of faces** 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

- either empty,
- or a joint k-dimensional hyper-simplex of both T and T', i.e., $T \cap T' = \operatorname{conv}(M)$ with $M \subseteq \mathcal{K}_T \cap \mathcal{K}_{T'}$ and $\#M = k \leq d-1$.

According to this regularity assumption, a face $E \in \mathcal{E}$ with surface measure $|E \cap \Gamma| > 0$ automatically satisfies $E \subseteq \Gamma$, i.e., a face E is either a boundary face or an interior face. 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 face $E \in \mathcal{E}$ 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 faces

$$\mathcal{E}_D := \{E \in \mathcal{E} \,|\, E \subseteq \overline{\Gamma}_D\} \quad \text{and} \quad \mathcal{E}_N := \{E \in \mathcal{E} \,|\, E \subseteq \overline{\Gamma}_N\} \tag{5.25}$$

as well as the set of all interior faces

$$\mathcal{E}_\Omega := \mathcal{E} \backslash (\mathcal{E}_D \cup \mathcal{E}_N). \tag{5.26}$$

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

For a regular triangulation $\mathcal{T}$, the hat functions provide a basis of $\mathcal{S}^1(\mathcal{T})$, and all results of Section 3.1 hold accordingly.

5.4 Shape Regularity & Scaling Arguments

A regular triangulation $\mathcal{T}$ is γ-shape regular if

$$\sigma(\mathcal{T}) := \max_{T \in \mathcal{T}} \frac{h_T}{\rho_T} \leq \gamma < \infty. \tag{5.27}$$

According to Exercise 17, this new definition is (up to some generic constant) equivalent to the definition given in Section 3.2.

For a non-degenerate simplex $T = \operatorname{conv}\{z_0, \dots, z_d\} \subset \mathbb{R}^d$, we define

$$\Phi_T : T_{\text{ref}} \to T, \quad \Phi_T v := z_0 + B_T v, \quad \text{where } B_T := \begin{pmatrix} z_1 - z_0 & z_2 - z_0 & \dots & z_d - z_0 \end{pmatrix} \in \mathbb{R}^{d \times d}.$$

Arguing as in Lemma 3.9, we see that $\|B_T\|_F \simeq h_T$, since the diameter of a simplex is its longest edge. To employ scaling arguments, it remains to prove $\|B_T^{-1}\|_F \lesssim \rho_T$. This is done with the help of the following lemma.

Lemma 5.6. *Let $T_1, T_2 \subset \mathbb{R}^d$ be compact sets with $B(x_j, \rho_j) \subseteq T_j \subseteq B(y_j, r_j)$ for some $x_j, y_j \in T_j$ and $\rho_j, r_j > 0$. Let $\Phi : T_1 \to T_2$ be affine with $\Phi(v) := Bv + w$ and $B \in \mathbb{R}^{d\times d}$. Then, it holds $\|B\|_2 \le r_2/\rho_1$ for the Euclidean operator norm.*

Proof. **1. step.** For $x \in \mathbb{R}^d$ with $|x| \le 2\rho_1$, it holds $|Bx| \le 2r_2$: Since $B(x_1, \rho_1) \subseteq T_1$, we find $y, z \in T_1$ with $x = y - z$. Then, $\Phi(y), \Phi(z) \in T_2$. Since $T_2 \subseteq B(y_2, r_2)$, it follows $2r_2 \ge |\Phi(y) - \Phi(z)| = |B(y - z)| = |Bx|$.

2. step. For $x \in \mathbb{R}^d$, it holds $|Bx| \le (r_2/\rho_1)\,|x|$: Let $x \in \mathbb{R}^d\backslash\{0\}$. Define $v := (2\rho_1/|x|)x$. From $|v| = 2\rho_1$, we obtain $(2\rho_1/|x|)|Bx| = |Bv| \le 2r_2$. This concludes the proof. ∎

Corollary 5.7. *With the above notation, the matrix $B_T \in \mathbb{R}^{d\times d}$ is invertible with $|\det B_T| \simeq |T|$ and $\|B_T^{-1}\|_F \lesssim \rho_T^{-1}$.*

Proof. As for 2D, one obtains $|\det B_T| \simeq |T| > 0$, and hence B_T and Φ_T are invertible. Note that B_T^{-1} is the linear part of the affine mapping Φ_T^{-1}. Hence, Lemma 5.6 gives $\|B_F^{-1}\|_2 \le h_{\mathrm{ref}}/\rho_T \le \rho_T^{-1}$. Norm equivalence on $\mathbb{R}^{d\times d}$ concludes $\|B_T^{-1}\|_F \simeq \|B_T^{-1}\|_2 \le \rho_T^{-1}$. ∎

5.4.1 Conclusion

The analysis of the previous chapters transfers from $d = 2$ to general dimension $d \ge 2$.

- The whole Chapter 2 is stated for $\Omega \subset \mathbb{R}^d$, $d \ge 2$.
- All results of Section 3.1 now hold verbatim for $d \ge 2$.
- The Approximation Theorem 3.5 holds for $d = 2, 3$. For $d \ge 4$, one requires higher smoothness of u to ensure continuity (cf. the Sobolev Theorem 3.4).
- Bramble-Hilbert Lemma 3.7 and transformation formula (Lemma 3.8) have already been formulated for $d \ge 2$.
- The inverse estimate and its applications hold verbatim.
- The data approximation analysis of Section 5.1 in the frame of the first Strang lemma applies for $d = 2, 3$. For $d \ge 4$, it requires only higher regularity assumptions on f to ensure continuity.
- Technical auxiliary results like the trace inequality remain valid for general $d \ge 2$.
- The a posteriori analysis of Chapter 4 remains valid. Only the proof of Lemma 4.1 which provides the dual basis functions to define the Scott-Zhang projection, has to be adapted.
- Finally, the adaptive convergence analysis requires $h_T := |T|^{1/d} \simeq \mathrm{diam}(T)$. All arguments remain valid.

5.5 Higher-order FEM

Due to the Céa lemma, we observe that it can be advantageous to consider higher-order polynomial discrete spaces for FEM. For example if the exact solution is smooth, higher-order spaces will achieve a better rate of convergence of the FEM error.

5.5.1 Higher-order elements in 1D

We consider a "triangulation" $\mathcal{T}$ with nodes x_i, $i = 0, \dots, M$ on an interval $\Omega \subset \mathbb{R}$. Instead of piecewise linear functions, we may also use higher-order polynomials to construct our discrete spaces, i.e.

$$\mathcal{S}^p(\mathcal{T}) := \{u \in H^1(\Omega) \,|\, u|_T \circ \Phi_T \in \mathcal{P}^p(T_{\text{ref}}) \qquad \forall T \in \mathcal{T}\}, \tag{5.28a}$$

$$\mathcal{S}_0^p(\mathcal{T}) := \mathcal{S}^p(\mathcal{T}) \cap H_0^1(\Omega) \tag{5.28b}$$

Here, the we use the mappings from the reference element $T_{\text{ref}} = [-1, 1]$, $\Phi_T : T_{\text{ref}} \to T$ as defined above.

Remark. Since Φ_T is affine, $u|_T \circ \Phi_T$ is a polynomial of degree p if and only if $u|_T$ is a polynomial of degree p. This means, the definition above is equivalent to $\mathcal{S}^p(\mathcal{T}) = \{u \in H^1(\Omega) \,|\, u|_T \in \mathcal{P}_p \forall T \in \mathcal{T}\}$. However, for non-affine maps Φ_T (e.g., for curved elements) the Definition (5.28) is still valid, while the second definition above does not generalize. (Remember that the scaling arguments and inverse estimates in the previous chapters required $u|_T \circ \Phi_T$ to be polynomial.) □

We construct a basis of $\mathcal{S}^p(\mathcal{T})$ on the reference element. We choose a basis $\{N_i \,|\, i = 1, \dots, p+1\}$ of the polynomial space $\mathcal{P}^p(T_{\text{ref}})$ such that

$$N_1(\xi) = \frac{1}{2}(1 - \xi), \quad N_2(\xi) = \frac{1}{2}(1 + \xi), \quad N_i(\pm 1) = 0 \quad i \geq 3.$$

Remark. The functions N_i, $i \geq 3$ can be chosen quite freely. The simplest possibility is $N_i(\xi) = (1 - \xi^2)\xi^{i-3}$ for all $i \in \{3, \dots, p + 1\}$. For small $p = 2, 3, 4$, this choice is fine. However, for higher p, the choice leads to very badly conditioned stiffness matrices and hence to numerical instabilities. It is better to choose more "orthogonal" basis functions as for example:

$$N_i(\xi) = \int_{-1}^{\xi} L_{i-2}(t)\, dt, \tag{5.29}$$

where $L_i \in \mathcal{P}_i$ is the i-th Legendre polynomial. Due to the orthogonality properties of Legendre polynomials, we have $N_i(\pm 1) = 0$ for $i \geq 3$. For the practical implementation, it is important to be able to quickly evaluate the basis functions. On one hand, there holds $(2i + 1) \int_{-1}^{\xi} L_i(t)\, dt = L_{i+1}(\xi) - L_{i-1}(\xi)$ and on the other hand, the Legendre polynomials can be computed very efficiently via three-term recurrences. □

Since the basis functions vanish for $i \geq 3$, it is easy to construct a basis of $\mathcal{S}^p(\mathcal{T})$ from these local definitions, i.e.,

$$\mathcal{B} = \mathcal{B}^{lin} \cup \left(\cup_{T \in \mathcal{T}} \mathcal{B}^T\right), \tag{5.30}$$

where $\mathcal{B}^{lin} = \{\varphi_i \,|\, i = 0, \dots, M\}$ are the hat-functions corresponding to x_i, $i = 0, \dots, M$ and $\mathcal{B}^T = \{\varphi_{T,i} \,|\, i = 3, \dots, p + 1\}$ with

$$\varphi_{T,i}(x) = \begin{cases} N_i(\Phi_T^{-1}(x)) & x \in T \\ 0 & x \in \Omega \setminus T \end{cases}$$

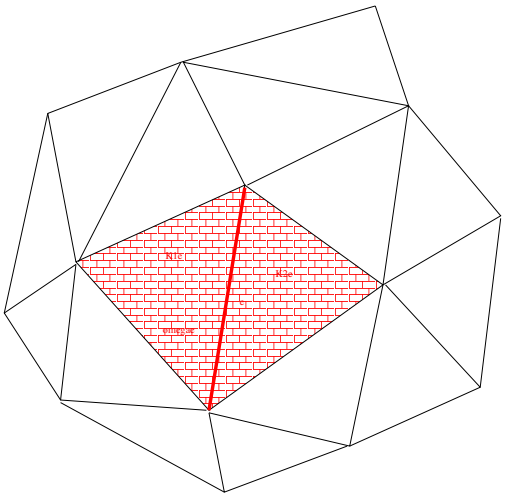


FIGURE 5.1. *The edge E with elements T_E^1, T_E^2 and $\Omega_E = T_E^1 \cup T_E^2$.*

We note that the construction of the basis above followed a typical recipe in FEM: The local basis function (the form functions) are associated with geometrical objects, e.g., the hat-functions are associated with nodes, whereas the bubble functions $\varphi_{T,i}$ are associated with elements $T \in \mathcal{T}$. Moreover, we observe that $\varphi|_T \circ \Phi_T \in \{0, N_1, \dots, N_{p+1}\}$, i.e., a basis function vanishes on an element, or it is exactly one of the local basis functions N_i.

5.5.2 Higher-order elements in 2D

Analogously to the 1D case, we may define higher-order basis functions in 2D. Let $\mathcal{T}$ denote a regular triangulation and define

$$\mathcal{S}^p(\mathcal{T}) := \{u \in H^1(\Omega) \mid u_{|K} \circ \Phi_T \in \mathcal{P}^p(T_{\text{ref}})\},$$

and $\mathcal{S}_0^p(\mathcal{T}) := \mathcal{S}^p(\mathcal{T}) \cap H_0^1(\Omega)$.

When constructing the basis functions for the FEM-spaces, we implicitly obeyed the following rules:

1. The basis functions $\varphi \in \mathcal{B}$ have a simple structure on the reference element T_{ref}, i.e., for all $T \in \mathcal{T}$ the function satisfies $\varphi|_T \circ \Phi_T \in \{0, N_1, N_2, \dots, \}$, where $\{0, N_1, \dots, \}$ is known explicitly.

2. The support $\operatorname{supp} \varphi$ of the basis functions $\varphi \in \mathcal{B}$ is small. This leads to sparse stiffness matrices and hence more efficient solvers.

3. For implementation, it is often advantageous to associate certain basis functions with geometrical objects, e.g., nodes, edges, elements, ...

The case $p = 2$

Idea: Construct $\mathcal{B}$ as a union of hat functions $\mathcal{B}^{lin}$ and edges-bubble functions. The latter functions φ_E are supported on Ω_E (see Figure 5.1). On the edge E, φ_E is a quadratic function as shown in Figure 5.2.

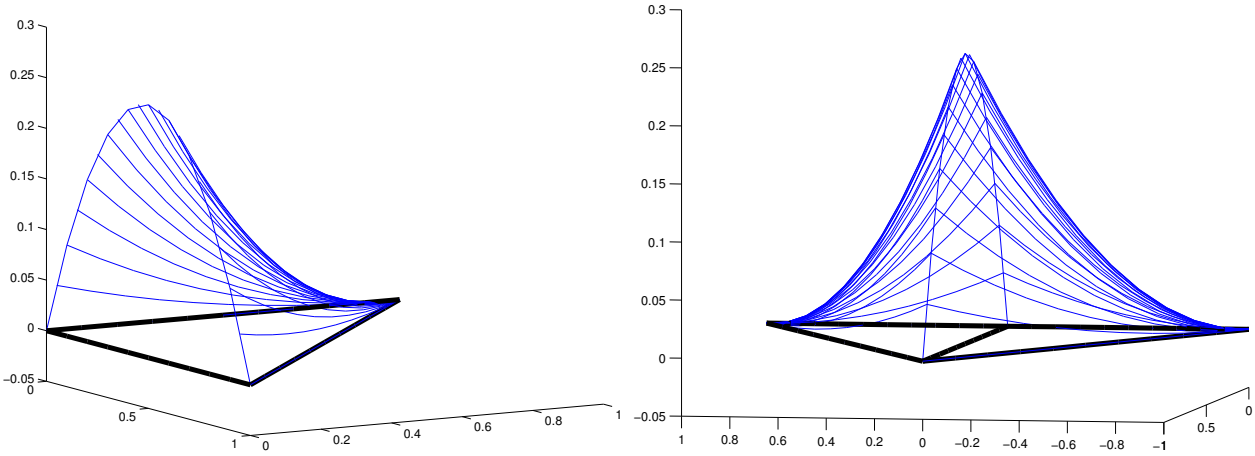
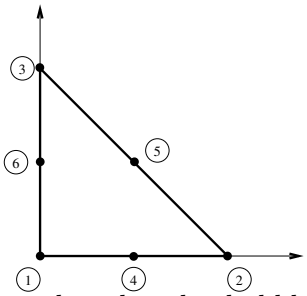


FIGURE 5.2. *Left:* N_4 *on* T_{ref}. *Right:* φ_E.

In engineering literature, the basis functions N_i are often illustrated with a diagram in which each dot represents a form-function:



N_1, N_2, N_3 are the hat-functions,

$$\begin{aligned} N_4(\xi,\eta) &:= \xi(1-\xi-\eta) \\ N_5(\xi,\eta) &:= \xi\eta \\ N_6(\xi,\eta) &:= \eta(1-\xi-\eta) \end{aligned}$$

We note that the edge bubbles $N_4, \ldots, N_6$ are chosen such that they vanish on two edges of T_{ref}. Hence, we may associate each of those functions with one edge where it is non-zero. We write the basis $\mathcal{B}$ of $\mathcal{S}^2(\mathcal{T})$ as

$$\mathcal{B} = \mathcal{B}^{lin} \cup \left(\cup_{e\in\mathcal{E}} \mathcal{B}^E\right),$$

where $\mathcal{B}^{lin}$ is again the set of hat-functions associated with the nodes $\mathcal{N}$. The one-element sets $\mathcal{B}^E = \{\varphi_E\}$, $E \in \mathcal{E}$ contain the edge bubble functions, which are characterized as follows:

$$\varphi_E \in H^1(\Omega), \qquad \operatorname{supp}\varphi_E \subset \overline{\Omega_E}, \qquad \varphi_E|_T \circ \Phi_T \in \{N_4, N_5, N_6\} \qquad \forall T \in \Omega_E. \tag{5.31}$$

Remark. If we restrict them to an edge of T_{ref}, the functions N_i ($i \in \{4,5,6\}$) are symmetric with respect to the midpoint of the edge. Hence, the above definition of φ_E leads to a continuous basis function. To see this, let $E \in \mathcal{E}$ with two elements T_E^1, $T_E^2 \in \Omega_E$. Let $\Gamma_4 = \{(x,0)\,|\,x \in (0,1)\}$, $\Gamma_5 = \{(x,y)\,|\,x \in (0,1)\, 1-x-y=0\}$, $\Gamma_6 = \{(0,y)\,|\,y \in (0,1)\}$ denote the three edges of the reference element T_{ref}. Let $i,\ j \in \{4,5,6\}$ denote the edge numbers corresponding to E, i.e.,

$\Phi_{T_E^1}(\Gamma_i) = E$ and $\Phi_{T_E^2}(\Gamma_j) = E$. Then, the above definition of φ_E is equivalent to

$$\varphi_E(x) := \begin{cases} N_i \circ \Phi_{T_E^1}^{-1}(x) & x \in \overline{T_E^1} \\ N_j \circ \Phi_{T_E^2}^{-1}(x) & x \in \overline{T_E^2} \\ 0 & else \end{cases}$$

The symmetry of N_i on the edges shows that this is well-defined since the two cases coincide on the edge E.

□

Chapter 6

Mixed Problems

6.1 Abstract Analysis of Petrov-Galerkin Schemes

Recall that for a continuous linear operator $T \in L(X,Y)$, the **adjoint operator** $T^* : Y^* \to X^*$ is formally defined by

$$T^*y^* \in X^* \quad \text{with} \quad (T^*y^*)(x) := y^*(Tx) \quad \text{for all } y^* \in Y^* \text{ and } x \in X. \tag{6.1}$$

It is an easy application of the Hahn-Banach extension theorem that $T^* \in L(Y^*, X^*)$ even with the same operator norm $\|T\| = \|T^*\|$. We start this section with some easy, but later on important, observations.

Lemma 6.1. *Let X and Y be normed spaces and $T \in L(X,Y)$. Then, T is an isomorphism between X and* range(T) *if and only if*

$$\tau := \inf_{x \in X\setminus\{0\}} \frac{\|Tx\|_Y}{\|x\|_X} > 0. \tag{6.2}$$

In this case, there holds $\|T^{-1} : \text{range}(T) \to X\| = 1/\tau$. Moreover, the range$(T)$ *is closed provided that X is a Banach space.*

Proof. Clearly, $T^{-1} : \text{range}(T) \to X$ is well-defined (and hence an isomorphism in the sense of Linear Algebra) if and only if T is injective. If T is not injective, there exists some $x \neq 0$ with $Tx = 0$, and hence it follows $\tau = 0$. In particular, $\tau > 0$ implies that T is injective. By elementary calculations, we see for $y = Tx \in \text{range}(T)$ that

$$\|T^{-1}y\|_X = \|x\|_X = \|y\|_Y \frac{\|x\|_X}{\|Tx\|_Y} \leq \|y\|_Y \sup_{x \in X\setminus\{0\}} \frac{\|x\|_X}{\|Tx\|_Y} = \|y\|_Y \frac{1}{\inf_{x \in X\setminus\{0\}} \frac{\|Tx\|_Y}{\|x\|_X}} = \|y\|_Y \frac{1}{\tau}.$$

Hence, $\tau > 0$ implies $\|T^{-1} : \text{range}(T) \to X\| = 1/\tau < \infty$, i.e., T^{-1} is even continuous. Conversely, if T^{-1} is well-defined and bounded, there holds

$$\tau = \inf_{x \in X\setminus\{0\}} \frac{\|Tx\|_Y}{\|x\|_X} = \inf_{y \in \text{range}(T)\setminus\{0\}} \frac{\|y\|_Y}{\|T^{-1}y\|_X} \geq \|T^{-1} : \text{range}(T) \to X\|^{-1} > 0.$$

Finally, suppose that X is a Banach space and $\tau > 0$. Then, range(T) is a Banach space as well and hence, in particular, a closed subspace of Y. ■

Exercise 51. For each operator $T \in L(X, Y)$ between normed spaces X and Y holds

$$\overline{\text{range}(T)} = (\ker T^*)_\circ := \big\{ y \in Y \,\big|\, \forall y^* \in \ker T^* \quad y^*(y) = 0 \big\}. \tag{6.3}$$

Hint: The inclusion range$(T) \subseteq (\ker T^*)_\circ$ can be shown directly, which leads to $\overline{\text{range}(T)} \subseteq \overline{(\ker T^*)_\circ} = (\ker T^*)_\circ$. The converse inclusion follows by use of the Hahn-Banach separation theorem. □

According to the Hahn-Banach extension theorem, the **Hahn-Banach embedding**

$$I_X : X \to X^{**}, \quad (I_X x)(x^*) := x^*(x) \quad \text{for } x \in X \text{ and } x^* \in X^* \tag{6.4}$$

is an isometric linear operator, whence injective and continuous. A normed space X is **reflexive** provided that I_X is also surjective and thus an isometric isomorphism between X and X^{**}. We stress that

- reflexive spaces are, in particular, complete and thus Banach spaces,
- finite dimensional spaces are reflexive,
- all Hilbert spaces are reflexive,
- closed subspaces of reflexive spaces are also reflexive.

All of these facts are simple exercises left to the reader.

