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 implementation are sufficiently accurate. For P1-FEM we will show that it is sufficient to approximate 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 $E \in \mathcal{E}_N$. Such an approximation can be considered as 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*

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

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

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

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

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

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,

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

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

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

Finally, the combination with a triangle inequality yields that

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

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

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

This concludes the proof. ∎

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

Exercise 33. 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 the Poisson problem and a P1-FEM, the bilinear form $\langle\!\langle v_h\,;w_h \rangle\!\rangle$ can be computed analytically. Nevertheless, for the right-hand side in most cases quadrature formulas are needed. 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

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

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 34. 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

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

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 35. 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,

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

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 $h \in L^\infty(\Gamma_N)$, $h|_E := h_E = \mathrm{diam}(E)$, it then holds*

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

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,

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

This concludes the proof. ■

Exercise 36. 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 $\lim\limits_{j\to\infty} x_j$. □

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 9 on page 18, one shows that

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

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*

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

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

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

Second, the triangle inequality gives

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

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 $u_{Dh} \in \mathcal{S}^1(\mathcal{T}|_{\Gamma_D}) := \{ v_h|_{\Gamma_D} \,|\, v_h \in \mathcal{S}^1(\mathcal{T}) \}$.
- 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

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

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 37. 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

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

where the right-hand side is independent of how u_D is actually discretized; see also Exercise 39–40 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

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

Conversely, it holds that

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

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 38. 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 39 (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) $\|P_h v\|_{H^1(\Omega)} \le C_{\rm stab} \|v\|_{H^1(\Omega)}$ 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

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

where $C > 0$ depends only on the stability constant $C_{\rm stab}$, 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

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

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 40. (a) Suppose that $P_h : H^1(\Omega) \to \mathcal{S}^1(\mathcal{T})$ satisfies the properties (i)–(iii) of Exercise 39 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

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

where $C > 0$ depends only on Ω and the stability constant $C_{\rm stab}$. **Hint.** Argue along the lines of Exercise 39.
(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 39. **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 36 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. □

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 $T = \mathrm{conv}\{z_0, \dots, z_d\}$ 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**

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

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

The **diameter** of T is denoted by

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

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., $\mathcal{E}_T := \big\{\mathrm{conv}(M) \,\big|\, M \subseteq \mathcal{K}_T \text{ with } \#M = d\big\}$. 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., $T \cap T' = \mathrm{conv}(M)$ 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

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

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.

5.3.1 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 14, this new definition is (up to some generic constant) equivalent to the definition given in Section 3.2.

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

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

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 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 $\|B_T^{-1}\|_2 \le h_{\mathrm{ref}}/\rho_T \le \rho_T^{-1}$. Norm equivalence on $\mathbb{R}^{d\times d}$ concludes $\|B_T^{-1}\|_F \simeq \|B_T^{-1}\|_2 \le \rho_T^{-1}$. ∎

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.

5.4 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.

5.4.1 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.

$$\mathcal{S}^p(\mathcal{T}) := \{u \in H^1(\Omega) \,|\, u|_T \circ \Phi_T \in \mathcal{P}^p(T_{\mathrm{ref}}) \qquad \forall T \in \mathcal{T}\}, \tag{5.28a}$$
$$\mathcal{S}_0^p(\mathcal{T}) := \mathcal{S}^p(\mathcal{T}) \cap H_0^1(\Omega) \tag{5.28b}$$

Here, the we use the mappings from the reference element $T_{\mathrm{ref}} = [-1, 1]$, $\Phi_T : T_{\mathrm{ref}} \to T$ 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 $\mathcal{P}^p(T_{\mathrm{ref}})$ 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

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

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.

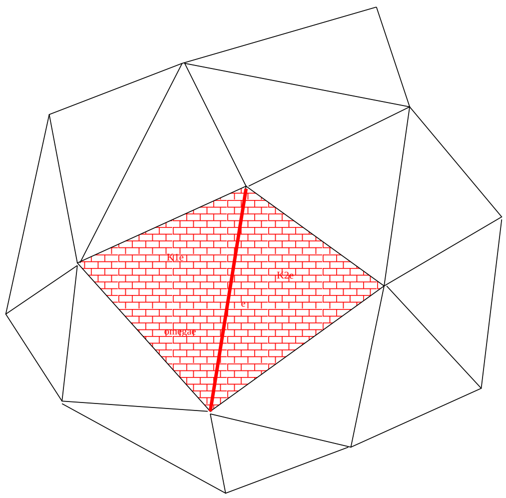


FIGURE 5.1. *The edge E with elements T_E^1, T_E^2 and $\Omega_E = T_E^1 \cup T_E^2$.*

5.4.2 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

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

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.

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

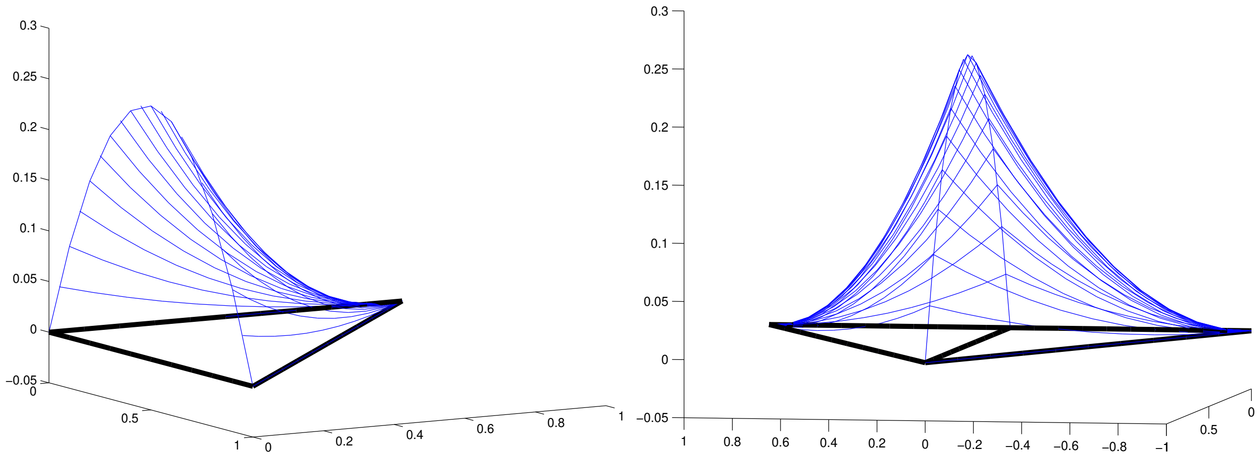
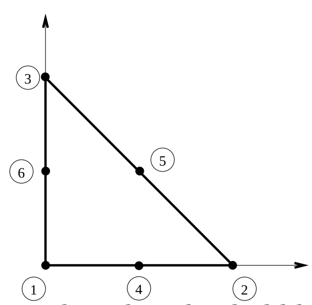


FIGURE 5.2. *Left:* N_4 *on* T_{ref}*. Right:* φ_E*.*



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.

$\square$