Numerical Treatment of Partial Differential Equations

Ralf Hiptmair Christoph Schwab *

Lecture notes for course held by R. Hiptmair in WS03/04

*Seminar for Applied Mathematics, ETH Zürich, {hiptmair,schwab}@sam.math.ethz.ch

Mistakes

We would like you the report if you discover any mistakes in the lecture notes to Patrick Meury by Email. The mistakes which are already known are published on the web on `http://www.math.ethz.ch/undergraduate/lectures/ws0304/math/numpde1/serien`.

If you report a mistake, please include the precise location and description of the mistake and (if possible) the correct version as well. **Please do not send in any attachments!**

Example 0.1 (Report of a Mistake).

```
From: Thomas Wihler <twihler@math.ethz.ch>
To: Patrick Meury <meury@sam.math.ethz.ch>
Subject: Mistake in the lecture notes, p. 25

Hi,

I found a mistake on page 25 of the lecture notes. In equation (0.12)
it says

a = 1

and the correct version would be

a = 2

Regards
Thomas Wihler
```

Every reported mistake is noted on the web page along with the reporting date and (if already done) the date of the correction. The date of the correction refers to the date of a printed release. If the date of a known mistake is younger than the date of your lecture notes, then the mistake is known but not yet corrected in your release.

Version

This release was printed on March 17, 2004.

Contents

1 Abstract Linear Variational Problems

This chapter establishes an abstract framework for the analysis of an important class of numerical schemes for the discretization of boundary value problems for partial differential equations. This framework heavily relies on tools provided by modern *functional analysis*. The study of functional analysis is strongly recommended to anyone who wants to specialize in the numerics of partial differential eqautions.

1.1 Fundamental concepts

From linear algebra we recall the abstract definition of a **vector space** over a field $\mathbb{K}$. Throughout, we will only consider real vector spaces, *i. e.* $\mathbb{K} = \mathbb{R}$. The results can easily be extended to complex vector spaces, *i. e.* the case $\mathbb{K} = \mathbb{C}$. For the remainder of this section the symbols V and W will denote real vector spaces. Please keep in mind that $\mathbb{R}$ can also be regarded as a real vector space.

More precisely, the relevant vector spaces will be **function spaces**. The underlying sets contain functions $\Omega \mapsto \mathbb{R}$, where Ω is a subset of $\mathbb{R}^n$, or equivalence classes of such functions. Addition and scalar multiplication are defined in a pointwise sense.

Example 1.1. Well known are the function spaces $C^m(I)$ of m-times continuously differentiable functions, $m \in \mathbb{N}$, on an interval $I \subset \mathbb{R}$.

Another example, known from elementary measure theory, is the space $L^2(\Omega)$ of square integrable functions (w.r.t. to Lebesgue measure $\mathrm{d}\xi$) on a "domain" $\Omega \subset \mathbb{R}^n$. In this case the elements of the function space are *equivalence classes* of functions that agree dx-almost everywhere in Ω. A generalization are the function spaces $L^p(\Omega)$, $1 \leq p < \infty$, see [26, Ch. 3], and the space $L^\infty(\Omega)$ of essentially bounded functions on Ω.

A subset of a vector space, which is closed with respect to addition and scalar multiplication is called a **subspace**.

A key role in the theory of vector spaces is played by the associated homomorphisms, which are called linear mappings in this particular context.

Definition 1.2. *Let V, W be real vector spaces. A mapping*

- $\mathsf{T} : V \mapsto W$ *is called a* ***(linear) operator****, if $\mathsf{T}(\lambda v + \mu w) = \lambda\,\mathsf{T}\,v + \mu\,\mathsf{T}\,w$ for all $v, w \in V$, $\lambda, \mu \in \mathbb{R}$. If $W = \mathbb{R}$, then T is a* ***linear form****.*

- $\mathsf{b} : V \times V \mapsto \mathbb{R}$ *is called a* ***bilinear form****, if for every* $w \in V$ *both* $v \mapsto \mathsf{b}(w, v)$ *and* $v \mapsto \mathsf{b}(v, w)$ *are linear forms on* V*.*

Obviously linear operators $V \mapsto W$ and bilinear forms $V \times V \mapsto \mathbb{R}$ form vector spaces themselves.

Notation: Symbols for operators and bilinear forms will be set in sans serif fonts.

Terminology distinguishes a few special linear operators

Definition 1.3. *A linear operator* $\mathsf{P} : V \mapsto V$ *on a vector space* V *is called a* ***projection****, if* $\mathsf{P}^2 = \mathsf{P}$*.*

Definition 1.4. *If* U *is a subspace of a vector space* V*, then the operator* $\mathsf{I}_U : U \mapsto V$ *defined by* $\mathsf{I}_U(u) = u$ *for all* $u \in U$ *is called the* ***(canonical) injection****.*

We will need to measure the "size" of an element v of a vector space V. To this end, we introduce a **norm** on V.

Definition 1.5 (Norm on a function space). *A mapping* $\|.\|_V : V \to \mathbb{R}_0^+ := \{\xi \in \mathbb{R}, \xi \geq 0\}$ *is a* ***norm*** *on the vector space* V*, if it satisfies*

$$\|v\|_V = 0 \Longleftrightarrow v = 0 \,, \tag{N1}$$

$$\|\lambda v\|_V = |\lambda| \, \|v\|_V \quad \forall \lambda \in \mathbb{R} \text{ (or } \lambda \in \mathbb{C}), \, \forall v \in V \,, \tag{N2}$$

$$\|w + v\|_V \leq \|u\|_V + \|v\|_V \quad \forall w, v \in V. \tag{N3}$$

Notation: Usually, the norm on a vector space X will be denoted by $\|\cdot\|_X$.

The property (N1) is called definiteness, (N2) is known as homogeneity, and (N3) is the so-called triangle inequality. One readily concludes the inverse triangle inequality

$$|\|v\|_V - \|w\|_V| \leq \|v - w\|_V \quad \forall v, w \in V \,. \tag{1.1}$$

If V is equipped with a norm $\|\cdot\|_V$, it is referred to as a normed vector space $(V, \|\cdot\|_V)$. Every subspace of a normed vector space is understood as a normed vector space with the *same* norm.

Example 1.6. For the particular function spaces introduced above suitable norms are

$$\|u\|_{C^m(I)} := \sum_{l=0}^{m} \sup_{x \in I} \left| \frac{d^k u}{dx^k}(x) \right| \,, \tag{1.2}$$

$$\|u\|_{L^2(\Omega)} := \left(\int_\Omega |u(x)|^2 \, \mathrm{d}\xi \right)^{\frac{1}{2}} \,, \tag{1.3}$$

$$\|u\|_{L^p(\Omega)} := \left(\int_\Omega |u(x)|^p \, \mathrm{d}\xi \right)^{\frac{1}{p}} \,, \quad 1 \leq p < \infty \,, \tag{1.4}$$

$$\|u\|_{L^\infty(\Omega)} := \sup_{x \in \Omega} \operatorname{ess} |f(x)| \,. \tag{1.5}$$

The norm $\|\cdot\|_V$ induces a metric in which the basic operations of addition and scalar multiplications are continuous, and, thus, V becomes a **topological vector space**, see [33, Ch. 1]. This makes it possible to define the convergence of a sequence $\{v_k\}_{k=1}^{\infty} \subset V$:

$$v_k \xrightarrow[k\to\infty]{} v \quad :\Leftrightarrow \quad \lim_{k\to\infty} \|v_k - v\|_V = 0 \,,$$

and closed subsets of $(V, \|\cdot\|_V)$:

Definition 1.7. *A subset $X_0 \subset V$ is **closed**, if every sequence $\{u_n\}_{k=1}^{\infty} \subset X_0$ which converges in V has its limit in X_0:*

$$u_k \xrightarrow[k\to\infty]{} u \quad \Longrightarrow \quad u \in X_0.$$

The topological structure also implies a notion of continuity.

Definition 1.8. *A linear operator $\mathsf{T} : V \mapsto W$ on normed vector spaces V, W is called **continous** or **bounded**, if*

$$\exists \gamma > 0 : \quad \|\mathsf{T}\, v\|_W \leq \gamma \, \|v\|_V \quad \forall v \in V \,.$$

*Then, the **operator norm** of T is given by*

$$\|T\|_{V\mapsto W} := \sup_{v\in V\setminus\{0\}} \frac{\|\mathsf{T}\, v\|_W}{\|v\|_V} \,.$$

*A bilinear form $b : V \times V \mapsto \mathbb{R}$ is **continuous**, if*

$$\exists \gamma > 0 : \quad |\mathsf{b}(v, w)| \leq \gamma \, \|v\|_V \, \|w\|_V \quad \forall v, w \in V \,.$$

The operator norm of b is defined by

$$\|\mathsf{b}\|_{V\times V\mapsto\mathbb{R}} := \sup_{v,w\in V\setminus\{0\}} \frac{|\mathsf{b}(v, w)|}{\|v\|_V \, \|w\|_V} \quad \forall v, w \in V$$

The operator norm is *sub-multiplicative* in the sense that, if U, V, W are normed vector spaces and $\mathsf{S} : U \mapsto V$, $\mathsf{T} : V \mapsto W$ bounded linear operators, then

$$\|\mathsf{T}\,\mathsf{S}\|_{U\mapsto W} \leq \|\mathsf{S}\|_{U\mapsto V} \, \|\mathsf{T}\|_{V\mapsto W} \,.$$

Using the operator norms of Def. 1.8, the vector spaces of linear operators $\mathsf{T} : V \mapsto W$ and bilinear forms $\mathsf{b} : V \times V \mapsto \mathbb{R}$ become normed vector spaces themselves. For them we write $L(V, W)$ and $L(V \times V, \mathbb{R})$, respectively.

Linear forms play a crucial role in our investigations of variational problems:

Definition 1.9. *The **dual** V^* of a normed vector space V is the normed vector space $L(V, \mathbb{R})$ of continuous linear forms on V.*

Notation: For $f \in V^*$ we will usually write $\langle f, v\rangle_{V^*\times V}$ instead of $f(v)$, $v \in V$ ("duality pairing"). The notation $\langle\cdot,\cdot\rangle$ is reserved for the Euklidean inner product in $\mathbb{R}^n$ and $|\cdot|$ will designate the Euklidean norm.

Example 1.10. Some well-known dual spaces of function spaces on $\Omega \subset \mathbb{R}^n$, for which the duality pairing is based on the Lebesgue integral:

- for $1 < p < \infty$ we have $(L^p(\Omega))^* = L^q(\Omega)$, with $p^{-1} + q^{-1} = 1$, see [26, § 19].
- $(L^1(\Omega))^* = L^\infty(\Omega)$.
- $(C^0(\Omega))^*$ is the space of regular complex Borel measures on Ω, see [34, Ch. 6].

We can also consider the dual V^{**} of V^*. In particular we find that

$$\iota : \left\{ \begin{array}{lcl} V & \mapsto & V^{**} \\ v & \mapsto & \{g \in V^* \mapsto \langle g, v\rangle_{V^*\times V}\} \end{array} \right. \tag{1.6}$$

defines a continuous linear operator.

Definition 1.11. *A normed vector space V is called **reflexive**, if the mapping ι defined in* (1.6) *satisfies*

$$\iota(V) = V^{**} \quad \textit{and} \quad \|\iota(v)\|_{V^{**}} = \|v\|_V \quad \forall v \in V\ .$$

Example 1.12. From Ex. 1.10 we see that the space $L^p(\Omega)$ is reflexive for any $1 < p < \infty$. However, $L^\infty(\Omega)$ and $L^1(\Omega)$ are not reflexive.

Exercise 1.1. Show that for a reflexive normed space V

$$\|v\|_V = \sup_{g\in V^*\setminus\{0\}} \frac{\left|\langle g, v\rangle_{V^*\times V}\right|}{\|g\|_{V^*}}\ . \tag{1.7}$$

Definition 1.13. *For a linear operator $\mathsf{T} : V \mapsto W$ on vector real spaces V, W we introduce its **adjoint operator** T^* by*

$$\mathsf{T}^* : \left\{ \begin{array}{lcl} W^* & \mapsto & V^* \\ h & \mapsto & \left\{ \begin{array}{lcl} V & \mapsto & \mathbb{R} \\ v & \mapsto & \langle h, \mathsf{T}\, v\rangle_{W^*\times W} \end{array} \right. \end{array} \right. .$$

Example 1.14. Linear operators $\mathbb{R}^N \mapsto \mathbb{R}^M$, $N, M \in \mathbb{N}$, can be expressed through $M \times N$-matrices. Furthermore, the dual space of $\mathbb{R}^N$ has the same dimension N and, thus, can be identified with $\mathbb{R}^N$. The duality pairing is supplied by the usual Euklidean inner product

$$\langle \boldsymbol{\xi}, \boldsymbol{\mu}\rangle_{(\mathbb{R}^N)^*\times\mathbb{R}} := \boldsymbol{\xi}^T\boldsymbol{\mu} = \sum\nolimits_{k=1}^N \xi_k\mu_k \quad \forall \boldsymbol{\xi} = (\xi_1,\ldots,\xi_N)^T, \boldsymbol{\mu} = (\mu_1,\ldots,\mu_N)^T \in \mathbb{R}^N\ .$$

If $\mathbf{T} \in \mathbb{R}^{M,N}$ is the matrix belonging to a linear operator $\mathsf{T} : \mathbb{R}^N \mapsto \mathbb{R}^M$, then the transposed matrix $\mathbf{T}^T$will describe the adjoint operator T^*:

$$\langle \boldsymbol{\xi}, \mathsf{T}\,\boldsymbol{\mu}\rangle_{(\mathbb{R}^N)^*\times\mathbb{R}} = \langle \boldsymbol{\xi}, \mathbf{T}\boldsymbol{\mu}\rangle_{(\mathbb{R}^N)^*\times\mathbb{R}} = \left\langle \mathbf{T}^T\boldsymbol{\xi}, \boldsymbol{\mu}\right\rangle_{(\mathbb{R}^N)^*\times\mathbb{R}} = \langle \mathsf{T}^*\boldsymbol{\xi}, \boldsymbol{\mu}\rangle_{(\mathbb{R}^N)^*\times\mathbb{R}}\ .$$

Exercise 1.2. Let U, V, W be reflexive normed vector spaces and $\mathsf{S} \in L(U,V)$, $\mathsf{T} \in L(V,W)$. Then the following holds true:

1. $(\mathsf{T}\,\mathsf{S})^* = \mathsf{S}^* \cdot \mathsf{T}^*$.
2. if T is surjective, then T^* is injective.
3. if T is an isomorphism with bounded inverse T^{-1}, then this also holds for T^* and we have $(\mathsf{T}^{-1})^* = (\mathsf{T}^*)^{-1}$
4. $\|\mathsf{T}\|_{V\mapsto W} = \|\mathsf{T}^*\|_{W^*\mapsto V^*}$

Definition 1.15. *A normed vector space V is* complete*, if every Cauchy sequence $\{v_k\}_k \subset V$ has a limit v in V. A complete normed vector space is called a* ***Banach space****.*

Example 1.16. The function spaces $L^p(\Omega)$, $1 \le p \le \infty$, and $C^m(\Omega)$, $m \in \mathbb{N}_0$, are Banach spaces.

It is known that for Banach spaces V, W also $L(V,W)$ and $L(V \times V, \mathbb{R})$ will become Banach spaces.

Bibliographical notes. Further information can be found in any textbook about functional analysis, e.g. in [26, Ch. 2], [23, Ch. 2], and [41, Chs. 1&2]. An advanced, but excellent, discussion of duality is given in [33, Ch. 4].

1.2 Generic theory

Throughout this section V will stand for a reflexive Banach space with norm $\|\cdot\|_V$, and b will designate a continous bilinear form on V, that is $\mathsf{b} \in L(V \times V, \mathbb{R})$.

Given some $f \in V^*$ a **linear variational problem** seeks $u \in V$ such that

$$\mathsf{b}(u,v) = \langle f, v\rangle_{V^*\times V} \quad \forall v \in V\ . \tag{LVP}$$

Theorem 1.17. *The following statements are equivalent:*

(i) For all $f \in V^*$ the linear variational problem (LVP) *has a unique solution $u_f \in V$ that satisfies*

$$\|u_f\|_V \le \frac{1}{\gamma_s}\ \|f\|_{V^*}\ , \tag{1.8}$$

with $\gamma_s > 0$ independent of f.

(ii) The bilinear form b *satisfies the* inf-sup conditions

$$\exists \gamma_s > 0 : \quad \sup_{v \in V \setminus \{0\}} \frac{|\mathsf{b}(w,v)|}{\|v\|_V} \geq \gamma_s \|w\|_V \quad \forall w \in V \,, \tag{IS1}$$

$$\sup_{v \in V \setminus \{0\}} \frac{|\mathsf{b}(v,w)|}{\|v\|_V} > 0 \quad \forall w \in V \setminus \{0\} \,. \tag{IS2}$$

Proof. (i) $\Rightarrow$ (ii): Fix some $w \in V$ and denote by $g_w \in V^*$ the continuous functional $v \mapsto \mathsf{b}(w,v)$. Let $u_g \in V$ be the unique solution of

$$\mathsf{b}(u_g, v) = \langle g_w, v \rangle_{V^* \times V} \quad \forall v \in V \quad \Rightarrow \quad u_g = w \,,$$

from which we conclude (IS1) by (1.8)

$$\|w\|_V = \|u_g\|_V \leq \frac{1}{\gamma_s} \|g_w\|_{V^*} = \frac{1}{\gamma_s} \sup_{v \in V \setminus \{0\}} \frac{|\mathsf{b}(w,v)|}{\|v\|_V} \,.$$

By the reflexivity of V and (1.7)

$$\|w\|_V = \sup_{g \in V^* \setminus \{0\}} \frac{\left|\langle g, w \rangle_{V^* \times V}\right|}{\|g\|_{V^*}} = \sup_{g \in V^* \setminus \{0\}} \frac{|\mathsf{b}(u_g, w)|}{\|g\|_{V^*}} \leq \frac{1}{\gamma_s} \sup_{u \in V} \frac{|\mathsf{b}(u,w)|}{\|u\|_V} \,, \tag{1.9}$$

where (1.8) has been used in the final step. This amounts to (IS2).

(ii) $\Rightarrow$ (i): Let $u_1, u_2 \in V$ be two solutions of (LVP) for the same $f \in V^*$. Then $\mathsf{b}(u_1 - u_2, v) = 0$ for all $v \in V$, and from (IS1) we immediately infer $u_1 = u_2$. This shows uniqueness of u_f.

To prove existence of solutions of (LVP) we define the following subspace of V^*:

$$V_b^* := \left\{ g \in V^* : \; \exists w \in V : \; \mathsf{b}(w,v) = \langle g, v \rangle_{V^* \times V} \; \forall v \in V \right\} \,.$$

Let $\{g_k\}_{k=1}^\infty$ be a Cauchy-sequence in V_b^*. Since V^* is complete, it will converge to some $g \in V^*$. By definition

$$\forall k \in \mathbb{N} : \quad \exists w_k \in V : \quad b(w_k, v) = \langle g_k, v \rangle_{V^* \times V} \quad \forall v \in V \,. \tag{1.10}$$

Thanks to the inf-sup condition (IS1), we have for any $k, m \in \mathbb{N}$

$$\|w_k - w_m\|_V \leq \frac{1}{\gamma_s} \sup_{v \in V \setminus \{0\}} \frac{\left|\langle g_k - g_m, v \rangle_{V^* \times V}\right|}{\|v\|_V} = \frac{1}{\gamma_s} \|g_k - g_m\|_{V^*} \,.$$

Hence, $\{w_k\}_{k=1}^\infty$ is a Cauchy-sequence, too, and will converge to some $w \in V$. The continuity of b and of the duality pairing makes it possible to pass to the limit on both sides of (1.10)

$$\mathsf{b}(w,v) = \langle g, v \rangle_{V^* \times V} \quad \forall v \in V \,,$$

which reveals that $g \in V_b^*$. Recall Def. 1.7 to see that V_b^* is a *closed* subspace of V^*.

Now, assume that $V_b^* \neq V^*$. As $V_b^* \subset V^*$ is closed, a corollary of the *Hahn-Banach theorem*, see [33, Thm. 3.5], confirms the existence of $z \in V^{**} = V$ (V reflexive!) such that

$$\langle g, z \rangle_{V^* \times V} = 0 \quad \forall g \in V_b^* \ .$$

By definition of V_b^* this means $\mathsf{b}(v, z) = 0$ for all $v \in V$ and contradicts (IS2).

Eventually, (1.8) is a simple consequence of (IS1), of the definition of u_f, and of the definition of the norm on V^*. □

The estimate (1.9) instantly confirms that the inf-sup conditions (IS1) and (IS2) involve

$$\sup_{v \in V \setminus \{0\}} \frac{|\mathsf{b}(v, w)|}{\|v\|_V} \geq \gamma_s \|w\|_V \ .$$

Remark 1.18. In order to verify the inf-sup condition (IS1) we can either tackle

$$\sup_{v \in V \setminus \{0\}} \frac{|\mathsf{b}(w, v)|}{\|v\|_V} \geq \gamma_s \|w\|_V \quad \forall w \in V \quad \text{or} \quad \sup_{v \in V \setminus \{0\}} \frac{|\mathsf{b}(v, w)|}{\|v\|_V} \geq \gamma_s \|w\|_V \quad \forall w \in V \ .$$

Then one fixes an arbitrary $w \in V$ and strives to find a "candidate" $v = v(w) \in V$ such that, for $\gamma_1, \gamma_2 > 0$

$$\mathsf{b}(w, v) \geq \gamma_1 \|w\|_V^2 \quad \text{and} \quad \|v\|_V \leq \gamma_2 \|w\|_V \ .$$

This will imply (IS1) with $\gamma_s \geq \gamma_1 \gamma_2^{-1}$. Of course the corresponding inf-sup condition (IS2) needs to be verified as well, but usually this is very easy.

Remark 1.19. The assertion (i) of Thm. 1.17 amounts to the statement that (LVP) is *well posed* in the sense that a unique solution exists and depends continuously on the data of the problem. The right hand side functional f is regarded as input data.

Obviously, for fixed $w \in V$ the mapping $v \mapsto \mathsf{b}(w, v)$ belongs to V^*. Hence we can define a mapping $\mathsf{B} : V \mapsto V^*$ associated with b by

$$\langle \mathsf{B}\, w, v \rangle_{V^* \times V} := \mathsf{b}(w, v) \quad \forall v, w \in V \ . \tag{1.11}$$

Exercise 1.3. The B from (1.11) provides a linear continuous operator with norm

$$\|\mathsf{B}\|_{V \mapsto V^*} = \|\mathsf{b}\|_{V \times V \mapsto \mathbb{R}} \ . \tag{1.12}$$

Using this operator B associated with b we can concisely state (LVP) as

$$u \in V : \quad \mathsf{B}\, u = f \ . \tag{1.13}$$

The equation (1.13) is called the operator notation for (LVP). Thm. 1.17 asserts that the inf-sup condition (IS1) and (IS2) guarantee that B is an isomorphism and that $\left\|\mathsf{B}^{-1}\right\|_{V^* \mapsto V} \leq \gamma_s^{-1}$.

There is a simple sufficient condition for the inf-sup conditions of Thm. 1.17:

Definition 1.20. *A continuous bilinear form* b *on a normed space* V *is called* V***-elliptic*** *with ellipticity constant* $\gamma_e > 0$*, if*

$$|\mathsf{b}(v,v)| \geq \gamma_e \|v\|_V^2 \quad \forall v \in V \ .$$

It is evident that a V-ellipticity of the bilinear form implies both inf-sup conditions (IS1) and (IS2) with $\gamma_s = \gamma_e$.

Exercise 1.4. On the Banach space $C^0([0,1])$ (equipped with the supremum norm) consider the bilinear form

$$\mathsf{a}(v,w) := \int_0^1 v(x)w(x)\,\mathrm{d}\xi \ .$$

Show that the inf-sup condition (IS1) is not satisfied.

Exercise 1.5. First of all let $\Omega :=]-\pi,\pi[$, then show that the following holds true:

(i) Define the subspace $U \subset L^2(\Omega)$ by

$$U := \big\{ u \in L^2(\Omega); \quad u(x) = u(-x) \quad \forall x \in \Omega \big\} \ .$$

Show that $V := U \times U$ is a closed subspace of $L^2(\Omega) \times L^2(\Omega)$.

(ii) The bilinear form

$$\mathsf{b}((u,p),(v,q)) := \int_{-\pi}^{\pi} u(x)v(x)\,\mathrm{d}\xi + \int_{-\pi}^{\pi} \sin(x)p(x)v(x)\,\mathrm{d}\xi - \int_{-\pi}^{\pi} \sin(x)u(x)q(x)\,\mathrm{d}\xi$$

is continuous on the Banach space V. Show that a variational problem based on this bilinear form will in general fail to possess a unique solution.

Bibliographical notes. For further information and details the reader is referred to [22, Sect. 6.5] or to [36, Sect. 1.3].