Theorem 6.2. *Let X and Y be reflexive Banach spaces over $\mathbb{R}$, and $T \in L(X, Y^*)$. Then, T is an isomorphism if and only if the following two conditions hold:*

- ***inf-sup condition*** $\tau := \inf\limits_{x \in X \setminus \{0\}} \sup\limits_{y \in Y \setminus \{0\}} \dfrac{(Tx)(y)}{\|x\|_X \|y\|_Y} > 0,$
- ***non-degeneracy condition*** $\forall y \in Y \setminus \{0\} \exists x \in X \quad (Tx)(y) \neq 0.$

*In this case, there holds $\|T^{-1}\| = 1/\tau$ for the operator norm of the inverse. The combination of inf-sup condition and non-degeneracy condition is called **LBB condition** in the literature, named after Ladyshenskaja, Babuška, and Brezzi.*

Proof. **1. step.** According to Lemma 6.1, $\tau > 0$ is equivalent to $T : X \to \text{range}(T)$ being an isomorphism with closed range. It thus remains to show that range$(T) = Y^*$ is equivalent to the non-degeneracy condition (ND). Assume there exists $y^* \in Y^* \setminus \text{range}(T)$. The Hahn-Banach separation theorem implies the existence of a functional $\psi \in Y^{**}$ such that $\psi(y^*) = 1$ and $\psi|_{\text{range}(T)} = 0$. With the identification of Y^{**} and Y, we obtain some $y \in Y \setminus \{0\}$ with $I_Y y = \psi$ and

$$0 = \psi(Tx) = (Tx)(y).$$

This contradicts the non-degeneracy condition (ND). We showed that ND implies range$(T) = Y^*$. For the converse direction assume range$(T) = Y^*$ and $y \in Y \setminus \{0\}$. There exists $y^* \in Y^*$ with $y^*(y) \neq 0$. Hence, we find $x \in X$ with $Tx = y^*$ and hence $(Tx)(y) \neq 0$. This concludes (ND) and hence the proof. ■

The following simple exercise proves that the assumptions on X in Theorem 6.2 are sharp.

Exercise 52. Let X be a normed space and Y be a reflexive Banach space over $\mathbb{R}$. Let $T \in L(X, Y^*)$ be an isomorphism. Prove that X is also a reflexive Banach space. **Hint:** It is known that a Banach space Z is reflexive, if and only if Z^* is reflexive. Moreover, Z is reflexive, if and only if each bounded sequence has a weakly convergent subsequence (i.e., the unit ball of Z is weakly compact). □

We now turn to continuous bilinear forms $a : X \times Y \to \mathbb{R}$ on normed spaces X and Y. So far, we only considered weak formulations of the type: Find $x \in X$ such that

$$a(x, \cdot) = x^* \in X^*, \tag{6.5}$$

where $a(\cdot, \cdot)$ is a continuous bilinear form on $X = Y$. For the classical Galerkin scheme, we assumed that $a(\cdot, \cdot)$ is even elliptic. Note that the last theorem provides a mathematical framework for weak formulations of the following type: Find $x \in X$ such that

$$a(x, \cdot) = y^* \in Y^*, \tag{6.6}$$

where $a(\cdot, \cdot)$ now is a continuous bilinear form $a : X \times Y \to \mathbb{R}$. In the literature, this approach is named after Petrov-Galerkin.

Corollary 6.3. *Let X and Y be real Banach spaces, where Y is reflexive. Let $a : X \times Y \to \mathbb{R}$ be bilinear and continuous. Then, the following statements* (i)–(ii) *are equivalent:*

(i) *For each $y^* \in Y^*$, exists a unique $x \in X$ with $a(x, \cdot) = y^*$.*

(ii) *The bilinear form satisfies the* **LBB condition***:*

- ***inf-sup condition*** $\alpha := \inf\limits_{x \in X \setminus \{0\}} \sup\limits_{y \in Y \setminus \{0\}} \dfrac{a(x,y)}{\|x\|_X \|y\|_Y} > 0,$
- ***non-degeneracy condition*** $\forall y \in Y \setminus \{0\} \exists x \in X \quad a(x, y) \neq 0.$

In this case, it holds

$$\alpha \, \|x\|_X \leq \|y^*\|_{Y^*} \leq \|a\| \, \|x\|_X, \tag{6.7}$$

where $\|a\| := \sup\limits_{\substack{x \in X \setminus \{0\} \\ y \in Y \setminus \{0\}}} \dfrac{a(x,y)}{\|x\|_X \|y\|_Y}$ *denotes the continuity bound of $a(\cdot, \cdot)$.*

Proof. We associate with $a(\cdot, \cdot)$ the operator $T \in L(X, Y^*)$ given by $Tx = a(x, \cdot)$. Note that (i) is equivalent to the fact that T is an isomorphism (according to the open mapping theorem).

According to Theorem 6.2, the latter is characterized by the LBB condition for T which, in fact, coincides with that for $a(\cdot,\cdot)$. For given $y^* \in Y^*$ and $x \in X$ with $a(x,\cdot) = y^* \in Y^*$, it holds $Tx = y^*$. With $\|T : X \to Y^*\| = \|a\|$, we see $\|y^*\|_{Y^*} \leq \|a\|\,\|x\|_X$. With $x = T^{-1}y^*$ and $\|T^{-1} : Y^* \to X\| \leq 1/\alpha$, we derive $\|x\|_X \leq \|y^*\|_{Y^*}/\alpha$. This concludes the proof. ■

One important difference to the elliptic framework now is, that we may not simply replace X and Y by discrete spaces X_h and Y_h, respectively. Instead, Corollary 6.3 states that we need to satisfy the inf-sup condition and the non-degeneracy condition not only for the pairing (X,Y) of continuous spaces, but also for any pairing (X_h,Y_h) of discrete spaces. To underline this, note that

$$T = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

is an isomorphism on $Y = X = \mathbb{R}^3$. For $X_h = Y_h = \mathbb{R}^2$ and the canonical embedding, i.e., $x \in \mathbb{R}^2$ is identified with $(x,0) \in \mathbb{R}^3$, the restricted matrix is

$$T_h = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

which is clearly singular. We finally note that in the discrete setting the inf-sup condition and the non-degeneracy condition are equivalent.

Proposition 6.4. *Let X and Y be real Banach spaces with $\dim X < \infty$ and $\dim Y < \infty$. Let $a : X \times Y \to \mathbb{R}$ be bilinear. Then, there holds the following:*

(i) *The inf-sup condition $\alpha := \inf_{x\in X\setminus\{0\}} \sup_{y\in Y\setminus\{0\}} \frac{a(x,y)}{\|x\|_X\|y\|_Y} > 0$ implies $\dim X \leq \dim Y$.*

(ii) *The non-degeneracy condition $\big(\forall y \in Y\setminus\{0\} \exists x \in X \quad a(x,y) \neq 0\big)$ implies $\dim Y \leq \dim X$.*

(iii) *For $\dim X = \dim Y$, the inf-sup condition is satisfied if and only if the non-degeneracy condition is satisfied.*

Proof. We define the operators $A_1 \in L(X,Y^*)$ and $A_2 \in L(Y,X^*)$ by $A_1x := a(x,\cdot)$ and $A_2y := a(\cdot,y)$. According to Linear Algebra, finite dimension implies

$$\dim X = \dim\ker(A_1) + \dim\operatorname{range}(A_1) \leq \dim\ker(A_1) + \dim Y^* = \dim\ker(A_1) + \dim Y,$$
$$\dim Y = \dim\ker(A_2) + \dim\operatorname{range}(A_2) \leq \dim\ker(A_2) + \dim X^* = \dim\ker(A_2) + \dim X.$$

1. step. If $\dim X > \dim Y$, we obtain $\dim\ker(A_1) > 0$. Hence, there exists $x \in X\setminus\{0\}$ with $A_1x = 0$. This implies $a(x,y) = 0$ for all $y \in Y$ and hence $\alpha = 0$ for the inf-sup constant. By contraposition, this shows that the inf-sup condition implies $\dim\ker(A_1) = 0$ and hence $\dim X \leq \dim Y$. This proves (i).

2. step. If $\dim Y > \dim X$, we obtain $\dim\ker(A_2) > 0$. Hence, there exists $y \in Y\setminus\{0\}$ with $A_2y = 0$. This implies $a(x,y) = 0$ for all $x \in X$, and hence the non-degeneracy condition fails. By contraposition, this shows that the non-degeneracy condition implies $\dim\ker(A_2) = 0$ and hence $\dim Y \leq \dim X$. This proves (ii).

3. step.

In Step (ii), we have shown that (ND) implies injectivity of A_2. Since $\dim X = \dim Y = \dim Y^*$, this proves that A_2 is bijective. The converse implication is obvious, i.e., A_2 is bijective if and only if (ND) holds. In Step (i), we showed that the inf-sup condition implies injectivity of A_1. Since $\dim X = \dim Y = \dim Y^*$, this proves that A_1 is bijective. Again, the converse implication is easy, i.e., A_1 is bijective if and only if the inf-sup condition holds. To conlcude (iii), we only have to show that bijectivity of A_1 and A_2 are equivalent. To that end, let $\{x_1, \dots, x_n\} \subset X$ and $\{y_1, \dots, y_n\} \subset Y$ be bases. We define the matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\boldsymbol{A}_{jk} := a(x_k, y_j)$ and note that $(A_1 x_k)(y_j) = a(x_k, y_j) = \boldsymbol{A}_{jk}$ as well as $(A_1 y_j)(x_k) = a(x_k, y_j) = \boldsymbol{A}_{jk}$. Therefore, $\boldsymbol{A}$ is the Petrov-Galerkin matrix corresponding to A_1 and its transpose $\boldsymbol{A}^T$ is the Petrov-Galerkin matrix corresponding to A_2. Therefore, Linear Algebra proves the equivalence

$$A_1 \text{ is bijective} \iff \boldsymbol{A} \text{ is regular} \iff \boldsymbol{A}^T \text{ is regular} \iff A_2 \text{ is bijective}$$

This concludes the proof. ∎

Exercise 53. Prove that a bilinear form $a : X \times Y \to \mathbb{R}$ on normed spaces X and Y is continuous if and only if $\|a\| := \sup\limits_{\substack{x\in X\setminus\{0\}\\ y\in Y\setminus\{0\}}} \dfrac{a(x,y)}{\|x\|_X \|y\|_Y} < \infty$. □

The following exercise states the quasi optimality of Petrov-Galerkin schemes. We stress, however, that the quasi-optimality constant depends on the discrete inf-sup condition.

Exercise 54 (Céa's Lemma for Petrov-Galerkin Schemes). We consider the weak form (6.6) with a continuous bilinear form $a : X \times Y \to \mathbb{R}$ on Banach spaces X and Y. Let $y^* \in Y^*$. Let X_h and Y_h be finite dimensional subspaces of X resp. Y with $\dim X_h = \dim Y_h$. We assume the

- **discrete inf-sup condition** $\alpha_h := \inf\limits_{x_h\in X_h\setminus\{0\}} \sup\limits_{y_h\in Y_h\setminus\{0\}} \dfrac{a(x_h,y_h)}{\|x_h\|_X \|y_h\|_Y} > 0,$

Then, there is a unique $x_h \in X_h$ with

$$a(x_h, \cdot) = y^* \in Y_h^*. \tag{6.8}$$

If $x \in X$ solves the weak form (6.6), we have quasi optimality

$$\|x - x_h\|_X \le \big(1 + \|a\|/\alpha_h\big) \min_{v_h\in X_h} \|x - v_h\|_X, \tag{6.9}$$

where $\|a\| := \sup\limits_{\substack{x\in X\setminus\{0\}\\ y\in Y\setminus\{0\}}} \dfrac{a(x,y)}{\|x\|_X \|y\|_Y}$ denotes the continuity bound of $a(\cdot,\cdot)$. □

A simple observation is that the LBB theory allows a generalization of the Lax-Milgram lemma to the case of reflexive Banach spaces.

Exercise 55 (Lax-Milgram Lemma for Reflexive Spaces). Let $a : X \times X \to \mathbb{R}$ be a continuous and elliptic bilinear form on the reflexive Banach space X. Prove that $a(\cdot, \cdot)$ satisfies the inf-sup condition

$$\tau := \inf_{x \in X \setminus \{0\}} \sup_{y \in X \setminus \{0\}} \frac{a(x,y)}{\|x\|_X \|y\|_X} > 0$$

as well as the non-degeneracy condition

$$\forall y \in X \setminus \{0\} \exists x \in X \quad a(x,y) \neq 0.$$

For each given right-hand side $x^* \in X^*$, the weak form (6.5) thus has a unique solution $x \in X$. □

Another observation is that for reflexive spaces, it is immaterial whether the LBB condition is stated for the first or the second component.

Exercise 56. Let X, Y be reflexive Banach spaces and $a : X \times Y \to \mathbb{R}$ be a continuous bilinear form. Prove that the following statements (i)–(ii) are equivalent:

(i) The bilinear form satisfies the **LBB condition for the first argument**:

- $\alpha_1 := \inf_{x \in X \setminus \{0\}} \sup_{y \in Y \setminus \{0\}} \frac{a(x,y)}{\|x\|_X \|y\|_Y} > 0,$
- $\forall y \in Y \setminus \{0\} \exists x \in X \quad a(x,y) \neq 0.$

(ii) The bilinear form satisfies the **LBB condition for the second argument**:

- $\alpha_2 := \inf_{y \in Y \setminus \{0\}} \sup_{x \in X \setminus \{0\}} \frac{a(x,y)}{\|x\|_X \|y\|_Y} > 0,$
- $\forall x \in X \setminus \{0\} \exists y \in Y \quad a(x,y) \neq 0.$

Moreover, in this case there holds $\alpha_1 = \alpha_2$. □

6.2 Abstract Analysis of Mixed Formulations

Instead of the general mixed formulation (6.6), we consider linear problems with side constraints in the following. These arise, for instance, for the Stokes problem.

Before we focus on the abstract solution theory, we explain why these problems are called *saddle point problems*: Plotting a function $f : \mathbb{R}^2 \to \mathbb{R}$ over the two-dimensional plane, we call a point (x, y) saddle point of f if the real function $f(x + t, y)$ has a minimum at $t = 0$ and the function $f(x, y + t)$ has a maximum for $t = 0$. This is, what is stated in the following proposition for the so-called *Lagrange functional*.

Proposition 6.5.
Let $a : X \times X \to \mathbb{R}$ and $b : X \times Y \to \mathbb{R}$ be bilinear forms on normed spaces X and Y. Assume that $a(\cdot,\cdot)$ is positive semidefinite, i.e., $a(x,x) \geq 0$ and symmetric. Then, given $(x^*, y^*) \in X^* \times Y^*$, $(x,y) \in X \times Y$ is a solution of the saddle point problem

$$\begin{array}{rcl} a(x,\cdot) + b(\cdot,y) & = & x^* \in X^* \\ b(x,\cdot) & = & y^* \in Y^*. \end{array} \tag{6.10}$$

if and only if the Lagrange functional $\mathcal{L}(v,w) := \frac{1}{2}\, a(v,v) - x^*(v) + b(v,w) - y^*(w)$ satisfies

$$\mathcal{L}(x,w) \leq \mathcal{L}(x,y) \leq \mathcal{L}(v,y) \quad \text{for all } (v,w) \in X \times Y, \tag{6.11}$$

i.e., (x,y) is a saddle point of $\mathcal{L}(\cdot,\cdot)$. In this case, the first estimate in (6.11) holds with equality.

Proof. First, assume that $(x,y) \in X \times Y$ is a solution of the saddle point problem (6.10). For $w \in Y$, the second equality in (6.10) implies

$$\mathcal{L}(x,y) - \mathcal{L}(x,w) = b(x, y-w) - y^*(y-w) = 0.$$

This proves the lower estimate of (6.11) even with equality. For $v \in X$, symmetry of $a(\cdot,\cdot)$ and the first equality in (6.10) prove

$$\mathcal{L}(v,y) - \mathcal{L}(x,y) = \frac{1}{2} a(x-v, x-v) + \underbrace{a(x, v-x) - x^*(v-x) + b(v-x, y)}_{=0} \geq 0,$$

and we obtain the upper estimate. Altogether, (x,y) is a saddle point of the Lagrange functional. The proof of the converse implication follows from a classical argument from the calculus of variations: Let $(x,y) \in X \times Y$ satisfy (6.11). For fixed $v \in X$, the real function $f(t) := \mathcal{L}(x+tv, y)$ has a global minimum at $t = 0$. There holds

$$f(t) = \frac{1}{2}\, a(x,x) - x^*(x) + b(x,y) - y^*(y) + \frac{t^2}{2}\, a(v,v) + t\{a(x,v) - x^*(v) + b(v,y)\}.$$

Hence $0 = f'(0) = a(x,v) - x^*(v) + b(v,y)$ for all $v \in X$. This proves the first equality in (6.10). To prove the second equality, consider, for fixed $w \in Y$, the real function $g(t) := \mathcal{L}(x, y+tw)$ which has a global maximum at $t = 0$. There holds

$$g(t) = \frac{1}{2}\, a(x,x) - x^*(x) + b(x,y) - y^*(y) + t\{b(x,w) - y^*(w)\}$$

and thus $0 = g'(0) = b(x,w) - y^*(w)$ for all $w \in Y$, i.e., $b(x,\cdot) = y^* \in Y^*$. ∎

The following theorem of Brezzi provides existence and uniqueness of the solution of saddle point problems.

Theorem 6.6 (Brezzi). *Let X be a Hilbert space and Y be a reflexive Banach space. Let $a : X \times X \to \mathbb{R}$ and $b : X \times Y \to \mathbb{R}$ be continuous bilinear forms. We define $X_0 := \{x \in X \mid b(x,\cdot) = 0 \in Y^*\}$ and assume*

- $\alpha := \inf\limits_{v \in X_0 \setminus \{0\}} \dfrac{a(v,v)}{\|v\|_X^2} > 0$, *i.e., $a(\cdot,\cdot)$ is elliptic on X_0,*
- $\beta := \inf\limits_{y \in Y \setminus \{0\}} \sup\limits_{x \in X \setminus \{0\}} \dfrac{b(x,y)}{\|x\|_X \|y\|_Y} > 0.$

Then, for any $(x^*, y^*) \in X^* \times Y^*$, there is a unique solution $(x,y) \in X \times Y$ of

$$\begin{array}{rcl} a(x,\cdot) + b(\cdot,y) & = & x^* \in X^* \\ b(x,\cdot) & = & y^* \in Y^*. \end{array} \tag{6.12}$$

Moreover, we have the stability estimates

$$\|x\|_X \le \frac{1}{\alpha}\|x^*\|_{X^*} + \frac{1}{\beta}\Big(1 + \frac{\|a\|}{\alpha}\Big)\|y^*\|_{Y^*} \tag{6.13}$$

and

$$\|y\|_Y \le \frac{1}{\beta}\Big(1 + \frac{\|a\|}{\alpha}\Big)\Big(\|x^*\|_{X^*} + \frac{\|a\|}{\beta}\|y^*\|_{Y^*}\Big) \tag{6.14}$$

Remark. (i) Note that one can identify $X^* \times Y^* = (X \times Y)^*$ as follows: For $x^* \in X^*$ and $y^* \in Y^*$, the definition $z^*(x,y) := x^*(x) + y^*(y)$ yields $z^* \in (X \times Y)^*$. Conversely, $z^* \in (X \times Y)^*$ gives rise to $x^*(x) := z^*(x,0)$ and $y^*(y) := z^*(0,y)$ with $(x^*, y^*) \in X^* \times Y^*$.

(ii) If we define operators $A_1 \in L(X, X^*)$, $B_1 \in L(X, Y^*)$, and $B_2 \in L(Y, X^*)$ by

$$A_1 x := a(x,\cdot), \quad B_1 x := b(x,\cdot), \quad \text{and} \quad B_2 y := b(\cdot,y),$$

Equation (6.12) can be written in the form

$$\begin{pmatrix} A_1 & B_2 \\ B_1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x^* \\ y^* \end{pmatrix}. \tag{6.15}$$

In this form, the Brezzi theorem states that this operator matrix is an isomorphism from $X \times Y$ to $X^* \times Y^* = (X \times Y)^*$ and so fits into the abstract framework given above.

(iii) We stress that the original proof of Brezzi works for reflexive Banach spaces X and Y. Therein, it is proved directly that the operator matrix from (6.15) satisfies the inf-sup condition as well as the non-degeneracy condition. Our stronger assumption that X is not only a reflexive Banach space, but even a Hilbert space, reduces the technical difficulties and leads to a much simpler proof. □

Sketch of Proof of Theorem 6.6. Let $(x,y) \in X \times Y$. With the orthogonal decomposition $X = X_0 \oplus X_0^\perp$, we write $x = x_1 + x_2$ with $x_1 \in X_0$ and $x_2 \in X_0^\perp$. Note that (6.12) is equivalent to the following three identities:

- $b(x_2, \cdot) = y^* \in Y^*$,
- $a(x_1, \cdot) = x^* - a(x_2, \cdot) \in X_0^*$,
- $b(\cdot, y) = x^* - a(x_1 + x_2, \cdot) \in X^*$.

For the proof of Theorem 6.6 we are going to show that these three equations — proved in the stated order — admit unique solutions $x_2 \in X_0^\perp$, $x_1 \in X_0$, and $y \in Y^*$. This proves existence and uniqueness of the solution $(x, y) = (x_1 + x_2, y) \in X \times Y$ of (6.12). ■

The main ingredient of the proof of Theorem 6.6 is the closed range theorem:

Theorem 6.7 (Banach's Closed Range Theorem). *For an operator $T \in L(X, Y)$ between Banach spaces X and Y, the following is equivalent:*
(i) $\operatorname{range}(T) \subseteq Y$ *is closed,*
(ii) $\operatorname{range}(T) = (\ker T^*)_\circ = \big\{y \in Y \,\big|\, \forall y^* \in \ker T^* \quad y^*(y) = 0\big\}$,
(iii) $\operatorname{range}(T^*) \subseteq X^*$ *is closed,*
(iv) $\operatorname{range}(T^*) = (\ker T)^\circ = \big\{x^* \in X^* \,\big|\, \forall x \in \ker T \quad x^*(x) = 0\big\}$. ■

Proof of Theorem 6.6. The essential steps of the proof are based on operator arguments for the operators defined by $B_1 x := b(x, \cdot)$ and $B_2 y := b(\cdot, y)$. We are going to consider the four operators

$$\begin{aligned} B_1 &\in L(X, Y^*), & B_1^* &\in L(Y^{**}, X^*), \\ B_2 &\in L(Y, X^*), & B_2^* &\in L(X^{**}, Y^*). \end{aligned}$$

More precisely, the first three steps state the essential observations about these operators, whereas the remaining proof follows the line of the sketch given before.