1.3 Symmetric positive definite variational problems

Definition 1.21. *A bilinear form* b *on a vector space* V *is called* ***symmetric****, if*

$$\mathsf{b}(v,w) = \mathsf{b}(w,v) \quad \forall v,w \in V \ .$$

A special class of symmetric bilinear forms often occurs in practical variational problems:

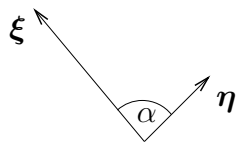


Figure 1.1: Angle $\alpha(\boldsymbol{\xi}, \boldsymbol{\eta})$.

Definition 1.22. *A bilinear form* b *on a real vector space is* ***positive definite****, if for all* $v \in V$

$$\mathsf{b}(v,v) > 0 \quad \Leftrightarrow \quad v \neq 0 \, .$$

A symmetric and positive definite bilinear form is called an ***inner product****.*

In the sequel V will stand for a Banach space with norm $\|\cdot\|_V$. We also adopt the symbol a for a generic inner product on V.

First, observe that every V-elliptic bilinear form is positive definite. Second, recall that an inner product a induces a norm through

$$\|v\|_a := \mathsf{a}(v,v)^{\frac{1}{2}} \quad v \in V \, .$$

The fundamental *Cauchy-Schwarz-inequality*

$$\mathsf{a}(v,w) \leq \|v\|_a \, \|w\|_a \quad \forall v, w \in V \tag{CSI}$$

ensures that a will always be continuous with norm 1 with respect to the energy norm. Moreover, we have Pythagoras' theorem

$$\mathsf{a}(v,w) = 0 \quad \Leftrightarrow \quad \|v\|_a^2 + \|w\|_a^2 = \|v+w\|_a^2 \, .$$

In the context of elliptic partial differential equations a norm that can be derived from a V-elliptic bilinear form is often dubbed **energy norm**. Vector spaces that yield Banach spaces when endowed with an energy norm offer rich structure.

Definition 1.23. *A* ***Hilbert space*** *is a Banach space whose norm is induced by an inner product.*

Exercise 1.6. If an inner product a is V-elliptic and continuous in Banach space V then $(V, \|\cdot\|_a)$ is a Hilbert space.

Notation: In the sequel, the symbol H is reserved for Hilbert spaces. When H is a Hilbert space, we often write $(\cdot,\cdot)_H$ to designate its inner product.

Example 1.24. Let $H = \mathbb{R}^2$ be the set of all vectors in the plane. H is a Hilbert space with the Euclidean norm

$$|\boldsymbol{\xi}|_2 = \left(\xi_1^2 + \xi_2^2\right)^{1/2} \text{ for } \boldsymbol{\xi} \in \mathbb{R}^2$$

and the inner product

$$\langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle = \xi_1\eta_1 + \xi_2\eta_2 \,.$$

For $\boldsymbol{\xi}, \boldsymbol{\eta} \in \mathbb{R}^2$, the angle $\alpha(\boldsymbol{\xi}, \boldsymbol{\eta})$ between $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ (cf. Figure 1.1) is given by

$$\cos\left(\alpha(\boldsymbol{\xi}, \boldsymbol{\eta})\right) = \frac{\langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle}{|\boldsymbol{\xi}|\,|\boldsymbol{\eta}|} \,.$$

This can be used as a general definition of an angle between elements of a Hilbert space.

Example 1.25. Another well known Hilbert space is $L^2(\Omega)$, $\Omega \subset \mathbb{R}^d$, with inner product

$$(u, v) := \int_\Omega u(x)\, v(x)\, \mathrm{d}\xi \,, \quad u, v \in L^2(\Omega) \,.$$

This is the definition for real-valued u, v. In the case of complex valued functions, complex conjugation has to be applied to v.

Based on a continuous bilinear form, we can always indentify a normed vector space with a subspace of its dual space. If we deal with a Hilbert space and this identification is based on the inner product, it becomes an isomorphism, the so-called **Riesz-isomorphism**.

Theorem 1.26. *For a Hilbert space H the mapping $H \mapsto H^*$, $v \mapsto \{w \mapsto (v, w)_H\}$, is a norm-preserving isomorphism (Riesz-isomorphism).*

This means that we need not distinguish H and H^* for Hilbert spaces. However, often it is wise to keep the distinction between "functions" and "linear functionals" though "$H = H^*$".

Corollary 1.27. *All Hilbert spaces are reflexive.*

A Hilbert space framework permits us to generalize numerous geometric notions from Euklidean space:

Definition 1.28. *If H is a Hilbert space we call two subspaces $V, W \subset H$ **orthogonal**, and write $V \perp W$, if $(v, w)_H = 0$ for all $v \in V$, $w \in W$. A linear operator $P : H \mapsto H$ is an **orthogonal projection**, if $P^2 = P$ and $\mathrm{Ker}(P) \perp P(H)$.*

Next, we consider the following variational problem on a Hilbert space H with inner product $\mathsf{a}(\cdot,\cdot)$ and induced norm $\|\cdot\|_a$: for $f \in H^*$ seek $u \in H$ such that

$$\mathsf{a}(u,v) = \langle f, v\rangle_{H^*\times H} \quad \forall v \in H \ . \tag{1.14}$$

Thanks to Thm. 1.17 this variational problem has a unique solution for all $f \in H^*$.

Exercise 1.7. The unique solution u of (1.14) satisfies $\|u\|_a = \|f\|_{H^*}$.

There is an intimate relationship of (1.14) with minimization problems for a coercive quadratic functional. Remember that a mapping $J : H \mapsto \mathbb{R}$, a so-called functional, is coercive, if $\lim_{\|v\|_H\to\infty} \|v\|_H^{-1} J(v) = +\infty$ uniformly in $\|v\|_H$.

Theorem 1.29. *The solution $u \in H$ of (1.14) can be characterized by*

$$u = \arg\min_{v\in H} J(v) \quad \textit{with} \quad J(v) = \tfrac{1}{2}\mathsf{a}(v,v) - \langle f, v\rangle_{H^*\times H} \ .$$

Proof. If u denotes the solution of (1.14), a simple calculation shows

$$J(v) - J(u) = \tfrac{1}{2}\|v-u\|_a^2 \quad \forall v \in H \ . \tag{1.15}$$

This shows that u will be the unique global minimizer of J.

It is easy to establish that J is strictly convex and coercive, which implies existence and uniqueness of a global minimizer u. Now, consider the function

$$h_v(\tau) := J(u+\tau v) \quad \tau \in \mathbb{R},\ v \in H \ .$$

h is smooth and, since u is a global minimizer

$$\frac{d}{d\tau} h_v(\tau)_{|\tau=0} = 0 \ ,$$

which is equivalent to (1.14), since any v can be chosen. □

Exercise 1.8. Show that the functional J from Thm. 1.29 is coercive on H.

1.4 Discrete variational problems

A first step towards finding a practical algorithm for the approximate solution of (LVP) is to convert it into a *discrete variational problem*. We use the attribute "discrete" in the sense that the solution can be characterized by a finite number of real (or complex) numbers.

Given a real Banach space V with norm $\|\cdot\|_V$ and a bilinear form $\mathsf{b} \in L(V\times V, \mathbb{R})$ we pursue a **Galerkin discretization** of (LVP). Its gist is to replace V in (LVP) by *finite dimensional subspaces*. The most general approach relies on two subspaces of V:

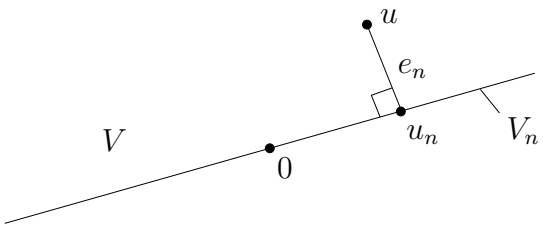


Figure 1.2: b-Orthogonality of the error $e_n = u - u_n$ with respect to V_n, if b is an inner product

$$\begin{array}{lll} W_n \subset V : & \text{“trial space”}, & \dim W_n = N \\ V_n \subset V : & \text{“test space”}, & \dim V_n = N \end{array} \quad , \quad N \in \mathbb{N} \, .$$

Notation: Throughout a subscript n will be used to label “discrete entities” like the above finite dimensional trial and test spaces, and their elements. Often we will consider sequences of such spaces; in this case n will assume the role of an index.

Given the two spaces W_n and V_n and some $f \in V^*$ the **discrete variational problem** corresponding to (LVP) reads: seek $u_n \in W_n$ such that

$$\mathsf{b}(u_n, v_n) = \langle f, v_n \rangle_{V^* \times V} \quad \forall v_n \in V_n \, . \tag{DVP}$$

This most general approach, where $W_n \neq V_n$ is admitted, is often referred to as **Petrov-Galerkin method**. In common parlance, the classical Galerkin discretization implies that trial and test space agree. If, moreover, b provides an inner product on V, the method is known as **Ritz-Galerkin scheme**.

If, for given b and f both (LVP) and (DVP) have unique solutions $u \in V$ and $u_n \in W_n$, respectively, then a simple subtraction reveals

$$\mathsf{b}(u - u_n, v_n) = 0 \quad \forall v_n \in V_n \, . \tag{1.16}$$

Abusing terminology, this property is called **Galerkin orthogonality**, though the term orthogonality is only appropriate, if b is an inner product on V. Sloppily speaking, the **discretization error** $e_n := u - u_n$ is “orthogonal” to the test space V_n, see Fig. 1.2.

Theorem 1.30. *Let V be a Banach space and $\mathsf{b} \in L(V \times V, \mathbb{R})$ satisfy the inf-sup conditions* (IS1) *and* (IS2) *from Thm. 1.17. Further, assume that*

$$\exists \gamma_n > 0 : \quad \sup_{v_n \in V_n \setminus \{0\}} \frac{|\mathsf{b}(v_n, w_n)|}{\|v_n\|_V} \geq \gamma_n \|w_n\|_V \quad \forall w_n \in W_n \, . \tag{DIS}$$

Then, for every $f \in V^*$ the discrete variational problem (DVP) *has a unique solution*

u_n that satisfies

$$\|u_n\|_V \leq \frac{1}{\gamma_n} \|f\|_{V_n^*} = \frac{1}{\gamma_n} \sup_{v_n \in V_n} \frac{|f(v_n)|}{\|v_n\|_V} , \tag{1.17}$$

$$\|u - u_n\|_V \leq \left(1 + \frac{\|\mathsf{b}\|_{V \times V \mapsto \mathbb{R}}}{\gamma_n}\right) \inf_{w_n \in W_n} \|u - w_n\|_V , \tag{1.18}$$

where $u \in V$ solves (LVP)*.*

Proof. It is clear that (DIS) implies the uniqueness of u_n. Since $\dim V_n = \dim W_n$, in the finite dimensional setting this implies existence of u_n.

The remainder of the proof combines the triangle inequality and (1.17):

$$\begin{aligned} \|u - u_n\|_V &\leq \|u - w_n\|_V + \|w_n - u_n\|_V \\ &\leq \|u - w_n\|_V + \frac{1}{\gamma_n} \sup_{v_n \in V \setminus \{0\}} \frac{|\mathsf{b}(w_n - u + u - u_n, v_n)|}{\|v_n\|_V} \quad \forall w_n \in W_n . \end{aligned}$$

Eventually, use the Galerkin orthogonality (1.16) and the continuity of b to finish the proof. □

Remark 1.31. One can not conclude (DIS) from (IS1) and (IS2) because the supremum is taken over a much smaller set.

Remark 1.32. It goes without saying that for a V-elliptic bilinear form b, *cf.* Def. 1.20, the assumptions of Thm. 1.30 are trivially satisfied. Moreover, we can choose γ_n equal to the ellipticity constant γ_e in this case.

Thm. 1.30 provides an **a-priori estimate** for the norm of the discretization error e_n. It reveals that the Galerkin solution will be **quasi-optimal**, that is, for arbitrary f the norm of the discretization can be bounded by a constant times the **best approximation error**

$$\inf_{w_n \in W_n} \|u - v_n\|_V ,$$

of the exact solution u w.r.t. W_n. It is all important that this constant must not depend on f.

Remark 1.33. Many of our efforts will target **asymptotic a-priori estimates** that involve sequences $\{V_n\}_{n=1}^{\infty}$, $\{W_n\}_{n=1}^{\infty}$ of test and trial spaces. Then it will be the principal objective to ensure that the constant γ_n is bounded away from zero uniformly in n. This will guarantee **asymptotic quasi-optimality** of the Galerkin solution: the estimate (1.17) will hold with a constant independent of n. Notice that the norm of b that also enters (1.18) does not depend on the finite dimensional trial and test spaces.

In the sequel we will take for granted that b and V_n, W_n meet the requirements of Thm. 1.30. The "Galerkin orthogonality" (1.16) suggests that we examine the so-called **Galerkin projection** $\mathsf{P}_n : V \mapsto W_n$ defined by

$$\mathsf{b}(\mathsf{P}_n w, v_n) = \mathsf{b}(w, v_n) \quad \forall v_n \in V_n \; . \tag{1.19}$$

Proposition 1.34. *Under the assumption of Thm. 1.30, the equation* (1.19) *defines a continuous projection* $\mathsf{P}_n : V \mapsto V$ *with norm* $\|\mathsf{P}_n\|_{V\mapsto V} \leq \gamma_n^{-1} \|\mathsf{b}\|_{V\times V\mapsto\mathbb{R}}$.

Proof. As a consequence of Thm. 1.30, P_n is well defined. Its linearity is straightforward and the norm bound can be inferred from (1.17). Also $\mathsf{P}_n^2 = \mathsf{P}_n$ is immediate from the definition. □

The Galerkin projection connects the two solutions u and u_n of (LVP) and (DVP), respectively, through

$$u_n = \mathsf{P}_n u \; . \tag{1.20}$$

Proposition 1.35. *If* $\in L(V \times V, \mathbb{R})$ *is an inner product on* V *and the assumptions of Thm. 1.30 are satisfied, then the Galerkin projection associated with* b *is an orthogonal projection with respect to the inner product* b.

Proof. According to Def. 1.28 and the previous proposition, we only have to check that $\mathrm{Ker}(\mathsf{P}_n) \perp W_n$. This is clear from (1.20) and the Galerkin orthogonality (1.16), because

$$v \in \mathrm{Ker}(\mathsf{P}_n) \;\Leftrightarrow\; \mathsf{P}_n v = 0 \;\Leftrightarrow\; (Id - \mathsf{P}_n)v = v \;\Leftrightarrow\; v \in \mathrm{Im}(Id - \mathsf{P}_n) \; .$$

□

As in the previous section, we can cast (DVP) into operator form by introducing $\mathsf{B}_n : W_n \mapsto V_n^*$, $f_n \in V_n^*$ through

$$\langle \mathsf{B}_n \, w_n, v_n \rangle_{V_n^* \times V_n} = \mathsf{b}(w_n, v_n) \quad \forall w_n \in W_n, v_n \in V_n \quad , \quad \langle f_n, v_n \rangle_{V_n^* \times V_n} = \langle f, v_n \rangle_{V^* \times V}$$

Using these operators, we can rewrite (DVP) as

$$\mathsf{B}_n \, u_n = f_n \; . \tag{1.21}$$

Exercise 1.9. There is a canonical embedding $V^* \subset V_n^*$ and the estimate $\|f_n\|_{V_n^*} \leq \|f\|_{V^*}$ holds true.

Exercise 1.10. Let $\mathsf{I}_n := \mathsf{I}_{V_n} : V_n \mapsto V$ stand for the canonical injection, *cf.* Def. 1.4. Then we have

$$\mathsf{B}_n = \mathsf{I}_n^* \circ \mathsf{B} \circ \mathsf{I}_n \; . \tag{1.22}$$

Now, let us consider the special case that V is a Hilbert space. To hint at this, we write H instead of V. It is surprising that under exactly the same assumptions on b, W_n, and V_n as have been stated in Thm. 1.30, the mere fact that the norm of $V = H$ arises from an inner product, permits us to get a stronger a-priori error estimate [44].

Theorem 1.36. *If V is a Hilbert space we obtain the sharper a-priori error estimate*

$$\|u - u_n\|_V \leq \frac{\|\mathsf{b}\|_{V\times V\mapsto\mathbb{R}}}{\gamma_n} \inf_{v_n\in V_n} \|u - v_n\|_V \ , \tag{1.23}$$

if the assumptions of Thm. 1.30 are satisfied.

The proof of this result requires a lemma due to Kato [27]:

Lemma 1.37. *Let H be a Hilbert space. If $\mathsf{P} \in L(H, H)$ is a non-trivial projection, i. e. $Id \neq \mathsf{P} = \mathsf{P}^2 \neq 0$, then*

$$\|\mathsf{P}\|_{H\mapsto H} = \|Id - \mathsf{P}\|_{H\mapsto H} \ .$$

Proof. (i) First we consider the case $\dim H = 2$, that is $H \cong \mathbb{R}^2$ after a choice of an orthonormal basis. Then both P and $Id - \mathsf{P}$ have rank 1, so that they possess a representation

$$\begin{aligned} \mathsf{P}\,\boldsymbol{\nu} &= \langle \boldsymbol{\beta}, \boldsymbol{\nu}\rangle\, \boldsymbol{\alpha} \quad \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\nu} \in \mathbb{R}^2 \ , \\ (Id - \mathsf{P})\boldsymbol{\nu} &= \langle \boldsymbol{\delta}, \boldsymbol{\nu}\rangle\, \boldsymbol{\gamma} \quad \boldsymbol{\delta}, \boldsymbol{\gamma}, \boldsymbol{\nu} \in \mathbb{R}^2 \ . \end{aligned}$$

Since $\mathsf{P}^2 = \mathsf{P}$ and $(Id - \mathsf{P})^2 = (Id - \mathsf{P})$ we conclude that

$$\langle \boldsymbol{\alpha}, \boldsymbol{\beta}\rangle = 1 \quad \text{and} \quad \langle \boldsymbol{\delta}, \boldsymbol{\gamma}\rangle = 1 \ .$$

Now, observe that

$$\boldsymbol{\nu} = \mathsf{P}\,\boldsymbol{\nu} + (Id - \mathsf{P})\boldsymbol{\nu} = \langle \boldsymbol{\beta}, \boldsymbol{\nu}\rangle\, \boldsymbol{\alpha} + \langle \boldsymbol{\delta}, \boldsymbol{\nu}\rangle\, \boldsymbol{\gamma} \ .$$

Setting $\nu = \boldsymbol{\beta}, \boldsymbol{\delta}$ and forming the inner product with $\boldsymbol{\alpha}, \boldsymbol{\gamma}$, respectively, we see

$$\begin{aligned} 1 &= |\boldsymbol{\alpha}|^2\, |\boldsymbol{\beta}|^2 + \langle \boldsymbol{\delta}, \boldsymbol{\beta}\rangle\, \langle \boldsymbol{\gamma}, \boldsymbol{\alpha}\rangle \ , \\ 1 &= \langle \boldsymbol{\beta}, \boldsymbol{\delta}\rangle\, \langle \boldsymbol{\alpha}, \boldsymbol{\gamma}\rangle + |\boldsymbol{\gamma}|^2\, |\boldsymbol{\delta}|^2 \ , \end{aligned}$$

which implies $|\boldsymbol{\alpha}|\, |\boldsymbol{\beta}| = |\boldsymbol{\gamma}|\, |\boldsymbol{\delta}|$. Note that

$$\|\mathsf{P}\|_{H\mapsto H} = \sup_{\boldsymbol{\nu}\neq 0} \frac{|\langle \boldsymbol{\beta}, \boldsymbol{\nu}\rangle|}{|\boldsymbol{\nu}|} = |\boldsymbol{\alpha}| \sup_{\boldsymbol{\nu}\neq 0} \frac{|\langle \boldsymbol{\beta}, \boldsymbol{\nu}\rangle|}{|\boldsymbol{\nu}|} = |\boldsymbol{\alpha}|\, |\boldsymbol{\beta}| \ ,$$

and similarly for $Id - \mathsf{P}$. This proves the assertion for $\dim H = 2$.

(ii) For the general case, take into account that for any $v \in H$ the space $U := \operatorname{span}\{v, \mathsf{P}\,v\}$ is an invariant w.r.t. both P and $Id - \mathsf{P}$. $\dim U = 1$ boils down to $\mathsf{P}\,v = 0$ or $\mathsf{P}\,v = v$. As $\|\mathsf{P}\|_{H\mapsto H} \geq 1$ and $\|Id - \mathsf{P}\|_{H\mapsto H} \geq 1$ we immediately conclude

$$\|(Id - \mathsf{P})v\|_V \leq \|\mathsf{P}\|_{H\mapsto H}\,\|v\|_V \quad , \quad \|\mathsf{P}\,v\|_V \leq \|(Id - \mathsf{P})\|_{H\mapsto H}\,\|v\|_V \ .$$

If $\dim U = 2$, we can apply the result of (i) and get

$$\begin{aligned}\|(Id - \mathsf{P})v\|_V &\leq \|Id - \mathsf{P}\|_{U\mapsto U}\,\|v\|_V = \|\mathsf{P}\|_{U\mapsto U}\,\|v\|_V \leq \|\mathsf{P}\|_{H\mapsto H}\,\|v\|_V \ , \\ \|\mathsf{P}\,v\|_V &\leq \|\mathsf{P}\|_{U\mapsto U}\,\|v\|_V = \|Id - \mathsf{P}\|_{U\mapsto U}\,\|v\|_V \leq \|Id - \mathsf{P}\|_{H\mapsto H}\,\|v\|_V \ ,\end{aligned}$$

Combining these estimates yields the assertion. □

Proof (of Thm. 1.36). Using (1.20) we can appeal to the previous lemma to estimate

$$\begin{aligned}\|u - u_n\|_V &= \|(Id - \mathsf{P}_n)u\|_V = \|(Id - \mathsf{P}_n)(u - w_n)\|_V \\ &\leq \|Id - \mathsf{P}_n\|_{V\mapsto V}\,\|u - w_n\|_V = \|\mathsf{P}_n\|_{V\mapsto V}\,\|u - w_n\|_V \ ,\end{aligned}$$

where $w_n \in W_n$ is arbitrary. From

$$\|\mathsf{P}_n w\|_V \leq \frac{1}{\gamma_n}\sup_{v_n\neq 0}\frac{|b(\mathsf{P}_n w, v_n)|}{\|v_n\|_V} \leq \frac{\|b\|_{V\times V\mapsto\mathbb{R}}}{\gamma_n}\,\|w\|_V$$

we conclude $\|\mathsf{P}_n\|_{V\mapsto V} \leq \gamma_n^{-1}\,\|b\|_{V\times V\mapsto\mathbb{R}}$, which finishes the proof. □

The other special case is that of Ritz-Galerkin discretization aimed at a symmetric, positive definite bilinear form b. Then the Ritz-Galerkin method will furnish an *optimal* solution in the sense that u_n is the best approximation of u in V_n.

Corollary 1.38. *If* b *is an inner product in* V*, with which* V *becomes a Hilbert space* H*, and* $V_n = W_n$*, then* (DVP) *will have a unique solution* u_n *for any* $f \in H^*$*. It satisfies*

$$\|u - u_n\|_b \leq \inf_{v_n\in V_n}\|u - v_n\|_b \ ,$$

where $\|\cdot\|_b$ *is the energy norm derived from* b*.*

Proof. Existence and uniqueness are straightforward. It is worth noting that the estimate can be obtained in a simple fashion from Galerkin orhtogonality (1.16) and the Cauchy-Schwarz inequality

$$\|u - u_n\|_b^2 = b(u - u_n, u - v_n) \leq \|u - u_n\|_b\,\|u - v_n\|_b \ ,$$

for any $v_n \in V_n$. □

Bibliographical notes. We refer to [21, Sects. 8.1 & 8.2] and [36, Sect. 2.3].

1.5 The algebraic setting

The variational problem (DVP) may be discrete, but it is by no means amenable to straightforward computer implementation, because an abstract concept like a finite dimensional vector space has no algorithmic representation. In short, a computer can only handle vectors (arrays) of finite length and little else.

We adopt the setting of Sect. 1.4. The trick to convert (DVP) into a problem that can be solved on a computer is to introduce **ordered bases**

$$\begin{array}{rcll} \mathfrak{B}_V & := & \{p_n^1, \ldots, p_n^N\} & \text{of } V_n \ , \\ \mathfrak{B}_W & := & \{q_n^1, \ldots, q_n^N\} & \text{of } W_n \ , \end{array} \qquad N := \dim V_n = \dim W_n \ .$$

Remember that a basis of a finite dimensional vector space is a maximal set of linearly indepedent vectors. By indexing the basis vectors with consecutive integers we indicate that the order of the basis vectors will matter.

Lemma 1.39. *The following is equivalent:*

(i) The discrete variational problem (DVP) *has a unique solution* $u_n \in W_n$.