1. step. B_2 is injective with closed range and $\|B_2^{-1} : \operatorname{range}(B_2) \to Y\| = 1/\beta$, which follows from Lemma 6.1 and

$$\beta = \inf_{y \in Y \setminus \{0\}} \frac{\|B_2 y\|_{X^*}}{\|y\|_Y}.$$

2. step. There holds $B_2 = B_1^* I_Y$, which follows from

$$(B_2 y)(x) = b(x, y) = (B_1 x)(y) = (I_Y y)(B_1 x) = (B_1^* I_Y y)(x) \quad \text{for all } x \in X, y \in Y.$$

3. step. Since Y is reflexive, B_1^* is injective with closed $\operatorname{range}(B_1^*) = \operatorname{range}(B_2)$. Moreover, the closed range theorem even proves

$$\operatorname{range}(B_2) = \operatorname{range}(B_1^*) = (\ker B_1)^\circ = (X_0)^\circ \quad \text{as well as} \quad \operatorname{range}(B_1) = (\ker B_1^*)_\circ = Y^*.$$

4. step. There is a unique $x_2 \in X_0^\perp$ with $b(x_2, \cdot) = y^* \in Y^*$: According to step 3, there is at least one $x \in X$ with $b(x, \cdot) = B_1 x = y^*$. The decomposition $x = x_1 + x_2$ with $x_1 \in X_0$ and $x_2 \in X_0^\perp$ proves $b(x_2, \cdot) = b(x, \cdot) = y^* \in Y^*$, which concludes existence. To prove uniqueness, let $\widetilde{x}_2 \in X_0^\perp$ with $b(\widetilde{x}_2, \cdot) = y^* \in Y^*$. Then, $b(x_2 - \widetilde{x}_2, \cdot) = 0 \in Y^*$, whence $x_2 - \widetilde{x}_2 \in \ker B_1 = X_0$. From $x_2 - \widetilde{x}_2 \in X_0^\perp$, we thus obtain $x_2 = \widetilde{x}_2$.

5. step. There is a unique element $x_1 \in X_0$ with $a(x_1, \cdot) = x^* - a(x_2, \cdot) \in X_0^*$ which immediately follows from the Lax-Milgram lemma and the observation that X_0 is a closed subspace of a Hilbert space and hence a Hilbert space as well.

6. step. There is a unique element $y \in Y$ with $b(\cdot, y) = x^* - a(x, \cdot)$, where $x := x_1 + x_2 \in X$: By construction in step 5, there holds

$$x^* - a(x, \cdot) \in (X_0)^\circ = \big\{ v^* \in X^* \,\big|\, \forall v \in X_0 \quad v^*(v) = 0 \big\}.$$

According to step 1 and step 3, B_2 is injective with $\operatorname{range}(B_2) = (X_0)^\circ$. Thus, there is a unique $y \in Y$ with $b(\cdot, y) = B_2 y = x^* - a(x, \cdot)$.

7. step. There holds $\|x_2\|_X \leq \|y^*\|_{Y^*}/\beta$: From $x_2 \in X_0^\perp$ follows $(x_2 \,;\, \cdot)_X \in (X_0)^\circ = \operatorname{range}(B_2)$. Thus, we may choose $\widetilde{y} \in Y$ with $B_2 \widetilde{y} = (x_2 \,;\, \cdot)_X$. From $\|B_2^{-1} : (X_0)^\circ \to Y\| = 1/\beta$, we infer $\|\widetilde{y}\|_Y \leq \|(x_2 \,;\, \cdot)_X\|_{X^*}/\beta = \|x_2\|_X/\beta$. Together with $b(x_2, \cdot) = y^*$, we conclude

$$\|x_2\|_X^2 = (x_2 \,;\, x_2)_X = (B_2 \widetilde{y})(x_2) = b(x_2, \widetilde{y}) = y^*(\widetilde{y}) \leq \|y^*\|_{Y^*} \|\widetilde{y}\|_Y \leq \frac{\|y^*\|_{Y^*}}{\beta} \, \|x_2\|_X.$$

8. step. There holds $\|x_1\|_X \leq \alpha^{-1} \big(\|x^*\|_{X^*} + \|a\| \|x_2\|_X \big)$: Note that $A_1 \in L(X_0, X_0^*)$ is an isomorphism with $\|A_1^{-1} : X_0^* \to X_0\| \leq 1/\alpha$. From $A_1 x_1 = a(x_1, \cdot) = x^* - a(x_2, \cdot)$, we thus infer

$$\|x_1\|_X \leq \frac{1}{\alpha} \, \|x^* - a(x_2, \cdot)\|_{X_0^*} \leq \frac{1}{\alpha} \big(\|x^*\|_{X^*} + \|a\| \|x_2\|_X \big).$$

9. step. The triangle inequality leads to

$$\|x\|_X \leq \|x_1\|_X + \|x_2\|_X \leq \frac{1}{\alpha} \, \|x^*\|_{X^*} + \Big(\frac{\|a\|}{\alpha} + 1 \Big) \, \|x_2\|_X \leq \frac{1}{\alpha} \, \|x^*\|_{X^*} + \frac{1}{\beta} \Big(\frac{\|a\|}{\alpha} + 1 \Big) \, \|y^*\|_{Y^*}.$$

10. step. It finally remains to dominate $\|y\|_Y$, where $B_2 y = b(\cdot, y) = x^* - a(x, \cdot) \in (X_0)^\circ$. We use $\|B_2^{-1} : (X_0)^\circ \to Y\| = 1/\beta$ to see

$$\begin{aligned} \|y\|_Y \leq \frac{1}{\beta} \, \|x^* - a(x, \cdot)\|_{X^*} &\leq \frac{1}{\beta} \, \|x^*\|_{X^*} + \frac{\|a\|}{\beta} \, \|x\|_X \\ &\leq \frac{1}{\beta} \, \|x^*\|_{X^*} + \frac{\|a\|}{\beta} \, \frac{1}{\alpha} \, \|x^*\|_{X^*} + \frac{\|a\|}{\beta^2} \Big(1 + \frac{\|a\|}{\alpha} \Big) \, \|y^*\|_{Y^*} \\ &= \frac{1}{\beta} \Big(1 + \frac{\|a\|}{\alpha} \Big) \Big(\|x^*\|_{X^*} + \frac{\|a\|}{\beta} \, \|y^*\|_{Y^*} \Big). \end{aligned}$$

This concludes the proof. ∎

Remark. (i) Let $B_1 \in L(X, Y^*)$ and $B_2 \in L(Y, X^*)$ be defined as in the proof of Theorem 6.6. In the proof, we have seen that $\beta > 0$ implies surjectivity of B_1. We note that even the converse implication holds, i.e.,

$$\beta := \inf_{y \in Y \setminus \{0\}} \sup_{x \in X \setminus \{0\}} \frac{b(x, y)}{\|x\|_X \|y\|_Y} > 0 \quad \Longleftrightarrow \quad B_1 \text{ is surjective.} \tag{6.16}$$

Suppose that B_1 is surjective. As in step 3 of the preceding proof, the closed range theorem proves that B_1^* is injective with closed range. Moreover, $B_2 = B_1^* I_Y$ proves that B_2 is injective with closed

$\operatorname{range}(B_2) = \operatorname{range}(B_1^*) = (\ker B_1)^\circ = (X_0)^\circ$, i.e., $B_2 : Y \to \operatorname{range}(B_2)$ is continuous and bijective between the Banach spaces Y and $\operatorname{range}(B_2) \subseteq X^*$. According to the open mapping theorem, $B_2 : Y \to \operatorname{range}(B_2)$ even is an isomorphism, i.e., $\beta^{-1} = \|B_2 : Y \to \operatorname{range}(B_2)\| < \infty$, whence $\beta > 0$.

(ii) Altogether, the two main assumptions on $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ can equivalently be stated as follows:

- The bilinear form $a(\cdot,\cdot)$ is elliptic on $X_0 = \ker B_1$.
- The operator $B_1 \in L(X, Y^*)$ is surjective.

We hope that the reader may keep this (abstract) formulation in mind much easier. For the statement of Theorem 6.6, we used the definition of α and β instead, to provide the stability estimates (6.13)–(6.14) with explicit constants. □

Going through the proof of Theorem 6.6, one realizes that ellipticity of $a(\cdot,\cdot)$ on X_0 is only used to provide a unique $x_1 \in X_0$ with $a(x_1,\cdot) = x_0^* \in X_0^*$ in step 5. To prove unique existence of x_1, it is, however, sufficient to assume that the operator $A_1 : X_0 \to X_0^*$ defined by $A_1 x := a(x,\cdot)$ is an isomorphism. This is done in the following exercise.

Exercise 57. Let X, Y, $a(\cdot,\cdot)$, and $b(\cdot,\cdot)$ be as in Theorem 6.6. Then, the following statements are equivalent:

(i) For all $(x^*, y^*) \in X^* \times Y^*$, there exists a unique solution $(x, y) \in X \times Y$ of the saddle point problem (6.12).

(ii) The bilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ satisfy the following three assumptions:

- $\alpha := \inf\limits_{v \in X_0\setminus\{0\}} \sup\limits_{w \in X_0\setminus\{0\}} \dfrac{a(v,w)}{\|v\|_X \|w\|_X} > 0,$
- $\forall w \in X_0\setminus\{0\} \exists v \in X_0 \quad a(v,w) \neq 0,$
- $\beta := \inf\limits_{y \in Y\setminus\{0\}} \sup\limits_{x \in X\setminus\{0\}} \dfrac{b(x,y)}{\|x\|_X \|y\|_Y} > 0.$

The first two assumptions state that $A_1 : X_0 \to X_0^*$ is an isomorphism, cf. Theorem 6.2. The assumption on β is the same as in the above statement of the Brezzi theorem. □

The following corollary provides the relation between saddle point problems and the abstract Petrov-Galerkin scheme from Section 6.1.

Corollary 6.8. *Suppose that X is a Hilbert space, Y is a reflexive Banach space, and $a : X \times X \to \mathbb{R}$ and $b : X \times Y \to \mathbb{R}$ are continuous bilinear forms. Then, $Z := X \times Y$ is a reflexive Banach space, and $c((x,y),(\widetilde{x},\widetilde{y})) := a(x,\widetilde{x}) + b(\widetilde{x},y) + b(x,\widetilde{y})$ defines a continuous bilinear form $c : Z \times Z \to \mathbb{R}$. Moreover, for $(x,y) \in X \times Y$ and $(x^*,y^*) \in X^* \times Y^*$, the saddle point problem (6.12) is equivalent to*

$$c((x,y),(\widetilde{x},\widetilde{y})) = x^*(\widetilde{x}) + y^*(\widetilde{y}) \quad \text{for all } (\widetilde{x},\widetilde{y}) \in X \times Y. \tag{6.17}$$

Finally, the following three statements are equivalent:

(i) *$a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ satisfy the assumptions of the Brezzi theorem, i.e.,*

- $\alpha := \inf\limits_{v\in X_0\setminus\{0\}} \sup\limits_{w\in X_0\setminus\{0\}} \dfrac{a(v,w)}{\|v\|_X\|w\|_X} > 0$ *with* $X_0 := \big\{x\in X \,\big|\, b(x,\cdot)=0\in Y^*\big\}$,
- $\forall w\in X_0\setminus\{0\}\exists v\in X_0 \quad a(v,w)\neq 0$,
- $\beta := \inf\limits_{y\in Y\setminus\{0\}} \sup\limits_{x\in X\setminus\{0\}} \dfrac{b(x,y)}{\|x\|_X\|y\|_Y} > 0$.

(ii) *$c(\cdot,\cdot)$ satisfies the LBB conditions*

- $\gamma := \inf\limits_{z\in Z\setminus\{0\}} \sup\limits_{w\in Z\setminus\{0\}} \dfrac{c(z,w)}{\|z\|_Z\|w\|_Z} > 0$,
- $\forall w\in Z\setminus\{0\}\exists z\in Z \quad c(z,w)\neq 0$.

(iii) *For all $(x^*,y^*)\in X^*\times Y$, the variational formulation (6.17) has a unique solution $(x,y)\in X\times Y$.*

In particular, it holds $\|c\|\le\|a\|+2\|b\|$ for the corresponding norms and there exists a constant $C>0$ such that

$$\gamma \ge C\Big[\frac{1}{\alpha}+\frac{1}{\beta}\Big(1+\frac{\|a\|}{\alpha}\Big)\Big(1+\frac{\|a\|}{\beta}\Big)\Big]^{-1}. \tag{6.18}$$

Proof. **1. step.** Since X and Y are reflexive, their closed unit balls $B_X\subset X$ and $B_Y\subset Y$ are weakly compact. According to the Tychonov theorem, $B_X\times B_Y$ and hence B_Z are weakly compact as well. Consequently, Z is reflexive. Moreover, it is obvious that $c(\cdot,\cdot)$ is bilinear and continuous with $\|c\|\le\|a\|+2\|b\|$.

2. step. Summing the equations of (6.12), we obtain the variational form (6.17). Testing (6.17) with test functions of the type $(\widetilde{x},0)$ or $(0,\widetilde{y})$, we see that (6.12) and (6.17) are, in fact, equivalent.

3. step. The equivalence of (ii) and (iii) is stated in Corollary 6.3. The equivalence of (i) and (iii) follows from step 2 and Exercise 57.

4. step. It remains to prove (6.18): From (6.13)–(6.14), we obtain

$$\begin{aligned}
\|x\|_X+\|y\|_Y &\le \frac{1}{\alpha}\|x^*\|_{X^*}+\frac{1}{\beta}\Big(1+\frac{\|a\|}{\alpha}\Big)\|y^*\|_{Y^*}+\frac{1}{\beta}\Big(1+\frac{\|a\|}{\alpha}\Big)\Big(\|x^*\|_{X^*}+\frac{\|a\|}{\beta}\|y^*\|_{Y^*}\Big)\\
&= \Big[\frac{1}{\alpha}+\frac{1}{\beta}\Big(1+\frac{\|a\|}{\alpha}\Big)\Big]\|x^*\|_{X^*}+\frac{1}{\beta}\Big(1+\frac{\|a\|}{\alpha}\Big)\Big(1+\frac{\|a\|}{\beta}\Big)\|y^*\|_{Y^*}\\
&\le \Big[\frac{1}{\alpha}+\frac{1}{\beta}\Big(1+\frac{\|a\|}{\alpha}\Big)\Big(1+\frac{\|a\|}{\beta}\Big)\Big]\big[\|x^*\|_{X^*}+\|y^*\|_{Y^*}\big].
\end{aligned}$$

With the operator $Tz := c(z,\cdot)$, this proves that the solution operator $T^{-1}: X^*\times Y^*\to X\times Y$ has operator norm $\|T^{-1}\|\le C\left[\frac{1}{\alpha}+\frac{1}{\beta}\left(1+\frac{\|a\|}{\alpha}\right)\frac{1}{\beta}\left(1+\frac{\|a\|}{\beta}\right)\right]$, where $C>0$ depends only on the norms chosen on $Z=X\times Y$ and $Z^*=X^*\times Y^*$. According to Theorem 6.2, it holds $\|T^{-1}\|=1/\gamma$. This concludes the proof. ∎

Exercise 58. Give a direct proof that $c(\cdot,\cdot)$ from Corollary 6.8 satisfies the LBB condition, i.e., prove directly that (i) implies (ii). **Hint.** For $(x,y) \neq 0$ use the orthogonal decomposition $x = x_1 + x_2 \in X_0 + X_0^\perp$ and estimate $\|x_1\|_X$, $\|x_2\|_X$, and $\|y\|_Y$ separately. □

Corollary 6.8 together with Exercise 54 provides a solvability theory and the Céa lemma for Galerkin discretizations of saddle point problems.

Corollary 6.9 (Céa Lemma for Saddle Point Problems, Version I). *Let $a : X \times X \to \mathbb{R}$ and $b : X \times Y \to \mathbb{R}$ be continuous bilinear forms on a Hilbert space X and a reflexive Banach space Y. Given $(x^*, y^*) \in X^* \times Y^*$, let $(x,y) \in X \times Y$ be a solution of the saddle point problem (6.12). Let $X_h \subset X$ and $Y_h \subset Y$ be finite dimensional subspaces and define $X_{0h} := \{x_h \in X_h \,|\, b(x_h,\cdot) = 0 \in Y_h^*\}$. Suppose that*

- $\alpha_h := \inf\limits_{v_h \in X_{0h}\setminus\{0\}} \sup\limits_{w_h \in X_{0h}\setminus\{0\}} \dfrac{a(v_h,w_h)}{\|v_h\|_X \|w_h\|_X} > 0,$
- $\beta_h := \inf\limits_{y_h \in Y_h\setminus\{0\}} \sup\limits_{x_h \in X_h\setminus\{0\}} \dfrac{b(x_h,y_h)}{\|x_h\|_X \|y_h\|_Y} > 0.$

Then, there is a unique solution $(x_h, y_h) \in X_h \times Y_h$ of the discrete saddle point problem

$$\begin{array}{lcll} a(x_h,\cdot) & + & b(\cdot,y_h) & = & x^* \in X_h^*, \\ b(x_h,\cdot) & & & = & y^* \in Y_h^*, \end{array} \tag{6.19}$$

and there holds

$$\|x - x_h\|_X + \|y - y_h\|_Y \leq C \left(\min_{\widetilde{x}_h \in X_h} \|x - \widetilde{x}_h\|_X + \min_{\widetilde{y}_h \in Y_h} \|y - \widetilde{y}_h\|_Y \right)$$

The constant $C > 0$ depends only on $(\|a\|+\|b\|)/\gamma_h$ with $\gamma_h := \left[\frac{1}{\alpha_h} + \frac{1}{\beta_h}\left(1+\frac{\|a\|}{\alpha_h}\right)\frac{1}{\beta_h}\left(1+\frac{\|a\|}{\beta_h}\right)\right]$.

Proof. The existence and uniqueness of (x_h, y_h) follows from the abstract Brezzi theorem; see Corollary 6.8. For Petrov-Galerkin schemes, the constant in the Céa lemma depends only on the quotient of the continuity bound and the discrete inf-sup constant; see Exercise 54. Both constants have been estimated in Corollary 6.8. ■

Remark. The Galerkin discretization of saddle point problems is structurally much more difficult than for problems of the Lax-Milgram lemma:

(i) Note that $X_{0h} \not\subseteq X_0 := \{v \in X \,|\, b(v,\cdot) = 0 \in Y^*\}$. There may be even no relation between X_0 and X_{0h} besides the trivial $X_0 \cap X_h \subseteq X_{0h}$. In particular, there is no relation between α and α_h even if $a(\cdot,\cdot)$ is elliptic on X_0.

(ii) However, if $a(\cdot,\cdot)$ is already elliptic on X, i.e., $\tau := \inf_{x \in X\setminus\{0\}} \frac{a(x,x)}{\|x\|_X^2} > 0$ this implies $\alpha \geq \tau$ and $\alpha_h \geq \tau$ for the continuous and discrete inf-sup constant of $a(\cdot,\cdot)$.

(iii) Moreover, $\beta > 0$ from the continuous formulation does not imply $\beta_h > 0$ for the discrete formulation. Below, we introduce Fortin's criterium which provides some help on this matter.

(iv) Finally, we recall that $\beta_h > 0$ implies necessarily $\dim Y_h \leq \dim X_h$; see Proposition 6.4. □

Exercise 59. For a matrix $A \in \mathbb{R}^{m\times n}$ holds $\ker(A^T) = (\text{range}A)^\perp$ as well as $\text{range}(A^T) = (\ker A)^\perp$, where $(\cdot)^\perp$ denotes the orthogonal complement with respect to the usual Euclidean product in $\mathbb{R}^m$ resp. $\mathbb{R}^n$. □

The following two exercises consider the discretization of the mixed problem (6.12). We stress that a linear system similar to the one here, also appeared for the discretization of the Neumann problem, where we had to realize the linear side constraint $\int_\Omega u_h\, dx = 0$.

Exercise 60. Let $a : X \times X \to \mathbb{R}$ and $b : X \times Y \to \mathbb{R}$ be continuous bilinear forms on a Hilbert space X and a reflexive Banach space Y. We replace X and Y by finite dimensional subspaces X_h and Y_h, respectively. Show that the computation of a discrete solution $(x_h, y_h) \in X_h \times Y_h$ of

$$\begin{array}{rcl} a(x_h,\cdot) + b(\cdot,y_h) &=& x^* \in X_h^*, \\ b(x_h,\cdot) &=& y^* \in Y_h^*, \end{array} \tag{6.20}$$

is equivalent to the solution of a linear system with a matrix of the type $M := \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}$. □

Exercise 61. Let $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{m\times n}$, and $M := \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}$. Assume that A is positive definite on the kernel of B. Prove that M is regular if and only if $\text{range}(B) = \mathbb{R}^m$. □

We conclude this section with an improved Céa lemma for saddle point problems; cf. Corollary 6.9.

Theorem 6.10 (Céa Lemma for Saddle Point Problems, Version II). *Let $a : X \times X \to \mathbb{R}$ and $b : X \times Y \to \mathbb{R}$ be continuous bilinear forms on a Hilbert space X and a reflexive Banach space Y. Given $(x^*, y^*) \in X^* \times Y^*$, let $(x, y) \in X \times Y$ be a solution of the saddle point problem (6.12). Let $X_h \subset X$ and $Y_h \subset Y$ be finite dimensional subspaces and define $X_{0h} := \{x_h \in X_h \,|\, b(x_h,\cdot) = 0 \in Y_h^*\}$. Suppose that*

- $\alpha_h := \inf\limits_{v_h \in X_{0h}\setminus\{0\}} \dfrac{a(v_h,v_h)}{\|v_h\|_X^2} > 0,$
- $\beta_h := \inf\limits_{y_h \in Y_h\setminus\{0\}} \sup\limits_{x_h \in X_h\setminus\{0\}} \dfrac{b(x_h,y_h)}{\|x_h\|_X \|y_h\|_Y} > 0.$

Then, there is a unique solution $(x_h, y_h) \in X_h \times Y_h$ of the discrete saddle point problem

$$\begin{array}{rcl} a(x_h,\cdot) + b(\cdot,y_h) &=& x^* \in X_h^*, \\ b(x_h,\cdot) &=& y^* \in Y_h^*, \end{array} \tag{6.21}$$

and there holds

$$\|x - x_h\|_X \leq \Big(1 + \frac{\|a\|}{\alpha_h}\Big)\Big(1 + \frac{\|b\|}{\beta_h}\Big) \min_{\widetilde{x}_h \in X_h} \|x - \widetilde{x}_h\|_X + \frac{\|b\|}{\alpha_h} \min_{\widetilde{y}_h \in Y_h} \|y - \widetilde{y}_h\|_Y \tag{6.22}$$

and

$$\|y - y_h\|_Y \leq \Big(1 + \frac{\|b\|}{\beta_h}\Big) \min_{\widetilde{y}_h \in Y_h} \|y - \widetilde{y}_h\|_Y + \frac{\|a\|}{\beta_h} \|x - x_h\|_X. \tag{6.23}$$

Sketch of Proof of Theorem 6.10. The unique existence of a discrete solution $(x_h, y_h) \in X_h \times Y_h$ follows from the Brezzi Theorem 6.6 applied for $X_h \times Y_h$. The quasioptimality is proven in three steps:

- First, we prove estimate (6.23).
- Second, we prove quasioptimality of $\|x - x_h\|_X$ with respect to the affine space $Z_h := \big\{\widetilde{x}_h \in X_h \,\big|\, b(\widetilde{x}_h, \cdot) = y^* \in Y_h^*\big\}$.
- In a final step, we estimate the bestapproximation error with respect to Z_h by the bestapproximation error with respect to the entire discrete space X_h which then leads to (6.22).