(ii) The linear system of equations

$$\mathbf{B}\boldsymbol{\mu} = \boldsymbol{\varphi} \tag{LSE}$$

with

$$\mathbf{B} := \left(\mathsf{b}(q_n^k, p_n^j)\right)_{j,k=1}^N \in \mathbb{R}^{N,N} \ , \tag{1.24}$$

$$\boldsymbol{\varphi} := \left(\left\langle f, p_n^k \right\rangle_{V^* \times V}\right)_{k=1}^N \in \mathbb{R}^N \ , \tag{1.25}$$

has a unique solution $\boldsymbol{\mu} = (\mu_k)_{k=1}^N \in \mathbb{R}^N$.

Then

$$u_n = \sum_{k=1}^N \mu_k q_n^k \ .$$

Proof. Due to the basis property we can set

$$u_n = \sum_{k=1}^N \mu_k q_n^k \quad , \quad v_n = \sum_{k=1}^N \nu_k p_n^k, \quad \mu_k, \nu_k \in \mathbb{R} \ ,$$

in (DVP). Hence, (DVP) becomes: seek $\mu_1, \ldots, \mu_N$ such that

$$\mathsf{b}(\textstyle\sum_{k=1}^N \mu_k q_n^k, \sum_{j=1}^N \nu_j p_n^j) = \left\langle f, \sum_{j=1}^N \nu_j p_n^j \right\rangle_{V^* \times V}$$

for all $\nu_1, \dots, \nu_N \in \mathbb{R}$. We can now exploit the linearity of b and f:

$$\sum_{j=1}^{N} \sum_{k=1}^{N} \mu_k \nu_j b(q_n^k, p_n^j) = \sum_{j=1}^{N} \nu_j \left\langle f, p_n^j \right\rangle_{V^* \times V} \ . \tag{1.26}$$

Next, plug in special test vectors given by $(\nu_1, \dots, \nu_N) = \boldsymbol{\epsilon}_l$, $l \in \{1, \dots, N\}$, where $\boldsymbol{\epsilon}_l$ is the l-th unit vector in $\mathbb{R}^N$. This gives us

$$\sum_{k=1}^{N} \mu_k \mathsf{b}(q_n^k, p_n^l) = \left\langle f, p_n^l \right\rangle_{V^* \times V}, \quad l = 1, \dots, N \ . \tag{1.27}$$

As the special test vectors span all of $\mathbb{R}^N$ and thanks to the basis property, we conclude that (1.26) and (1.27) are equivalent. On the other hand, (1.27) corresponds to (LSE), as is clear by recalling the rules of matrix×vector multiplication. □

Note that in (1.24) j is the row index, whereas k is the column index. Consequently, the element in the j-th row and k-th column of the matrix $\mathbf{B}$ in (LSE) is given by $\mathsf{b}(q_n^k, p_n^j)$.

Notation: Throughout, bold greek symbols will be used for vectors in some Euklidean vector space $\mathbb{R}^n$, $n \in \mathbb{N}$, whereas bold capital roman font will designate matrices. The entries of a matrix $\mathbf{M}$ will either be written in small roman letters tagged by two subscripts: m_{ij} or will be denoted by $(\mathbf{M})_{ij}$.

Corollary 1.40. *If and only if the bilinear form* b *and trial/test space* W_n/V_n *satisfy the assumptions of 1.30, then the matrix* $\mathbf{B}$ *of* (LSE) *will be regular.*

Thus, we have arrived at the final "algebraic problem" (LSE) through the two stage process outlined in Fig. 1.3. It is important to realize that the choice of basis does not affect the discretization error at all: the latter solely depends on the choice of trial and test spaces. Also, Cor. 1.40 teaches that some properties of $\mathbf{B}$ will only depend on V_n, too.

However, the choice of basis may have a big impact on other properties of the resulting matrix $\mathbf{B}$ in (LSE).

Example 1.41. If b induces an inner product on $V_n = W_n$, then a theorem from linear algebra (Gram-Schmidt-orthogonalisation) tells us that we can always find a b-orthonormal basis of V_n. Evidently, with respect to this basis the matrix associated with b according to (1.24) will be the $N \times N$ identity matrix $\mathbf{I}$.

Lemma 1.42. *Consider a fixed bilinear form* b *and finite dimensional trial/test space* $W_n = V_n$ *for the discrete variational problem* (DVP). *We choose two different bases* $\mathfrak{B} := \{p_n^1, \dots, p_n^N\}$ *and* $\widetilde{\mathfrak{B}} := \{\widetilde{p}_n^1, \dots, \widetilde{p}_n^N\}$ *of* $W_n = V_n$, *for which*

$$\widetilde{p}_n^j = \sum_{k=1}^{N} s_{jk} p_n^k \quad \textit{with } \mathbf{S} = (s_{jk})_{j,k=1}^{N} \in \mathbb{R}^{N,N} \textit{ regular.}$$

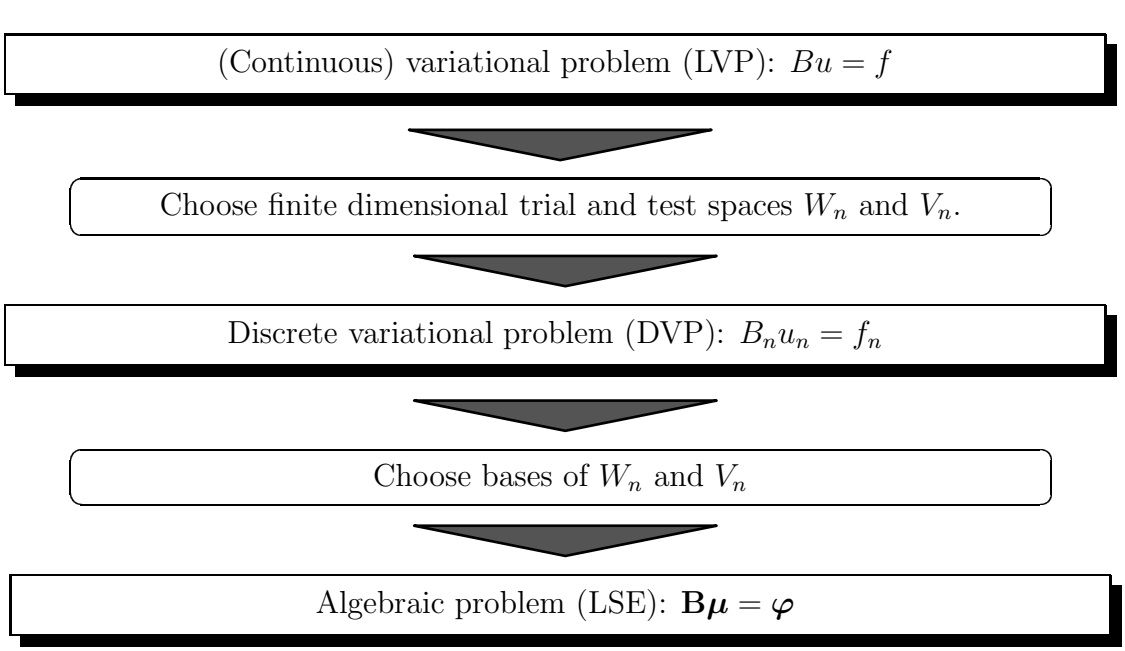


Figure 1.3: Overview of stages involved in the complete Galerkin discretization of an abstract variational problem

Relying on these bases we convert (DVP) *into two linear systems of equations* $\mathbf{B}\boldsymbol{\mu} = \boldsymbol{\varphi}$ *and* $\widetilde{\mathbf{B}}\widetilde{\boldsymbol{\mu}} = \widetilde{\boldsymbol{\varphi}}$*, respectively.*

If the discrete variational problem (DVP) *possesses a unique solution, then the two linear systems and their respective solutions are related by*

$$\widetilde{\mathbf{B}} = \mathbf{S}\mathbf{B}\mathbf{S}^T \quad , \quad \widetilde{\boldsymbol{\varphi}} = \mathbf{S}\boldsymbol{\varphi} \quad , \quad \widetilde{\boldsymbol{\mu}} = \mathbf{S}^{-T}\boldsymbol{\mu} \, . \tag{1.28}$$

Proof. Using (1.24) and the linearity of b we get

$$\widetilde{b}_{lm} = \mathsf{b}(\widetilde{p}_n^m, \widetilde{p}_n^l) = \sum_{k=1}^{N}\sum_{j=1}^{N} s_{mk}\mathsf{b}(p_n^k, p_n^j)s_{lj} = \sum_{k=1}^{N}\underbrace{\Big(\sum_{j=1}^{N} s_{lj}b_{jk}\Big)}_{(\mathbf{SB})_{lk}} s_{mk} = (\mathbf{S}\mathbf{B}\mathbf{S}^T)_{lm} \, ,$$

which gives the relationship between $\mathbf{B}$ and $\widetilde{\mathbf{B}}$. The other relationships are as straightforward. □

Remark 1.43. The lemma reveals that all possible Galerkin matrices $\mathbf{B}$ from (LSE) that we can obtain for a given discrete variational problem (DVP) form a *congruence class* of matrices. It is exactly the invariants of congruence classes that are invariants of Galerkin matrices: symmetry, regularity, positive definiteness, and the total dimensions of eigenspaces belonging to positive, negative, and zero eigenvalues.

Exercise 1.11. For the case $\mathfrak{B}_V = \mathfrak{B}_W$, discuss the impact of reshuffling the basis vectors in $\mathfrak{B}_V$ on the matrix $\mathbf{B}$ of (LSE).

Exercise 1.12. The linear system of equations (LSE) has been built using the bases $\mathfrak{B}_V = \mathfrak{B}_W = \{p_1, \ldots, p_N\}$. How will the matrix $\mathbf{B}$ be affected, when we switch to scaled bases

$$\widetilde{\mathfrak{B}}_V = \widetilde{\mathfrak{B}}_W = \{\alpha_1 p_1, \ldots, \alpha_N p_N\} \,, \quad \alpha_i \in \mathbb{R} \setminus \{0\},\, i = 1, \ldots, N \,?$$

Exercise 1.13. Let V be a Banach space and $f \in V^*$. Let $V_n \subset V$ be a finite dimensional subspace of $\dim V_n = n$. Now let $\mathfrak{B}_n := \{p_1, \ldots, p_n\}$ be a basis of V_n. Find a basis ${\mathfrak{B}_n}^*$ of V_n^* such that the coefficient vector of f in V_n^* is given by $\left(\langle f, p_k\rangle_{V^*\times V}\right)_{k=1}^n$.

Exercise 1.14. The right hand side of (DVP) is mapped to the vector $\boldsymbol{\varphi}$ in (LSE). Which basis of V_n^* renders $\boldsymbol{\varphi}$ according to (1.25) the coefficient vector of f_n.

Remark 1.44. The constant γ_n in (DIS) is independent of the choice of bases. Nevertheless, its computation for concrete W_n, V_n has to employ bases of W_n and V_n.

Let us assume that V is a Hilbert space with inner product $(\cdot,\cdot)_V$. After endowing W_n, V_n with bases $\mathfrak{B}_W := \{q_n^1, \ldots, p_n^N\}$, $\mathfrak{B}_V := \{p_n^1, \ldots, p_n^N\}$ we can express

$$\|v_n\|_V^2 = \boldsymbol{\mu}^T \mathbf{M}_V \boldsymbol{\mu} \quad , \quad \|w_n\|_V^2 = \boldsymbol{\xi}^T \mathbf{M}_W \boldsymbol{\xi} \,,$$

where

$$v_n = \sum_{k=1}^N \mu_k p_n^k \quad , \quad w_n = \sum_{k=1}^N \xi_k p_n^k \,,$$
$$\mathbf{M}_V := \left((p_n^k, p_n^j)_V\right)_{k,j=1}^N \in \mathbb{R}^{N,N} \quad , \quad \mathbf{M}_W := \left((q_n^k, q_n^j)_V\right)_{k,j=1}^N \in \mathbb{R}^{N,N} \,.$$

Reusing (1.24) we can convert (DIS) into

$$\sup_{\boldsymbol{\mu}\in\mathbb{R}^N\setminus\{0\}} \frac{|\boldsymbol{\mu}^T \mathbf{B} \boldsymbol{\xi}|}{(\boldsymbol{\mu}^T \mathbf{M}_V \boldsymbol{\mu})^{\frac{1}{2}}} \geq \gamma_n (\boldsymbol{\xi}^T \mathbf{M}_W \boldsymbol{\xi})^{\frac{1}{2}} \quad \forall \boldsymbol{\xi} \in \mathbb{R}^N \,. \tag{1.29}$$

Since $\mathbf{M}_V$ and $\mathbf{M}_W$ are symmetric and positive definite, there are "square roots" $\mathbf{X}_V, \mathbf{X}_W \in \mathbb{R}^{N,N}$ of $\mathbf{M}_V, \mathbf{M}_W$ such that, e.g., $\mathbf{X}_V^2 = \mathbf{M}_V$ and $\mathbf{X}_V$ is s.p.d. itself. Thus, in (1.29) we can replace

$$\boldsymbol{\mu} \leftarrow \mathbf{X}_V^{-1} \boldsymbol{\mu} \quad , \quad \boldsymbol{\xi} \leftarrow \mathbf{X}_W^{-1} \boldsymbol{\xi} \,,$$

and we end up with

$$\sup_{\boldsymbol{\mu}\in\mathbb{R}^N\setminus\{0\}} \frac{|\boldsymbol{\mu}^T (\mathbf{X}_V^{-1} \mathbf{B} \mathbf{X}_W^{-1}) \boldsymbol{\xi}|}{|\boldsymbol{\mu}|} \geq \gamma_n |\boldsymbol{\xi}| \quad \forall \boldsymbol{\xi} \in \mathbb{R}^N \,. \tag{1.30}$$

From numerical linear algebra we recall the *singular value decomposition*

$$\mathbf{X}_V^{-1}\mathbf{B}\mathbf{X}_W^{-1} = \mathbf{U}\mathbf{D}\mathbf{V}^T ,$$

where $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{N,N}$ are orthogonal matrices and $\mathbf{D} \in \mathbb{R}^{N,N}$ is diagonal. The diagonal entries of $\mathbf{D} = \operatorname{diag}(\sigma_1, \ldots, \sigma_N)$, $\sigma_k \geq 0$, are called the singular values of $\mathbf{X}_V\mathbf{B}\mathbf{X}_W$. Since multiplication with orthogonal matrices leaves the Euklidean norm of a vector invariant and $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$, the replacement

$$\boldsymbol{\mu} \leftarrow \mathbf{U}\boldsymbol{\mu} \quad , \quad \boldsymbol{\xi} \leftarrow \mathbf{V}\boldsymbol{\xi}$$

gives the equivalent inequality

$$\sup_{\boldsymbol{\mu}\in\mathbb{R}^N\setminus\{0\}} \frac{|\boldsymbol{\mu}^T\mathbf{D}\boldsymbol{\xi}|}{|\boldsymbol{\mu}|} \geq \gamma_n|\boldsymbol{\xi}| \quad \forall \boldsymbol{\xi} \in \mathbb{R}^N .$$

Hence, the smallest singular value of $\mathbf{X}_V\mathbf{B}_W\mathbf{X}$ is the largest possible value for γ_n.

If b is symmetric and $\mathfrak{B}_V = \mathfrak{B}_W$, then the singular values of $\mathbf{X}_V\mathbf{B}\mathbf{X}_W$ agree with the eigenvalues of this matrix. These can be determined by solving the generalized eigenvalue problem

$$\boldsymbol{\xi} \neq 0, \lambda \in \mathbb{R} : \quad \mathbf{B}\boldsymbol{\xi} = \lambda\mathbf{M}_V\boldsymbol{\xi} .$$

For the remainder of this section we assume that trial and test spaces and their respective bases agree. Any choice of basis $\mathfrak{B} := \{p_n^1, \ldots, p_n^N\}$ for V_n spawns a **coefficient isomorphism** $\mathsf{C}_n : \mathbb{R}^N \mapsto V_n$ by

$$\mathsf{C}_n\boldsymbol{\mu} = \sum_{k=1}^N \mu_k p_n^k .$$

It can be used to link the operator form (1.21) of the discrete variational problem and the associated linear system of equations (LSE) w.r.t. $\mathfrak{B}$.

Lemma 1.45. *We have*

$$\mathbf{B} = \mathsf{C}_n^* \circ \mathsf{B}_n \circ \mathsf{C}_n ,$$

where $\mathsf{B}_n : V_n \mapsto V_n^*$ *is the operator related to the bilinear form* b *on* V_n*, cf.* (1.21)*, and* $\mathbf{B}$ *is the coefficient matrix for* b *w.r.t. the basis* $\mathfrak{B}$ *of* V_n*.*

Proof. Pick any $\boldsymbol{\mu}, \boldsymbol{\xi} \in \mathbb{R}^N$ and remember that $\mathbb{R}^N$ is its own dual with the duality pairing given by the Euklidean inner product, that is

$$\langle \boldsymbol{\mu}, \boldsymbol{\xi} \rangle_{(\mathbb{R}^N)^*\times\mathbb{R}^N} = \boldsymbol{\mu}^T\boldsymbol{\xi} .$$

Owing to Def. 1.13 and (1.24) we find

$$\boldsymbol{\mu}^T\mathbf{B}\boldsymbol{\xi} = \mathsf{b}(\mathsf{C}_n\boldsymbol{\xi},\mathsf{C}_n\boldsymbol{\mu}) = \langle \mathsf{B}_n\,\mathsf{C}_n\boldsymbol{\xi},\mathsf{C}_n\boldsymbol{\mu}\rangle_{V_n^*\times V_n} = \langle \mathsf{C}_n{}^*\,\mathsf{B}_n\,\mathsf{C}_n\boldsymbol{\xi},\boldsymbol{\mu}\rangle_{(\mathbb{R}^N)^*\times\mathbb{R}^N} = \boldsymbol{\mu}^T(\mathsf{C}_n{}^*\,\mathsf{B}_n\,\mathsf{C}_n)\boldsymbol{\xi}\ .$$

□

An important property of a regular matrix $\mathbf{M}\in\mathbb{R}^{N,N}$, $N\in\mathbb{N}$, is its **spectral condition number**

$$\kappa(\mathbf{M}) = |\mathbf{M}|\left|\mathbf{M}^{-1}\right|\ ,$$

where, by default, $\mathbb{R}^N$ is equipped with the Euklidean vector norm, and $|\mathbf{M}|$ will always stand for the operator norm associated with the Euklidean vector norm.

Lemma 1.46. *If* $\mathbf{B}$ *is a matrix according to* (1.24) *based on a bilinear form* b *and a trial/test space* V_n *(equipped with basis* $\mathfrak{B}$*) that satisfy the assumptions of Thm. 1.30, then*

$$\kappa(\mathbf{B}) \le \|\mathsf{C}_n\|^2_{\mathbb{R}^N\mapsto V}\left\|\mathsf{C}_n^{-1}\right\|^2_{V_n\mapsto\mathbb{R}^N}\frac{\|b\|_{V\times V\mapsto\mathbb{R}}}{\gamma_n}\ ,$$

where C_n *is the coefficient isomorphism belonging to basis* $\mathfrak{B}$*, and* γ_n *is the if-sup constant from* (DIS)*.*

Proof. By the definition of the operator norm, *cf.* Def. 1.8,

$$\|\mathsf{B}_n\,w_n\|_{V_n^*} = \sup_{v_n\in V_n}\frac{\mathsf{b}(w_n,v_n)}{\|v_n\|_V} \ge \gamma_n\,\|w_n\|_V \quad \forall w_n\in V_n\ ,$$

which means

$$\gamma_n\left\|\mathsf{B}_n^{-1}\,f_n\right\|_V \le \|f_n\|_{V_n^*} \quad \forall f_n\in V_n^* \quad\Rightarrow\quad \left\|\mathsf{B}_n^{-1}\right\|_{V_n^*\mapsto V_n}\le\gamma_n^{-1}\ .$$

By Lemma 1.45 we have

$$\mathbf{B} = \mathsf{C}_n^*\circ\mathsf{B}_n\circ\mathsf{C}_n \quad,\quad \mathbf{B}^{-1} = \mathsf{C}^{-1}\circ\mathsf{B}_n^{-1}\circ(\mathsf{C}_n^*)^{-1}\ .$$

Then the submultiplicativity of operator norms and (1.12) finish the proof. □

Exercise 1.15. Assume that b induces an inner product on V, which renders V a Hilbert space. Given a finite dimensional subspace $V_n\subset V$ denote by C_n the coefficient isomorphism associated with a basis $\mathfrak{B}$ of V_n. Then

$$\lambda_{\max}(\mathbf{B}) = \|\mathsf{C}_n\|^2_{\mathbb{R}^N\mapsto V} \quad,\quad \lambda_{\min}(\mathbf{B}) = \left\|\mathsf{C}_n^{-1}\right\|^2_{V_n\mapsto\mathbb{R}^N}\ ,$$

where, for some matrix $\mathbf{M}\in\mathbb{R}^{N,N}$

$$\lambda_{\max}(\mathbf{M}) := \max\{|\lambda|\,,\ \lambda \text{ is eigenvalue of } \mathbf{M}\}\ ,$$
$$\lambda_{\min}(\mathbf{M}) := \min\{|\lambda|\,,\ \lambda \text{ is eigenvalue of } \mathbf{M}\}\ .$$

Exercise 1.16. Let b be a bilinear form on V and V_h, V_H be two **nested** subspaces of V, that is $V_H \subset V_h$. Equip both V_h and V_H with bases $\mathfrak{B}_h, \mathfrak{B}_H$ and write $\mathsf{C}_h : \mathbb{R}^n \mapsto V_h$, $n := \dim V_h$, and $\mathsf{C}_H : \mathbb{R}^N \mapsto V_H$, $N := \dim V_H$, for the associated coefficient isomorphisms.

(i) Based on C_h and C_H determine the matrix representations of the inclusion mappings $I_H^h : V_H \mapsto V_h$ and $(I_H^h)^* : V_h^* \mapsto V_H^*$, when dual bases are used for the dual spaces.

(ii) How can the matrix $\mathbf{B}_H \in \mathbb{R}^{N,N}$ associated with b w.r.t. $\mathfrak{B}_H$ be computed from the matrix $\mathbf{B}_h \in \mathbb{R}^{n,n}$ that we get as representation of b w.r.t. $\mathfrak{B}_h$?

Exercise 1.17. Given a finite dimensional subspace $V_n \subset V$, $N := \dim V_n$, endowed with a basis $\mathfrak{B}$, let C_n be the associated coefficient isomorphism.

(i) Give a bound for the norm of the "matrix" $\mathbf{M} := \mathsf{C}_n^* \mathsf{C}_n : \mathbb{R}^N \mapsto \mathbb{R}^N$.

(ii) We can view C_n also as a mapping $\overline{\mathsf{C}}_n : \mathbb{R}^N \mapsto V$ into V. What is $\mathrm{Ker}(\overline{\mathsf{C}}_n^*)$?

Exercise 1.18. Consider the quadratic functional on $L^2(]0,1[)$

$$J(v) := \int_0^1 (v(x) - \exp(x))^2 \, \mathrm{d}\xi \; .$$

(i) Formulate the linear variational problem related to the minimization of this functional. Discuss existence and uniqueness of solutions.

(ii) This variational problem is to be discretized in a Ritz-Galerkin fashion based on the following choices of trial/test spaces

a) $V_n = \mathrm{span}\{x^k,\, k = 0, \ldots, n-1\}$,

b) $V_n = \mathrm{span}\{\sin(k\pi x),\, k = 1, \ldots, n\}$,

c) $V_n = \mathrm{span}\{\chi_{[\frac{k-1}{n}, \frac{k}{n}]},\, k = 1, \ldots, n\}$, χ_I the characteristic function of the interval I,

where $n \in \mathbb{N}$ is the **discretization parameter** that serves as index for a family of trial/test spaces and also tells us $\dim V_n$. The characteristic function of an interval is defined as

$$\chi_{]a,b[}(x) = \begin{cases} 1 & \text{if } x \in]a,b[\; , \\ 0 & \text{elsewhere} \; . \end{cases}$$

Compute the resulting linear systems of equations that correspond to this trial/test spaces, when the functions occurring in the above definitions are used as bases.

(iii) Compute the condition numbers of the coefficient matrices of the linear systems of equations from the sub-task (ii) for the schemes (b) and (c).

(iv) For the scheme (a) compute the condition numbers of the linear systems of equations from the sub-task (ii) with the use of MATLAB.

(v) For scheme (c) compute the $L^2(]0,1[)$-norm of the discretization error as a function of the discretization parameter N.

Exercise 1.19. Consider the bilinear form $\mathsf{a}(u,v) := \int_0^1 u(x)v(x)\,\mathrm{d}\xi$ on $L^2(]0,1[)$. Test and trial spaces V_n, W_n are indexed by $n \in \mathbb{N}$ and specified through their bases

$$\mathfrak{B}_V := \{x \mapsto x^k, k = 0, \ldots, n-1\} \quad , \quad \mathfrak{B}_W := \{\chi_{]t_{k-1},t_k[}, k = 1, \ldots, n\} \,,$$

where $t_k = k/n$ and $\chi_{]t_{k-1},t_k[}$ denotes the characteristic function of $]t_{k-1}, t_k[$, see Ex. 1.18. Write a MATLAB code that computes the inf-sup constant γ_n in (DIS) for this choice of b, V_n, and W_n for a certain range of n.