This general concept even works for nonlinear problems with linear side constraint. ■

Proof. We first note the Galerkin orthogonality, which now reads

$$\begin{array}{lcll} a(x - x_h, \cdot) & + & b(\cdot, y - y_h) & = & 0 \in X_h^*, \\ b(x - x_h, \cdot) & & & = & 0 \in Y_h^*, \end{array} \tag{6.24}$$

1. step. There holds

$$\|y - y_h\|_Y \leq \Big(1 + \frac{\|b\|}{\beta_h}\Big) \|y - \widetilde{y}_h\|_Y + \frac{\|a\|}{\beta_h} \|x - x_h\|_X \quad \text{for all } \widetilde{y}_h \in Y_h :$$

According to the definition of β_h, there holds

$$\beta_h \|\widetilde{y}_h - y_h\|_Y \leq \sup_{\widetilde{x}_h \in X_h \setminus \{0\}} \frac{b(\widetilde{x}_h, \widetilde{y}_h - y_h)}{\|x_h\|_X}.$$

With the Galerkin orthogonality, the nominator may be written as

$$\begin{aligned} b(\widetilde{x}_h, \widetilde{y}_h - y_h) &= -\big(a(x - x_h, \widetilde{x}_h) + b(\widetilde{x}_h, y - y_h)\big) + b(\widetilde{x}_h, \widetilde{y}_h - y_h) \\ &= -a(x - x_h, \widetilde{x}_h) + b(\widetilde{x}_h, \widetilde{y}_h - y) \end{aligned}$$

Therefore, continuity of $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ lead to

$$\beta_h \|\widetilde{y}_h - y_h\|_Y \leq \|a\| \|x - x_h\|_X + \|b\| \|\widetilde{y}_h - y\|_Y.$$

Altogether, a triangle inequality $\|y - y_h\|_Y \leq \|y - \widetilde{y}_h\|_Y + \|\widetilde{y}_h - y_h\|_Y$ yields step 1.

2. step. With the affine space $Z_h := \{\widetilde{x}_h \in X_h \,|\, b(\widetilde{x}_h, \cdot) = y^* \in Y_h^*\}$, there holds

$$\|x - x_h\|_X \le \Big(1 + \frac{\|a\|}{\alpha_h}\Big) \|x - z_h\|_X + \frac{\|b\|}{\alpha_h} \|y - \widetilde{y}_h\|_Y \quad \text{for all } z_h \in Z_h \text{ and } \widetilde{y}_h \in Y_h :$$

Since $x_h, z_h \in Z_h$, there holds $x_h - z_h \in X_{0h}$. According to the definition of α_h, we see

$$\alpha_h \|x_h - z_h\|_X^2 \le a(x_h - z_h, x_h - z_h) = a(x_h - x, x_h - z_h) + a(x - z_h, x_h - z_h).$$

For the first term, the Galerkin orthogonality implies

$$a(x_h - x, x_h - z_h) = b(x_h - z_h, y - y_h) = b(x_h - z_h, \widetilde{y}_h - y_h) + b(x_h - z_h, y - \widetilde{y}_h),$$

where the first summand $b(x_h - z_h, \widetilde{y}_h - y_h) = 0$ drops out by use of $x_h - z_h \in X_{0h}$. By continuity of $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$, we see

$$\alpha_h \|x_h - z_h\|_X \le \|a\| \|x - z_h\|_X + \|b\| \|y - \widetilde{y}_h\|_Y.$$

Again, a triangle inequality $\|x - x_h\|_X \le \|x - z_h\|_X + \|x_h - z_h\|_X$ yields step 2.

3. step. There holds

$$\|x - z_h\|_X \le \Big(1 + \frac{\|b\|}{\beta_h}\Big) \|x - \widetilde{x}_h\|_X \quad \text{for all } \widetilde{x}_h \in X_h \text{ and some } z_h \in Z_h \text{ depending on } \widetilde{x}_h :$$

We define $W_h := (X_{0h})^\perp \subseteq X_h$ and consider the operators $B_1 \in L(W_h, Y_h^*)$ and $B_2 \in L(Y_h, W_h^*)$ defined by $B_1 w_h := b(w_h, \cdot)$ and $B_2 y_h := b(\cdot, y_h)$. Note that

$$0 < \beta_h = \inf_{\widetilde{y}_h \in Y_h \setminus \{0\}} \sup_{\widetilde{x}_h \in X_h \setminus \{0\}} \frac{b(\widetilde{x}_h, \widetilde{y}_h)}{\|\widetilde{x}_h\|_X \|\widetilde{y}_h\|_Y} = \inf_{\widetilde{y}_h \in Y_h \setminus \{0\}} \sup_{w_h \in W_h \setminus \{0\}} \frac{b(w_h, \widetilde{y}_h)}{\|w_h\|_X \|\widetilde{y}_h\|_Y}.$$

According to Lemma 6.1, the operator B_2 is injective with closed range and $1/\beta_h = \|B_2^{-1} : \operatorname{range}(B_2) \to Y_h\|$. From this, we derive that $B_1 = B_2^* \circ I_{Y_h}$ is surjective due to $\operatorname{range}(B_1) = \operatorname{range}(B_2^*) = (\ker B_2)^\circ = Y_h^*$. Note that by definition of $W_h := (X_{0h})^\perp \subseteq X_h$, the operator B_1 is injective and thus an isomorphism between W_h and Y_h^*. In particular, this yields bijectivity of B_2 as well as

$$\|B_1^{-1}\| = \|I_{Y_h}^{-1} (B_2^*)^{-1}\| = \|(B_2^{-1})^*\| = \|B_2^{-1}\| = 1/\beta_h.$$

In particular, there is a unique element $w_h \in W_h$ with $b(w_h, \cdot) = B_1 w_h = b(x - \widetilde{x}_h, \cdot) \in Y_h^*$ and there holds $\|w_h\|_X \le \beta_h^{-1} \|b(x - \widetilde{x}_h, \cdot)\|_{X^*} \le (\|b\|/\beta_h) \|x - \widetilde{x}_h\|_X$. The element $z_h := \widetilde{x}_h + w_h \in X_h$ satisfies $b(z_h, \cdot) = b(x, \cdot) = y^* \in Y_h^*$ and thus $z_h \in Z_h$. Now, we finally see

$$\|x - z_h\|_X \le \|x - \widetilde{x}_h\|_X + \|w_h\|_X \le \Big(1 + \frac{\|b\|}{\beta_h}\Big) \|x - \widetilde{x}_h\|_X.$$

This concludes step 3.

4. step. The proof of (6.23) follows by finite dimension: Note that step 1 implies

$$\|y - y_h\|_Y \le \Big(1 + \frac{\|b\|}{\beta_h}\Big) \inf_{\widetilde{y}_h \in Y_h} \|y - \widetilde{y}_h\|_Y + \frac{\|a\|}{\beta_h} \|x - x_h\|_X,$$

and it only remains to see that the infimum is, in fact, attained: To that end, choose an infimizing sequence (y_k) in Y_h, i.e.

$$\lim_{k\to\infty} \|y - y_k\|_Y = \inf_{\widetilde{y}_h \in Y_h} \|y - \widetilde{y}_h\|_Y.$$

According to the triangle inequality, there holds $\|y_k\|_Y \leq \|y\|_Y + \|y - y_k\|_Y$, i.e. the sequence (y_k) is a bounded sequence in the finite dimensional space Y_h. Thus, the Bolzano-Weierstrass theorem yields the existence of a convergent subsequence (y_{k_ℓ}) with limit $y_0 \in Y_h$. By continuity, we conclude

$$\inf_{\widetilde{y}_h \in Y_h} \|y - \widetilde{y}_h\|_Y = \lim_{\ell\to\infty} \|y - y_{k_\ell}\|_Y = \|y - y_0\|_Y.$$

5. step. The proof of (6.22) now follows from a combination of step 2 and step 3: For arbitrary $\widetilde{x}_h \in X_h$, choose $z_h \in Z_h$ by use of step 3. Let $\widetilde{y}_h \in Y_h$ and be arbitrary. We then infer

$$\begin{aligned}\|x - x_h\|_X &\leq \Big(1 + \frac{\|a\|}{\alpha_h}\Big) \|x - z_h\|_X + \frac{\|b\|}{\alpha_h} \|y - \widetilde{y}_h\|_Y \\ &\leq \Big(1 + \frac{\|a\|}{\alpha_h}\Big)\Big(1 + \frac{\|b\|}{\beta_h}\Big) \|x - \widetilde{x}_h\|_X + \frac{\|b\|}{\alpha_h} \|y - \widetilde{y}_h\|_Y.\end{aligned}$$

Now, we take the infimum over $\widetilde{x}_h$ and $\widetilde{y}_h$ and note that, according to finite dimension, this infimum is attained by independent minima. ∎

6.2.1 Discrete inf-sup conditions

Often, the continuous inf-sup condition is not that hard to prove, but the discrete one is the problem. The next two lemmata provide a tool to derive the discrete inf-sup condition from the continuous condition.

Lemma 6.11 (M. Fortin). *Let $b : X \times Y \to \mathbb{R}$ denote a continuous bilinearform which satisfies the continuous inf-sup condition*

$$\inf_{0\neq\lambda\in Y} \sup_{0\neq u\in X} \frac{b(u,\lambda)}{\|u\|_X \|\lambda\|_Y} \geq \gamma > 0. \tag{6.25}$$

Let $X_h \subset X$ and $Y_h \subset Y$ denote closed subspaces and let $\Pi : X \to X_h$ denote a linear mapping with

$$\begin{aligned} b(u - \Pi u, \lambda) &= 0 \qquad \forall \lambda \in Y_h && (6.26)\\ \|\Pi u\|_X &\leq C_\Pi \|u\|_X \qquad \forall u \in X. && (6.27)\end{aligned}$$

Then, there holds

$$\inf_{0\neq\lambda\in Y_h} \sup_{0\neq u\in X_h} \frac{b(u,\lambda)}{\|u\|_X \|\lambda\|_Y} \geq \gamma_N := \frac{\gamma}{C_\Pi} > 0.$$

Proof. Let $\lambda \in Y_h$ and note

$$\begin{aligned}\gamma\|\lambda\|_Y &\overset{(6.25)}{\leq} \sup_{0\neq v\in X} \frac{b(v,\lambda)}{\|v\|_X} \overset{(6.26)}{=} \sup_{0\neq v\in X} \frac{b(\Pi v,\lambda)}{\|v\|_X} \\ &\overset{(6.27)}{\leq} C_\Pi \sup_{0\neq v\in X} \frac{b(\Pi v,\lambda)}{\|\Pi v\|_X} = C_\Pi \sup_{0\neq v\in \mathrm{range}\Pi} \frac{b(v,\lambda)}{\|v\|_X} \leq C_\Pi \sup_{0\neq v\in X_h} \frac{b(v,\lambda)}{\|v\|_X}\end{aligned}$$

■

Often, it is easier to generate the operator Π in two steps, as done in the following lemma.

Lemma 6.12. *Let $\Pi_i : X \to X_h$, $i = 1, 2$ denote linear mappings with*

$$\begin{aligned}\|\Pi_1 u\|_X &\leq C_1\|u\|_X \qquad \forall u \in X \\ \|\Pi_2(\boldsymbol{I} - \Pi_1)u\|_X &\leq C_2\|u\|_X \qquad \forall u \in X \\ b(u - \Pi_2 u, \lambda) &= 0 \qquad \forall \lambda \in Y_h.\end{aligned}$$

Then, (6.25) implies the discrete inf-sup condition

$$\inf_{0\neq\lambda\in Y_h} \sup_{0\neq u\in X_h} \frac{b(u,\lambda)}{\|u\|_X\|\lambda\|_Y} \geq \frac{\gamma}{C_1 + C_2}.$$

Proof. Let $\lambda \in Y_h$ and define $\Pi : X \to X_h$ via $\Pi u := \Pi_2(\boldsymbol{I} - \Pi_1)u + \Pi_1 u$. Then, we have

$$b(\Pi u, \lambda) = b(\Pi_2(u - \Pi_1 u), \lambda) + b(\Pi_1 u, \lambda) = b(u - \Pi_1 u, \lambda) + b(\Pi_1 u, \lambda) = b(u, \lambda).$$

Moreover, there holds

$$\|\Pi u\|_X \leq \|\Pi_2(\boldsymbol{I} - \Pi_1)u\|_X + \|\Pi_1 u\|_X \leq (C_1 + C_2)\|u\|_X.$$

This concludes the proof. ■

6.3 The Stokes problem

6.3.1 Setting

We apply the general theory of saddle point-problems from the previous section to the Stokes problem: Let $\Omega \subset \mathbb{R}^2$ be a Lipschitz domain. Find $u = (u_1, u_2) \in H^1_0(\Omega) \times H^1_0(\Omega)$ and $p \in L^2(\Omega)$ such that

$$-\Delta u + \nabla p = f \qquad \text{in } \Omega \tag{6.28a}$$

$$\nabla \cdot u = 0 \qquad \text{in } \Omega \tag{6.28b}$$

for given $f = (f_1, f_2)^\top \in L^2(\Omega) \times L^2(\Omega)$. Here, the operator $-\Delta$ is understood component wise, i.e. $\Delta u = (\Delta u_1, \Delta u_2)^\top$.

Remark. From a physical perspective, u denotes the velocity and p the pressure of a fluid in a case where an equilibrium has been reached and the quantities do not depend on time anymore. The

incompressibility condition $\nabla \cdot u = 0$ implies that the fluid can not be compressed (e.g. water). The equation $-\Delta u + \nabla p = f$ describes conservation of momentum. The stationary Stokes problem (6.28) stems from a severe simplification of the Navier-Stokes Equations and are physically meaningful only in slow flowing fluids with high viscosity, e.g., honey. □

A weak form can be formulated as

$$\int_\Omega \nabla u : \nabla v - \int_\Omega p \nabla \cdot v = \int_\Omega f v \qquad \forall v \in (H_0^1(\Omega))^2 \tag{6.29a}$$

$$-\int_\Omega q \nabla \cdot u = 0 \qquad \forall q \in L^2(\Omega). \tag{6.29b}$$

Obviously, the pressure is unique only up to an additive constant and hence one usually chooses to satisfy $\int_\Omega p = 0$. This motivates the choice of space

$$L^2_\star(\Omega) := \{p \in L^2(\Omega) \mid \int_\Omega p = 0\}. \tag{6.30}$$

With this side-constraint, (6.29) is equivalent to the problem: Find $(u, p) \in (H_0^1(\Omega))^2 \times L^2_\star(\Omega)$ such that

$$a(u, v) + b(v, p) = x^\star(v) \qquad \forall v \in (H_0^1(\Omega))^2 \tag{6.31a}$$

$$b(u, q) = 0 \qquad \forall q \in L^2_\star(\Omega), \tag{6.31b}$$

where

$$a(u, v) = \int_\Omega \nabla u : \nabla v \tag{6.32a}$$

$$b(v, p) = -\int_\Omega p \nabla \cdot v \tag{6.32b}$$

Existence of a unique solution for the Stokes problem results from Theorem 6.6 together with the following theorem. (Note that the bilinearform $a(u, v)$ satisfies $a(u, u) = \|\nabla u\|^2_{L^2(\Omega)}$ and is hence elliptic.)

Theorem 6.13 (deRham). *Let Ω be a Lipschitz domain and recall the bilinearform $b(\cdot, \cdot)$ from (6.32). Then, there exists $\gamma > 0$ such that*

$$\inf_{0 \neq p \in L^2_\star(\Omega)} \sup_{0 \neq u \in (H_0^1(\Omega))^2} \frac{|b(v, u)|}{\|p\|_{L^2(\Omega)} \|v\|_{H^1(\Omega)}} \geq \gamma > 0.$$

6.3.2 FEM for Stokes

The finite element method corresponding to (6.31) reads: For $X_h \subset (H_0^1(\Omega))^2$ and $Y_h \subset L^2_\star(\Omega)$ find $(u_h, p_h) \in X_h \times Y_h$ such that

$$a(u_h, v) + b(v, p_h) = x^\star(v) \qquad \forall v \in X_h \tag{6.33a}$$

$$b(u_h, q) = 0 \qquad \forall q \in Y_h. \tag{6.33b}$$

From the abstract theory of saddle-point problems (particularly Theorem 6.6) we know that the discrete spaces also need to satisfy an inf-sup condition, i.e.

$$\inf_{0\neq p\in Y_h}\sup_{0\neq v\in X_h}\frac{b(v,p)}{\|p\|_{L^2(\Omega)}\|v\|_{H^1(\Omega)}}\geq\gamma_h>0. \tag{6.34}$$

Particularly from the Céa Lemma for saddle-point problems (Theorem 6.10), we want to choose discrete spaces which lead to the same rate of convergence

$$\inf_{v\in X_h}\|u-v\|_{H^1(\Omega)},\qquad \inf_{q\in Y_h}\|p-q\|_{L^2(\Omega)}.$$

This motivates the choice $X_h=(\mathcal{S}^1_0(\mathcal{T}))^2$ and $Y_h=P^0(\mathcal{T})\cap L^2_\star(\Omega)$. However, this choice does not satisfy a discrete inf-sup condition as we will show using a version of Euler's formula for planar graphs.

Theorem 6.14. *Let $\mathcal{T}$ denote a regular triangulation of a simply connected domain Ω. Then, there holds*

$$\#\mathcal{T}=2\#(\mathcal{K}\cap\Omega)+\#(\mathcal{K}\cap\partial\Omega)-2.$$

With this, we see

$$\dim(Y_h)=\#\mathcal{T}-1=2\#(\mathcal{K}\cap\Omega)+\#(\mathcal{K}\cap\partial\Omega)-3=\dim(X_h)+\#(\mathcal{K}\cap\partial\Omega)-3.$$

Since $\#(\mathcal{K}\cap\partial\Omega)-3>0$ for all meshes with more than one element, we see $\dim(Y_h)>\dim(X_h)$. This contradicts the inf-sup condition (see also Proposition 6.4) and hence this discretization does not lead to regular linear systems.

In the following, we discuss a couple of valid choices of discrete spaces.

Theorem 6.15 (Taylor-Hood-type element). *Let $\mathcal{T}$ denote a regular triangulation of Ω. Let*

$$X_h:=\left(\mathcal{S}^2_0(\mathcal{T})\right)^2,\qquad Y_h:=P^0(\mathcal{T})\cap L^2_\star(\Omega).$$

Then, there exists a constant $\gamma>0$, which depends only on the shape regularity of $\mathcal{T}$ such that

$$\inf_{0\neq p\in Y_h}\sup_{0\neq u\in X_M}\frac{b(u,p)}{\|p\|_{L^2(\Omega)}\|u\|_{H^1(\Omega)}}\geq\gamma>0.$$

Proof. We apply Lemma 6.12. For that, we choose $\Pi_1:(H^1_0(\Omega))^2\to(\mathcal{S}^1_0(\mathcal{T}))^2\subset X_h$ as the Scott-Zhang interpolation operator (or any Clément operator). Particularly, this shows

$$\begin{aligned}\|u-\Pi_1u\|_{L^2(T)}&\leq Ch_T\|u\|_{H^1(\Omega_T)}\qquad\forall T\in\mathcal{T},\\ \|\Pi_1u\|_{H^1(\Omega)}&\leq C\|u\|_{H^1(\Omega)}.\end{aligned}$$

The operator Π_2 is defined elementwise via

$$\Pi_2u\in(\mathcal{S}^2_0(\mathcal{T}))^2 \tag{6.35a}$$

$$(\Pi_2u)(V)=0\qquad\forall V\in\mathcal{N}(\mathcal{T}) \tag{6.35b}$$

$$\int_e\Pi_2u-u=0\qquad\forall e\in\mathcal{E}(\mathcal{T}). \tag{6.35c}$$

Obviously, Π_2 is a well-defined linear operator. Moreover, a scaling argument shows for all $T \in \mathcal{T}$ that

$$\begin{aligned}\|\Pi_2 u\|^2_{L^2(T)} &\le Ch_T^2 \|(\Pi_2 u)\circ\Phi_T\|^2_{L^2(T_{\mathrm{ref}})} \le Ch_T^2 \|u\circ\Phi_T\|^2_{L^2(\partial T_{\mathrm{ref}})} \le Ch_T^2 \|u\circ\Phi_T\|^2_{H^1(T_{\mathrm{ref}})} \\ &= Ch_T^2 \left[\|u\circ\Phi_T\|^2_{L^2(T_{\mathrm{ref}})} + \|\nabla u\circ\Phi_T\|^2_{L^2(T_{\mathrm{ref}})}\right] \\ &\le C\|u\|^2_{L^2(T)} + Ch_T^2\|\nabla u\|^2_{H^1(T)}, \text{ as well as} \\ \|\nabla\Pi_2 u\|^2_{L^2(T)} &\le C\|\nabla(\Pi_2 u)\circ\Phi_T\|^2_{L^2(T_{\mathrm{ref}})} \le C\|\nabla(\Pi_2 u)\circ\Phi_T\|^2_{L^2(T_{\mathrm{ref}})} \le \cdots \le C\left[h_T^{-2}\|u\|^2_{L^2(T)} + \|\nabla u\|^2_{H^1(T)}\right].\end{aligned}$$

Altogether, this shows

$$\|\Pi_2 u\|_{H^1(T)} \le C\left[h_T^{-1}\|u\|_{L^2(T)} + |u|_{H^1(KT)}\right] \qquad \forall u \in (H^1(T))^2.$$

This implies

$$\|\Pi_2(\boldsymbol{I}-\Pi_1)u\|_{H^1(T)} \le Ch_T^{-1}\|u-\Pi_1 u\|_{L^2(T)} + C\|u-\Pi_1 u\|_{H^1(T)} \le C\|u\|_{H^1(\tilde{\Omega}_T)}.$$

Summing up over all $T \in \mathcal{T}$ shows $\|\Pi_2(\boldsymbol{I}-\Pi_1)u\|_{H^1(\Omega)} \le C\|u\|_{H^1(\Omega)}$.