2 Elliptic Boundary Value Problems

In this chapter we study a few elliptic boundary value problems on bounded domains. The focus will be on weak formulations, which fit the framework presented in the preceeding chapter. To this end we have to introduce appropriate function spaces that are known as Sobolev spaces.

2.1 Domains

We consider the partial differential equations of interest on bounded, connected, and open subsets of (the affine space) $\mathbb{R}^d$, $d = 1, 2, 3$. These are called the (spatial) **domains** of related boundary value problem and will be denoted by Ω. The topological closure $\overline{\Omega}$ of Ω will be compact and this is also true of its **boundary** $\Gamma := \partial\Omega := \overline{\Omega} \setminus \Omega$. A domain has an unbounded open complement $\Omega' := \mathbb{R}^d \setminus \overline{\Omega}$.

Example 2.1. For $d = 1$ admissible domains will be open intervals $]a, b[$, $a > b$, and $\Gamma = \{a, b\}$.

Meaningful boundary values for solutions of partial differential equations can only be imposed if we make addition assumptions on Γ. First, we recall that a function $f : U \subset \mathbb{R}^d \mapsto \mathbb{R}^m$, $d, m \in \mathbb{N}$, is Lipschitz continuous, if there is a $\gamma > 0$ such that

$$|f(\boldsymbol{\xi}) - f(\boldsymbol{\eta})| \leq \gamma |\boldsymbol{\xi} - \boldsymbol{\eta}| \quad \forall \boldsymbol{\xi}, \boldsymbol{\eta} \in \mathbb{R}^d \, .$$

Definition 2.2. *A domain $\Omega \subset \mathbb{R}^d$ is called a **Lipschitz** domain, if for every $x \in \Gamma$ we can find an open neighborhood U in $\mathbb{R}^d$ such that there is a bijective mapping*

$$\boldsymbol{\Phi} = (\Phi_1, \ldots, \Phi_d)^T : U \mapsto R := \{\boldsymbol{\xi} \in \mathbb{R}^d, \, |\xi_k| < 1\} \, ,$$

which satisfies

1. *Both $\boldsymbol{\Phi}$ and $\boldsymbol{\Phi}^{-1}$ are Lipschitz continuous.*
2. $\boldsymbol{\Phi}(U \cap \Gamma) = \{\boldsymbol{\xi} \in R : \, \xi_d = 0\}$.
3. $\boldsymbol{\Phi}(U \cap \Omega) = \{\boldsymbol{\xi} \in R : \, \xi_d < 0\}$.
4. $\boldsymbol{\Phi}(U \cap \Omega') = \{\boldsymbol{\xi} \in R : \, \xi_d > 0\}$.

If $\mathbf{\Phi}$ can be chosen to be k-times continuously differentiable, $k \in \mathbb{N}$, then Ω is said to be ***of class*** *C^k.*

The reader should be aware that there are a few examples of simple domains that do not qualify as Lipschitz domains.

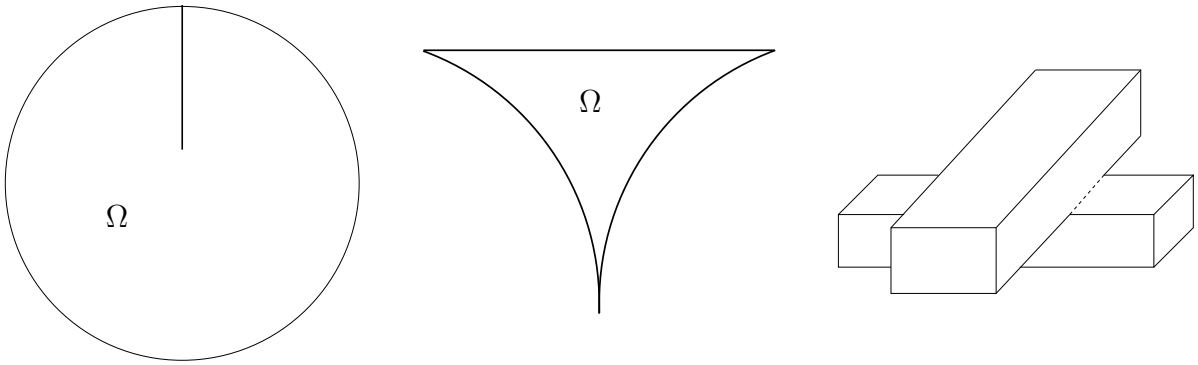


Figure 2.1: Domains that are not Lipschitz: slit domain (left), cusp domain (middle), crossing edges (right)

A profound result from measure theory asserts that a Lipschitz continuous function with values in $\mathbb{R}$ is differentiable almost everywhere with partial derivatives in L^∞. Therefore we can be define the **exterior unit vectorfield** $\mathbf{n} : \Gamma \mapsto \mathbb{R}^d$ by

$$\mathbf{n}(\boldsymbol{\xi}) := \frac{\left(\frac{\partial \Phi_d}{\partial \xi_1}(\boldsymbol{\xi}), \cdots, \frac{\partial \Phi_d}{\partial \xi_d}(\boldsymbol{\xi})\right)^T}{\left|\left(\frac{\partial \Phi_d}{\partial \xi_1}(\boldsymbol{\xi}), \cdots, \frac{\partial \Phi_d}{\partial \xi_d}(\boldsymbol{\xi})\right)^T\right|} \quad \text{for almost all } \boldsymbol{\xi} \in \Gamma \,, \tag{2.1}$$

where Φ_d is the d-th components of a $\mathbf{\Phi}$ from Def. 2.2 that belongs to a neighborhood of $\boldsymbol{\xi}$.

In almost all numerical computations only a special type of Lipschitz domains will be relevant, namely domains that can be described in the widely used CAD software packages.

Definition 2.3. *In the case $d = 2$ a connected domain Ω is called a* ***curvilinear Lipschitz polygon****, if Ω is a Lipschitz domain, and there are open subsets $\Gamma_k \subset \Gamma$, $k = 1, \ldots, P$, $P \in \mathbb{N}$, such that*

$$\Gamma := \overline{\Gamma}_1 \cup \cdots \cup \overline{\Gamma}_P \quad , \quad \Gamma_k \cap \Gamma_l = \emptyset \text{ if } k \neq l \,,$$

and for each $k \in \{1, \ldots, P\}$ there is a C^1-diffeomorphism $\mathbf{\Phi}_k : [0,1] \mapsto \overline{\Gamma}_k$.

The boundary segments are called edges, their endpoints are the vertices of Ω. A tangential direction can be defined for all points of an edge including the endpoints,

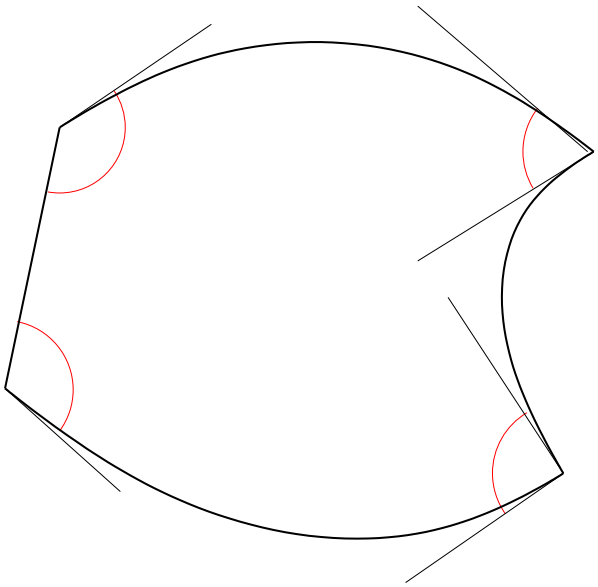


Figure 2.2: Curvilinear polygon with added angles at vertices

which gives rise to the concept of an angle at a vertex, see Fig. 2.2. The mappings $\mathbf{\Phi}_k$ can be regarded as smooth parametrizations of the edges. An analoguous notion exists in three dimensions. To give a rigorous definition we appeal to the intuitive concept of a closed **polyhedron** in $\mathbb{R}^3$, which can be obtained as a finite union of convex hulls of finitely many points in $\mathbb{R}^3$. The surface of a polyhedron can be split into flat faces. Moreover, it is clear what is meant by "edges" and "vertices".

Definition 2.4. *A connected domain $\Omega \subset \mathbb{R}^3$ is called a* ***curved Lipschitz polyhedron****, if*

1. *Ω is a Lipschitz domain.*
2. *there is a continuous bijective mapping $\mathbf{\Phi} : \partial\Pi \mapsto \Gamma$, where Π is a polyhedron with flat faces $F_1, \ldots, F_P$, $P \in \mathbb{N}$.*
3. *$\mathbf{\Phi} : \overline{F}_k \mapsto \mathbf{\Phi}(\overline{F}_k)$ is a C^1-diffeomorphism for every $k = 1, \ldots, P$.*

We call $\mathbf{\Phi}(F_k)$ the (open) face Γ_k of Ω. Further, $\mathbf{\Phi}$ takes edges and vertices of Π to edges and vertices of Ω.

Based on the parametrization we can introduce the surface measure $\mathrm{d}S$ and define integrals of measurable functions on Γ

$$\int_\Gamma f(\boldsymbol{\xi})\,\mathrm{d}S(\boldsymbol{\xi}) := \sum_{k=1}^{P} \int_{F_k=\mathbf{\Phi}^{-1}(\Gamma_k)} f(\mathbf{\Phi}(\widehat{\boldsymbol{\xi}}))\sqrt{\left|\det(D\mathbf{\Phi}(\widehat{\boldsymbol{\xi}})^T D\mathbf{\Phi}(\widehat{\boldsymbol{\xi}}))\right|}\,\mathrm{d}\widehat{\boldsymbol{\xi}}\,, \tag{2.2}$$

where $\mathrm{d}\widehat{\boldsymbol{\xi}}$ is the $d-1$-dimensional Lebesgue measure on the flat face $F_k = \mathbf{\Phi}^{-1}(\Gamma_k)$.

Definition 2.5. *A subset $\Omega \subset \mathbb{R}^d$ is called a* ***computational domain*** *if it is of class C^k, $k \geq 1$, or*

- *a bounded connected interval for $d = 1$.*
- *a curvilinear Lipschitz polygon for $d = 2$.*
- *a curved Lipschitz polyhedron for $d = 3$.*

Bibliographical notes. More information about classes of domains and their definition can be found in [43, § 2], [19, Ch. 1], and [36, Sect. A.1].

2.2 Linear differential operators

Let $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ be a multi-index, *i. e.* , a vector of d non-negative integers:

$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^n.$$

Set $|\boldsymbol{\alpha}| := \alpha_1 + \cdots + \alpha_n$ and denote by

$$\partial^{\boldsymbol{\alpha}} := \frac{\partial^{\alpha_1}}{\partial \xi_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_n}}{\partial \xi_n^{\alpha_n}}$$

the partial derivative of order $|\boldsymbol{\alpha}|$. Remember that for sufficiently smooth functions all partial derivatives commute. Provided that the derivatives exist, the **gradient** of a function $f : \Omega \subset \mathbb{R}^d \mapsto \mathbb{R}$ is the column vector

$$\mathbf{grad}\, f(\boldsymbol{\xi}) := \left(\frac{\partial f}{\partial \xi_1}(\boldsymbol{\xi}), \ldots, \frac{\partial f}{\partial \xi_1}(\boldsymbol{\xi}) \right)^T, \quad \boldsymbol{\xi} \in \Omega\,.$$

The **divergence** of a vector field $\mathbf{f} = (f_1, \ldots, f_d) : \Omega \subset \mathbb{R}^d \mapsto \mathbb{R}^d$ is the function

$$\operatorname{div} \mathbf{f}(\boldsymbol{\xi}) := \sum_{k=1}^{d} \frac{\partial f_k}{\partial \xi_k}(\boldsymbol{\xi})\,, \quad \boldsymbol{\xi} \in \Omega\,.$$

The differential operator $\Delta := \operatorname{div} \circ \mathbf{grad}$ is known as **Laplacian**. In the case $d = 3$ the **rotation** of a vectorfield $\mathbf{f} : \Omega \subset \mathbb{R}^3 \mapsto \mathbb{R}^3$ is given by

$$\mathbf{curl}\, \mathbf{f}(\boldsymbol{\xi}) := \begin{pmatrix} \dfrac{\partial f_3}{\partial \xi_2}(\boldsymbol{\xi}) - \dfrac{\partial f_2}{\partial \xi_3}(\boldsymbol{\xi}) \\ \dfrac{\partial f_1}{\partial \xi_3}(\boldsymbol{\xi}) - \dfrac{\partial f_3}{\partial \xi_1}(\boldsymbol{\xi}) \\ \dfrac{\partial f_2}{\partial \xi_1}(\boldsymbol{\xi}) - \dfrac{\partial f_1}{\partial \xi_2}(\boldsymbol{\xi}) \end{pmatrix}, \quad \boldsymbol{\xi} \in \Omega\,.$$

Notation: Bold roman typeface will be used for vector-valued functions, whereas plain print tags $\mathbb{R}(\mathbb{C})$-valued functions. For the k-th component of a vector valued function $\mathbf{f}$ we write f_k or, in order to avoid ambiguity, $(\mathbf{f})_k$.

For twice continuously differentiable functions/vectorfields we have

$$\mathbf{curl} \circ \mathbf{grad} = 0 \quad , \quad \mathrm{div} \circ \mathbf{curl} = 0 \; . \tag{2.3}$$

Transformations of vectorfields and functions under a change of variables will play a key role in both theory and implementation of numerical methods for partial differential equations. From another perspective, a change of variables can also be regarded as a transformation to another domain, but these points of view are perfectly dual.

The right transformation of a function crucially depends on the kind of differential operator that will act on the function in the actual model equations. The reason is that functions and vectorfields occur as mere representatives of certain tensors, which have their distinct transformation rules.

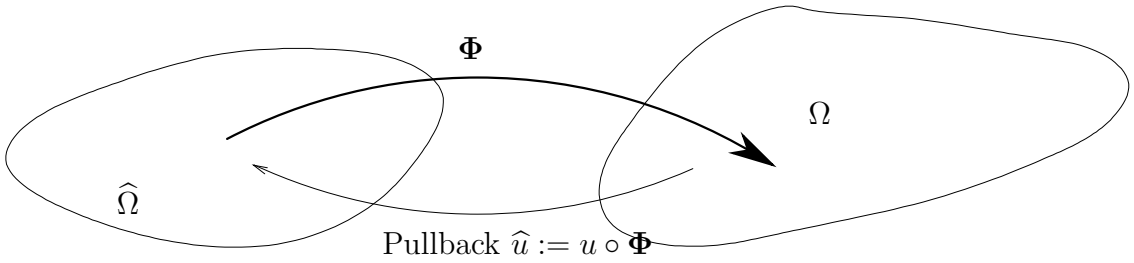


Figure 2.3: Mapping of the domain

Let us assume a diffeomorphic mapping $\mathbf{\Phi} : \widehat{\Omega} \mapsto \Omega$, $\widehat{\Omega}, \Omega \subset \mathbb{R}^d$. The transformation of a *potential type* (scalar) function $u : \Omega \mapsto \mathbb{R}$ is a straighforward **pullback**

$$\mathsf{FT}_{\mathbf{\Phi}}\, u := u \circ \mathbf{\Phi} : \widehat{\Omega} \mapsto \mathbb{R} \; . \tag{FT}$$

Notation: It it is clear from the context, what the underlying change of variables is, we will write $\widehat{u} := u \circ \mathbf{\Phi}$. In contrast to (FT) this does not really designate the transformation of a 0-tensor, but merely a plain change of variables (for functions and vectorfields alike).

A second class of functions are so-called densities $v : \Omega \mapsto \mathbb{R}$, whose transformation has to ensure that "mass", that is, the integral of v over a part of Ω is conserved. As a consequence of the transformation theorem for integrals, we find the correct transformation

$$\mathsf{DT}_{\mathbf{\Phi}}\, v = |\det D\mathbf{\Phi}|(v \circ \mathbf{\Phi}) \; . \tag{DT}$$

Here, $D\mathbf{\Phi} : \widehat{\Omega} \mapsto \mathbb{R}^{d,d}$ stands for the Jacobi matrix of $\mathbf{\Phi}$.

Another important class of functions comprises *gradient type* vectorfields $\mathbf{u} : \Omega \mapsto \mathbb{R}^d$. For them the appropriate transformation reads

$$\mathsf{GT}_{\boldsymbol{\Phi}}\,\mathbf{u} := D\boldsymbol{\Phi}^T(\mathbf{u} \circ \boldsymbol{\Phi})\ . \tag{GT}$$

The next result underscores that (GT) is the correct transformation of gradients.

Lemma 2.6. *For any continuously differentiable function $u : \Omega \mapsto \mathbb{R}$*

$$\mathsf{GT}_{\boldsymbol{\Phi}}(\mathbf{grad}\, u) = \widehat{\mathbf{grad}}(\mathsf{FT}_{\boldsymbol{\Phi}}\, u)\ ,$$

where $\widehat{\mathbf{grad}}$ indicates that the gradient is computed w.r.t. the $\widehat{\ }$-variables.

Proof. A straightforward application of the chain rules confirms

$$\frac{\partial}{\partial \widehat{\xi}_k} u(\boldsymbol{\Phi}(\widehat{\boldsymbol{\xi}})) = \left\langle \mathbf{grad}\, u(\boldsymbol{\Phi}(\widehat{\boldsymbol{\xi}})\, , \frac{\partial \boldsymbol{\Phi}}{\partial \widehat{\xi}_k} \right\rangle ,$$

which amounts to the assertion of the lemma written componentwise. □

Moreover, the transformation (GT) leaves path integrals invariant:

Lemma 2.7. *Let γ be a piecewise smooth oriented curve in $\widehat{\Omega}$. Then*

$$\int_\gamma \mathsf{GT}_{\boldsymbol{\Phi}}\,\mathbf{u} \cdot \mathrm{d}\boldsymbol{s} = \int_{\boldsymbol{\Phi}(\gamma)} \mathbf{u} \cdot \mathrm{d}\boldsymbol{s}$$

for all continuous vectorfields $\mathbf{u} \in (C^0(\widehat{\Omega}))^d$.

Proof. Assume a piecewise smooth continuous parameterization $\gamma : [0,1] \mapsto \gamma(t) \in \widehat{\Omega}$. Then, by the definition of the path integral,

$$\begin{aligned}\int_\gamma \mathsf{GT}_{\boldsymbol{\Phi}}\,\mathbf{u} \cdot \mathrm{d}\boldsymbol{s} &= \int_0^1 \left\langle D\boldsymbol{\Phi}(\gamma(\tau))^T \mathbf{u}(\boldsymbol{\Phi}(\gamma(\tau))), \frac{d\gamma}{dt}(\tau) \right\rangle \mathrm{d}\tau \\ &= \int_0^1 \left\langle \mathbf{u}(\boldsymbol{\Phi}(\gamma(\tau))), \frac{d}{dt}(\boldsymbol{\Phi} \circ \gamma)(\tau) \right\rangle \mathrm{d}\tau = \int_{\boldsymbol{\Phi}(\gamma)} \mathbf{u} \cdot \mathrm{d}\boldsymbol{s}\ .\end{aligned}$$

□

Notation: The symbol $\cdot \mathrm{d}\boldsymbol{s}$ is used to indicate the path integral of vector field along a piecewise smooth curve with respect to the arclength measure. Similarly, we use $\cdot \mathbf{n} \mathrm{d}S$ to refer the integration of the (exterior) normal component of a vectorfield over an oriented piecewise smooth surface ($d-1$-dimensional manifold)

Another type of vector fields that we will often come across are *flux fields*. Their behavior under a change of variables is governed by the *Piola transformation*

$$\mathsf{PT}_{\boldsymbol{\Phi}}\,\mathbf{u} := |\det D\boldsymbol{\Phi}| D\boldsymbol{\Phi}^{-1}(\mathbf{u} \circ \boldsymbol{\Phi})\ . \tag{PT}$$

Of course the transformation of a flux field should leave the total flux through any oriented surface invariant:

Lemma 2.8. *Let $\widehat{\Sigma}$ be an oriented piecewise smooth $d-1$-dimensional submanifold ("surface") in $\widehat{\Omega}$ and $\Sigma \subset \Omega$ its image under $\mathbf{\Phi}$. Then*

$$\int_{\widehat{\Sigma}} \mathsf{PT}_{\mathbf{\Phi}}\, \mathbf{u} \cdot \mathbf{n} \mathrm{d}\widehat{S} = \int_{\Sigma} \mathbf{u} \cdot \mathbf{n} \mathrm{d}S$$

for any continuous vectorfield $\mathbf{u} \in (C^0(\widehat{\Omega}))^d$.

Taking the divergence of a flux field produces a density, which is reflected by the following commuting diagram:

Lemma 2.9. *For any continuously differentiable vectorfield $\mathbf{u}$ on Ω we have*

$$\mathsf{DT}_{\mathbf{\Phi}}(\operatorname{div} \mathbf{u}) = \widehat{\operatorname{div}}(\mathsf{PT}_{\mathbf{\Phi}}\, \mathbf{u}) \,.$$

Proof. Pick some bounded subdomain $\widehat{U} \subset \widehat{\Omega}$ with smooth boundary. Thanks to Lemma 2.8, applying Gauss' theorem (Thm. 2.17 in Sect. 2.4) twice shows

$$\int_{\widehat{U}} \widehat{\operatorname{div}}(\mathsf{PT}_{\mathbf{\Phi}}\, \mathbf{u}) \,\mathrm{d}\widehat{\boldsymbol{\xi}} = \int_{\partial\widehat{U}} \mathsf{PT}_{\mathbf{\Phi}}\, \mathbf{u} \cdot \mathbf{n} \mathrm{d}\widehat{S} = \int_{\partial U} \mathbf{u} \cdot \mathbf{n} \mathrm{d}S = \int_{U} \operatorname{div} \mathbf{u} \,\mathrm{d}\boldsymbol{\xi} = \int_{\widehat{U}} \mathsf{DT}_{\mathbf{\Phi}}(\operatorname{div} \mathbf{u}) \,\mathrm{d}\widehat{\boldsymbol{\xi}} \,,$$

where $U := \mathbf{\Phi}(\widehat{U}) \subset \Omega$. Since, $\widehat{U}$ is arbitrary and $\mathbf{u}$ continuously differentiable, the assertion follows immediately. □

For $d = 3$, the transformations $\mathsf{GT}_{\mathbf{\Phi}}$ and $\mathsf{PT}_{\mathbf{\Phi}}$ fit the **curl**-operator in the sense of the following commuting diagram:

Lemma 2.10. *For any $\mathbf{u} \in (C^1(\Omega))^3$ holds true*

$$\mathsf{PT}_{\mathbf{\Phi}}(\mathbf{curl}\, \mathbf{u}) = \widehat{\mathbf{curl}}(\mathsf{GT}_{\mathbf{\Phi}}\, \mathbf{u}) \,.$$

Proof. Let $\widehat{\Sigma} \subset \widehat{\Omega}$ denote some smooth oriented two-dimensional submanifold of $\widehat{\Omega}$. Let Σ be its image under the transformation $\mathbf{\Phi}$. Recall Stokes' theorem

$$\int_{\partial\Sigma} \mathbf{u} \cdot \mathrm{d}\boldsymbol{s} = \int_{\Sigma} \mathbf{curl}\, \mathbf{u} \cdot \mathbf{n} \mathrm{d}S \quad \forall \mathbf{u} \in C^1(\Omega) \,,$$

where $\partial\Sigma$ bears the induced orientation.

Similar to the proof of Lemma 2.9 can exploit Lemmas 2.8 and 2.7

$$\begin{aligned} \int_{\widehat{\Sigma}} \widehat{\mathbf{curl}}(\mathsf{GT}_{\mathbf{\Phi}}\, \mathbf{u}) \cdot \mathbf{n} \mathrm{d}\widehat{S} &= \int_{\partial\widehat{\Sigma}} \mathsf{GT}_{\mathbf{\Phi}}\, \mathbf{u} \cdot \mathrm{d}\boldsymbol{s} = \int_{\partial\Sigma} \mathbf{u} \cdot \mathrm{d}\boldsymbol{s} \\ &= \int_{\Sigma} \mathbf{curl}\, \mathbf{u} \cdot \mathbf{n} \mathrm{d}S = \int_{\widehat{\Sigma}} \mathsf{PT}_{\mathbf{\Phi}}(\mathbf{curl}\, \mathbf{u}) \,\cdot \mathbf{n} \mathrm{d}\widehat{S}. \end{aligned}$$

The "test surface" Σ being arbitrary, the assertion is proved, because $\mathbf{u}$ has been assumed to be smooth. □

2.3 Second-order boundary value problems

Second order boundary value problems (BVPs) provide the most important class of models for stationary (equilibrium) phenomena: on a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$, we consider the partial differential equations

$$\mathbf{j} = -\mathbf{A}(\boldsymbol{\xi})\ \mathbf{grad}\, u\ , \tag{FL}$$
$$\operatorname{div} \mathbf{j} = f - c(\boldsymbol{\xi})\, u\ . \tag{EL}$$

The scalar function $u : \Omega \mapsto \mathbb{R}$ is called a **potential**, whereas we refer to the vectorfield $\mathbf{j} : \Omega \mapsto \mathbb{R}^d$ as **flux**. The function $f : \Omega \mapsto \mathbb{R}$ plays the role of a **source term** and will represent one of the input data of the problem. The first equation may be labelled a **flux law**, while the second is the mathematical way to express an **equilibrium** or **conservation law**.

Remark 2.11. Boundary value problems for (FL) and (EL) are pervasive in science and engineering. They are used in various models of **stationary equilibrium states**.

- Heat conduction:
  $u \to$ temperature, $[u] = 1\mathrm{K}$
  $\mathbf{j} \to$ heat flux, $[\mathbf{j}] = 1\frac{\mathrm{W}}{\mathrm{m}^2}$
  $f \to$ heat source/sink, $[f] = 1\frac{\mathrm{W}}{\mathrm{m}^3}$

  In this case $\mathbf{A}$ represents the heat conductivity tensor and $c \equiv 0$. The equation (FL) is known as Fourier's law, while (EL) ensures the conservation of energy.

- Electrostatics:
  $u \to$ electric potential, $[u] = 1\mathrm{V}$
  $\mathbf{j} \to$ displacement current ($\mathbf{D}$), $[\mathbf{j}] = 1\frac{\mathrm{As}}{\mathrm{m}^2}$
  $f \to$ charge density (ρ), $[f] = 1\frac{\mathrm{As}}{\mathrm{m}^3}$

  Here, $\mathbf{A}$ stands for the dielectric tensor, which is usually designated by ϵ, whereas $c \equiv 0$. The relationship (EL) is Gauss' law, and (FL) arises from Faraday's law $\mathbf{curl}\, \mathbf{E} = 0$ and the linear constitutive law $\mathbf{D} = \epsilon \mathbf{E}$.

- Stationary electric currents:
  $u \to$ electric potential, $[u] = 1\mathrm{V}$
  $\mathbf{j} \to$ electric current, $[\mathbf{j}] = 1\frac{\mathrm{A}}{\mathrm{m}^2}$

  In this case, the source term f usually vanishes and excitation is solely provided by non-homogeneous boundary conditions. The tensor $\mathbf{A}$ represents the conductivity and $c \equiv 0$. In this context (FL) can be regarded as Ohm's law and (EL) is a consequence of the conservation of charge.

- Molecular diffusion:
  $u \to$ concentration, $[u] = 1\frac{\mathrm{mol}}{\mathrm{m}^3}$
  $\mathbf{j} \to$ flux, $[\mathbf{j}] = 1\frac{\mathrm{mol}}{\mathrm{m}^2\mathrm{s}}$
  $f \to$ production/consumption rate, $[f] = 1\frac{\mathrm{mol}}{\mathrm{m}^3\mathrm{s}}$

  Here $\mathbf{A}$ stands for the diffusion constant and, if non-zero, c denotes a so-called reaction coefficient. The equation (EL) guarantees the conservation of total mass of the relevant species.

- Linear elasticity: $u \to$ vertical displacement $[u] = 1\mathrm{m}$
  $\mathbf{j} \to$ stress $[\mathbf{j}] = 1\frac{\mathrm{N}}{\mathrm{m}^2}$
  $f \to$ external load $[f] = 1\frac{\mathrm{N}}{\mathrm{m}^3}$

  For $d = 2$, the equations (FL) and (EL) provide a reduced model for computing the shape of an elastic membrane under small external forces. This model arises from the fundamental equations of elasticity after linearization and reduction to two dimensions. Then, $\mathbf{A}$ is a tensor describing the elastic properties of the membrane and c will usually vanish.

Aware of these physical models and the laws of thermodynamics, we make the following natural assumptions on the **coefficient functions** $\mathbf{A} : \Omega \mapsto \mathbb{R}^{d,d}$, $c : \Omega \mapsto \mathbb{R}$:

- The $d \times d$-matrix $\mathbf{A}(\boldsymbol{\xi})$ is symmetric almost everywhere in Ω, and features entries in $L^\infty(\Omega)$. Moreover, there are to be constants $0 < \underline{\gamma} < \overline{\gamma} < \infty$ such that

$$\underline{\gamma}|\boldsymbol{\mu}|^2 \leq \boldsymbol{\mu}^T \mathbf{A}(\boldsymbol{\xi})\boldsymbol{\mu} \leq \overline{\gamma}|\boldsymbol{\mu}|^2 \quad \forall \boldsymbol{\mu} \in \mathbb{R}^d \text{ and almost all } \boldsymbol{\xi} \in \Omega\,. \tag{UPD}$$

  In words, $\mathbf{A}$ is supposed to be uniformly positive definite and bounded on Ω.

- The function c belongs to $L^\infty(\Omega)$ and $c \geq 0$ almost everywhere in Ω.

Without loss of generality we may even restrict ourselves to coefficient functions $\mathbf{A}$ and c that are piecewise smooth with respect to a partition of Ω into a few sub-intervals ($d = 1$), Lipschitz polygons ($d = 2$), or Lipschitz polyhedra ($d = 3$). This corresponds to the widely encountered case of a computational domain filled with different materials that themselves are fairly homogeneous.

Notation: For the sake of brevity we will often supress the dependence of coefficient functions from the spatial variable $\boldsymbol{\xi} \in \Omega$, that is we write $\mathbf{A}$ and c instead of $\mathbf{A}(\boldsymbol{\xi})$ and $c(\boldsymbol{\xi})$, respectively. However, unless mentioned explicitly, the coefficient functions must not be assumed to be constant.

The partial differential equations (FL) and (EL) have to be supplemented with **boundary conditions**. We distinguish several types of them:

- **Dirichlet boundary conditions.** They amount to fixing the value of u on (a part Γ_D of) the boundary $\Gamma := \partial\Omega$.

- **Neumann boundary conditions.** They prescribe the normal flux $\langle \mathbf{j}, \mathbf{n} \rangle$ on (a part Γ_N of) Γ.

- **Robin boundary conditions.** They impose a linear relationship between the normal flux and the values of the potentials on (a part Γ_R of) Γ:

$$\langle \mathbf{j}, \mathbf{n} \rangle = q\,u \quad \text{almost everywhere on } \Gamma_R\,,$$

  where $q \in L^\infty(\Gamma)$ is uniformly positive almost everywhere on Γ.

Mixed boundary conditions prevail, if different types of the above boundary conditions are prescribed on different parts of Γ. The most general situation assumes that Γ is partitioned into a Dirichlet, Neumann, and Robin part: $\Gamma = \overline{\Gamma}_D \cup \overline{\Gamma}_N \cup \overline{\Gamma}_R$, $\Gamma_D, \Gamma_N, \Gamma_R$ mutually disjoint, where both Γ_N and Γ_D have non-zero measure. In this case the mixed boundary conditions read

$$u = g \quad \text{on } \Gamma_D \quad , \quad -\langle \mathbf{j}, \mathbf{n} \rangle = h \quad \text{on } \Gamma_N \quad , \quad \langle \mathbf{j}, \mathbf{n} \rangle = q\, u \quad \text{on } \Gamma_R \, ,$$

with suitable functions $g, h, q : \Gamma \mapsto \mathbb{R}$.

Remark 2.12. For the particular physical models listed in Remark 2.11 the different boundary conditions have the following meaning:

- Heat conduction: Dirichlet boundary conditions model a prescribed temperature, Neumann boundary conditions a fixed heat flux through Γ_N. Robin boundary conditions amount to "cooling conditions" for which the heat flux is proportional to a temperature difference.
- Electrostatics and stationary currents: On Γ_D the potential is prescribed, whereas the current density through Γ_N is fixed. Robin boundary conditions are known as impedance boundary conditions.
- Molecular diffusion: The concentration is kept constant at Γ_D (for instance, on the surface of a sediment layer), whereas release/absorption of the compound at a constant rate occurs at Γ_N.
- Linear elasticity: The displacement of the membrane is enforced on Γ_D. On Γ_N we usually have $\langle \mathbf{j}, \mathbf{n} \rangle = 0$, which describes an edge of the membrane, on which no force is exerted.

Definition 2.13. *Assume that* $\mathbf{A}$ *and* c *are bounded and continuously differentiable on* Ω. *If* $u \in C^1(\Omega) \cap C^0(\overline{\Omega})$, $\mathbf{j} \in (C^1(\Omega))^d \cap (C^0(\overline{\Omega}))^d$ *satisfy* (FL), (EL) *in a pointwise sense, and the prescribed boundary conditions, then these functions are called a* ***classical solution*** *of the boundary value problem.*

Remark 2.14. In general it is impossible to establish existence and uniqueness of classical solutions. This can be achieved for constant coefficients and pure Dirichlet or Neumann boundary conditions, see [13, Vol. I].

Remark 2.15. Boundary value problems that are closely related to (FL) and (EL) arise in electromagnetism in the context of the eddy current model. Then the various quantities have the following meaning

$$\mathbf{curl}\, \mathbf{u} = -\mathbf{A}\, \mathbf{j} \, , \tag{FAL}$$

$$\mathbf{curl}\, \mathbf{j} = \mathbf{f} + \mathbf{C}\, \mathbf{u} \, . \tag{AML}$$

$$\begin{array}{ll} \mathbf{u} \to \text{electric field } (\mathbf{E}) & [\mathbf{u}] = 1\frac{\mathrm{V}}{\mathrm{m}} \\ \mathbf{j} \to \text{magnetic field } (\mathbf{H}) & [\mathbf{j}] = 1\frac{\mathrm{A}}{\mathrm{m}} \\ \mathbf{f} \to \text{“generator current”} & [\mathbf{f}] = 1\mathrm{A}\mathrm{m}^{-2} \end{array}$$

Here, $\mathbf{A}$ is the tensor of the magnetic permeability, whereas $\mathbf{C}$ designates the conductivity[1]. Both satisfy the assumptions on $\mathbf{A}$ made for (FL) and (EL).

The boundary value problem is posed for $d = 3$ and the following boundary conditions are appropriate:

- **Dirichlet/Neumann boundary conditions.** They prescribe the *tangential components* of either $\mathbf{u}$ or $\mathbf{j}$ on (parts of) Γ. For instance, this can be expressed by

  $$\mathbf{u} \times \mathbf{n} = \mathbf{g} \quad \text{on } \Gamma \, .$$

  In contrast to (FL) and (EL), in the case of (FAL) and (AML) $\mathbf{u}$ and $\mathbf{j}$ are completely symmetric. A distinction between Dirichlet and Neumann boundary conditions is no longer possible.

- **Impedance boundary conditions.** This counterpart of the Robin boundary conditions reads

  $$(\mathbf{n} \times \mathbf{u}) \times \mathbf{n} = q\,(\mathbf{j} \times \mathbf{n}) \quad \text{on } \Gamma \, .$$

Of course, mixed boundary conditions that impose the tangential components of either $\mathbf{u}$ or $\mathbf{j}$ on different parts of Γ are also possible.

So far, we have written the boundary value problems as **first-order systems** of partial differential equations. By various elimination strategies, they can be converted into equations involving second-order differential operators.

Plugging (FL) into (EL) and the boundary conditions, we get the following **primal version** of a general scalar second-order elliptic boundary value problem with the underlying differential operator written in **divergence form**

$$\begin{array}{rcll} -\operatorname{div}(\mathbf{A}\,\mathbf{grad}\,u) + cu &=& f & \text{in } \Omega \, , \\ u &=& g & \text{on } \Gamma_D \, , \\ \langle \mathbf{A}\,\mathbf{grad}\,u, \mathbf{n} \rangle &=& h & \text{on } \Gamma_N \, , \\ \langle \mathbf{A}\,\mathbf{grad}\,u, \mathbf{n} \rangle + q\,u &=& 0 & \text{on } \Gamma_R \, . \end{array} \tag{E2P}$$

Here we take for granted an underlying partition $\Gamma = \overline{\Gamma}_D \cup \overline{\Gamma}_N \cup \overline{\Gamma}_R$ of the boundary Γ.

However, if c is unformly positive almost everywhere on Ω, then u can be eliminated from (FL), which leads to the **dual version** of the second-order elliptic boundary value

[1] More precisely, the equations are related to the eddy current model in frequency domain, for which c is the conductivity scaled by a purely imaginary factor.

problem

$$\begin{aligned} -\mathbf{grad}(c^{-1}\operatorname{div}\mathbf{j}) + \mathbf{A}^{-1}\mathbf{j} &= -\mathbf{grad}(c^{-1}f) && \text{in } \Omega\,, \\ c^{-1}(\operatorname{div}\mathbf{j} - f) &= -g && \text{on } \Gamma_D\,, \\ \langle \mathbf{j}, \mathbf{n}\rangle &= -h && \text{on } \Gamma_N\,, \\ c\,\langle \mathbf{j}, \mathbf{n}\rangle + q\,\operatorname{div}\mathbf{j} &= qf && \text{on } \Gamma_R\,. \end{aligned} \tag{E2D}$$

Remark 2.16. Similar manipulations can be aimed at (FAL) and (AML) and will yield the partial differential equations

$$\mathbf{curl}\,\mathbf{A}^{-1}\,\mathbf{curl}\,\mathbf{u} + \mathbf{C}\,\mathbf{u} = -\mathbf{f} \quad \text{in } \Omega\,, \tag{2.4}$$

$$\mathbf{curl}\,\mathbf{C}^{-1}\,\mathbf{curl}\,\mathbf{j} + \mathbf{A}\,\mathbf{j} = \mathbf{curl}(\mathbf{C}^{-1}\mathbf{f}) \quad \text{in } \Omega\,. \tag{2.5}$$

It is clear how to incorporate the tangential boundary conditions.

2.4 Integration by parts

The boundary conditions proposed in the previous section have a clear physical significance. *cf.* Remark 2.12. The mathematical link between the boundary conditions and the differential operators is provided by integration by parts formulas.

Below we assume that $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$, is bounded and an interval for $d = 1$, a Lipschitz polygon for $d = 2$, and a Lipschitz polyhedron for $d = 3$. Throughout we adopt the notation $\mathbf{n} = (n_1, \ldots, n_d)^T$ for the exterior unit normal vectorfield that is defined almost everywhere on $\Gamma := \partial\Omega$, see Sect. 2.1.

Integration by parts formulas have their roots in the calculus of differential forms [11], but here we will pursue a classical treatment that sets out from **Gauss' theorem**.

Theorem 2.17 (Gauss' theorem). *If* $\mathbf{f} \in (C^1(\Omega))^d \cap (C^0(\overline{\Omega}))^d$, *then*

$$\int_\Omega \operatorname{div}\mathbf{f}\,\mathrm{d}\boldsymbol{\xi} = \int_\Gamma \langle \mathbf{f}, \mathbf{n}\rangle\,\mathrm{d}S\,.$$

Proof. Please consult [15, § 15] and [28]. □

By the product rule

$$\operatorname{div}(u\,\mathbf{f}) = u\,\operatorname{div}\mathbf{f} + \langle \mathbf{grad}\,u, \mathbf{f}\rangle$$

for $u \in C^1(\Omega)$, $\mathbf{f} \in (C^1(\Omega))^d$, we deduce the **first Green formula**

$$\int_\Omega \langle \mathbf{f}, \mathbf{grad}\,u\rangle + \operatorname{div}\mathbf{f}\,u\,\mathrm{d}\boldsymbol{\xi} = \int_\Gamma \langle \mathbf{f}, \mathbf{n}\rangle\,u\mathrm{d}S \tag{FGF}$$

for all $\mathbf{f} \in (C^1(\Omega))^d \cap (C^0(\overline{\Omega}))^d$ and all $u \in C^1(\Omega) \cap C^0(\overline{\Omega})$. Plugging in the special $\mathbf{f} = f\boldsymbol{\epsilon}_k$, $k = 1, \ldots, d$, $\boldsymbol{\epsilon}_k$ the k-th unit vector, we get

$$\int_\Omega f\,\frac{\partial u}{\partial \xi_k} + \frac{\partial f}{\partial \xi_k}\,u\,\mathrm{d}\boldsymbol{\xi} = \int_\Gamma f u\,n_k\,\mathrm{d}S \tag{IPF}$$

for $f, u \in C^1(\Omega) \cap C^0(\overline{\Omega})$. We may also plug $\mathbf{f} = \mathbf{grad}\, v$ into (FGF), which yields

$$\int_\Omega \langle \mathbf{grad}\, v, \mathbf{grad}\, u \rangle + \Delta v\, u \,\mathrm{d}\boldsymbol{\xi} = \int_\Gamma \langle \mathbf{grad}\, v, \mathbf{n} \rangle \; u \,\mathrm{d}S \tag{2.6}$$

for all $v \in C^2(\Omega) \cap C^1(\overline{\Omega})$, $u \in C^1(\Omega) \cap C^0(\overline{\Omega})$.

In three dimensions, $d = 3$, another product rule ($\times$ stands for the antisymmetric vector product)

$$\operatorname{div}(\mathbf{u} \times \mathbf{f}) = \langle \mathbf{curl}\, \mathbf{u}, \mathbf{f} \rangle - \langle \mathbf{u}, \mathbf{curl}\, \mathbf{f} \rangle$$

for continuously differentiable vectorfields $\mathbf{u}, \mathbf{f} \in (C^1(\Omega))^3$ can be combined with Gauss' theorem, and we arrive at

$$\int_\Omega \langle \mathbf{curl}\, \mathbf{u}, \mathbf{f} \rangle - \langle \mathbf{u}, \mathbf{curl}\, \mathbf{f} \rangle \; \mathrm{d}\boldsymbol{\xi} = \int_\Gamma \langle \mathbf{u} \times \mathbf{f}, \mathbf{n} \rangle \; \mathrm{d}S \tag{CGF}$$

Remark 2.18. For $d = 1$ Gauss' theorem boils down to the fundamental theorem of calculus, and (FGF) becomes the ordinary integration by parts formula.

Exercise 2.1. Show that for $\mathbf{f} \in (C^1(\Omega))^d \cap (C^0(\overline{\Omega}))^d$

$$\int_\Omega \mathbf{curl}\, \mathbf{f} \,\mathrm{d}\boldsymbol{\xi} = \int_\Gamma \mathbf{f} \times \mathbf{n} \,\mathrm{d}S \,.$$

Applying suitable integration by parts formulas reveals the canonical boundary conditions associated with a linear partial differential operator. Let $\mathcal{L}$ be a general $p \times p$ ($p \in \mathbb{N}$) system of linear partial differential operators acting on p real valued functions.

Setting $\mathbf{u} = (u_1, \ldots, u_p)^T$, we multiply $\mathcal{L}\mathbf{u}$ by $\mathbf{v} = (v_1, \ldots, v_p)^T$, $u_k, v_k \in C^\infty(\Omega)$, and integrate over Ω. By successive application of (IPF) all derivations are shifted onto $\mathbf{v}$, which yields

$$\int_\Omega \langle \mathcal{L}\,\mathbf{u}, \mathbf{v} \rangle - \langle \mathbf{u}, \mathcal{L}^*\,\mathbf{v} \rangle \; \mathrm{d}\boldsymbol{\xi} = \int_\Gamma \{\text{boundary terms}\} \,\mathrm{d}S \,. \tag{2.7}$$

Here, $\mathcal{L}^*$ is the so-called adjoint differential operator, and the boundary terms will tell us the boundary conditions pertinent to $\mathcal{L}$: they are composed of bilinear pairings (of derivatives) of components of $\mathbf{u}$ and $\mathbf{v}$. Sloppily speaking, the pairings will always comprise a Dirichlet and a Neumann boundary value.

Example 2.19. The system (FL) and (EL) can be rewritten as

$$\mathcal{L} \begin{pmatrix} \mathbf{j} \\ u \end{pmatrix} = \begin{pmatrix} \mathbf{A}^{-1} & \mathbf{grad} \\ -\operatorname{div} & -c \end{pmatrix} \begin{pmatrix} \mathbf{j} \\ u \end{pmatrix} = \begin{pmatrix} 0 \\ -f \end{pmatrix} \,.$$

Let us set $\mathbf{u} = (\mathbf{j}, u)^T$, $\mathbf{v} = (\mathbf{q}, v)^T$ and carry out formal integration by parts using (FGF)

$$\begin{aligned}\int_\Omega \langle \mathcal{L}\mathbf{u}, \mathbf{v}\rangle \,\mathrm{d}\boldsymbol{\xi} &= \int_\Omega \left\langle \mathbf{A}^{-1}\mathbf{j}, \mathbf{q}\right\rangle + \langle \mathbf{grad}\, u, \mathbf{q}\rangle - \operatorname{div}\mathbf{j}\, v - c\, uv \,\mathrm{d}\boldsymbol{\xi} \\ &= \int_\Omega \left\langle \mathbf{j}, \mathbf{A}^{-1}\mathbf{q}\right\rangle - \langle u, \operatorname{div}\mathbf{q}\rangle + \langle \mathbf{j}, \mathbf{grad}\, v\rangle - c\, uv \,\mathrm{d}\boldsymbol{\xi} \\ &\quad + \int_\Gamma u\, \langle \mathbf{q}, \mathbf{n}\rangle - \langle \mathbf{j}, \mathbf{n}\rangle\, v \,\mathrm{d}S\,.\end{aligned}$$

This shows that in this case $\mathcal{L} = \mathcal{L}^*$ and that the Dirichlet and Neumann boundary conditions introduced above fit the first order system of partial differential equations.

Example 2.20. We investigate the formal procedure in the case of scalar second-order differential operators in divergence form: if $\mathcal{L} = -\operatorname{div}(\mathbf{A}(\boldsymbol{\xi})\,\mathbf{grad})$ then we have to apply formula (FGF) twice:

$$\begin{aligned}\int_\Omega \mathcal{L}\, u\, v \,\mathrm{d}\boldsymbol{\xi} &= \int_\Omega \langle \mathbf{A}\,\mathbf{grad}\, u, \mathbf{grad}\, v\rangle \,\mathrm{d}\boldsymbol{\xi} - \int_\Gamma \langle \mathbf{A}\,\mathbf{grad}\, u, \mathbf{n}\rangle\, v \,\mathrm{d}S \\ &= \int_\Omega u\, \mathcal{L}\, v \,\mathrm{d}\boldsymbol{\xi} - \int_\Gamma \langle \mathbf{A}\,\mathbf{grad}\, u, \mathbf{n}\rangle\, v - u\, \langle \mathbf{A}\,\mathbf{grad}\, v, \mathbf{n}\rangle \,\mathrm{d}S\,.\end{aligned}$$

Pairings of the Dirichlet and Neumann boundary conditions imposed in problem (E2P) constitute the boundary terms.

Exercise 2.2. Determine the formal adjoint and boundary terms for the second order differential operator of (2.4).

Exercise 2.3. The **Stokes system** in d dimensions reads

$$\begin{pmatrix} \boldsymbol{\Delta} & \mathbf{grad} \\ -\operatorname{div} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ p \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ 0 \end{pmatrix}\,,$$

where $\mathbf{u}$ is a vectorfield with d components (velocity) and p a real valued function (pressure). The symbol $\boldsymbol{\Delta}$ means the Laplacian applied to the components of a vector field. Find the formal adjoint of the Stokes operator and the associated boundary conditions.

2.5 Formal weak formulations

We retain setting and notations of the previous section. In order to determine the formal adjoint acording to (2.7) we moved all derivatives onto the *test function* $\mathbf{v}$. The formal weak formulation emerges by carrying this procedure half way through. We call it formal, because we are going to take for granted that all functions occurring in the various expressions are sufficiently smooth to render them well defined.

The derivation of the formal weak formulation is best explained in the case of second order operator, because in this case the construction of the formal adjoint already involves two similar steps, one of which can simply be omitted. Consider a second order scalar partial differential equation in divergence form, *cf.* (E2P) and Ex. 2.20,

$$\mathcal{L}\,u = -\operatorname{div}(\mathbf{A}\,\mathbf{grad}\,u) + c\,u = f \quad \text{in } \Omega\,.$$

Multiplying by a test function $v \in C^1(\Omega) \cap C^0(\overline{\Omega})$ and applying the first Green formula (FGF) leads to

$$\int_\Omega \langle \mathbf{A}\,\mathbf{grad}\,u, \mathbf{grad}\,v\rangle + c\,uv\,\mathrm{d}\boldsymbol{\xi} - \int_\Gamma \langle \mathbf{A}\,\mathbf{grad}\,u, \mathbf{n}\rangle\; v\,\mathrm{d}S = \int_\Omega fv\,\mathrm{d}\boldsymbol{\xi} \qquad \text{(FWP)}$$

for all $v \in C^1(\Omega) \cap C^0(\overline{\Omega})$.