For $p \in Y_h$ und $u \in (H^1_0(\Omega))^2$, we obtain

$$b(u-\Pi_2 u, p) \;=\; \sum_{T\in\mathcal{T}}\int_p \nabla\cdot(u-\Pi_2 u) = \sum_{T\in\mathcal{T}} \underbrace{\int_{\partial T} p(u-\Pi_2 u)\cdot n_T}_{=0 \text{by construction of } \Pi_2} - \int_T \underbrace{\nabla p}_{=0}(u-\Pi_2 u) = 0.$$

■

Theorem 6.16 (MINI-Element). *Let $\mathcal{T}$ a regular triangulation of Ω. Let $B_3 := \{u \in H^1(\Omega) \,|\, u|_T \circ \Phi_T \in \operatorname{span}\{b_3\}\}$, where b_3 is the cubic element bubble function $b_3(x,y) := xy(1-x-y)$ on the reference triangle T_{ref}. Let*

$$X_h := \left(\mathcal{S}^1_0(\mathcal{T}) + B_3\right)^2, \qquad Y_h := \mathcal{S}^1(\mathcal{T}) \cap L^2_\star(\Omega).$$

Then, there exists a constant $\gamma > 0$, which depends only on the shape regularity of $\mathcal{T}$ such that

$$\inf_{0\neq p\in Y_h}\sup_{0\neq u\in X_M} \frac{b(u,p)}{\|p\|_{L^2(\Omega)}\|u\|_{H^1(\Omega)}} \ge \gamma > 0.$$

Proof. Again, we use 6.12. Let Π_1 denote again the Scott-Zhang operator. The operator Π_2 is defined elementwise as follows: The bubble-functions $b_T := b_3 \circ \Phi_T^{-1}$ satisfy $\operatorname{supp} b_T \subset T$ und $B_3 = \operatorname{span}\{b_T \,|\, T \in \mathcal{T}\}$. We define

$$\Pi_2 u|_T := \frac{1}{\int_T b_T} b_T \begin{pmatrix} \int_T u_1 \\ \int_T u_2 \end{pmatrix},$$

with $u = (u_1, u_2)$. Then, there holds

$$\begin{aligned}\Pi_2 : (H^1_0(\Omega))^2 &\to B_3^2 \qquad \text{is a linear operator} \\ \|\Pi_2 u\|_{L^2(T)} &\le C\|u\|_{L^2(T)} \qquad \forall T \in \mathcal{T} \\ \|\Pi_2 u\|_{H^1(T)} &\le Ch_T^{-1}\|u\|_{L^2(T)} \qquad \forall T \in \mathcal{T}.\end{aligned}$$

Analogously to the proof of Theorem 6.15, we obtain $\|\Pi_2(\boldsymbol{I}-\Pi_1)u\|_{H^1(\Omega)} \leq C\|u\|_{H^1(\Omega)}$. Moreover, for $p \in \mathcal{S}^1(\mathcal{T})$:

$$\begin{aligned} b(u-\Pi_2 u, p) &= \int_\Omega p\nabla\cdot(u-\Pi_2 u) = \underbrace{\int_{\partial\Omega} p(u-\Pi_2 u)}_{=0 \text{ due to boundary condition}} - \int_\Omega \nabla p\cdot(u-\Pi_2 u) \\ &= \sum_{T\in\mathcal{T}} \int_K \underbrace{\nabla p|_T}_{=\text{constant}} \cdot(u-\Pi_2 u) \overset{\text{by construction of } \Pi_2}{=} 0 \end{aligned}$$

■

The most widely used discretization for Stokes is the following Tayler-Hood element.

Theorem 6.17 (Taylor-Hood). *Let $\mathcal{T}$ denote a regular triangulation such that each element $T \in \mathcal{T}$ has at most one edge on $\partial\Omega$. Define*

$$X_h := \left(\mathcal{S}_0^2(\mathcal{T})\right)^2, \qquad Y_h := \mathcal{S}^1(\mathcal{T}) \cap L^2_\star(\Omega).$$

Then, there exists a constant $\gamma > 0$ which depends only on the shape regularity of $\mathcal{T}$ such that

$$\inf_{0\neq p\in Y_h} \sup_{0\neq u\in X_M} \frac{b(u,p)}{\|p\|_{L^2(\Omega)}\|u\|_{H^1(\Omega)}} \geq \gamma > 0.$$

6.4 Further remarks on mixed methods

Mixed methods can be useful if a direct discretization of a problem is difficult. We demonstrate this for the biharmonic equation:

$$\Delta^2 u = f \qquad \text{in } \Omega, \tag{6.36a}$$
$$u = 0 \qquad \text{on } \partial\Omega \tag{6.36b}$$
$$\partial_n u = 0 \qquad \text{on } \partial\Omega \tag{6.36c}$$

The classical weak form of the biharmonic equation is

$$\text{Find } u \in H^2_0(\Omega) \text{ such that} \qquad \int_\Omega \Delta u \Delta v = \int_\Omega fv \qquad \forall v \in H^2_0(\Omega). \tag{6.37}$$

To derive a FEM for the above problem, we need to choose discrete subspaces $X_h \subseteq H^2_0(\Omega)$. Note that the standard spaces $\mathcal{S}^p_0 \not\subset H^2(\Omega)$ do not work. By ensuring C^1-regularity over element interfaces, it is possible to construct piecewise polynomial spaces which are subspaces of $H^2(\Omega)$ (for example the Argyris-element or the Hsieh-Clough-Tocher-element). However, such an implementation is complicated and not very popular among users. It is easier to change the weak form. To that end, we introduce a new variable $\sigma = -\Delta u$. This leads to the following problem: Find $(u,\sigma) \in H^1_0(\Omega) \times H^1(\Omega)$ such that

$$\int_\Omega \nabla\sigma\cdot\nabla w = \int_\Omega fw \qquad \forall w \in H^1_0(\Omega), \tag{6.38a}$$
$$\int_\Omega \nabla u\cdot\nabla v - \int_\Omega \sigma v = 0 \qquad \forall v \in H^1(\Omega) \tag{6.38b}$$

Without looking into the solution theory of the mixed FEM, we note that we want to find $(u, \sigma) \in H_0^1(\Omega) \times H^1(\Omega)$. Hence, we only need to choose finite dimensional subspaces of $H_0^1(\Omega) \times H^1(\Omega)$, which can be done by using classical polynomial spaces.

6.5 The Gårding inequality

Often, an elliptic problem is perturbed by a lower order term such that the resulting problem is no longer elliptic but satisfies a Gårding inequality.

Definition. Let X_0, X_1 denote Hilbert spaces with compact embedding $X_1 \subset X_0$. A bilinearform $a : X_1 \times X_1 \to \mathbb{R}$ satisfies a Gårding inequality if there exist constants C_0, $C_1 > 0$ with

$$a(u,u) \geq C_1 \|u\|_{X_1}^2 - C_0 \|u\|_{X_0}^2 \qquad \forall u \in X_1.$$

Problems which satisfy a Gårding inequality arise for example if one considers PDEs with lower order terms.

Exercise 62. Consider

$$\begin{aligned} -\Delta u - b(x) \cdot \nabla u + c(x) u &= f \quad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \partial\Omega \end{aligned}$$

Show that the corresponding bilinear form satisfies a Gårding inequality with $X_1 = H_0^1(\Omega)$ and $X_0 = L^2(\Omega)$. □

We use the following result from functional analysis:

Exercise 63. Let X, Y denote Banach spaces and let $K : X \to Y$ be a compact operator. Let $(\Pi_N)_{N\in\mathbb{N}}$ a sequence of linear operators $\Pi_N : Y \to Y$ with $\|\Pi_N\| \leq 1$ and $\Pi_N \to \mathrm{Id}$ pointwise (i.e. $\lim_{N\to\infty} \Pi_N u = u$ for all $u \in Y$). Then, there holds

$$\lim_{N\to\infty} \|(\mathrm{Id} - \Pi_h) K\|_{Y \leftarrow X} = 0.$$

□

For well-posed problems (i.e. the continuous equation has a unique solution) with Gårding inequality, the following result shows that their discretization is asymptotically quasi-optimal.

Theorem 6.18. *Let X_1, X_0 be Hilbert spaces with compact embedding $X_1 \subset X_0$. Let $a(\cdot,\cdot)$ satisfy a Gårding inequality and let the induced operator $\mathbf{A} : X_1 \to X_1'$, $\mathbf{A}u := a(u,\cdot)$ be bijective. Let $(X_h)_{h>0} \subset X_1$ denote a sequence of closed subspaces such that*

$$\lim_{h\to 0} \inf_{v\in X_h} \|u - v\|_{X_1} = 0 \qquad \forall u \in X_h.$$

Then, there exists $h_0 > 0$ and $\gamma > 0$ such that for all $0 < h < h_0$

$$\inf_{u\in X_h} \sup_{v\in X_h} \frac{a(u,v)}{\|u\|_{X_1}\|v\|_{X_1}} \geq \gamma > 0.$$

In particular, there holds for the FEM error

$$\|u-u_h\|_{X_1} \le \left(1+\frac{\|a\|}{\gamma}\right) \inf_{v\in X_h} \|u-v\|_{X_1}$$

Proof. *Step 1:* We show that there exists $\widetilde{C}>0$ such that for all $u\in X_1$ we find $v\in X_1$ of the form $v=u+z$ such that

$$\begin{aligned} a(u,v) &= a(u,u+z) \ge C_1\|u\|_{X_1}^2 && (6.39)\\ \|v\|_{X_1} &\le \widetilde{C}\|u\|_{X_1}. && (6.40)\end{aligned}$$

The choice of z is motivated by the Gårding inequality $a(u,u)\ge C_1\|u\|_{X_1}^2 - C_0\|u\|_{X_0}^2$, i.e

$$a(u,u+z) = a(u,u)+a(u,z) \ge C_1\|u\|_{X_1}^2 - C_0\|u\|_{X_0}^2 + a(u,z).$$

Hence, we choose $z\in X_1$ as solution of the (adjoint) problem

$$\text{Find } z\in X_1 \text{ s.t.} \quad a(w,z) = C_0\langle w,u\rangle_{X_0} \qquad \forall w\in X_1$$

In operator notation, this reads as

$$\mathbf{A}^\top z = \mathbf{K}u,$$

where $\mathbf{A}: X_1\to X_1'$ is induced by the bilinearform $a(\cdot,\cdot)$ and $\mathbf{K}: X_1\to X_1'$ is defined via

$$\langle \cdot, \mathbf{K}u\rangle_{X_1\times X_1'} = \langle \cdot, C_0 u\rangle_{X_0}.$$

We note that

(i) Since $\mathbf{A}$ is bijective, also $\mathbf{A}^\top$ bijective and $\|\mathbf{A}^{-\top}\| = \|\mathbf{A}^{-1}\|$.

(ii) Since $X_1\subset X_0$ is compact, also $\mathbf{K}: X_1\to X_1'$ is a compact operator.

The operator $\mathbf{A}^{-\top}\mathbf{K}: X_1\to X_1$, which maps u to z is compact. We obtain

$$\begin{aligned} a(u,v) &= a(u,u+z) \ge C_1\|u\|_{X_1}^2\\ \|v\|_{X_1} &\le \|u\|_{X_1}+\|z\|_{X_1} \le \left(1+\|\mathbf{A}^{-\top}\mathbf{K}\|\right)\|u\|_{X_1}.\end{aligned}$$

Step 2: For given $u\in X_h$, we construct $v\in X_h$ such that

$$\begin{aligned} a(u,v) &\ge \frac{C_1}{2}\|u\|_{X_1}^2\\ \|v\|_{X_1} &\le \widetilde{C}\|u\|_{X_1}\end{aligned}$$

Let $\Pi_h: X_1\to X_h$ denote the orthogonal projection (in X_1). Following Step 1, we define $v = u+\Pi_h z\in X_h$ for given $u\in X_h$. This shows

$$\begin{aligned} a(u,v) &= a(u,u+z)+a(u,\Pi_h z - z) \ge C_1\|u\|_{X_1}^2 - \|a\|\|u\|_{X_1}\|(\mathrm{Id}-\Pi_h)z\|_{X_1}\\ &\ge C_1\|u\|_{X_1}^2 - \|a\|_{X_1}\|(\mathrm{Id}-\Pi_h)\mathbf{A}^{-\top}\mathbf{K}\|\|u\|_{X_1}^2\end{aligned}$$

Since $\mathbf{K}$ is compact and $\mathbf{A}^{-\top}$ bounded, also their composition is compact. Since $\mathrm{Id} - \Pi_h$ converges to zero pointwise (according to the assumption), we obtain with Exercise 63 that $\lim_{h\to 0} \|(\mathrm{Id} - \Pi_h)\mathbf{A}^{-\top}\mathbf{K}\| = 0$. Hence, there exists $h_0 > 0$ (independently of u) such that all $0 < h < h_0$ satisfy

$$a(u, u + \Pi_h z) \geq \frac{C_1}{2}\|u\|_{X_1}^2.$$

Furthermore, $v = u + \Pi_h z \in X_h$ satisfies

$$\|v\|_{X_1} \leq \|u\|_{X_1} + \|\Pi_h z\|_{X_1} \leq \|u\|_{X_1} + \|z\|_{X_1} \leq (1 + \|\mathbf{A}^{-\top}\mathbf{K}\|)\|u\|_{X_1}.$$

Step 3: This shows the discrete inf-sup condition. Quasi-optimality follows from the Céa lemma. ■

A bilinear form which satisfies a Gårding inequality does not necessarily induce a bijective operator. A famous result from functional analysis states however, that injectivity implies already bijectivity.

Theorem 6.19 (Fredholmalternative). *Let a denote a bounded bilinearform on the Hilbertspace X_1, which satisfies a Gårding inequality. Let the induced operator $\mathbf{A} : X_1 \to X_1'$ be injective, i.e.*

$$a(u, v) = 0 \qquad \forall v \in X_1 \qquad \Longrightarrow u = 0.$$

Then, $\mathbf{A}$ is already bijective.

Proof. The Gårding inequality states

$$a(u, u) \geq C_1\|u\|_{X_1}^2 - C_0\|u\|_{X_0}^2$$

Consider $\widetilde{a} : X_1 \times X_1 \to \mathbb{R}$ defined by

$$\widetilde{a}(u, v) := a(u, v) + C_0\langle u, v\rangle_{X_0}$$

Due to the Lax-Milgram Lemma $\widetilde{a}(\cdot, \cdot)$ induces a bijective operator $\widetilde{\mathbf{A}} : X_1 \to X_1'$. The difference $\mathbf{K} := \widetilde{\mathbf{A}} - \mathbf{A} : X_1 \to X_1'$ is compact since

$$\langle \mathbf{K}u, v\rangle_{X_1' \times X_1} = C_0\langle u, v\rangle_{X_0}$$

and $X_1 \subset X_0$ is compact. Hence $\mathbf{A}$ reads as

$$\mathbf{A} = \widetilde{\mathbf{A}} - \mathbf{K} = \widetilde{\mathbf{A}}\left(\mathrm{Id} - \widetilde{\mathbf{A}}^{-1}\mathbf{K}\right).$$

The injectivity of $\mathbf{A}$ implies that 1 is not an Eigenvalue of the compact operator $\widetilde{\mathbf{A}}^{-1}\mathbf{K}$. The theory of compact operators shows that this implies that $\mathrm{Id} - \widetilde{\mathbf{A}}^{-1}\mathbf{K}$ is invertible and hence $\mathbf{A}$ is bijective. ■

Chapter 7

High-dimensional problems

The number of elements in a regular mesh $\mathcal{T}_h$ in which each element $T \in \mathcal{T}_h$ satisfies $\operatorname{diam}(T) \simeq |T|^{1/d} \simeq h$ scales roughly like $\mathcal{O}(h^{-d})$, i.e., exponentially in the dimension. We remember the a priori convergence of FEM which states

$$\|u - u_h\|_{H^1(\Omega)} = \mathcal{O}(h)$$

in case u is sufficiently smooth. To compute u_h, we need to solve a linear system with $\#\mathcal{T}_h$ elements. The cost for this is at least $\mathcal{O}(\#\mathcal{T}_h) = \mathcal{O}(h^{-d})$. Thus, in terms of cost, we get the estimate

$$\|u - u_h\|_{H^1(\Omega)} = \mathcal{O}(N^{-1/d})$$

with $N = \#\mathcal{T}_h$.

This shows that the convergence rate with respect to cost goes down in higher dimensions. For $d = 2$, halving the error requires four times as many elements. For $d = 4$, the same error reduction requires 16-times as many elements. This is usually called the *curse of dimensionality* and we will look into two methods which are able to overcome this curse to some extent.

7.1 Sparse grids

To illustrate the idea, we first look at standard tensor interpolation: For a given set of intervals $\mathcal{T}$, let $\mathcal{Q}^1(\mathcal{T})$ denote the continuous functions which are affine on each interval in $\mathcal{T}$. Let $I_\ell \colon C([0,1]) \to \mathcal{Q}^1(\{[k2^{-\ell}, (k+1)2^{-\ell}] \,|\, k = 0, \dots, 2^\ell - 1\})$ denote the nodal interpolation operator in 1D, i.e.,

$$I_\ell v(t_k) = v(t_k) \quad \text{for all } t_k = k2^{-\ell},\ k = 0, \dots, 2^\ell.$$

Note that $I_0 v(x) = x(v(1) - v(0)) + v(0)$. Hence, we have by Theorem 3.4 that $\|I_0 v\|_{L^2([0,1])} \le \|v\|_{C([0,1])} \lesssim \|v\|_{H^1([0,1])}$. Moreover, we get $\partial_x I_0 v = v(1) - v(0) = \int_0^1 \partial_x v(s)\, ds$ and hence the estimate $\|\partial_x I_0 v\|_{L^2([0,1])} \le \|v\|_{H^1([0,1])}$. Altogether, this shows

$$\|I_0 v\|_{H^1([0,1])} \le C \|v\|_{H^1([0,1])} \quad \text{for all } v \in H^1([0,1]). \tag{7.1}$$

We denote with I_ℓ^x that the interpolation operator is applied in dimension x. The approximation on the d-dimensional tensor mesh

$$\mathcal{T}_\ell^\otimes := \Big\{ \prod_{i=1}^d [k_i 2^{-\ell}, (k_i+1)2^{-\ell}] \,\Big|\, k_1, \dots, k_d \in \{0, \dots, 2^\ell - 1\} \Big\}$$

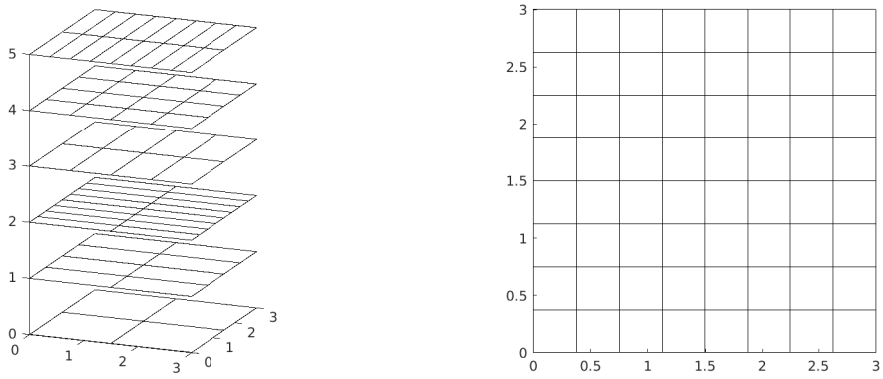


FIGURE 7.1. *The different sparse grid contributions on the left stacked on top of each other combine to the full grid on the right. The interpolation operator $I_{\boldsymbol{\ell}} = I^{x_1}_{\ell_1} I^{x_2}_{\ell_2}$ corresponds to one of the grids on the left-hand side (e.g., grid number 1 for $\boldsymbol{\ell} = (1,2)$ or grid number 5 for $\boldsymbol{\ell} = (3,1)$).*

is given for $v \in C^0([0,1]^d)$ by

$$(I^{\otimes}_{\ell} v)(\boldsymbol{x}) := I^{x_1}_{\ell}(I^{x_2}_{\ell} \dots (I^{x_d}_{\ell} v) \dots)(x_1, \dots, x_d) \in \mathcal{Q}^1(\mathcal{T}^{\otimes}_{\ell}),$$

where

$$\mathcal{Q}^1(\mathcal{T}^{\otimes}_{\ell}) := \big\{ v \in C^0([0,1]^d) \,\big|\, \forall 1 \le i \le d,\ (x_i \mapsto v(x))|_T \text{ is a polynomial of degree } \le 1 \big\}.$$

Similarly to the proof of the approximation theorem (Theorem 3.5), one can show

$$\|v - I^{\otimes}_{\ell} v\|_{L^\infty([0,1]^d)} \le C 2^{-\ell} \|v\|_{C^1([0,1]^d)}$$

for $v \in C^1([0,1]^d)$. As we see, the computation of $I^{\otimes}_{\ell} v$ requires the evaluation of $2^{d\ell}$ points in $[0,1]^d$ and hence is impractical for many purposes (if $d = 100$, and $\ell = 1$, we would need 2^{100} points).