The same approach also succeeds with the partial differential equation from (E2D)

$$\mathcal{L}\,\mathbf{j} = -\,\mathbf{grad}(c^{-1}\operatorname{div}\mathbf{j}) + \mathbf{A}^{-1}\mathbf{j} = \mathbf{f} \quad \text{in } \Omega\,,$$

that is related to (E2D), and gives us

$$\int_\Omega c^{-1}\operatorname{div}\mathbf{j}\;\operatorname{div}\mathbf{q} + \left\langle \mathbf{A}^{-1}\mathbf{j}, \mathbf{q}\right\rangle\;\mathrm{d}\boldsymbol{\xi} - \int_\Gamma c^{-1}\operatorname{div}\mathbf{j}\;\langle \mathbf{q}, \mathbf{n}\rangle\;\mathrm{d}S = \int_\Omega \langle \mathbf{f}, \mathbf{q}\rangle\;\mathrm{d}\boldsymbol{\xi} \qquad \text{(FWD)}$$

for all $\mathbf{q} \in (C^1(\Omega))^d \cap (C^0(\overline{\Omega}))^d$.

Exercise 2.4. Derive the formal variational problem for the partial differential equation

$$\mathbf{curl}\,\mathbf{A}^{-1}\,\mathbf{curl}\,\mathbf{u} + \mathbf{C}\,\mathbf{u} = -\mathbf{f} \quad \text{in}\Omega\,.$$

In the case of the first order system comprised of (FL) and (EL) the derivation of the formal weak formulation is more subtle. The idea is to test both equations with smooth vectorfields, for (FL), and functions, for (EL), respectively, and integrate over Ω, but apply integration by parts to only one of the two equations: this equation is said to be **cast in weak form**, whereas the other is retained **in strong form**.

We restrict ourselves to the case $c \equiv 0$. If we cast (EL) in weak form and use the strong form of (FL) we get

$$\begin{aligned} \int_\Omega \langle \mathbf{j}, \mathbf{q}\rangle\;\mathrm{d}\boldsymbol{\xi} &= -\int_\Omega \langle \mathbf{A}\,\mathbf{grad}\,u, \mathbf{q}\rangle\;\mathrm{d}\boldsymbol{\xi} && \forall \mathbf{q}\,, \\ -\int_\Omega \langle \mathbf{j}, \mathbf{grad}\,v\rangle + \int_\Gamma \langle \mathbf{j}, \mathbf{n}\rangle\; v\,\mathrm{d}S &= \int_\Omega fv\,\mathrm{d}\boldsymbol{\xi} && \forall v\,. \end{aligned} \qquad (2.8)$$

Obviously, we can merge both equations, and the result will coincide with (FWP). This formal weak formulation is called **primal**.

The alternative is to cast (FL) in weak form and keep (EL) strongly, which results in the **dual** formal weak formulation:

$$\begin{aligned} -\int_\Omega \left\langle \mathbf{A}^{-1}\mathbf{j}, \mathbf{q} \right\rangle + u \operatorname{div} \mathbf{q} \, \mathrm{d}\boldsymbol{\xi} &= \int_\Gamma u \, \langle \mathbf{q}, \mathbf{n} \rangle \, \mathrm{d}S \qquad \forall \mathbf{q} \, , \\ \int_\Omega \operatorname{div} \mathbf{j} \, v \, \mathrm{d}\boldsymbol{\xi} &= \int_\Omega f v \, \mathrm{d}\boldsymbol{\xi} \qquad \forall v \, . \end{aligned} \tag{2.9}$$

Here, elimination of u is not possible, which leaves us with a so-called **saddle point problem**. We will examine these more closely in Sect. 5.2.

The above variational problems (FWP), (FWD), (2.8), and (2.9) are by no means satisfactory, because they elude any attempt to establish well-posedness if considered in spaces of continuously differentiable functions. First of all these spaces are not reflexive, so that they fail to provide the framework supplied in Ch. 1. Secondly, any attempt to establish an inf-sup condition like (IS1) will be doomed:

Example 2.21. Consider $\Omega =]-1,1[$ and the bilinear form

$$\mathsf{a}(u,v) := \int_{-1}^{1} u' \, v' + u \, v \, \mathrm{d}\xi$$

on the Banach space $C^1([-1,1])$, *cf.* Example 1.6. Based on the family of functions

$$q_\delta(\xi) = \delta^{-1} \frac{1}{(\xi/\delta)^2 + 1} \, , \quad \xi \in \Omega, \ \delta > 0 \, ,$$

we choose $u_\delta(\xi) := \int_{-1}^{\xi} q_\delta(\tau) \, \mathrm{d}\tau$. By simple calculations

$$\int_{-1}^{1} |u_\delta'| \, \mathrm{d}\xi = \tfrac{1}{2}\pi \quad , \quad \int_{-1}^{1} |u_\delta| \, \mathrm{d}\xi = \tfrac{1}{2}\pi \, .$$

On the other hand $\|u_\delta\|_{C^1(\Omega)} \geq \delta^{-1}$. We conclude

$$\sup_{v \in C^1(\overline{\Omega}) \setminus \{0\}} \frac{|\mathsf{a}(u_\delta, v)|}{\|v\|_{C^1(\Omega)}} \leq \int_{-1}^{1} |u_\delta'| + |u_\delta| \, \mathrm{d}\xi = \pi \leq \pi\delta \, \|u_\delta\|_{C^1(\Omega)} \, .$$

Letting $\delta \to 0$ rules out (IS1).

2.6 The Dirichlet principle

Consider the homogeneous Neumann boundary value problem for (E2P), i.e. $\Gamma_N = \Gamma$ and $h \equiv 0$, and assume that c is strictly positive. Then the boundary term in (FWP) can be dropped and we get the variational problem: seek $u \in C^1(\overline{\Omega})$ such that

$$\int_\Omega \langle \mathbf{A}\,\mathbf{grad}\, u, \mathbf{grad}\, v\rangle + c\,uv\,\mathrm{d}\boldsymbol{\xi} = \int_\Omega fv\,\mathrm{d}\boldsymbol{\xi} \quad \forall v \in C^1(\Omega)\cap C^0(\overline{\Omega})\ . \tag{2.10}$$

Evidently, its associated bilinear form is symmetric positive definite. The argument in the proof of Thm. 1.29 shows that a solution of (2.10) will be a global minimizer of the **energy functional**

$$J : C^1(\overline{\Omega}) \mapsto \mathbb{R} \quad , \quad J(v) := \int_\Omega \langle \mathbf{A}\,\mathbf{grad}\, v, \mathbf{grad}\, v\rangle + c\,|v|^2\,\mathrm{d}\boldsymbol{\xi} - \int_\Omega fv\,\mathrm{d}\boldsymbol{\xi}\ .$$

This hints at the general fact that

> *selfadjoint elliptic boundary value problems are closely related to minimization problems for convex functionals on a function space.*

This accounts for their prevasive presence in mathematical models, because the state of many physical systems is characterized by some quantity (energy, entropy) achieving a minimum.

Let us try to elaborate the connection for second-order scalar elliptic boundary value problems. As in Sect. 2.3 let $\Gamma := \partial\Omega$ be partitioned into Γ_D (Dirichlet boundary), Γ_N (Neumann boundary), and Γ_R, with $|\Gamma_D| > 0$. For $g \in C^0(\Gamma)$ define the affine subset of $C^1(\Omega)\cap C^0(\overline{\Omega})$

$$X_g := \{u \in C^1(\Omega)\cap C^0(\overline{\Omega}) :\ u = g \text{ on } \Gamma_D\}\ .$$

Consider the *strictly convex* functional $J : X_g \mapsto \mathbb{R}$

$$J(v) := \frac{1}{2}\int_\Omega \langle \mathbf{A}\,\mathbf{grad}\, v, \mathbf{grad}\, v\rangle + c\,|u|^2\,\mathrm{d}\boldsymbol{\xi} - \int_\Omega fv\,\mathrm{d}\boldsymbol{\xi} + \frac{1}{2}\int_{\Gamma_R} q\,v^2\,\mathrm{d}S - \int_{\Gamma_N} hv\,\mathrm{d}S\ .$$

A necessary and sufficient criterium for u to be a global minimum of J, is

$$\frac{d}{d\tau}J(u+\tau v)_{|\tau=0} = 0 \quad \forall v \in X_0\ ,$$

which is equivalent to the linear variational problem: seek $u \in X_g$ such that

$$\int_\Omega \langle \mathbf{A}\,\mathbf{grad}\, u, \mathbf{grad}\, v\rangle + c\,uv\,\mathrm{d}\boldsymbol{\xi} + \int_{\Gamma_R} q\,uv = \int_\Omega fv\,\mathrm{d}\boldsymbol{\xi} + \int_{\Gamma_N} hv\,\mathrm{d}S \quad \forall v \in X_0\ . \tag{2.11}$$

Remark 2.22. Strictly speaking, (2.11) does not match the definition (LVP) of a linear variational problem, because the unknown is sought in an affine space rather than a vector space. Yet, (2.11) can easily be converted into the form (LVP) by using an extension $u_g \in C^1(\Omega) \cap C^0(\overline{\Omega})$ of the Dirichlet data, that is, $u_g = g$ on Γ_D, and plugging $u := u_g + \delta u$ into (2.11), where, now, the **offset** $\delta u \in X_0$ assumes the role of the unknown function and u_g will show up in an extra contribution to the right hand side functional.

Assuming extra smoothness of u, more precisely, that $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$, we can resort to (FGF) and arrive at

$$\int_\Omega (-\operatorname{div}(\mathbf{A}\,\mathbf{grad}\, u) + cu)\, v \,\mathrm{d}\boldsymbol{\xi} + \int_{\Gamma_N} (\langle \mathbf{A}\,\mathbf{grad}\, u, \mathbf{n}\rangle - h)\, v \,\mathrm{d}S + \int_{\Gamma_R} (\langle \mathbf{A}\,\mathbf{grad}\, u, \mathbf{n}\rangle + qu)\, v \, dS = \int_\Omega f v \,\mathrm{d}\boldsymbol{\xi} \quad (2.12)$$

for all $v \in X_0$. Notice that contributions to the boundary terms from Γ_D turn out to be zero, because $v_{|\Gamma} = 0$. Of course, in (2.12) we can use a test function that is compactly supported in Ω, i.e., $v_{|\Gamma} = 0$. Then we conclude

$$-\operatorname{div}(\mathbf{A}\,\mathbf{grad}\, u) + cu = f \quad \text{in } \Omega\,.$$

This implies that (2.12) can be condensed into

$$\int_{\Gamma_N} (\langle \mathbf{A}\,\mathbf{grad}\, u, \mathbf{n}\rangle - h)\, v \,\mathrm{d}S + \int_{\Gamma_R} (\langle \mathbf{A}\,\mathbf{grad}\, u, \mathbf{n}\rangle + qu)\, v \, dS = 0 \quad \forall v \in X_0\,,$$

from which we infer the boundary conditions

$$\langle \mathbf{A}\,\mathbf{grad}\, u, \mathbf{n}\rangle = h \quad \text{on } \Gamma_N \quad , \quad \langle \mathbf{A}\,\mathbf{grad}\, u, \mathbf{n}\rangle = -qu \quad \text{on } \Gamma_R\,.$$

Summing up, by these formal manipulations we have discovered that a smooth minimizer of J will solve the second order elliptic boundary value problem (E2P). Moreover, we derived the PDE via the variational problem (2.11), which emerges as the link between the PDE and the minimization problem, see Fig. 2.4. Therefore, *numerical methods for variational problems are also suitable for solving a certain class of optimization problems.*

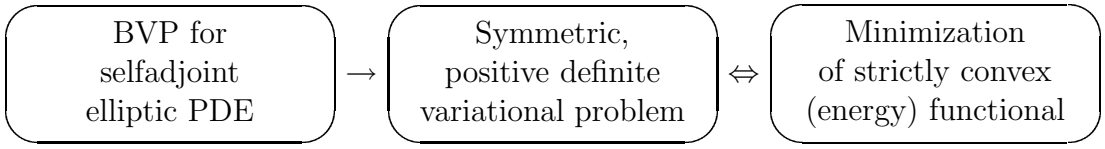


Figure 2.4: Relationship between minimization problems, variational problems, and elliptic boundary value problems

Another important hint about the proper variational formulations of elliptic BVPs is offered by the connection with minimization problems. From the above considerations

we instantly see how to incorporate Dirichlet boundary conditions into a variational formulation of (E2P): they have to be imposed on the trial space, whereas the test functions have to vanish on Γ_D. Conversely, Neumann boundary conditions show up neither in the test nor in the trial space, but are taken into account by an extra term on the right hand side of the variational problem. Eventually, Robin boundary condition affect the bilinear form.

This motivates the following distinction of boundary conditions with respect to a variational formulation of a boundary value problem:

Essential boundary conditions	: boundary conditions that involve a constraint on trial and test functions
Natural boundary conditions	: boundary conditions that are reflected in the variational equation.

The classification clearly hinges on the variational formulation. Consider the boundary value problem (E2D) with $\Gamma_N = \Gamma$. This will be related to the global minimization of the functional

$$J(\mathbf{v}) := \frac{1}{2}\int_\Omega c^{-1}|\operatorname{div}\mathbf{v}|^2 + \mathbf{A}^{-1}|\mathbf{v}|^2\,\mathrm{d}\boldsymbol{\xi} - \int_\Omega c^{-1} f \operatorname{div}\mathbf{v}\,\mathrm{d}\boldsymbol{\xi}$$

over the space

$$X_h := \{\mathbf{v} \in (C^1(\Omega))^3 \cap (C^0(\overline{\Omega}))^3,\ \langle \mathbf{v}, \mathbf{n}\rangle = h\}\ .$$

This gives rise to the variational problem: find $\mathbf{j} \in X_{-h}$ such that

$$\int_\Omega c^{-1}\operatorname{div}\mathbf{j}\ \operatorname{div}\mathbf{v} + \left\langle \mathbf{A}^{-1}\mathbf{j}, \mathbf{v}\right\rangle \mathrm{d}\boldsymbol{\xi} = \int_\Omega c^{-1} f \operatorname{div}\mathbf{v}\,\mathrm{d}\boldsymbol{\xi} \quad \forall \mathbf{v} \in X_0\ . \tag{FVD}$$

Obviously, the Neumann boundary conditions turn out to be essential boundary conditions in this case. Dirichlet boundary conditions will become natural boundary conditions.

Exercise 2.5. Consider (E2D) in the case $\Gamma = \Gamma_D$, find the related minimization problem and derive the associated formal variational problem

However, recasting the boundary value problems as minimization problems does not cure the fundamental problems inherent in the use of the spaces $C^m(\Omega)$: in general it is not possible to establish existence of minimizing functions in these spaces.

2.7 Sobolev spaces

Example 2.23. Consider heat conduction in a plane wall composed of two layers of equal thickness and with heat conductivity coefficients κ_1 and κ_2. The inside of the wall

is kept at temperature $u = u_1$, the outside at $u = 0$. Provided that the width and the height of the wall are much greater than its thickness, a one-dimensional model can be used. After spatial scaling it boils down to

$$j = -\kappa(\xi)\frac{d}{d\xi}u \quad , \quad \frac{d}{d\xi}j = 0 \quad , \quad u(0) = 0 \, , \quad u(1) = u_1 \, , \tag{2.13}$$

where

$$\kappa(\xi) = \begin{cases} \xi_1 & \text{for } 0 < \xi < \frac{1}{2} \, , \\ \xi_2 & \text{for } \frac{1}{2} < \xi < 1 \, . \end{cases}$$

An obvious "physical solution" that guarantees the continuity of the heat flux is

$$u(\xi) = \begin{cases} \dfrac{2u_1\kappa_2\xi}{\kappa_1 + \kappa_2} & \text{for } 0 < \xi < \frac{1}{2} \, , \\ \dfrac{2u_1\kappa_1(\xi - 1)}{\kappa_1 + \kappa_2} + u_1 & \text{for } \frac{1}{2} < \xi < 1 \, . \end{cases} \tag{2.14}$$

Evidently, this solution is not differentiable and can not be a "classical solution".

> The concept of classical solutions sought in spaces of continuously differentiable functions and satisfying the partial differential equations in a pointwise sense is inadequate for elliptic boundary value problems describing physical phenomena. The right approach is to *reinterpret the boundary value problem* in terms of a suitable variational formulation or of an underlying minimization problem, respectively.

Once we have accepted that the variational problem is the principal problem, the issue of suitable function spaces is turned upside down.

> The function spaces that offer the framework for considering the variational problems are chosen to *fit the variational problem* in the sense that (for linear variational problems) the bilinear form is continuous and, if possible, satisfies certain inf-sup conditions.

2.7.1 Distributional derivatives

If we want to invoke the variational framework to investigate the elliptic boundary value problem from Example 2.23 we still need a concept of the derivative u' of u from (2.14). To this end we introduce the space of **test functions**

$$C_0^\infty(\Omega) := \{v \in C^\infty(\overline{\Omega}) : \ \operatorname{supp} v := \overline{\{\boldsymbol{\xi} \in \Omega : \ v(\boldsymbol{\xi}) \neq 0\}} \subset \Omega\}$$

on an open domain $\Omega \subset \mathbb{R}^d$.

Definition 2.24. *Let $u \in L^2(\Omega)$ and $\boldsymbol{\alpha} \in \mathbb{N}_0^n$. A function $w \in L^2(\Omega)$ is called the **weak derivative** or **distributional derivative** $\partial^{\boldsymbol{\alpha}} u$ (of order $|\boldsymbol{\alpha}|$) of u, if*

$$\int_\Omega w v \,\mathrm{d}\boldsymbol{\xi} = (-1)^{|\boldsymbol{\alpha}|} \int_\Omega u\, \partial^{\boldsymbol{\alpha}} v \,\mathrm{d}\boldsymbol{\xi} \quad \forall v \in C_0^\infty(\Omega) \,.$$

Based on this definition, all linear differential operators introduced in Sect. 2.2 can be given a weak/distributional interpretation. For example, the "weak" gradient $\mathbf{grad}\, u$ of a function $u \in L^2(\Omega)$ will be vectorfield $\mathbf{w} \in (L^2(\Omega))^d$ with

$$\int_\Omega \langle \mathbf{w}, \mathbf{v} \rangle \,\mathrm{d}\boldsymbol{\xi} = -\int_\Omega u \,\operatorname{div} \mathbf{v} \,\mathrm{d}\boldsymbol{\xi} \quad \forall \mathbf{v} \in (C_0^\infty(\Omega))^d \,. \tag{2.15}$$

This can be directly concluded from (FGF). The same is true of the "weak divergence" $w \in L^2(\Omega)$ of a vectorfield $\mathbf{u} \in (L^2(\Omega))^d$

$$\int_\Omega w\, v \,\mathrm{d}\boldsymbol{\xi} = -\int_\Omega \langle \mathbf{u}, \mathbf{grad}\, v \rangle \,\mathrm{d}\boldsymbol{\xi} \quad \forall v \in C_0^\infty(\Omega) \,. \tag{2.16}$$

Exercise 2.6. Provide a definition of the "weak" **curl** analoguous to (2.15).

The term "weak derivatives" is justified, because this concept is a genuine generalization of the classical derviative.

Theorem 2.25. *If $u \in C^m(\overline{\Omega})$, then all weak derivatives of order $\leq m$ agree in $L^2(\Omega)$ with the corresponding classical derviatives.*

Proof. Clear by a straightforward application of (IPF). □

Hence, without changing notations, all derivatives will be understood as weak derivatives in the sequel. Straight from the definition we also infer that all linear differential operators in weak sense commute.

Remark 2.26. If u has a continuous m-th classical derivative, $m \in \mathbb{N}_0$, in Ω except on a piecewise smooth q-dimensional submanifold, $q < d$, of Ω, and u has a weak m-th derivative in Ω, then the latter agrees with the pointwise classical derivative almost everywhere in Ω.

The following lemmata settle when "piecewise derivatives" of piecewise smooth functions can be regarded as their weak derivative.

Lemma 2.27. *Let $\Omega \subset \mathbb{R}^d$ be bounded with Lipschitz boundary and assume a partition $\overline{\Omega} = \overline{\Omega}_1 \cup \overline{\Omega}_2$, $\Omega_1 \cap \Omega_2 = \emptyset$, where both sub-domains are supposed to have a Lipschitz boundary, too. Assume that the restriction of the function $u \in L^2(\Omega)$ to Ω_l, $l = 1, 2$, belongs to $C^1(\Omega_l)$ and that $u_{|\Omega_l}$ can be extended to a function in $C^1(\overline{\Omega_l})$.*

Then u possesses weak derivatives $\frac{\partial u}{\partial \xi_k}$, $k = 1, \ldots, d$, if and only if $u \in C^0(\overline{\Omega})$. In this case

$$\frac{\partial u}{\partial \xi_k}(\boldsymbol{\xi}) = \begin{cases} \frac{\partial}{\partial \xi_k} u_{|\Omega_1} & \text{if } \boldsymbol{\xi} \in \Omega_1 \,, \\ \frac{\partial}{\partial \xi_k} u_{|\Omega_2} & \text{if } \boldsymbol{\xi} \in \Omega_2 \,. \end{cases} \tag{2.17}$$

Proof. Using lcoally supported test functions in the definition of the weak derivative, it is clear that (2.17) supplies the only meaningful candidate for the weak derivative of u. Then we appeal to (FGF) and the fact that any crossing direction $\mathbf{n}_\Sigma$ of the interface $\Sigma := \partial\Omega_1 \cap \partial\Omega_2$ will be parallel to the exterior unit normal of one sub-domain, and anti-parallel to that of the other. Thus, we get the identity

$$\int_{\Omega_1} \langle \mathbf{grad}_{\mathrm{cl}}\, u, \mathbf{v}\rangle \, \mathrm{d}\boldsymbol{\xi} + \int_{\Omega_2} \langle \mathbf{grad}_{\mathrm{cl}}\, u, \mathbf{v}\rangle \, \mathrm{d}\boldsymbol{\xi} = -\int_\Omega u \,\mathrm{div}\,\mathbf{v}\,\mathrm{d}\boldsymbol{\xi} + \int_\Sigma [u]_\Sigma \, \langle \mathbf{v}, \mathbf{n}_\Sigma\rangle \cdot \mathbf{n}\mathrm{d}S \,,$$

where $[u]_\Sigma \in C^0(\Sigma)$ stands for the jump of u across Σ and $\mathbf{grad}_{\mathrm{cl}}$ denotes the "classical gradient" of a sufficiently smooth function. Thanks to the assumptions on u this will be a continuous function on Σ. As

$$\int_\Sigma [u]_\Sigma \, \langle \mathbf{v}, \mathbf{n}_\Sigma\rangle \cdot \mathbf{n}\mathrm{d}S = 0 \quad \forall \mathbf{v} \in C_0^\infty(\Omega) \quad \Leftrightarrow \quad [u]_\Sigma = 0 \,,$$

the assertion of the lemma follows from the definition of the weak gradient. □

Example 2.28. The weak derivative of the temperature distribution from Example 2.23 is given by

$$u'(\xi) = \begin{cases} 2u_1\kappa_2(\kappa_1+\kappa_2)^{-1} & \text{if } 0 < \xi < \frac{1}{2} \,, \\ 2u_1\kappa_1(\kappa_1+\kappa_2)^{-1} & \text{if } \frac{1}{2} < \xi < 1 \,. \end{cases}$$

Exercise 2.7. For $\Omega =]0;1[^2$ give an example of a function $u \in C^1(\Omega)$ that does not possess a gradient in $L^2(\Omega)$.

Corollary 2.29. *Under the geometric assumptions of the previous lemma let $u_{|\Omega_l}$ belong to $C^m(\Omega_l)$ with possible extension to $C^{m-1}(\overline{\Omega_l})$. Then*

$$\partial^{\boldsymbol{\alpha}} u \in L^2(\Omega) \quad \forall \boldsymbol{\alpha} \in \mathbb{N}_0^d,\, |\boldsymbol{\alpha}| \le m \quad \Leftrightarrow \quad u \in C^{m-1}(\overline{\Omega}) \,.$$

Lemma 2.30. *We retain the assumptions of Lemma 2.27 with the exception that u is replaced by a vectorfield $\mathbf{u} \in (L^2(\Omega))^d$ with restrictions $\mathbf{u}_{|\Omega_l} \in (C^1(\Omega_l))^d$ that can be extended to continuously differentiable functions on $\overline{\Omega}_l$, $l = 1, 2$.*

Then $\mathbf{u}$ has a weak divergence $\mathrm{div}\,\mathbf{u} \in L^2(\Omega)$, if and only if the normal component of $\mathbf{u}$ is continuous across $\Sigma := \partial\Omega_1 \cap \partial\Omega_2$. Its divergence agrees with the classical divergence on the sub-domains.

If $d = 3$, $\mathbf{u}$ has a weak rotation $\mathbf{curl}\,\mathbf{u} \in (L^2(\Omega))^3$, if and only if the tangential components of $\mathbf{u}$ are continuous across Σ. The combined rotations on the sub-domains yield the weak rotation.

Exercise 2.8. Prove Lemma 2.30.

Remark 2.31. A meaningful divergence marks vectorfields of flux type, *cf.* Sect. 2.2. Hence the tangential continuity asserted in Lemma 2.30 just reflects that such a vectorfields must have meaningful flux through a surface that is only slightly affected by minute perturbations of the surface.

Exercise 2.9. Specify a vectorfield $\mathbf{u} \in L^2(]-1;1[^2)$ with $\operatorname{div}\mathbf{u} = 0$, $|\mathbf{u}| = 1$ almost everywhere in $]-1/2, 1/2[$ and $\operatorname{supp}\mathbf{u} = [-1/2, 1/2]$. Define

$$\mathbf{curl}_{2\mathrm{D}}\, v = \big(-\frac{\partial v}{\partial \xi_2}, \frac{\partial v}{\partial \xi_1}\big)^T$$

and find $v \in L^2(]-1;1[^2)$ such that $\mathbf{u} = \mathbf{curl}_{2\mathrm{D}}\, v$.

Bibliographical notes. Weak derivatives can be motivated through the theory of distributions. For details the reader is referred to [33, Ch. 6].

2.7.2 Definition of Sobolev spaces

In Sect. 2.5 we learned that the formal variational problem associated with the pure homogeneous Neumann problem for (E2P) is: seek $u : \Omega \mapsto \mathbb{R}$ such that

$$\int_\Omega \langle \mathbf{A}\,\mathbf{grad}\, u, \mathbf{grad}\, v\rangle + c\,uv\,\mathrm{d}\boldsymbol{\xi} = \int_\Omega f v\,\mathrm{d}\boldsymbol{\xi} \quad \forall v\ . \tag{2.18}$$

We already know that **grad** has to be used in distributional sense. The concrete spaces have deliberately been omitted in (2.18), because we want to heed the guideline formulated in the context of Example 2.23 and set out from (2.18) and design the "ideal" space. It goes without saying that the investigation of (2.18) is easiest, when the underlying Banach space V features the energy norm

$$\|v\|_V^2 := \int_\Omega \langle \mathbf{A}\,\mathbf{grad}\, v, \mathbf{grad}\, v\rangle + c|v|^2\,\mathrm{d}\boldsymbol{\xi} \tag{2.19}$$

as its norm. So we arrive at the preliminary "definition"

$$V := \{v : \Omega \mapsto \mathbb{R} :\ \text{energy norm (2.19) of } v < \infty\}\ .$$

This has led us straight to a pivotal concept in the modern theory of elliptic boundary value problems.

Definition 2.32. *For $m \in \mathbb{N}_0$ and $\Omega \subset \mathbb{R}^d$ we define the* ***Sobolev space*** *of order m as*

$$H^m(\Omega) := \{v \in L^2(\Omega) :\ \partial^{\boldsymbol{\alpha}} v \in L^2(\Omega),\, \forall\, |\boldsymbol{\alpha}| \le m\}\ ,$$

equipped with the norm

$$\|v\|_{H^m(\Omega)} := \Big(\sum_{|\boldsymbol{\alpha}|\le m} \|\partial^{\boldsymbol{\alpha}} v\|_{L^2(\Omega)}^2\Big)^{\frac{1}{2}}\ .$$

A vector field is said to belong to $H^m(\Omega)$, if this is true of each of its components.

Notation: For all $m \in \mathbb{N}_0$ and $\Omega \subset \mathbb{R}^d$

$$|v|_{H^1(\Omega)} := \Big(\sum_{|\boldsymbol{\alpha}|=m} \|\partial^{\boldsymbol{\alpha}} v\|^2_{L^2(\Omega)} \Big)^{\frac{1}{2}}$$

denotes a semi-norm on $H^m(\Omega)$.
The Sobolev spaces are a promising framework for variational problems, see [43, Thm. 3.1]:

Theorem 2.33. *The Sobolev spaces $H^m(\Omega)$, $m \in \mathbb{N}_0$, are Hilbert spaces with the inner product*

$$(u, v)_{H^m(\Omega)} := \sum_{|\boldsymbol{\alpha}|\le m} (\partial^{\boldsymbol{\alpha}} u, \partial^{\boldsymbol{\alpha}} v)_{L^2(\Omega)} \quad u, v \in H^m(\Omega) \; .$$

The above Sobolev spaces are based on all partial derivatives up to a fixed order. We can as well rely on some partial derivatives or any linear differential operator in the definition of a Sobolev-type space.

Definition 2.34. *If $\mathsf{D} : (C^\infty(\Omega))^l \mapsto (C^\infty(\Omega))^k$, $l, k \in \mathbb{N}$, is a linear differential operator of order m, $m \in \mathbb{N}$, we write*

$$H(\mathsf{D}; \Omega) := \{\mathbf{u} \in (H^{m-1}(\Omega))^l : \; \mathsf{D}\,\mathbf{u} \in (L^2(\Omega))^k\} \; ,$$

where the corresponding norm on this space is given by

$$\|\mathbf{u}\|_{H(\mathsf{D};\Omega)} := \Big(\|\mathbf{u}\|^2_{H^{m-1}(\Omega)} + \|\mathsf{D}\,\mathbf{u}\|^2_{L^2(\Omega)} \Big)^{\frac{1}{2}} \; .$$

The kernel of D in $H(\mathsf{D}; \Omega)$ will be denoted by

$$H(\mathsf{D}\,0; \Omega) := \{\mathbf{u} \in H(\mathsf{D}; \Omega) : \; \mathsf{D}\,\mathbf{u} = 0\} \; .$$

An analogue of Thm. 2.33 holds true for such spaces $H(\mathsf{D}; \Omega)$.

Example 2.35. The most important representatives of spaces covered by Def. 2.34 are

$$H(\mathrm{div}; \Omega) := \{\mathbf{u} \in (L^2(\Omega))^d : \; \mathrm{div}\,\mathbf{u} \in L^2(\Omega)\} \; , \tag{2.20}$$

$$H(\mathbf{curl}; \Omega) := \{\mathbf{u} \in (L^2(\Omega))^3 : \; \mathbf{curl}\,\mathbf{u} \in (L^2(\Omega))^3\} \; , \tag{2.21}$$

$$H(\Delta, \Omega) := \{v \in H^1(\Omega) : \; \Delta u \in L^2(\Omega)\} \; . \tag{2.22}$$

and, derived from them, $H(\mathrm{div}\,0; \Omega)$ and $H(\mathbf{curl}\,0; \Omega)$.

Exercise 2.10. Show that $H(\mathsf{D}\,0; \Omega)$ is a closed subspace of $H(\mathsf{D}; \Omega)$.

Remark 2.36. On an intersection of Hilbert spaces we use the product norm:

$$\|u\|_{V\cap W}^2 := \|u\|_V^2 + \|u\|_W^2 \ , \quad u \in V \cap W \ . \tag{2.23}$$

For instance, this can be used to introduce the Hilbert space $H(\mathrm{div};\Omega) \cap H(\mathbf{curl};\Omega)$.

>From now we confine $\Omega \subset \mathbb{R}^3$ to the class of computational domains according to Def. 2.5 from Sect. 2.1. We recall a definition from functional analysis

Definition 2.37. *A subspace U of a normed space V is called **dense**, if*

$$\forall \epsilon > 0,\, v \in V : \quad \exists u \in U : \quad \|v - u\|_V \leq \epsilon \ .$$

This means that elements of a dense subspace can arbitrarily well approximate elements of a normed vector space.

Then we can state a key result in the theory of Sobolev spaces, the famous Meyers-Serrin theorem, whose proof is way beyond the scope of these lecture notes, see [43, Thm. 3.6]:

Theorem 2.38. *The space $C^\infty(\overline{\Omega})$ is a dense subspace of $H^m(\Omega)$ for all $m \in \mathbb{N}_0$. Moreover, the space $(C^\infty(\overline{\Omega}))^d$ is a dense subspace of $H(\mathrm{div};\Omega)$ and $H(\mathbf{curl};\Omega)$ ($d = 3$ in the latter case).*

The first fundamental insight gleaned from this theorem is that, putting it bluntly,

> the Sobolev spaces $H^m(\Omega)$, $H(\mathrm{div};\Omega)$, and $H(\mathbf{curl};\Omega)$ themselves are immaterial. It is only their norms that matter.

How can we make such a bold claim. The answer is offered by the procedure of **completion**, by which for every normed space one can construct a Banach space, of which the original space will become a dense subspace, see [26, Thm. 2.3]. In addition, the completion of a normed space is *unique*, which means that the completion of a space is completely determined by the normed space itself: the procedure of completion adds no extra particular properties.

Owing to Thm. 2.38 we can give an alternative definition of the Sobolev spaces.

Corollary 2.39. *The spaces $H^m(\Omega)$, $H(\mathrm{div};\Omega)$, and $H(\mathbf{curl};\Omega)$ (for $d = 3$) can be obtained by the completion of spaces of smooth functions with respect to the corresponding norms.*

> All assertions about Sobolev spaces will be assertions about properties and relationships of their norms.

Remark 2.40. Thm. 2.38 also paves the way for an important technique of proving relationships between norms. If we have an assertion that boils down to and (in)equality of the form

$$A(u) \leq B(u) \quad \text{or} \quad A(u) = B(u) \tag{2.24}$$

claimed for all functions u of a Sobolev space and involving *continuous* expressions A, B, then it suffices to prove (2.24) for the dense subspace of smooth functions.

Exercise 2.11. Show that there is a $\gamma > 0$ independent of u and q such that

$$\|qu\|_{H^1(\Omega)} \leq \gamma \, \|q\|_{C^1(\Omega)} \, \|u\|_{H^1(\Omega)} \quad \forall q \in C^1(\Omega),\, u \in H^1(\Omega) \, .$$

Give a bound for γ.

Example 2.41. For $\Omega := \{\boldsymbol{\xi} : |\boldsymbol{\xi}| < 1\}$ the function $s_\alpha : \Omega \mapsto \mathbb{R}$, $s_\alpha(\boldsymbol{\xi}) := |\boldsymbol{\xi}|^\alpha$, $\alpha \in \mathbb{R}$, belongs belongs to $L^2(\Omega)$ if and only if $\alpha > -d/2$, because

$$\int_\Omega |\boldsymbol{\xi}|^{2\alpha} \, \mathrm{d}\boldsymbol{\xi} = |B_1(0)| \int_0^1 r^{2\alpha+(d-1)} \, dr < \infty \quad \Leftrightarrow \quad 2\alpha + d - 1 > -1 \, ,$$

where $|B_1(0)|$ is the volume of the unit sphere in d dimensions. Hence,

$$s_\alpha \in H^m(\Omega) \quad \Leftrightarrow \quad a > -d/2 + m \quad \text{or} \quad \alpha \in \mathbb{N}_0 \, .$$

This example shows that Sobolev norms are a good probe for the strenght of **singularities**. By a singularity we mean the behavior of a function or of some of its derivatives to become unbounded on a lower dimensional submanifold of Ω.

Example 2.42. For the function $s_k :]0;1[\mapsto \mathbb{R}$, $s_k(x) = \sin(k\pi x)$, $k \in \mathbb{N}$, trivially holds

$$\|s_k\|_{L^2(]0;1[)} = \tfrac{1}{2}\sqrt{2} \quad , \quad \|s_k\|_{H^m(]0;1[)} = (k\pi)^m \, \tfrac{1}{2}\sqrt{2} \quad \Leftrightarrow \quad \frac{\|s_k\|_{H^m(]0;1[)}}{\|s_k\|_{L^2(]0;1[)}} = (k\pi)^m \, .$$

This demonstrates that ratios of Sobolev norms of different order can be used to gauge the smoothness of a function. The role of Sobolev norms as a very versatile tool is summarized in Fig. 2.5.

2.7.3 Embeddings

If two Banach spaces V and W both contain U as a dense subspace, and the norm of W is *stronger* than that of V on U, *i.e.*,

$$\exists \gamma > 0 : \quad \|u\|_V \leq \gamma \, \|u\|_W \quad \forall u \in U \, ,$$

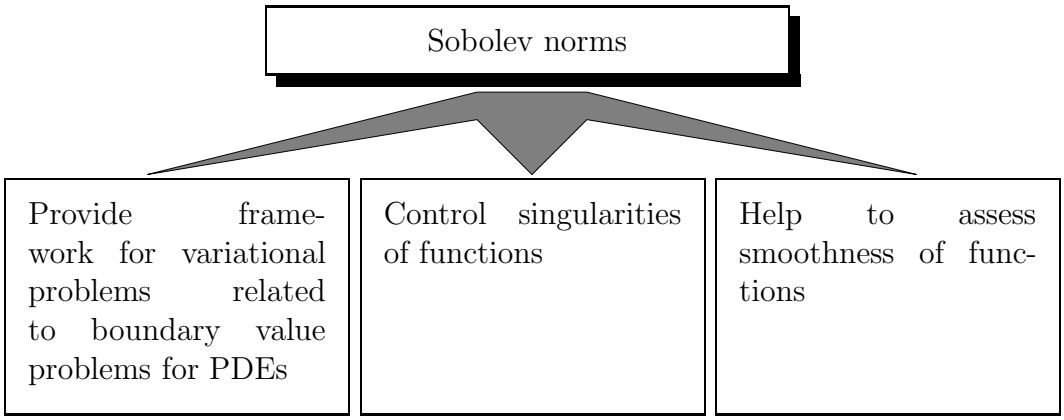


Figure 2.5: Use of Sobolev norms

then W can be regarded as a subspace of V, because both arise as completions of U.

To begin with, it is straightforward that

$$m \le n \quad \Rightarrow \quad \|u\|_{H^m(\Omega)} \le \|u\|_{H^n(\Omega)} \quad \forall u \in C^\infty(\overline{\Omega}) \, .$$

Following Remark 2.40 this amounts to the **continuous embedding** $H^n(\Omega)$ into $H^m(\Omega)$, that is, the canonical injection is continuous.

Exercise 2.12. Can $H^1(\Omega)$ be regarded as a closed subspace of $L^2(\Omega)$?

Example 2.43. Functions from Sobolev spaces do not necessarily take finite values in all points $\boldsymbol{\xi} \in \Omega$. Consider

$$u(\xi) = \left|\xi - \tfrac{1}{2}\right|^{-1/4} \in L^2(]0;1[) \, ,$$

but $u(\frac{1}{2}) = \infty$.

Example 2.44. A computation of the norm in polar coordinates shows that the function $u(\boldsymbol{\xi}) = \log|\log(|\boldsymbol{\xi}|/e)|$ belongs to $H^1(\{\boldsymbol{\xi} : |\boldsymbol{\xi}| < 1\})$. Obviously, this function is not bounded. In other words, for $d = 2$ the space $H^1(\Omega)$ is *not continuously embedded in* $C^0(\overline{\Omega})$.

Example 2.45. We consider $d = 1$ and use $\Omega =]0;1[$. Then we apply the trick from Remark 2.40: we pick an arbitrary $u \in C^\infty(\overline{\Omega})$ and find, by the fundamental theorem of calculus and the Cauchy-Schwarz inequality in $L^2(]0;1[)$,

$$|u(\xi) - u(\eta)| = \left| \int_\eta^\xi u'(\tau) \, \mathrm{d}\tau \right| \le \Big(\int_\eta^\xi 1 \, \mathrm{d}\tau \Big)^{\frac{1}{2}} \, \Big(\int_\eta^\xi (u'(\tau))^2 \, \mathrm{d}\tau \Big)^{\frac{1}{2}} \le |\xi - \eta|^{\frac{1}{2}} \, \|u'\|_{L^2(]0;1[)} \, .$$

By another application of the Cauchy-Schwarz inequality,

$$u(\xi) = \int_0^1 u(\xi) - u(\eta) + u(\eta)\,\mathrm{d}\eta \leq \int_0^1 |\xi-\eta|^{\frac{1}{2}}\ \|u'\|_{L^2(]0;1[)}\ \mathrm{d}\eta + \int_0^1 u(\eta)\,\mathrm{d}\eta$$
$$\leq \|u'\|_{L^2(]0;1[)} + \|u\|_{L^2(]0;1[)} \leq \sqrt{2}\,\|u\|_{H^1(]0;1[)}\ ,$$

we conclude that

$$\|u\|_{C^0(]0;1[)} \leq \sqrt{2}\,\|u\|_{H^1(]0;1[)} \quad \forall u \in C^\infty([0;1])\ . \tag{2.25}$$

Thanks to Thm. 2.38, for any $w \in H^1(\Omega)$ we can find a sequence $\{u_k\}_{k=1}^\infty \subset C^\infty(\overline{\Omega})$ such that $u_k \xrightarrow{k\to\infty} w$ in $H^1(\Omega)$. Since $\{u_k\}_{k=1}^\infty$ is a Cauchy sequence in $H^1(\Omega)$, (2.25) tells us that it will also be a Cauchy sequence in $C^0(\overline{\Omega})$. By completeness of $C^0(\overline{\Omega})$ its limit will be a continuous function. Uniqueness of the limit confirms that w will agree with a continuous functions almost everywhere. Further, the estimate (2.25) will extend to w.

Summing up, we infer the *continuous embedding* $H^1(\Omega) \subset C^0(\overline{\Omega})$ for $d = 1$.

Exercise 2.13. Determine a (reasonably small) bound for the norm of the canonical injection $H^1(\Omega) \mapsto C^0(\overline{\Omega})$ for $\Omega =]\alpha, \beta[$, $\alpha, \beta \in \mathbb{R}$, $\alpha < \beta$.

These examples fit a general pattern, see [43, Thm. 6.2].

Theorem 2.46. *If and only if $d/2 < m$, $m \in \mathbb{N}$, then $H^m(\Omega)$ is continuously embedded in $C^0(\overline{\Omega})$.*

2.7.4 Extensions and traces

Again, we suppose that $\Omega \subset \mathbb{R}^d$ is a computational domain. We call a function $\widetilde{u} \in L^2(\mathbb{R}^d)$ an extension of $u \in L^2(\Omega)$, if $\widetilde{u}_{|\Omega} = u$ in $L^2(\Omega)$. Here, the restriction has to be read in the sense of distributions, that is,

$$\int_\Omega \widetilde{u}_{|\Omega}\, v\,\mathrm{d}\boldsymbol{\xi} = \int_\Omega \widetilde{u}\, v\,\mathrm{d}\boldsymbol{\xi} \quad \forall v \in C_0^\infty(\Omega)\ .$$

An **extension operator** is a linear operator that maps functions from a subspace of $L^2(\Omega)$ into a subspace of $L^2(\mathbb{R}^d)$. It is a non-trivial task to establish the existence of extension operators that are continuous in certain Sobolev norms, *cf.* [43, Thm. 5.4].

Theorem 2.47. *For every $m \in \mathbb{N}$ there is a continuous extension operator $\mathsf{E}_m : H^m(\Omega) \mapsto H^m(\mathbb{R}^3)$.*

Similar extension theorems remain true for $H(\mathrm{div};\Omega)$ and $H(\mathbf{curl};\Omega)$.

Exercise 2.14. Give an explicit construction of E_1 in the case $\Omega =]0;1[$.

In order to give a meaning to essential boundary conditions in the context of Sobolev spaces, we have to investigate "restrictions" of their functions onto Γ or parts of it. By a **trace operator** R_m on a Sobolev space $H^m(\Omega)$ we mean linear mapping from $H^m(\Omega)$ into a subspace of $L^2(\Gamma)$, $\Gamma := \partial\Omega$, such that

$$(\mathsf{R}_m\, u)(\boldsymbol{\xi}) = u(\boldsymbol{\xi}) \quad \forall \boldsymbol{\xi} \in \Gamma,\ \forall u \in C^\infty(\overline{\Omega})\ .$$

In a sense, a trace operator is the extension to $H^m(\Omega)$ of the plain pointwise restriction $u_{|\Gamma}$ of a smooth function u onto Γ. It is by no means obvious that such trace operators exist (as continuous mappings $H^m(\Omega) \mapsto L^2(\Gamma)$).

Example 2.48. For $u \in L^2(\Omega)$ a continuous trace operator cannot be defined. In particular, a **trace inequality** of the form

$$\exists \gamma_t > 0: \quad \left\|u_{|\Gamma}\right\|_{L^2(\Gamma)} \leq \gamma_t \left\|u\right\|_{L^2(\Omega)} \quad \forall u \in C^\infty(\overline{\Omega}) \tag{2.26}$$

remains elusive. Indeed, let $\Omega =]0;1[^2$ and, for $0 < h < 1$, define

$$v(\boldsymbol{\xi}) := \begin{cases} 0 & \text{if } h \leq \xi_1 \leq 1, 0 \leq \xi_2 \leq 1\ , \\ 1 - \frac{\xi_1}{h} & \text{if } 0 \leq \xi_1 \leq h, 0 \leq \xi_2 \leq 1\ . \end{cases}$$

Then we can compute

$$1 = \int_0^1 |v(0,\xi_2)|^2 \,\mathrm{d}\xi_2 \leq \left\|v_{|\Gamma}\right\|^2_{L^2(\Gamma)}\ ,$$

and

$$\|v\|^2_{L^2(\Omega)} = \int_0^1 \left(1 - \frac{\xi_1}{h}\right) \mathrm{d}\xi_1 \overset{y=\xi_1/h}{=} h \int_0^1 (1-y)^2\, dy = {}^h\!/_3\ .$$

If (2.26) were true, there would exists a constant $\gamma_t > 0$ such that $1 \leq \frac{1}{3}\gamma_t h$. For $h \to 0$ we obtain a contradiction.

A continuous trace operator can only be found, if we have control of derivatives of the argument functions:

Theorem 2.49. *The trace operator R_1 is continous from $H^1(\Omega)$ into $L^2(\Gamma)$, that is,*

$$\exists \gamma_t > 0: \quad \left\|u_{|\Gamma}\right\|_{L^2(\Gamma)} \leq \gamma_t \left\|u\right\|_{H^1(\Omega)} \quad \forall u \in C^\infty(\overline{\Omega})\ .$$

More precisely, the following multiplicative trace inequality holds true:

$$\exists \gamma(\Omega) > 0: \quad \|\mathsf{R}_1\, u\|^2_{L^2(\Gamma)} \leq \gamma(\Omega) \left\|u\right\|_{L^2(\Omega)} \left\{\left\|u\right\|_{L^2(\Omega)} + \left\|\mathbf{grad}\, u\right\|_{L^2(\Omega)}\right\} \quad \forall u \in H^1(\Omega)\ .$$

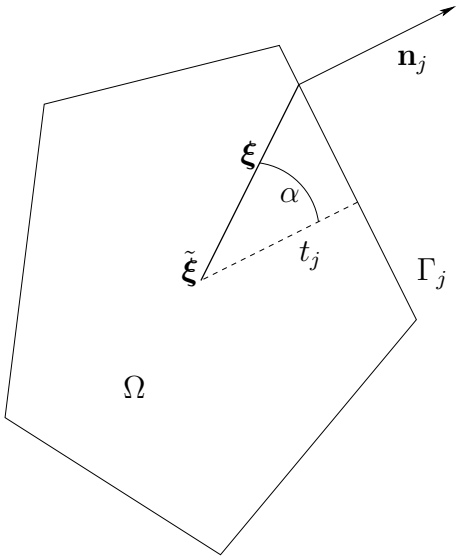


Figure 2.6: Distances t_j.

Notation: Writing $\gamma(\Omega)$ we hint that the "constant" γ may only depend on the domain Ω.