The sparse grid idea is as follows: With the definition $I_{-1} = 0$, we may rewrite

$$\begin{aligned}
(I^{\otimes}_{\ell} v)(\boldsymbol{x}) &= \sum_{\ell_1=0}^{\ell} (I^{x_1}_{\ell_1} - I^{x_1}_{\ell_1-1})(I^{x_2}_{\ell} \dots (I^{x_d}_{\ell} v) \dots)(x_1, \dots, x_d) \\
&= \sum_{\ell_1=0}^{\ell} \sum_{\ell_2=0}^{\ell} (I^{x_1}_{\ell_1} - I^{x_1}_{\ell_1-1})(I^{x_2}_{\ell_2} - I^{x_2}_{\ell_2-1})(I^{x_3}_{\ell} \dots (I^{x_d}_{\ell} v) \dots)(x_1, \dots, x_d) \\
&= \sum_{\boldsymbol{\ell}=(\ell_1,\dots,\ell_d) \in \{0,\dots,\ell\}^d} \underbrace{(I^{x_1}_{\ell_1} - I^{x_1}_{\ell_1-1})(I^{x_2}_{\ell_2} - I^{x_2}_{\ell_2-1}) \dots (I^{x_d}_{\ell_d} - I^{x_d}_{\ell_d-1})}_{=:\Delta_{\boldsymbol{\ell}}} (v)(\boldsymbol{x}).
\end{aligned}$$

Lemma 7.1. *For a subset $\mathfrak{u} \subseteq \{1, \dots, d\}$ let $\partial_{\boldsymbol{x}_\mathfrak{u}} := \prod_{i \in \mathfrak{u}} \partial_{x_i}$ denote the partial derivatives in all directions in $\mathfrak{u}$. For sufficiently smooth $v \in C^0([0,1]^d)$, there holds*

$$\|\Delta_{\boldsymbol{\ell}} v\|_{H^1([0,1]^d)} \le C^d 2^{-|\boldsymbol{\ell}|} \|\partial_{\boldsymbol{x}_\mathfrak{u}} v\|_{H^1([0,1]^d)},$$

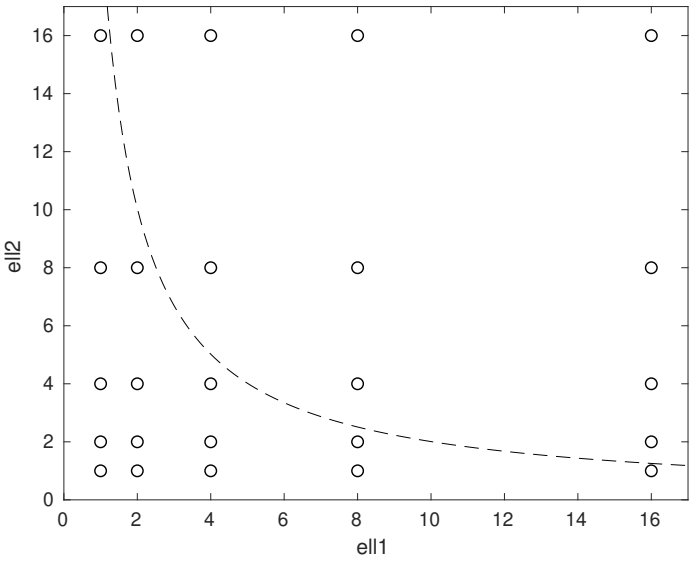


FIGURE 7.2. *The circles represent the number of degrees of freedom of $I_{\boldsymbol{\ell}}$ in each coordinate direction. The sparse grid approach only uses interpolation operators which correspond to circles below the dashed line. This shape is the upper right quadrant of the so-called* hyperbolic cross.

where $|\boldsymbol{\ell}| := \ell_1 + \ldots + \ell_d$ and $\mathfrak{u} \subseteq \{1, \ldots, d\}$ contains each dimension i with $\ell_i > 0$.

Proof. Let $x_0 = (x_{0,1}, \ldots, x_{0,d}) \in [0,1]^d$ and $i \in \{1, \ldots, d\}$. Choose $k \in \mathbb{N}$ such that $|k2^{-\ell} - x_{0,i}|$ is minimal. Without loss of generality, we assume $x_{0,i} \geq k2^{-\ell}$ (the other case works analogously). Rolle's theorem implies that there exists $\xi \in (k2^{-\ell}, (k+1)2^{-\ell})$ with

$$\partial_{x_i}(1 - I_\ell^{x_i})v(x_{0,1}, \ldots, x_{0,i-1}, \xi, x_{0,i+1}, \ldots, x_{0,d}) = 0.$$

With this, there holds

$$\begin{aligned}
(1 - I_\ell^{x_i})v(x_0) &= \int_{k2^{-\ell}}^{x_{0,i}} \partial_{x_i}(1 - I_\ell^{x_i})v(x_{0,1}, \ldots, x_{0,i-1}, z, x_{0,i+1}, \ldots, x_{0,d})\, dz \\
&= \int_{k2^{-\ell}}^{x_{0,i}} \int_\xi^z \partial_{x_i}^2 v(x_{0,1}, \ldots, x_{0,i-1}, t, x_{0,i+1}, \ldots, x_{0,d})\, dt\, dz \\
&\leq |(k+1)2^{-\ell} - k2^{-\ell}|^{3/2} \|\partial_{x_i}^2 v(x_{0,1}, \ldots, x_{0,i-1}, \cdot, x_{0,i+1}, \ldots, x_{0,d})\|_{L^2([k2^{-\ell}, (k+1)2^{-\ell}])} \\
&\leq 2^{-3\ell/2} \|\partial_{x_i}^2 v(x_{0,1}, \ldots, x_{0,i-1}, \cdot, x_{0,i+1}, \ldots, x_{0,d})\|_{L^2([k2^{-\ell}, (k+1)2^{-\ell}])}.
\end{aligned}$$

We define

$$\Omega_k := \otimes_{j=1}^{i-1}[0,1] \times [k2^{-\ell}, (k+1)2^{-\ell}] \times \otimes_{j=i+1}^{d}[0,1].$$

With $dx_{0,-i} := dx_{0,1} \dots dx_{0,i-1} dx_{0,i+1} \dots dx_{0,d}$, this results in

$$
\begin{aligned}
&\|(1 - I_\ell^{x_i})v\|^2_{L^2(\Omega_k)} \\
&\le 2^{-3\ell} \int_{\Omega_k} \|\partial^2_{x_i} v(x_{0,1}, \dots, x_{0,i-1}, \cdot, x_{0,i+1}, \dots, x_{0,d})\|^2_{L^2([k2^{-\ell},(k+1)2^{-\ell}])}\, dx_0 \\
&= 2^{-3\ell} \underbrace{\int_0^1 \cdots \int_0^1}_{i-\text{times}} \int_{k2^{-\ell}}^{(k+1)2^{-\ell}} \underbrace{\int_0^1 \cdots \int_0^1}_{(d-i-1)-\text{times}} \int_{k2^{-\ell}}^{(k+1)2^{-\ell}} |\partial^2_{x_i} v(x_{0,1}, \dots, x_{0,i-1}, z, x_{0,i+1}, \dots, x_{0,d})|^2\, dz\, dx_0 \\
&= 2^{-3\ell} \int_{k2^{-\ell}}^{(k+1)2^{-\ell}} \underbrace{\int_0^1 \cdots \int_0^1}_{(d-1)-\text{times}} \int_{k2^{-\ell}}^{(k+1)2^{-\ell}} |\partial^2_{x_i} v(x_{0,1}, \dots, x_{0,i-1}, z, x_{0,i+1}, \dots, x_{0,d})|^2\, dz\, dx_{0,-i}\, dx_{0,i} \\
&= 2^{-3\ell} \int_{k2^{-\ell}}^{(k+1)2^{-\ell}} \|\partial^2_{x_i} v\|^2_{L^2(\Omega_k)}\, dx_{0,i} \le 2^{-4\ell} \|\partial^2_{x_i} v\|^2_{L^2(\Omega_k)}.
\end{aligned}
$$

Since $[0,1]^d = \dot{\bigcup}_{k=0}^{2^\ell - 1} \Omega_k$, we obtain

$$\|(1 - I_\ell^{x_i})v\|_{L^2([0,1]^d)} \le 2^{-2\ell} \|\partial^2_{x_i} v\|_{L^2([0,1]^d)}.$$

Analogously, we show

$$\|\nabla(1 - I_\ell^{x_i})v\|_{L^2([0,1]^d)} \le 2^{-\ell} \|\partial_{x_i} \nabla v\|_{L^2([0,1]^d)}$$

and obtain

$$\|(1 - I_\ell^{x_i})v\|_{H^1([0,1]^d)} \le 2^{-\ell} \|\partial_{x_i} v\|_{H^1([0,1]^d)}.$$

If $\ell > 0$, the triangle inequality concludes

$$
\begin{aligned}
\|(I_\ell^{x_i} - I_{\ell-1}^{x_i})v\|_{H^1([0,1]^d)} &\le \|(1 - I_{\ell-1}^{x_i})v\|_{H^1([0,1]^d)} + \|(1 - I_\ell^{x_i})v\|_{H^1([0,1]^d)} \\
&\le 2^{-(\ell-1)} \|\partial_{x_i} v\|_{H^1([0,1]^d)} + 2^{-\ell} \|\partial_{x_i} v\|_{H^1([0,1]^d)} \\
&\le 2^{-\ell+2} \|\partial_{x_i} v\|_{H^1([0,1]^d)}.
\end{aligned} \tag{7.2}
$$

For $\ell = 0$, we use (7.1) to show

$$\|(I_0^{x_i} - I_{-1}^{x_i})v\|_{H^1([0,1]^d)} = \|I_0^{x_i} v\|_{H^1([0,1]^d)} \le C \|v\|_{H^1([0,1]^d)}. \tag{7.3}$$

Let $i_1, \dots, i_k$ be the dimensions with $\ell_{i_j} > 0$ for $j = 1, \dots, k$. There holds with (7.3)

$$
\begin{aligned}
&\|(I_{\ell_1}^{x_1} - I_{\ell_1-1}^{x_1}) \dots (I_{\ell_d}^{x_1} - I_{\ell_d-1}^{x_d})v\|_{H^1([0,1]^d)} \\
&\quad = \|\Big(\prod_{\substack{i=1 \\ \ell_i = 0}}^{d} (I_{\ell_i}^{x_i} - I_{\ell_i-1}^{x_i}) \Big) \Big(\prod_{\substack{i=1 \\ \ell_i > 0}}^{d} (I_{\ell_i}^{x_i} - I_{\ell_i-1}^{x_i}) \Big) v\|_{H^1([0,1]^d)} \\
&\quad \le C^d \|(I_{\ell_{i_1}}^{x_{i_1}} - I_{\ell_{i_1}-1}^{x_{i_1}}) \dots (I_{\ell_{i_k}}^{x_{i_k}} - I_{\ell_{i_k}-1}^{x_{i_k}})v\|_{H^1([0,1]^d)}.
\end{aligned}
$$

From here, we may iteratively apply (7.2) in each dimension to get

$$\begin{aligned}
&\|(I_{\ell_{i_1}}^{x_{i_1}} - I_{\ell_{i_1}-1}^{x_{i_1}}) \dots (I_{\ell_{i_k}}^{x_{i_k}} - I_{\ell_{i_k}-1}^{x_{i_k}}) v\|_{H^1([0,1]^d)} \\
&\quad \le 2^{-\ell_{i_1}+2} \|\partial_{x_{i_1}} (I_{\ell_{i_2}}^{x_{i_2}} - I_{\ell_{i_2}-1}^{x_{i_2}}) \dots (I_{\ell_{i_k}}^{x_{i_k}} - I_{\ell_{i_k}-1}^{x_{i_k}}) v\|_{H^1([0,1]^d)} \\
&\quad = 2^{-\ell_{i_1}+2} \|(I_{\ell_{i_2}}^{x_{i_2}} - I_{\ell_{i_2}-1}^{x_{i_2}}) \dots (I_{\ell_{i_k}}^{x_{i_k}} - I_{\ell_{i_k}-1}^{x_{i_k}}) \partial_{x_{i_1}} v\|_{H^1([0,1]^d)} \\
&\quad \dots \\
&\quad \le 4^k 2^{-\ell_{i_1} - \dots - \ell_{i_k}} \|\partial_{x_{i_1}} \partial_{x_{i_2}} \dots \partial_{x_{i_k}} v\|_{H^1([0,1]^d)}.
\end{aligned}$$

Since $k \le d$ and $\mathfrak{u} = \{i_1, \dots, i_k\}$, the proof is complete. ∎

With the last result to obtain an error of $2^{-\ell}$, we may ignore all $\Delta_{\boldsymbol{\ell}}$ with $|\boldsymbol{\ell}| > \ell$. This leads to the sparse grid interpolation operator I_ℓ^d defined by

$$I_\ell^d v := \sum_{\substack{\boldsymbol{\ell} \in \{0,\dots,\ell\}^d \\ |\boldsymbol{\ell}| \le \ell}} \Delta_{\boldsymbol{\ell}} v. \tag{7.4}$$

This truncation is illustrated in Figures 7.1–7.2. To analyze the error, we need the following nice combinatorial identity.

Lemma 7.2. *There holds*

$$\#\{\boldsymbol{\ell} \in \mathbb{N}_0^d \,|\, |\boldsymbol{\ell}| = j\} = \binom{j+d-1}{d-1}.$$

Proof. There are many proofs of this identity. A nice one goes like this: Imagine the index $\boldsymbol{\ell} \in \mathbb{N}_0^d$ as

$$\underbrace{1 \dots 1}_{\ell_1} \,|\, \underbrace{1 \dots 1}_{\ell_2} \,|\, \dots \dots \,|\, \underbrace{1 \dots 1}_{\ell_d}$$

This line contains the $|\ell| + d - 1$ symbols $z \in \{1, |\}$. Exactly $d-1$ of the symbols z must satisfy $z = |$. Hence there are $\binom{j+d-1}{d-1}$ possibilities. ∎

Theorem 7.3. *The sparse grid interpolation error satisfies*

$$\|(1 - I_\ell^d) v\|_{H^1([0,1]^d)} \le C 4^d (\ell + d)^{d-1} 2^{-\ell} \|v\|_{H^2_{\mathrm{mix}}([0,1]^d)},$$

where

$$\|v\|_{H^2_{\mathrm{mix}}([0,1]^d)} := \max_{\mathfrak{u} \subseteq \{1,\dots,d\}} \|\partial_{\boldsymbol{x}_\mathfrak{u}}^2 v\|_{L^2([0,1]^d)}.$$

Proof. Given $v \in H^2_{\mathrm{mix}}([0,1]^d)$ we have with convergence of tensor interpolation in $H^1([0,1]^d)$ that

$$v = \lim_{\ell \to \infty} I_\ell^{\otimes} v = \lim_{\ell \to \infty} \sum_{\boldsymbol{\ell} \in \{0,\dots,\ell\}^d} \Delta_{\boldsymbol{\ell}}.$$

As shown in Lemma 7.1, we have

$$\|\Delta_{\boldsymbol{\ell}} v\|_{H^1([0,1]^d)} \leq 4^d 2^{-|\boldsymbol{\ell}|} \|v\|_{H^2_{\mathrm{mix}}([0,1]^d)}.$$

This implies that the series above converges absolutely and hence $v = \sum_{\boldsymbol{\ell} \in \mathbb{N}_0^d} \Delta_{\boldsymbol{\ell}} v$. With this, we may write the approximation error as

$$v - I_\ell^d v = \sum_{\substack{\boldsymbol{\ell} \in \mathbb{N}_0^d \\ |\boldsymbol{\ell}| > \ell}} \Delta_{\boldsymbol{\ell}} v.$$

Altogether, we have

$$\|v - I_\ell^d v\|_{H^1([0,1]^d)} \leq \sum_{\substack{\boldsymbol{\ell} \in \mathbb{N}_0^d \\ |\boldsymbol{\ell}| > \ell}} \|\Delta_{\boldsymbol{\ell}} v\|_{H^1([0,1]^d)} \leq 4^d \|v\|_{H^2_{\mathrm{mix}}([0,1]^d)} \sum_{\substack{\boldsymbol{\ell} \in \mathbb{N}_0^d \\ |\boldsymbol{\ell}| > \ell}} 2^{-|\boldsymbol{\ell}|}.$$

The sum can be rewritten as

$$\sum_{\substack{\boldsymbol{\ell} \in \mathbb{N}_0^d \\ |\boldsymbol{\ell}| > \ell}} 2^{-|\boldsymbol{\ell}|} = \sum_{j=\ell+1}^{\infty} 2^{-j} \sum_{\substack{\boldsymbol{\ell} \in \mathbb{N}_0^d \\ |\boldsymbol{\ell}| = j}} 1 = \sum_{j=\ell+1}^{\infty} 2^{-j} \binom{j+d-1}{d-1},$$

where we used Lemma 7.2 for the last identity. There holds for $x \in (0,1)$

$$\begin{aligned}\sum_{j=\ell+1}^{\infty} x^j \binom{j+d-1}{d-1} &= \partial_x^{d-1} \sum_{j=\ell+1}^{\infty} x^{j+d-1}/(d-1)! = \partial_x^{d-1} \frac{x^{\ell+d}}{1-x}/(d-1)! \\ &= \sum_{k=0}^{d-1} \binom{d-1}{k} \partial_x^k x^{\ell+d} \partial_x^{d-1-k} (1-x)^{-1}/(d-1)!\end{aligned}$$

since the series converges absolutely. There holds

$$\begin{aligned}&\binom{d-1}{k} \partial_x^k x^{\ell+d} \partial_x^{d-1-k} (1-x)^{-1}/(d-1)! \\ &= \frac{(d-1)(d-2)\cdots(d-k)}{k!(d-1)!} \big((\ell+d)\cdots(\ell+d-k+1)\big)(d-1-k)! \frac{x^{\ell+d-k}}{(1-x)^{d-k}} \\ &= \frac{1}{k!}\big((\ell+d)\cdots(\ell+d-k+1)\big) \frac{x^{\ell+d-k}}{(1-x)^{d-k}} \leq \frac{(\ell+d)^{d-1}}{k!} \frac{x^{\ell+d-k}}{(1-x)^{d-k}}.\end{aligned}$$

Inserting $x = 1/2$, we end up with

$$\sum_{j=\ell+1}^{\infty} 2^{-j} \binom{j+d-1}{d-1} \lesssim (\ell+d)^{d-1} 2^{-\ell}.$$

This concludes the proof. ∎

The representation in (7.4) is not really good for implementation due to cancelation effects and the requirement to constantly transform coefficient vectors between different meshes. A better variant is provided by the inclusion-exclusion formula which is an interesting combinatorial fact in it self.

Lemma 7.4. *For $d \in \mathbb{N}$ and $r \leq d$, the binomial coefficient satisfies the identity*

$$\sum_{q=0}^{r} (-1)^q \binom{d}{q} = (-1)^r \binom{d-1}{r}.$$

Proof. The proof works by induction. For $r = 0$, there holds $\binom{d}{0} = \binom{d-1}{0} = 1$. Assume the statement holds for $r < d$. Then, we have

$$\sum_{q=0}^{r+1} (-1)^q \binom{d}{q} = (-1)^{r+1} \binom{d}{r+1} + \sum_{q=0}^{r} (-1)^q \binom{d}{q} = (-1)^{r+1} \left(\binom{d}{r+1} - \binom{d-1}{r} \right).$$

The well-known identity

$$\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1}$$

with $n = d-1$ and $k = r$ concludes the proof. ∎

Lemma 7.5. *With $I_{\boldsymbol{\ell}} := I_{\ell_1}^{x_1} I_{\ell_2}^{x_2} \dots I_{\ell_d}^{x_d}$ for $\boldsymbol{\ell} \in \mathbb{N}_0^d$, there holds*

$$I_\ell^d = \sum_{k=0}^{d-1} (-1)^k \binom{d-1}{k} \sum_{\substack{\boldsymbol{\ell} \in \mathbb{N}_0^d \\ |\boldsymbol{\ell}| = \ell - k}} I_{\boldsymbol{\ell}}.$$

Proof. We rewrite (7.4) by

$$I_\ell^d = \sum_{\substack{\boldsymbol{\ell} \in \mathbb{N}_0^d \\ |\boldsymbol{\ell}| \leq \ell}} \Delta_{\boldsymbol{\ell}} = \sum_{\substack{\boldsymbol{\ell}' \in \mathbb{N}_0^d \\ |\boldsymbol{\ell}'| \leq \ell}} \alpha_{\boldsymbol{\ell}'} I_{\boldsymbol{\ell}'} \tag{7.5}$$

for some $\alpha_{\boldsymbol{\ell}} \in \mathbb{R}$. Given the definition

$$\Delta_{\boldsymbol{\ell}} = (I_{\ell_1}^{x_1} - I_{\ell_1 - 1}^{x_1})(I_{\ell_2}^{x_2} - I_{\ell_2 - 1}^{x_2}) \dots (I_{\ell_d}^{x_d} - I_{\ell_d - 1}^{x_d})$$

we note that a particular $I_{\boldsymbol{\ell}'}$ appears in (7.5) if and only if there exists $\boldsymbol{\ell} \in \mathbb{N}_0^d$ with

$$|\boldsymbol{\ell}| \leq \ell \quad \text{and} \quad \ell_i' \leq \ell_i \leq \ell_i' + 1 \quad \text{for all } i = 1, \dots, d. \tag{7.6}$$

Moreover, the sign of that $I_{\boldsymbol{\ell}'}$ is determined by the parity (odd or even) of the number of dimensions i with $\ell_i = \ell_i' + 1$. For $q = 0, \dots, d$, let

$$P_q(\boldsymbol{\ell}') := \left\{ \boldsymbol{\ell} \in \mathbb{N}_0^d \,\middle|\, \boldsymbol{\ell} \text{ satisfies (7.6) and } \ell_{i_k} = \ell_{i_k}' + 1,\ k = 1, \dots, q \right\}.$$

Then, we observe $P_q(\boldsymbol{\ell}') = \emptyset$ if $|\boldsymbol{\ell}'| > |\boldsymbol{\ell}| - q$. Moreover, since for each choice of q indices i_k we have an element of $P_q(\boldsymbol{\ell}')$, there holds

$$\# P_q(\boldsymbol{\ell}') = \binom{d}{q}.$$

This implies

$$\alpha_{\boldsymbol{\ell}'} = \sum_{q=0}^{d} (-1)^q \# P_q(\boldsymbol{\ell}') = \sum_{q=0}^{\ell-|\boldsymbol{\ell}'|} (-1)^q \# P_q(\boldsymbol{\ell}') = \sum_{q=0}^{\ell-|\boldsymbol{\ell}'|} (-1)^q \binom{d}{q}.$$

Lemma 7.4 shows for $r = \ell - |\boldsymbol{\ell}'| \leq d$

$$\sum_{q=0}^{\ell-|\boldsymbol{\ell}'|} (-1)^q \binom{d}{q} = (-1)^{\ell-|\ell'|} \binom{d-1}{\ell-|\ell'|}.$$

Altogether, we see with $k = \ell - |\ell'|$

$$I_\ell^d = \sum_{k=0}^{d} (-1)^k \binom{d-1}{k} \sum_{\substack{\boldsymbol{\ell}' \in \mathbb{N}_0^d \\ |\boldsymbol{\ell}'| = \ell - k}} I_{\boldsymbol{\ell}'}.$$

This concludes the proof. ■

Lemma 7.6. *The number of evaluations of v required for the computation of $I_\ell^d v$ is less than*

$$d \binom{\ell + d - 1}{d-1} 2^\ell \leq d(\ell + d)^{d-1} 2^\ell.$$

Proof. We use the representation from Lemma 7.5. Each $I_{\boldsymbol{\ell}} v$ requires $2^{|\boldsymbol{\ell}|}$ evaluations of v for computation. Lemma 7.2 concludes the proof. ■

The last result together with Theorem 7.3 shows the following: A sparse grid of size $h > 0$ (this means $2^{-\ell} = h$) allows an interpolation error of

$$\|(1 - I_\ell^d) v\|_{H^1([0,1]^d)} \lesssim (1 + |\log(h)|)^\alpha h$$

for some exponent $\alpha \in \mathbb{N}$ with a cost of computation of $I_\ell^d v$ less than

$$\mathcal{O}\big((1 + |\log(h)|)^\alpha h^{-1}\big).$$

This means that the error estimate with respect to cost reads

$$\|(1 - I_\ell^d) v\|_{H^1([0,1]^d)} \lesssim \mathrm{cost}^{-1}$$

(up to logarithmic factors). The convergence rate is independent of the dimension.

Instead of the sparse interpolation operator, we may also consider the sparse Galerkin projection. Define the (quad-)mesh

$$\mathcal{T}_{\boldsymbol{\ell}}^{\otimes} := \Big\{ \prod_{i=1}^{d} [k_i 2^{-\ell_i}, (k_i + 1) 2^{-\ell_i}] \,\Big|\, k_i \in \{0, \dots, 2^{\ell_i} - 1\},\ i = 1, \dots, d \Big\}$$

for $\boldsymbol{\ell} = (\ell_1, \dots, \ell_d) \in \mathbb{N}_0^d$. Note that we don't have a triangle mesh any more. However, the abstract theory just used the fact that

$$\mathcal{X}_\ell = \bigoplus_{\substack{\boldsymbol{\ell} \in \mathbb{N}_0^d \\ |\boldsymbol{\ell}| \le \ell}} \mathcal{Q}^1(\mathcal{T}_{\boldsymbol{\ell}}^{\otimes})$$

is a closed subspace of $H^1([0,1]^d)$. Hence, we may apply all the results of the previous sections.

Theorem 7.7. *We consider*

$$\begin{aligned} -\Delta u &= f \text{ in } [0,1]^d, \\ u &= 0 \text{ on } \partial[0,1]^d. \end{aligned}$$

Assume that $u \in H^2_{\mathrm{mix}}([0,1]^d)$ and let $u_\ell \in \mathcal{X}_\ell \cap H^1_0([0,1]^d)$ denote the unique Galerkin approximation. Then, there holds

$$\|u - u_\ell\|_{H^1([0,1]^d)} \le C 4^d (\ell + d)^{d-1} 2^{-\ell} \|u\|_{H^2_{\mathrm{mix}}([0,1]^d)}.$$

Proof. Note that $I_{\boldsymbol{\ell}} u \in \mathcal{Q}^1(\mathcal{T}_{\boldsymbol{\ell}}^{\otimes})$ by definition. This implies that $I_\ell^d u \in \mathcal{X}_\ell$. Thus, the Céa Lemma and Theorem 7.3 show the statement. ■

Analogously to the proof of Lemma 7.6, we obtain that

$$\dim(\mathcal{X}_\ell \cap H^1_0([0,1]^d)) \lesssim d(\ell + d)^{d-1} 2^\ell.$$

7.1.1 Examples of high-dimensional PDEs

There are many examples of high-dimensional PDEs in practical applications such as finance, physics, and chemistry.

Electron-Schrödinger equation

One notable example is the Schrödinger eigenvalue problem: Given $n \in \mathbb{N}$ electrons and $m \in \mathbb{N}$ nuclei, the goal is to to find the wave function $\psi \colon \mathbb{R}^{3n} \to \mathbb{C}$ which gives a probability density of the position $x_i \in \mathbb{R}^3$ of the i-th electron. The wave function is a solution of the problem

$$\begin{aligned} -\frac{1}{2} \underbrace{\sum_{i=1}^n \Delta_{x_i} \psi(x_1, \dots, x_n)}_{\text{Laplace in every dimension } x_i} + \Big(\underbrace{- \sum_{i=1}^n \sum_{j=1}^m \frac{Z_j}{|x_i - R_j|^2}}_{\text{force between electrons and nuclei}} + \underbrace{\sum_{i=1}^n \sum_{j=i+1}^n \frac{1}{|x_i - x_j|^2}}_{\text{force between electrons}} \Big) \psi(x_1, \dots, x_n) \\ = E \psi(x_1, \dots, x_n). \end{aligned}$$

The position of the nuclei of the atoms is given by $R_j \in \mathbb{R}^3$ and Z_j is the charge of the j-th nucleus. Finally, $E \in \mathbb{C}$ is the eigenvalue of the wave-function ψ. The first part of the operator (the Laplacian) is often abbreviated with T and the remaining part with V. This allows us to write the equation as

$$(T + V)\psi = E\psi.$$

We do not yet know how to solve eigenvalue problems, however in the simplified setting

$$(T + V)\psi = f$$

for some right-hand side f and $Z_j < 0$ for all $j = 1, \dots, m$, we can derive a weak formulation analogously to the previous chapters. (Note that a negative charge is not physical for a nucleous, however, for $Z_j > 0$ one needs Fredholm theory to show well-posedness of the weak form.) This results in a problem with $d = 3n$ and hence standard FEM is out of the question even for a moderate number of electrons. It thus makes sense to approximate the solution ψ in the sparse grid space $\mathcal{X}_\ell$.

Black-Scholes equation

Another example is the Black-Scholes equation in finance. The Black-Scholes equation is a PDE which describes the price $u(t, y_1, \dots, y_d)$ of a financial derivative (in our case a *European basket option*) as a function of the prices y_i, $i = 1, \dots, d$ of the underlying assets (e.g., some stocks) and time. A European basket option is a product which gives the owner the right to buy a collection of d stocks (the basket) at a fixed price $K > 0$ at a fixed time $T > 0$ in the future. Obviously, if the price of the d stocks in the basket at time T is larger than K, then the owner of the option makes a profit (as they buy it for price K). This profit (and hence the fair price of the option at time T) is given by

$$u(T, y_1, \dots, y_d) = \max\left(\sum_{i=1}^{d} y_i - K, 0\right).$$

Since we would like to buy the option at some time $t < T$ and hopefully make a profit, the goal is to find the fair price $u(t, y_1, \dots, y_d)$ (and hope to buy it cheaper). Since the prices y_i are always assumed to be positive, one may replace $x_i := \log(y_i) \in \mathbb{R}$ to obtain the (log-transformed) Black-Scholes equation. This is a parabolic PDE given by

$$\partial_t u(t, x) + \frac{1}{2}\mathrm{div}(C\nabla u(t, x)) + \boldsymbol{b} \cdot \nabla u(t, x) - r u(t, x) = 0 \quad \text{for } (t, x) \in [0, \infty] \times \mathbb{R}^d.$$

The parameters $r > 0$, $C \in \mathbb{R}^{d\times d}$, and $\boldsymbol{b} \in \mathbb{R}^d$ depend on specific assumptions on the financial market. For example C encodes the correlation between different stocks and $r > 0$ is the interest rate at which one can borrow cash to buy stocks.

While we don't know how to solve time-dependent PDEs yet, we see that the x-dependent part is a PDE in d dimensions. Hence, standard FEM is not applicable even for a moderate number of stocks. It thus makes sense to approximate $u(t, \cdot)$, for fixed $t > 0$, in the sparse grid space $\mathcal{X}_\ell$.

7.2 Neural Networks for solving high-dimensional problems

Artificial neural networks are extremely versatile and have been shown to posses almost all the approximation characteristics that classical methods (sparse grids, polynomial interpolation, FEM, ...) enjoy. On the downside, they are mathematically harder to study and often lead to non-linear, non-convex interpolation problems.

7.2.1 Definition of Neural Networks

Artificial Neural Networks (or just networks or neural networks in the following) are a class of functions $F\colon \mathbb{R}^s \to \mathbb{R}^{s'}$ which can be represented by a number of parameters (also often called *weights*). In that regard, neural networks are no different to the class of polynomial functions, or the class of rational functions.

Directed acyclic graphs

A directed acyclic graph (DAG) is a finite set of vertices $\mathcal{V}$ and edges $\mathcal{E}$ such that each edge $(v, w) \in \mathcal{E}$ connects a vertex $v \in \mathcal{V}$ with a vertex $w \in \mathcal{V}$. The graph is called *directed* if the edges have a direction, i.e., $(v, w) \neq (w, v)$ in general. A *path* in a graph is a sequence of vertices $v_0, \dots, v_n$ such that $(v_i, v_{i+1}) \in \mathcal{E}$ for all $i = 0, \dots, n-1$. A *cycle* is a path such that $v_0 = v_n$. A graph is called *acyclic* if it has no cycles.

We call

$$\mathcal{I} := \left\{ v \in \mathcal{V} \,\middle|\, (w, v) \notin \mathcal{E} \text{ for all } w \in \mathcal{V} \right\} \subseteq \mathcal{V}$$

the input of the DAG and

$$\mathcal{O} := \left\{ v \in \mathcal{V} \,\middle|\, (v, w) \notin \mathcal{E} \text{ for all } w \in \mathcal{V} \right\} \subseteq \mathcal{V}$$

the output.

Lemma 7.8. *The input and output of a DAG are nonempty and each vertex $v \in \mathcal{V}$ is connected to $\mathcal{I}$ and $\mathcal{O}$ by at least one path. If each vertex in $\mathcal{V}$ has at least one edge connected to it, $\mathcal{I}$ and $\mathcal{O}$ are also disjoint.*

Proof. Pick an arbitrary vertex $v \in \mathcal{V}$ and build the longest path $v_0, \dots, v_n$ that contains $v = v_i$ for some $0 \leq i \leq n$. Since the path is maximal, there is no edge $(w, v_0) \in \mathcal{E}$ for any $w \in \mathcal{V}$. Thus, $v_0 \in \mathcal{I}$. Similarly, there is no edge $(v_n, w) \in \mathcal{E}$ for any $w \in \mathcal{V}$, so $v_n \in \mathcal{O}$. If $\mathcal{I} \cap \mathcal{O} \neq \emptyset$, there exists $z \in \mathcal{I} \cap \mathcal{O}$ with $(z, w), (w, z) \notin \mathcal{E}$ for all $w \in \mathcal{V}$. This contradicts the assumption that each vertex has at least one edge connected to it. ■

As is shown in Lemma 7.8, each vertex $v \in \mathcal{V}$ is connected to the input by at least one path of some length $n \in \mathbb{N}$. Since $(\mathcal{V}, \mathcal{E})$ has no cycles, the longest possible path has length $\#\mathcal{V}$. Thus, there exists a maximal path length $m \leq \#\mathcal{V}$ with $v_0, \dots, v_m \in \mathcal{V}$ such that $v_0 \in \mathcal{I}$ and $v_m = v$. With this, we may define the layers of a DAG as follows

$$\mathcal{L}(\ell) := \left\{ v \in \mathcal{V} \,\middle|\, \max\left\{ n \in \mathbb{N} \,\middle|\, v_0, \dots, v_n \in \mathcal{V},\, v_0 \in \mathcal{I},\, v_n = v \right\} = \ell \right\}.$$

Lemma 7.9. *There holds $\mathcal{V} = \bigcup_{\ell \in \mathbb{N}} \mathcal{L}(\ell)$ and $\mathcal{L}(\ell) \cap \mathcal{L}(\ell') = \emptyset$ for $\ell \neq \ell'$. There exists $L \in \mathbb{N}$ with $\mathcal{L}(\ell) = \emptyset$ for all $\ell > L$ and $\mathcal{L}(L) \subseteq \mathcal{O}$.*

Proof. Since the length of the longest path is unique (even if the longest path itself might not be), the $\mathcal{L}(\ell)$ are disjoint by definition. As argued above, each vertex has a longest path connecting it to the input $\mathcal{I}$. Thus, $\mathcal{V} = \bigcup_{\ell \in \mathbb{N}} \mathcal{L}(\ell)$. Since $\mathcal{V}$ is finite, there must exist minimal $L \in \mathbb{N}$ with $\mathcal{L}(\ell) = \emptyset$ for all $\ell > L$. Assume that $v \in \mathcal{L}(L)$ and some $w \in \mathcal{V}$ satisfy $(v, w) \in \mathcal{E}$. Then, there would exist a path of length $L + 1$ connecting w to the input, which contradicts $\mathcal{L}(\ell) = \emptyset$ for all $\ell > L$. Hence, $\mathcal{L}(L) \subseteq \mathcal{O}$. ■

Artificial Neural Networks

An artificial neural network (or just neural network) is a DAG $(\mathcal{V}, \mathcal{E})$ together with a weight vector $\boldsymbol{W} \in \mathbb{R}^{\mathcal{E}}$, a bias vector $\boldsymbol{b} \in \mathbb{R}^{\mathcal{V}}$, and an activation function $\phi\colon \mathbb{R} \to \mathbb{R}$. Popular activation functions are

- *ReLU:* $\phi(x) = \max\{x, 0\}$
- *leaky-ReLU:* $\phi(x) = \max\{x, \delta x\}$ for some $0 < \delta \ll 1$.
- *sigmoid:* $\phi(x) = 1/(1 + e^x)$
- *swish:* $\phi(x) = x/(1 + e^{-x})$
- *softplus:* $\phi(x) = \log(1 + e^x)$

Note that the activation function must be non-linear, otherwise the network would be equivalent to an affine function as we will see below.

Remark. Note that in practical applications, all sorts of (and combinations of) activation functions have proven themselves useful. In the mathematical analysis of neural networks, the choice of ϕ often doesn't make a real difference and one sticks with simple choices such as ReLU. □

Since each layer is finite, it makes sense to introduce an enumeration of their elements, i.e.,

$$\mathcal{L}(\ell) = \{v_1^{(\ell)}, \dots, v_{\#\mathcal{L}(\ell)}^{(\ell)}\}.$$

The realization function $\mathcal{R}$ maps a neural network $\mathcal{N} = (\mathcal{V}, \mathcal{E}, \boldsymbol{W}, \boldsymbol{b}, \phi)$ to a function $\mathcal{R}(\mathcal{N})\colon \mathbb{R}^{\#\mathcal{I}} \to \mathbb{R}^{\#\mathcal{O}}$ as follows: Define $\mathcal{R}_\ell\colon \mathbb{R}^{\#\mathcal{I}} \to \mathbb{R}^{\#\mathcal{L}(\ell)}$ iteratively by $\mathcal{R}_0 := \mathrm{id}$ and for $x \in \mathbb{R}^{\#\mathcal{I}}$

$$(\mathcal{R}_\ell(x))_i := \phi\Big(\sum_{\substack{j=1 \\ (v_j^{(\ell-1)}, v_i^{(\ell)}) \in \mathcal{E}}}^{\#L(\ell-1)} \boldsymbol{W}_{(v_j^{(\ell-1)}, v_i^{(\ell)})} \mathcal{R}_{\ell-1}(x)_j + \boldsymbol{b}_{v_i^{(\ell)}} \Big) \quad \text{for all } \ell = 1, \dots, L-1.$$

Finally, we define

$$(\mathcal{R}(\mathcal{N})(x))_i := \sum_{\substack{j=1 \\ (v_j^{(L-1)}, v_i^{(L)}) \in \mathcal{E}}}^{\#L(L-1)} \boldsymbol{W}_{(v_j^{(L-1)}, v_i^{(L)})} \mathcal{R}_{L-1}(x)_j + \boldsymbol{b}_{v_i^{(L)}}.$$

Feed forward networks

A very common subset of neural networks are *feed forward* networks. They correspond to DAGs, where

$$\mathcal{E} = \bigcup_{\ell=0}^{L-1} \big\{(v, w) \in \mathcal{V} \times \mathcal{V} \,\big|\, v \in \mathcal{L}(\ell),\, w \in \mathcal{L}(\ell+1)\big\},$$

i.e., all nodes of subsequent layers are connected to each other, but there are no connections that skip layers. Due to this simpler structure, feed forward networks can be written as iterated matrix-vector products. We define the weight matrices

$$\boldsymbol{W}_\ell \in \mathbb{R}^{\#\mathcal{L}(\ell)\times\#\mathcal{L}(\ell-1)},\ (\boldsymbol{W}_\ell)_{i,j} := \boldsymbol{W}_{(v_j^{(\ell-1)},v_i^{(\ell)})} \quad \text{for all } \ell = 1,\dots,L$$

and the biases

$$\boldsymbol{b}_\ell \in \mathbb{R}^{\#\mathcal{L}(\ell)},\ (\boldsymbol{b}_\ell)_i := \boldsymbol{b}_{v_i^{(\ell)}} \quad \text{for all } \ell = 1,\dots,L.$$

The realization of a feed-forward network $\mathcal{N}$ simplifies to $\mathcal{R}_0 := \mathrm{id}$ and

$$\mathcal{R}_\ell(x) = \phi(\boldsymbol{W}_\ell \mathcal{R}_{\ell-1}(x) + \boldsymbol{b}_\ell) \quad \text{for all } \ell = 1,\dots,L-1$$

and $\mathcal{R}(\mathcal{N})(x) := \boldsymbol{W}_L \mathcal{R}_{L-1}(x) + \boldsymbol{b}_L$.

Training of neural networks

In the following, we will fix the DAG and just consider the weights and biases. Thus, we write $\mathcal{R}(\boldsymbol{W},\boldsymbol{b})$ instead of $\mathcal{R}(\mathcal{N})$. Note that fixing the DAG determines a set of admissible weights and biases, i.e.,

$$(\boldsymbol{W},\boldsymbol{b}) \in \mathcal{W}(\mathcal{N}) := \mathbb{R}^{\mathcal{E}} \times \mathbb{R}^{\mathcal{V}}.$$

As with interpolation in polynomial spaces, one can try to approximate given data with neural networks. Given $x_1,\dots,x_N \in \mathbb{R}^{\#\mathcal{I}}$ and $y_1,\dots,y_N \in \mathbb{R}^{\#\mathcal{O}}$, the approximation problem is to find weights $\boldsymbol{W}$ and biases $\boldsymbol{b}$ such that

$$\mathcal{L}(\boldsymbol{W},\boldsymbol{b}) := \sum_{i=1}^{N} |\mathcal{R}(\boldsymbol{W},\boldsymbol{b})(x_i) - y_i|^2 \to \min. \tag{7.7}$$

The function $\mathcal{L}(\boldsymbol{W},\boldsymbol{b})$ is called the *loss function* (note that there are many other useful definitions of loss, but the least-square loss is the most common).

Due to the non-linearity of $(\boldsymbol{W},\boldsymbol{b}) \mapsto \mathcal{R}(\boldsymbol{W},\boldsymbol{b})$, we have to use a non-linear optimization method. One of these methods is *Gradient descent*.

Algorithm 7.10. *Input: function $f\colon \mathbb{R}^d \to \mathbb{R}$, starting value $w_0 \in \mathbb{R}^d$ and step-size $\alpha > 0$. For $\ell = 0,1,2,\dots$ do:*

1. *Compute gradient $G_\ell := \nabla_w f(w_\ell) \in \mathbb{R}^d$.*
2. *Update $w_{\ell+1} = w_\ell - \alpha G_\ell$.*

Output: sequence of approximations w_ℓ of the minimizer of f.

Remark. Obviously, the gradient descent algorithm (Algorithm 7.10) can be applied to minimize $\mathcal{L}(\boldsymbol{W},\boldsymbol{b})$ by embedding $(\boldsymbol{W},\boldsymbol{b}) \in \mathbb{R}^{\#\mathcal{E}+\#\mathcal{V}}$ and setting $w_\ell = (\boldsymbol{W}_\ell,\boldsymbol{b}_\ell)$. □

Remark. Note that for non-linear, non-convex optimization, convergence of $(\boldsymbol{W}_\ell, \boldsymbol{b}_\ell)$ to the true minimizer is all but certain. This is the reason, why most results about approximation by neural networks fall into one of the following two categories: (1) Results that show that a neural network with certain approximation properties exists and (2) results that show that a certain optimization algorithm will find a network with certain approximation properties. Naturally, category 2 is much harder to prove than category 1. □

Remark. Note that practical implementations of machine learning often use *Stochastic Gradient Descent* instead of plain gradient descent. The only difference to Algorithm 7.10 is that instead of computing $G_\ell = \nabla_{(\boldsymbol{W},\boldsymbol{b})}\mathcal{L}(\boldsymbol{W},\boldsymbol{b})$, one randomly selects $m \ll n$ (the so-called *batch size*) indices $1 \le i_1, \dots, i_m \le N$ and computes

$$G_\ell^{\text{stoch}} := \frac{n}{m}\nabla_{(\boldsymbol{W}_\ell,\boldsymbol{b}_\ell)} \sum_{j=1}^{m} |\mathcal{R}(\boldsymbol{W}_\ell,\boldsymbol{b}_\ell)(x_{i_j}) - y_{i_j}|^2 \approx \nabla_{(\boldsymbol{W}_\ell,\boldsymbol{b}_\ell)}\mathcal{L}(\boldsymbol{W}_\ell,\boldsymbol{b}_\ell)$$

in Step (1) of Algorithm 7.10. Often, one uses $m = 32$. The algorithm has several practical advantages such as:

- more optimization steps for the same cost,
- often more efficient as the batch size can be optimized such that G_ℓ^{stoch} can be computed in the fast memory close to the processor,
- stochastic nature of G_ℓ^{stoch} can prevent getting stuck at local minima.

The mathematical analysis of stochastic gradient descent is very similar to plain gradient descent, since one can rely on the fact

$$\mathbb{E}G_\ell^{\text{stoch}} = \nabla_{(\boldsymbol{W}_\ell,\boldsymbol{b}_\ell)}\mathcal{L}(\boldsymbol{W}_\ell,\boldsymbol{b}_\ell) = G_\ell.$$

□

7.2.2 Approximation of PDEs with neural networks

The Deep Ritz method

Many partial differential equations are derived from energy minimization problems. For example, the Laplace equation is derived from the minimization of the Dirichlet energy

$$J(u) := \frac{1}{2}\int_\Omega |\nabla u|^2\,dx - \int_\Omega f u\,dx$$

for given $f\colon \Omega \to \mathbb{R}$. Instead of searching for the energy minimizer u in a FEM space, we can search for a neural network that minimizes the energy. This is the idea behind the *Deep Ritz method*. It trys to solve the problem

$$\mathcal{L}(\boldsymbol{W},\boldsymbol{b}) := \frac{1}{2}\int_\Omega |\nabla\mathcal{R}(\boldsymbol{W},\boldsymbol{b})(x)|^2\,dx - \int_\Omega f(x)\mathcal{R}(\boldsymbol{W},\boldsymbol{b})(x)\,dx \to \min.$$

If we restrict the minimization to $H^1(\Omega)$, minimizing $J(u)$ is equivalent to solving the Neumann problem

$$-\Delta u = f \quad \text{with} \quad \partial_n u = 0 \text{ on } \partial\Omega.$$

If instead we want to solve the Dirichlet problem, we have to include a penalty term for the deviation from the boundary condition, i.e., we have to minimize

$$\mathcal{L}(\boldsymbol{W}, \boldsymbol{b}) := \frac{1}{2}\int_\Omega |\nabla \mathcal{R}(\boldsymbol{W}, \boldsymbol{b})(x)|^2 \, dx - \int_\Omega f(x)\mathcal{R}(\boldsymbol{W}, \boldsymbol{b})(x)\, dx + \beta \|u\|^2_{L^2(\partial\Omega)} \to \min,$$

where $\beta > 0$ penalizes $u \neq 0$ on $\partial\Omega$.

Application to Eigenvalue problems

Eigenvalue problems are another class of PDEs that can be solved with neural networks. Given a Hilbert space V and a self-adjoint, positive definite operator $A\colon V \to V^\star$, we search for the minimal eigenvalue-eigenfunction pair $\lambda > 0$, $v_\lambda \in V$ with $Av_\lambda = \lambda v_\lambda$ and $\|v_\lambda\|_V = 1$. The idea is to search for a neural network that minimizes the Rayleigh quotient:

$$\mathcal{L}(\boldsymbol{W}, \boldsymbol{b}) := \frac{\langle A\mathcal{R}(\boldsymbol{W}, \boldsymbol{b}) \,;\, \mathcal{R}(\boldsymbol{W}, \boldsymbol{b})\rangle}{\|\mathcal{R}(\boldsymbol{W}, \boldsymbol{b})\|_V^2} \to \min.$$

Lemma 7.11. *Let H denote a Hilbert space with $V \subseteq H \subseteq V^\star$ such that $A^{-1}\colon H \to H$ is well-defined, compact, self-adjoint, and positive definite. Then, the minimizer of $J(v) := \langle Av \,;\, v\rangle / \|v\|_H^2$ is an eigenfunction of A with minimal eigenvalue.*

Proof. Since A is self-adjoint and compact, also $A^{-1}\colon H \to H$ is self-adjoint and compact. The spectral theorem of self-adjoint and compact operators gives an orthonormal basis $(b_i)_{i\in\mathbb{N}}$ of H such that $A^{-1}b_i = \mu_i b_i$ with $\mu_1 \geq \mu_2 \geq \ldots \geq 0$. Since A is invertible and positive definite, any eigenpair (x, λ) of A satisfies $\lambda > 0$ and $A^{-1}x = x/\lambda$. Conversely, any eigenpair (y, μ) of A^{-1} must satisfy $Ay = y/\mu$. Hence, the eigenvalues of A are given by $\lambda_i := 1/\mu_i$ and the eigenfunctions are the same. Given $x \in H$ with $x = \sum_{i=1}^\infty \alpha_i b_i$ and $\|x\|_H^2 = \sum_{i=1}^\infty \alpha_i^2 = 1$, we have

$$\langle Ax \,;\, x\rangle = \sum_{i=1}^\infty \lambda_i \alpha_i^2 \geq \lambda_1 \sum_{i=1}^\infty \alpha_i^2 = \lambda_1.$$

Moreover, $x := b_1$ attains the minimum $J(x) = \lambda_1$. ■

Physics informed neural networks (PINNs)

Given an (nonlinear) operator $L\colon X \to Y$ with a normed space Y, we can consider the equation: Given $f \in Y$, find $u \in X$ such that

$$L(x) = f.$$

If the solution is unique, we can equivalently search for the minimizer of $\|L(x) - f\|_Y$ for $x \in \mathbb{X}$. This corresponds to the optimization problem

$$\mathcal{L}(\boldsymbol{W}, \boldsymbol{b}) := \|L(\mathcal{R}(\boldsymbol{W}, \boldsymbol{b})) - f\|_Y^2 \to \min .$$

The problem with this approach is that in many applications Y is the dual space of some Banach or Hilbert space (e.g., $Y = H^{-1}(\Omega)$). In this case, the norm $\|\cdot\|_Y$ is not easily computable and often replaced by some other norm. This can introduce systematic errors in the approximation.

Weak adversarial method

This method circumvents the problem of the dual space norm by using a weak form of the PDE. We assume that we have normed spaces X and Y and a weak form of the equation given by: Find $u \in X$ such that

$$a(u, v) = f(v) \quad \text{for all } v \in Y. \tag{7.8}$$

Note that we do not assume that $a(\cdot, \cdot)$ is linear or elliptic, just that a unique solution $u \in X \setminus \{0\}$ exists. The weak adversarial method consists in searching for one neural network the represents the solution, and for another one that tries to find the error, i.e., we solve

$$\inf_{(\boldsymbol{W}, \boldsymbol{b}) \in \mathcal{W}(\mathcal{N})} \sup_{(\boldsymbol{W}', \boldsymbol{b}') \in \mathcal{W}(\mathcal{N})} \frac{\Big(a\big(\mathcal{R}(\boldsymbol{W}, \boldsymbol{b})(\cdot), \mathcal{R}(\boldsymbol{W}', \boldsymbol{b}')(\cdot)\big) - f\big(\mathcal{R}(\boldsymbol{W}', \boldsymbol{b}')(\cdot)\big)\Big)^2}{\|\mathcal{R}(\boldsymbol{W}, \boldsymbol{b})(\cdot)\|_X^2 \|\mathcal{R}(\boldsymbol{W}', \boldsymbol{b}')(\cdot)\|_Y^2}.$$

Lemma 7.12. *Let $(u, v) \in X \times Y$ solve*

$$\frac{(a(u, v) - f(v))^2}{\|u\|_X^2 \|v\|_Y^2} = \inf_{u' \in X} \sup_{v' \in \mathcal{Y}} \frac{(a(u', v') - f(v'))^2}{\|u'\|_X^2 \|v'\|_Y^2}.$$

Then, u is the solution of (7.8).

Proof. We know that (7.8) has a unique solution that we call u_0 for the moment. We need to show $u = u_0$. There holds

$$0 \le \inf_{u' \in X} \sup_{v' \in \mathcal{Y}} \frac{(a(u', v') - f(v'))^2}{\|u'\|_X^2 \|v'\|_Y^2} \le \sup_{v' \in \mathcal{Y}} \frac{(a(u_0, v') - f(v'))^2}{\|u_0\|_X^2 \|v'\|_Y^2} = 0.$$

This implies

$$\frac{(a(u, v') - f(v'))^2}{\|u\|_X^2 \|v'\|_Y^2} = 0 \quad \text{for all } v' \in Y,$$

which is equivalent to (7.8). Since the solution is unique, we have $u = u_0$ and conclude the proof. ■

7.2.3 The elephant in the room: Quadrature

We observe that all examples above require us to compute integrals over Ω of some form with the integrand being a neural network (norms, scalar products,...). This is a non-trivial task for several reasons:

- Classical quadrature works well in low dimensions and with regularity assumptions on the integrand. In both cases, we would rather use classical methods such as FEM to solve the PDEs, as they will be much faster.
- We could use Monte-Carlo quadrature without requiring any regularity of the integrand. But this gives only a very slow convergence rate of $n^{-1/2}$ on average. Sparse grid quadrature methods require a lot of regularity of the integrand to work in high dimensions and if this is the case, then one would rather use sparse grid FEM.
- The following theorem shows that the slow Monte-Carlo convergence rate cannot be improved by much in high dimensions and hence all PDE approximation methods based on neural networks will suffer from this slow convergence rate. Therefore, neural network based methods can only compete in very high dimensional problems where classical methods (even sparse grids) fail.

We require a geometric observation: Given a $(d+1)$-dimensional cube $\widehat{Q}$ of sidelength a with $2(d+1)$-hyperfaces $Q_1, \dots, Q_{2d+2}$, we define subsets

$$P_i := \big\{ x \in Q \,\big|\, \operatorname{dist}(x, Q_i) = \min_{j=1,\dots,2d+2} \operatorname{dist}(x, Q_j) \big\}.$$

The objects P_i are called (hyper-)pyramids over their respective bases Q_i. Analogously, given any hypercube $Q \subseteq \mathbb{R}^d$, we can construct the cube $\widehat{Q} \subseteq \mathbb{R}^{d+1}$ with rotated and translated copies of Q as faces. The corresponding pyramid with base Q is then denoted by P_Q.

Lemma 7.13. *Given a cube $Q \subseteq \mathbb{R}^d$ of sidelength $a > 0$, the $(d+1)$-dimensional volume of P_Q is given by $\frac{1}{2(d+1)} a^{d+1}$ and the height of the pyramid over a point $x' \in Q$ is given by* $\operatorname{dist}(x', \partial Q)$.

Proof. By construction, the P_i are disjoint and $\widehat{Q} = \bigcup_{i=1}^{2(d+1)} P_i$. This implies the volume formula. Let $x = (x', h) \in \mathbb{R}^{d+1}$ with $x' \in Q_i$ and let Q_j be any other face of $\widehat{Q}$ apart from the opposite one.Since $F := Q_i \cap Q_j \subseteq \partial Q_i$, we have $\operatorname{dist}(x', \partial Q_i) \leq \operatorname{dist}(x', F)$. Since Q_j is orthogonal to Q_i, it is parallel to the $(d+1)$st dimension. This implies

$$\operatorname{dist}(x, Q_j) = \operatorname{dist}(x', F) \geq \operatorname{dist}(x', \partial Q_i) \quad \text{and} \quad \operatorname{dist}(x, Q_i) = h. \tag{7.9}$$

Thus, the fact $h < \operatorname{dist}(x', \partial Q_i)$ implies that $\operatorname{dist}(x, Q_i)$ is minimal among the distances to the faces of $\widehat{Q}$ apart from the opposite one. The distance to the opposite face, however, is given by $a - h$. Since $h < \operatorname{dist}(x', \partial Q_i) \leq a/2$, there holds $a - h > h$, and we conclude $x \in P_i$.

On the other hand, if we choose Q_j such that $\operatorname{dist}(x', F) = \operatorname{dist}(x', \partial Q_i)$, we have equality in (7.9). Thus, $h > \operatorname{dist}(x', \partial Q_i)$ implies $x \notin P_i$ (since the distance to Q_j is definitely smaller). This shows that the height of the pyramid over x' is given by $\operatorname{dist}(x', \partial Q_i)$ and concludes the proof. ■

Theorem 7.14. *Let $\Omega = [0,1]^d$. Then, for each set of quadrature points $z_1, \dots, z_n \in \Omega$, there exist a neural network with activation function $\phi(x) = \max\{x, 0\}$ (ReLU), $\#\mathcal{I} = d$, $\#\mathcal{O} = 1$, and weights $(\boldsymbol{W}, \boldsymbol{b}) \in \mathcal{W}(\mathcal{N})$ such that $\|\mathcal{R}(\boldsymbol{W}, \boldsymbol{b})\|_{L^\infty(\Omega)} \leq 1/2$, $\|\nabla \mathcal{R}(\boldsymbol{W}, \boldsymbol{b})\|_{L^\infty(\Omega)} \leq 1$,*

$$\int_\Omega \mathcal{R}(\boldsymbol{W}, \boldsymbol{b})(x)\, dx \geq \frac{1}{2(d+1)} \lceil (n+1)^{1/d} \rceil^{-d-1} \quad \text{and} \quad \mathcal{R}(\boldsymbol{W}, \boldsymbol{b})(z_i) = 0 \quad \text{for all } i = 1, \dots, n.$$

Moreover, the number of parameters (weights and biases) is bounded my $\mathcal{O}(d^2)$. In particular, for fixed dimension d, any quadrature formula $\sum_{i=1}^n w_i \mathcal{R}(\boldsymbol{W}, \boldsymbol{b})(z_i)$ with arbitrary weights $w_i \in \mathbb{R}$ has an error of at least $\mathcal{O}(n^{-(d+1)/d})$.

Proof. For simplicity of presentation, we assume that $\log_2(d) \in \mathbb{N}$, i.e., that d is a power of two. The general case can be treated analogously, with a bit more bookkeeping.

Step 1: First, we observe that $\min\{a_1, a_2\} = -\max\{-a_1, -a_2\} = a_1 - \max\{a_1 - a_2, 0\}$. Given $a_1, \dots, a_m \in \mathbb{R}$, this implies the recursive formula

$$\begin{aligned}\min\{a_1, \dots, a_m\} &= \min\{\min\{a_1, \dots, a_{m/2}\}, \min\{a_{m/2+1}, \dots, a_m\}\} \\ &= \min\{a_1, \dots, a_{m/2}\} - \max\{\min\{a_1, \dots, a_{m/2}\} - \min\{a_{m/2+1}, \dots, a_m\}, 0\}.\end{aligned}$$

We can use this formula to construct a neural network that computes the minimum of $m = 2^k$ numbers. To that end, define the weights and biases

$$\boldsymbol{W}_{m,0} = \begin{pmatrix} 1 & 0 & 0 & & \\ -1 & 0 & 0 & & \\ 1 & -1 & 0 & & \\ & & \ddots & & \\ & & 0 & 1 & 0 \\ & & 0 & -1 & 0 \\ & & 0 & 1 & -1 \end{pmatrix} \in \mathbb{R}^{3m/2 \times m}, \quad \boldsymbol{b}_{m,0} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{3m/2},$$

$$\boldsymbol{W}_{m,1} = \begin{pmatrix} 1 & -1 & -1 & 0 & \cdots & & & \\ 0 & 0 & 0 & 1 & -1 & -1 & 0 & \cdots \\ & & \ddots & & & & & \end{pmatrix} \in \mathbb{R}^{m/2 \times 3m/2}, \quad \boldsymbol{b}_{m,1} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{m/2}.$$

Note that for all $x = (x_1, \dots, x_m) \in \mathbb{R}^m$, there holds

$$\begin{aligned}&\mathcal{R}((\boldsymbol{W}_{m,0}, \boldsymbol{W}_{m,1}), (\boldsymbol{b}_{m,0}, \boldsymbol{b}_{m,1}))(x) \\ &\quad = \boldsymbol{W}_{m,1} \phi((x_1, -x_1, x_1 - x_2, \dots, x_{m-1}, -x_{m-1}, x_{m-1} - x_m)) \\ &\quad = \big(\phi(x_1) - \phi(-x_1) - \phi(x_1 - x_2), \dots, \phi(x_{m-1}) - \phi(-x_{m-1}) - \phi(x_{m-1} - x_m)\big) \\ &\quad = (x_1 - \max\{x_1 - x_2, 0\}, \dots, x_{m-1} - \max\{x_{m-1} - x_m, 0\}) \\ &\quad = (\min\{x_1, x_2\}, \dots, \min\{x_{m-1}, x_m\}) \in \mathbb{R}^{m/2}.\end{aligned}$$

Thus, with $\widetilde{\boldsymbol{W}} := (\boldsymbol{W}_{m,0}, \boldsymbol{W}_{m,1}, \boldsymbol{W}_{m/2,0}, \boldsymbol{W}_{m/2,1}, \dots, \boldsymbol{W}_{2,0}, \boldsymbol{W}_{2,1})$ and $\widetilde{\boldsymbol{b}} := (\boldsymbol{b}_{m,0}, \boldsymbol{b}_{m,1}, \dots, \boldsymbol{b}_{2,0}, \boldsymbol{b}_{2,1})$, we have for all $x \in \mathbb{R}^m$

$$\mathcal{R}(\widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{b}})(x) = \min\{x_1, \dots, x_m\} \in \mathbb{R}.$$

Step 2: Let $n_0 = \lceil (n+1)^{1/d} \rceil^d \geq n+1$ denote the smallest number such that $n_0^{1/d} \in \mathbb{N}$. We partition Ω into n_0 cubes $I_1, \dots, I_{n_0}$ of sidelength $\alpha \geq n_0^{-1/d}$. The pigeonhole principle implies that there exists $j \in \{1, \dots, n_0\}$ with $I_j \cap \{z_1, \dots, z_n\} = \emptyset$. We define $\beta_i > 0$ such that $I_j = \prod_{i=1}^d [\beta_i, \beta_i + \alpha]$. We consider the function

$$G(x) = \min_{i=1,\dots,d} \min\{\max\{x_i - \beta_i, 0\}, \max\{\beta_i + \alpha - x_i, 0\}\}.$$

We construct $\widetilde{\boldsymbol{W}}$ and $\widetilde{\boldsymbol{b}}$ as in Step 1 with $m = 2d$. Moreover, we define

$$\boldsymbol{W}_0 := \begin{pmatrix} 1 & 0 & \cdots & & \\ -1 & 0 & \cdots & & \\ & & \cdots & 0 & 1 \\ & & \cdots & 0 & -1 \end{pmatrix}, \qquad \boldsymbol{b}_0 := \begin{pmatrix} -\beta_1 \\ \beta_1 + \alpha \\ \vdots \\ -\beta_d \\ \beta_d + \alpha \end{pmatrix}.$$

With this, we define $\boldsymbol{W} := (\boldsymbol{W}_0, \boldsymbol{W}_{m,0}, \boldsymbol{W}_{m,1}, \dots, \boldsymbol{W}_{2,0}, \boldsymbol{W}_{2,1})$ and $\boldsymbol{b} := (\boldsymbol{b}_0, \boldsymbol{b}_{m,0}, \boldsymbol{b}_{m,1}, \dots, \boldsymbol{b}_{2,0}, \boldsymbol{b}_{2,1})$ and observe

$$\mathcal{R}(\boldsymbol{W}, \boldsymbol{b})(x) = G(x) \quad \text{for all } x \in \Omega.$$

Step 3: Lets investigate the function G further. For $x \notin I_j$, there holds $G(x) = 0$ and hence $\mathcal{R}(\boldsymbol{W}, \boldsymbol{b})(z_j) = 0$ for all $j = 1, \dots, n$. For $x \in I_j$, we note that $G(x)$ measures the minimal distance of x to the boundary of the cube I_j. According to Lemma 7.13, $G(x)$ measures the height of the pyramid P_{I_j} over $x \in I_j$ (see Figure 7.3 for an illustration). Hence, we have

$$\int_\Omega \mathcal{R}(\boldsymbol{W}, \boldsymbol{b})(x)\, dx = \int_\Omega G(x)\, dx = \operatorname{vol}(P_{I_j}) = \frac{1}{2(d+1)} \alpha^{d+1} = \frac{1}{2(d+1)} n_0^{\frac{d+1}{d}}.$$

By definition, we have $0 \leq G(x) \leq \alpha/2 \leq 1/2$ for all $x \in \Omega$. This implies $\|\mathcal{R}(\boldsymbol{W}, \boldsymbol{b})\|_{L^\infty(\Omega)} \leq 1/2$. Moreover, apart from a set of measure zero, we have

$$\nabla G(x) = 0 \quad \text{or} \quad \nabla G(x) = (0, \dots, 0, \pm 1, 0, \dots, 0),$$

where the position of the non-zero entry depends on which term of the minimum is smallest for the given x. This shows $\|\nabla \mathcal{R}(\boldsymbol{W}, \boldsymbol{b})\|_{L^\infty(\Omega)} \leq 1$. Finally, the number of network parameters is

$$(2d)d + 2d + \sum_{i=1}^{\log_2(2d)} \left(32 \cdot 2^{2i}/2 + 32 \cdot 2^i/2 + 32 \cdot 2^{2i}/4 + 2^i/2 \right) = \mathcal{O}(d^2).$$

This concludes the proof. ∎

Remark. A much more detailed and general version of the above theorem can be found in [Gro]. This shows that, in general, the training of neural networks will suffer from the curse of dimensionality, even if there exist networks that approximate the exact solution very well in theory. □

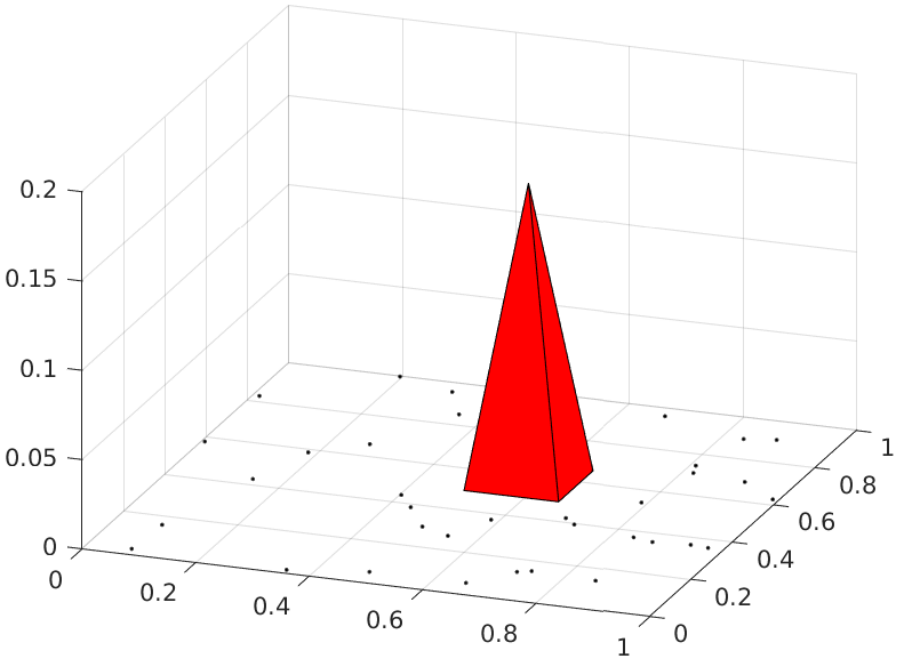


FIGURE 7.3. *Schematic illustration of* $G(x)$ *in red and quadrature nodes* z_j, $j = 1, \ldots, 35$ *in black.*

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.

[Gro] Philipp Grohs and Felix Voigtlaender, *Proof of the theory-to-practice gap in deep learning via sampling complexity bounds for neural network approximation spaces*, Found. Comput. Math., 24, 1085–1143, 2024.

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.* ■