Proof. The proof will be presented for a convex polygon Ω only: denote by $\tilde{\boldsymbol{\xi}}$ the center of the largest $d-$dimensional ball inscribed into Ω and by ρ_Ω its radius. Without loss of generality, we suppose that $\tilde{\boldsymbol{\xi}}$ is the origin of the coordinate system. We start from the following relation

$$\int_{\partial\Omega} v^2 \boldsymbol{\xi} \cdot \mathbf{n}\, dS = \int_\Omega \operatorname{div}(v^2 \boldsymbol{\xi})\, \mathrm{d}\boldsymbol{\xi}, \quad v \in H^1(\Omega). \tag{2.27}$$

Let $\mathbf{n}_j$ be the outer unit normal to Ω on the edge Γ_j, $j \in S$. Then

$$\langle \boldsymbol{\xi}, \mathbf{n}_j \rangle = |\boldsymbol{\xi}||\mathbf{n}_j| \cos\alpha = |\boldsymbol{\xi}| \cos\alpha = t_j, \quad \boldsymbol{\xi} \in \Gamma_j,\ j \in S, \tag{2.28}$$

where t_j is the distance from $\tilde{\boldsymbol{\xi}}$ to Γ_j, see Figure 2.6. Obviously,

$$t_j \geq \rho_\Omega \quad \forall j \in S. \tag{2.29}$$

From (2.28) and (2.29) we have

$$\begin{aligned} \int_{\partial\Omega} v^2 \boldsymbol{\xi} \cdot \mathbf{n}\, dS &= \sum_{j\in S} \int_{\Gamma_j} v^2 \boldsymbol{\xi} \cdot \mathbf{n}_j\, dS = \sum_{j\in S} t_j \int_{\Gamma_j} v^2\, dS \\ &\geq \rho_\Omega \sum_{j\in S} \int_{\Gamma_j} v^2\, dS = \rho_\Omega \|v\|^2_{L^2(\partial\Omega)}. \end{aligned} \tag{2.30}$$

Moreover,

$$\begin{aligned}\int_\Omega \operatorname{div}(v^2\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} &= \int_\Omega v^2 \operatorname{div}\boldsymbol{\xi} + \langle \boldsymbol{\xi}, \mathbf{grad}(v^2)\rangle\,\mathrm{d}\boldsymbol{\xi}\\ = d\int_\Omega v^2\,\mathrm{d}\boldsymbol{\xi} + 2\int_\Omega v\,\langle\boldsymbol{\xi}, \mathbf{grad}\, v\rangle\,\mathrm{d}\boldsymbol{\xi} &\le d\|v\|^2_{L^2(\Omega)} + 2\int_\Omega |v\,\langle\boldsymbol{\xi}, \mathbf{grad}\, v\rangle|\,\mathrm{d}\boldsymbol{\xi}.\end{aligned} \tag{2.31}$$

With the Cauchy inequality the second term in the right hand side of (2.31) is estimated as

$$2\int_\Omega |v\,\langle\boldsymbol{\xi}, \mathbf{grad}\, v\rangle|\,\mathrm{d}\boldsymbol{\xi} \le 2\sup_{\boldsymbol{\xi}\in\Omega}|\boldsymbol{\xi}|\int_\Omega |v||\mathbf{grad}\, v|\,\mathrm{d}\boldsymbol{\xi} \le 2h_\Omega\|v\|_{L^2(\Omega)}|v|_{H^1(\Omega)}. \tag{2.32}$$

Then $h_\Omega/\rho_\Omega \le C_1$, (2.27), (2.30), (2.31) and (2.32) give

$$\begin{aligned}\|v\|^2_{L^2(\partial\Omega)} &\le \frac{1}{\rho_\Omega}\left[2h_\Omega\|v\|_{L^2(\Omega)}|v|_{H^1(\Omega)} + d\|v\|^2_{L^2(\Omega)}\right]\\ &\le C_1\left[2\|v\|_{L^2(\Omega)}|v|_{H^1(\Omega)} + \frac{d}{h_\Omega}\|v\|^2_{L^2(\Omega)}\right],\end{aligned} \tag{2.33}$$

which yields the destined inequality with $C = C_1\max\left\{2, dh_\Omega^{-1}\right\}$. □

Corollary 2.50. *The* ***Neumann trace*** $u \mapsto \langle\mathbf{grad}\, u, \mathbf{n}\rangle_{|\Omega}$, $u \in C^1(\overline{\Omega})$ *can be extended to a continuous mapping* $H^2(\Omega) \mapsto L^2(\Gamma)$.

Notation: The trace operator R_1 is often suppressed in expressions like $\int_\Gamma v \ldots \mathrm{d}S$, when it is clear that the restriction of the function $v \in H^1(\Omega)$ to a boundary Γ is used.

If Γ_0 denotes a part of Γ with positive measure, we can restrict $\mathsf{R}_1 u$, $u \in H^1(\Omega)$, to Γ_0, write R_{Γ_0} for the resulting operator, and trivially have the continuity

$$\exists\gamma > 0: \quad \|\mathsf{R}_{\Gamma_0} u\|_{L^2(\Gamma_0)} \le \gamma\,\|u\|_{H^1(\Omega)} \quad \forall u \in H^1(\Omega)\ .$$

Given the continuity of the trace operator, we can introduce the following closed subspace of $H^1(\Omega)$.

Definition 2.51. *For* $m \in \mathbb{N}$ *and any part* Γ_0 *of the boundary* Γ *of* Ω *with* $|\Gamma_0| > 0$ *we define*

$$H^m_{\Gamma_0}(\Omega) := \{v \in H^1(\Omega):\ \mathsf{R}_{\Gamma_0}(\partial^{\boldsymbol{\alpha}} v) = 0 \text{ in } L^2(\Gamma_0),\ \forall\boldsymbol{\alpha} \in \mathbb{N}_0^d, |\boldsymbol{\alpha}| < m\}\ .$$

If $\Gamma_0 = \Gamma$ *we write* $H^m_0(\Omega) = H^m_{\Gamma_0}(\Omega)$.

Obviously, the spaces $H^1_{\Gamma_0}(\Omega)$ are closed subspaces of $H^1(\Omega)$. Another important density result holds true, see [43, Thm. 3.7].

Theorem 2.52. *The functions in* $C^\infty(\overline{\Omega})$ *whose support does not intersect* Γ_0 *form a dense subspace of* $H^m_{\Gamma_0}(\Omega)$, $m \in \mathbb{N}$. *In particular,* $C^\infty_0(\Omega)$ *is a dense subspace of* $H^m_0(\Omega)$.

By Thm. 2.49 the trace operator R_1 maps continuosly into $L^2(\Gamma)$. This raises the issue, whether it is also *onto*. The answer is negative.

Example 2.53. On $\Omega =]0;1[^2$ consider the functions

$$u_k(\boldsymbol{\xi}) = \sin(k\pi\xi_1)\frac{\sinh(k\pi\xi_2)}{\sinh(k\pi)} , \quad k \in \mathbb{N} .$$

Its restriction to the upper boundary segment $\Gamma_u := \{(\xi_1, 1)^T,\, 0 < \xi_1 < 1\}$ is $v_k(\xi_1) = \sin(k\pi\xi_1)$ and a straightforward calculation yields

$$\left\| \frac{\partial u_k}{\partial \xi_{1,2}} \right\|^2_{L^2(\Omega)} : \|v_k\|_{L^2(]0;1[)} = k\pi \int_0^1 \left(\frac{\cosh(k\pi\xi_2)}{\sinh(k\pi)} \right)^2 \mathrm{d}\xi_2 = O(k) \quad \text{for } k \to \infty .$$

Since, $\Delta u_k = 0$ the function u_k is that extension of v_k with minimal $H^1(\Omega)$-seminorm and zero boundary condition on $\Gamma' := \Gamma \setminus \Gamma_u$.

Next consider the series

$$v_N(\xi) = \sum_{k=1}^{N} \frac{1}{k} \sin(k\pi\xi) ,$$

which converges in $L^2(]0;1[)$. The extension of v_N to $H^1_{\Gamma'}(\Omega)$ with minimal $H^1(\Omega)$-seminorm is given by

$$u_N(\boldsymbol{\xi}) := \sum_{k=1}^{N} \frac{1}{k} \sin(k\pi\xi_1)\frac{\sinh(k\pi\xi_2)}{\sinh(k\pi)} .$$

However, this series diverges in $H^1(\Omega)$.

The question is, how we can characterize the range of the trace operator R_1. Let Γ temporarily stand for a connected component of the boundary of Ω. We start by introducing a norm

$$\|v\|_{H^{1/2}(\Gamma)} = \inf\{\|w\|_{H^1(\Omega)} :\ w \in C^\infty(\overline{\Omega}),\, w_{|\Gamma} = v\} \tag{2.34}$$

on the space of restrictions of smooth functions to Γ. It is highly desirable that this norm is *intrinsic* to Γ, that is, switching to another domain $\widetilde{\Omega}$, for which Γ is also a connected component of the boundary, and using (2.34) produces an *equivalent* norm. It turns out that this is true as a consequence of the extension theorem Thm. 2.47.

Exercise 2.15. Show that (2.34) defines a norm on $C^\infty(\overline{\Omega})_{|\Gamma}$ that arises from an inner product.

Exercise 2.16. Assume that Ω is a computational domain with connected boundary. Create another computational domain $\widetilde{\Omega} \subset \Omega$ by punching out a hole of the interior of Ω. Show that (2.34) spawns equivalent norms for Ω and $\widetilde{\Omega}$.

Definition 2.54. *The completion of $C^\infty(\overline{\Omega})_{|\Gamma}$ with respect to the norm $\|\cdot\|_{H^{1/2}(\Gamma)}$ is designated by $H^{1/2}(\Gamma)$.*

The next theorem shows that the definition of the $H^{1/2}(\Gamma)$-norm is really intrinsic, see [43, § 3].

Theorem 2.55. *The space $H^{1/2}(\Gamma)$ is a Hilbert space and can be equipped with the (equivalent)* ***Sobolev-Slobodeckij-norm***

$$\|v\|^2_{H^{1/2}(\Gamma)} := \int_\Gamma \int_\Gamma \frac{|v(\boldsymbol{\xi}) - v(\boldsymbol{\eta})|^2}{|\boldsymbol{\xi} - \boldsymbol{\eta}|^d} \, \mathrm{d}S(\boldsymbol{\xi}) \mathrm{d}S(\boldsymbol{\eta}) \, .$$

Exercise 2.17. For $\Gamma := \{\boldsymbol{\xi} \in \mathbb{R}^2 : \, |\boldsymbol{\xi}| = 1\}$ show that the Sobolev-Slobodeckij-norm $\|\cdot\|_{H^{1/2}(\Gamma)}$ of the function

$$g : \Gamma \mapsto \mathbb{R} \quad , \quad g(\boldsymbol{\xi}) := \begin{cases} 1 & \text{for } \xi_1 > 0 \, , \\ 0 & \text{for } \xi_1 < 0 \, , \end{cases}$$

is not bounded. This means that this functions does not belong to $H^{1/2}(\Gamma)$.

Theorem 2.56. *The trace operator $\mathsf{R}_1 : H^1(\Omega) \mapsto H^{1/2}(\Gamma)$ is continuous and surjective and has a bounded right inverse $\mathsf{F}_1 : H^{1/2}(\Gamma) \mapsto H^1(\Omega)$, ie., $\mathsf{R}_1 \circ \mathsf{F}_1 = Id$.*

Exercise 2.18. Show the following *Hardy inequality* in one dimension

$$\int_0^1 \left| \frac{u(\xi)}{\xi} \right|^2 \, \mathrm{d}\xi \leq 2 \, |u|^2_{H^1(]0;1[)} \quad \forall u \in H^1_0(]0;1[) \, .$$

2.7.5 Dual spaces

It is easy to verify that

$$\|u\|_{H^{-1}(\Omega)} := \sup_{v \in H^1_0(\Omega) \setminus \{0\}} \frac{\int_\Omega uv \, \mathrm{d}\boldsymbol{\xi}}{\|v\|_{H^1(\Omega)}} \tag{2.35}$$

defines a norm on $L^2(\Omega)$. Thanks to the density theorem Thm. 2.52 we have the equivalent definition

$$\|u\|_{H^{-1}(\Omega)} := \sup_{v \in C^\infty_0(\Omega) \setminus \{0\}} \frac{\int_\Omega uv \, \mathrm{d}\boldsymbol{\xi}}{\|v\|_{H^1(\Omega)}} \, .$$

It can be shown that this norm arises from some inner product on $L^2(\Omega)$.

Definition 2.57. *The completion of $L^2(\Omega)$ with respect to the norm given by 2.35 is called $H^{-1}(\Omega)$.*

Lemma 2.58. *The space $H^{-1}(\Omega)$ is a Hilbert space, which is isometrically isomorphic to $(H_0^1(\Omega))^*$.*

Proof. For any $u \in L^2(\Omega)$ the mapping $f_u : v \in H_0^1(\Omega) \mapsto \int_\Omega uv \,\mathrm{d}\boldsymbol{\xi}$ is a continuous functional on $H_0^1(\Omega)$ with norm

$$\|f_u\|_{(H_0^1(\Omega))^*} = \sup_{v\in H_0^1(\Omega)} \frac{\langle f_u, v\rangle_{(H_0^1(\Omega))^*\times H_0^1(\Omega)}}{\|v\|_{H^1(\Omega)}} = \sup_{v\in H_0^1(\Omega)} \frac{\int_\Omega uv \,\mathrm{d}\boldsymbol{\xi}}{\|v\|_{H^1(\Omega)}} = \|u\|_{H^{-1}(\Omega)} \ . \qquad (2.36)$$

This establishes an isometric imbedding of $L^2(\Omega)$ into $(H_0^1(\Omega))^*$. Hence, $L^2(\Omega)$ can be regarded as a subspace of $(H_0^1(\Omega))^*$. If it was not dense (see Def. 2.37) than the Hahn-Banach theorem would guarantee the existence of a non-zero $w \in H_0^1(\Omega) = (H_0^1(\Omega))^{**}$, such that

$$\langle f_u, w\rangle_{(H_0^1(\Omega))^*\times H_0^1(\Omega)} = \int_\Omega uw \,\mathrm{d}\boldsymbol{\xi} = 0 \quad \forall u \in L^2(\Omega) \ .$$

This cannot hold for $w \neq 0$ and provides the desired contradiction. By the uniqueness of completion and (2.36) we get the assertion. □

Remark 2.59. By construction we have the continuous and dense embeddings

$$H^{-1}(\Omega) = (H_0^1(\Omega))^* \subset L^2(\Omega) \subset H_0^1(\Omega) \ .$$

Sometimes, such an arrangement is called a Gelfand triple.

Notation: Often the integral $\int_\Omega \ldots \mathrm{d}\boldsymbol{\xi}$ is used to denote the duality pairing of $H^{-1}(\Omega)$ and $H_0^1(\Omega)$.

Exercise 2.19. By the Riesz respresentation theorem Thm. 1.26 $(H_0^1(\Omega))^*$ can be identified with $H_0^1(\Omega)$. On the other hand it can also be identified with $H^{-1}(\Omega)$. Regard a function $u \in L^2(\Omega)$ as an element of $H^{-1}(\Omega)$ and formulate a boundary value problem, whose solution will yield the function from $H_0^1(\Omega)$ that the Riesz-isomorphism will associate with u.

Exercise 2.20. For $\Omega =]0;1[$ find a functional in $H^{-1}(\Omega)$ that does not belong to $(H^1(\Omega))^*$.

The same considerations that above targeted $H_0^1(\Omega)$ can be applied to $H^{1/2}(\Gamma)$ on a surface without boundary. This will yield the Hilbert space $H^{-1/2}(\Gamma)$, which contains $L^2(\Gamma)$ and is (isometrically isomorphic to the) dual to $H^{1/2}(\Gamma)$. As before, the integral $\int_\Gamma \ldots \mathrm{d}S$ is often used to indicate the corresponding duality pairing.

Exercise 2.21. Show that for fixed $\boldsymbol{\xi} \in \Gamma$ the functional $d : C^\infty(\overline{\Omega})_{|\Gamma} \mapsto \mathbb{R}$, $d(u) := u(\boldsymbol{\xi})$, does not belong to $H^{-1/2}(\Gamma)$.

Dual space play a crucial role when it comes to defining traces of vectorfields.

Lemma 2.60. *The normal components trace* $\mathsf{R}_\mathbf{n}$ *for* $\mathbf{u} \in (C^\infty(\overline{\Omega}))^d$*, defined by* $\mathsf{R}_\mathbf{n}\,\mathbf{u}(\boldsymbol{\xi}) := \langle \mathbf{u}(\boldsymbol{\xi}), \mathbf{n}(\boldsymbol{\xi})\rangle$ *for all* $\boldsymbol{\xi} \in \Gamma$*, can be extended to a continuous and surjective operator* $\mathsf{R}_\mathbf{n} : H(\mathrm{div};\Omega) \mapsto H^{-1/2}(\Gamma)$*.*

Proof. Pick some $\mathbf{u} \in (C^\infty(\overline{\Omega}))^d$. By (FGF) and the Cauchy-Schwarz inequality in $L^2(\Omega)$ we find

$$\begin{aligned}
\|\langle \mathbf{u}, \mathbf{n}\rangle\|_{H^{-1/2}(\Gamma)} &= \sup_{v\in H^{1/2}(\Gamma)\setminus\{0\}} \frac{\int_\Gamma \langle \mathbf{u}, \mathbf{n}\rangle\, v \,\mathrm{d}S}{\|v\|_{H^{1/2}(\Gamma)}} \\
&= \sup_{v\in H^{1/2}(\Gamma)\setminus\{0\}} \frac{1}{\|v\|_{H^{1/2}(\Gamma)}} \int_\Omega \langle \mathbf{u}, \mathbf{grad}\,\mathsf{F}_1\, v\rangle + \mathrm{div}\,\mathbf{u}\ \mathsf{F}_1\, v \,\mathrm{d}\boldsymbol{\xi} \\
&\le \frac{\|\mathsf{F}_1\, v\|_{H^1(\Omega)}}{\|v\|_{H^{1/2}(\Gamma)}}\ \|\mathbf{u}\|_{H(\mathrm{div};\Omega)} \le \|\mathsf{F}_1\|_{H^{1/2}(\Gamma)\mapsto H^1(\Omega)}\ \|\mathbf{u}\|_{H(\mathrm{div};\Omega)}\ ,
\end{aligned}$$

where $\widetilde{v} := \mathsf{F}_1\, v$, which means that $\mathsf{R}_1\,\widetilde{v} = v$, see Thm. 2.56. Applying the principle explained in Remark 2.40 shows the continuity of $\mathsf{R}_\mathbf{n}$ asserted in the Lemma.

To confirm that $\mathsf{R}_\mathbf{n}$ is onto $H^{-1/2}(\Gamma)$, we rely on the symmetric positive definite linear variational problem: seek $u \in H^1(\Omega)$ such that

$$\int_\Omega \langle \mathbf{grad}\, u, \mathbf{grad}\, v\rangle + uv \,\mathrm{d}\boldsymbol{\xi} = \int_\Gamma h\ \mathsf{R}_1\, v \,\mathrm{d}S \quad \forall v \in H^1(\Omega)\ . \tag{2.37}$$

Here, h is an arbitrary function from $H^{-1/2}(\Gamma)$ and, clearly, the boundary integral has to be understood in the sense of a duality pairing. By the results of Ch. 1 the variational problem (2.37) has a unique solution in $H^1(\Omega)$.

Testing with $v \in C_0^\infty(\Omega)$ and recalling the definition of weak derivatives, see Def. 2.24, we find that

$$-\,\mathrm{div}(\mathbf{grad}\, u) + u = 0 \quad \text{in } L^2(\Omega)\ .$$

Note that div has to be understood as differential operator in the sense of distributions according to (2.16). This shows $\mathrm{div}(\mathbf{grad}\, u) \in L^2(\Omega)$, which allows to apply (FGF) to (2.37):

$$\int_\Omega \underbrace{\big(-\,\mathrm{div}(\mathbf{grad}\, u) + u\big)}_{=0}\, v \,\mathrm{d}\boldsymbol{\xi} + \int_\Gamma \langle \mathbf{grad}\, u, \mathbf{n}\rangle\ \mathsf{R}_1\, v \,\mathrm{d}S = \int_\Gamma h\ \mathsf{R}_1\, v \,\mathrm{d}S$$

for all $v \in C^\infty(\Omega)$. By a density argument, this amounts to $h = \langle \mathbf{grad}\, u, \mathbf{n}\rangle$ in $H^{-1/2}(\Gamma)$ □

Based on this trace theorem, we can introduce the following closed subspace of $H(\mathrm{div};\Omega)$

$$H_0(\mathrm{div};\Omega) := \{\mathbf{v} \in H(\mathrm{div};\Omega) :\ \mathsf{R}_\mathbf{n}\,\mathbf{v} = 0\}\ .$$

Moreover, the trace theorems for $H^1(\Omega)$ and $H(\mathrm{div};\Omega)$ enable us to apply the argument elaborated in Remark 2.40 to (FGF). Hence this integration by parts formula is seen to hold for all $\mathbf{f} \in H(\mathrm{div};\Omega)$ and $u \in H^1(\Omega)$:

$$\int_\Omega \langle \mathbf{f}, \mathbf{grad}\, u\rangle + \mathrm{div}\,\mathbf{f}\, u \,\mathrm{d}\boldsymbol{\xi} = \int_\Gamma \langle \mathbf{f}, \mathbf{n}\rangle \; u \mathrm{d}S \quad \forall \mathbf{f} \in H(\mathrm{div};\Omega), u \in H^1(\Omega)\;. \tag{FGF}$$

Exercise 2.22. Show that

$$H_0(\mathrm{div}\,0;\Omega) := \{\mathbf{v} \in (L^2(\Omega))^d : \int_\Omega \langle \mathbf{v}, \mathbf{grad}\, w\rangle \; \mathrm{d}\boldsymbol{\xi} = 0 \quad \forall w \in H^1(\Omega)\;\}.$$

Exercise 2.23. Show that for $d = 3$ the tangential trace $\mathsf{R}_\mathbf{t}$, for $\mathbf{u} \in C^\infty(\overline{\Omega})$ defined by $(\mathsf{R}_\mathbf{t}\,\mathbf{u})(\boldsymbol{\xi}) := \mathbf{u}(\boldsymbol{\xi}) \times \mathbf{n}(\boldsymbol{\xi})$, $\boldsymbol{\xi} \in \Gamma$, can be extended to a continuous operator $\mathsf{R}_\mathbf{t} : H(\mathbf{curl};\Omega) \mapsto (H^{-1/2}(\Gamma))^3$, where $(H^{-1/2}(\Gamma))^3$ is understood as the dual of $(H^{1/2}(\Gamma))^3$.

2.8 Weak variational formulations

With Sobolev spaces on hands, we are ultimately in a position to put the variational problems derived in Sects. 2.5 and 2.6 into the framework of Ch. 1.

The considerations of the two previous sections show that

- the Dirichlet data g have to be a restriction of a function from $H^{1/2}(\Gamma)$ to Γ_D.
- the Neumann data h on Γ_N must be chosen such that their extension by zero to all of Γ belongs to $H^{-1/2}(\Gamma)$.

First, we examine (2.11) for $\Gamma_R = \emptyset$. We take the cue from Remark 2.22 and aim to implement the "offset policy" to get a problem of the form (LVP): we appeal to Thm. 2.56 to introduce $u_g := \mathsf{F}_1\,\widetilde{g} \in H^1(\Omega)$, whence $\widetilde{g} \in H^{1/2}(\Gamma)$ is a suitable extension of the Dirichlet data g. Then the "rigorous" formulation of (2.11) reads: seek $u \in H^1_{\Gamma_D}(\Omega)$ such that

$$\int_\Omega \langle \mathbf{A}\,\mathbf{grad}(u+u_g), \mathbf{grad}\, v\rangle + c\,(u+u_g)v \,\mathrm{d}\boldsymbol{\xi} = \int_\Omega f v \,\mathrm{d}\boldsymbol{\xi} + \int_{\Gamma_N} h v \,\mathrm{d}S \tag{EVP}$$

for all $v \in H^1_{\Gamma_D}(\Omega)$. The surface integral over Γ_N has to be viewed as a duality pairing on $H^{-1/2}(\Gamma) \times H^{1/2}(\Gamma)$ after an extension of h by zero.

Evidently, (EVP) is a linear variational problem on the Hilbert space $V := H^1_{\Gamma_D}(\Omega)$. The bilinear form associated with (EVP) is

$$\mathsf{a}(u,v) := \int_\Omega \langle \mathbf{A}\,\mathbf{grad}(u+u_g), \mathbf{grad}\, v\rangle + c\,(u+u_g)v \,\mathrm{d}\boldsymbol{\xi} \quad u, v \in H^1_{\Gamma_D}(\Omega)\;. \tag{PBF}$$

The right hand side functional is given by

$$\langle f, v\rangle_{V^*\times V} := \int_\Omega fv \,\mathrm{d}\boldsymbol{\xi} + \int_{\Gamma_N} hv \,\mathrm{d}S \quad v \in H^1_{\Gamma_D}(\Omega) \,.$$

By virtue of the boundedness of the coefficient functions $\mathbf{A}$ and c, see (UPD), the continuity of a is an immediate consequence of the Cauchy-Schwarz inequality (CSI) in $L^2(\Omega)$. The continuity of the right hand side functional can be concluded from (CSI) and the trace theorem Thm. 2.56. Thus, the setting of Sect. 1.2 is established.

If $c(\boldsymbol{\xi}) \geq \gamma_c > 0$ for almost all $\boldsymbol{\xi} \in \Omega$ existence and uniqueness of solutions of (EVP) is immediate, because, by virtue of the assumption (UPD) on $\mathbf{A}$, the associated bilinear form is $H^1_{\Gamma_D}(\Omega)$-elliptic. If $c \equiv 0$, the case is settled by the following key lemma.

Lemma 2.61 (Poincaré-Friedrichs inequality). *If $\Gamma_D \subset \Gamma$ with positive measure, then there is $\gamma_F = \gamma_F(\Omega, \Gamma_D)$ such that*

$$\|u\|_{L^2(\Omega)} \leq \gamma_F \, |u|_{H^1(\Omega)} \quad \forall u \in H^1_{\Gamma_D}(\Omega) \,.$$

If $\Gamma = \Gamma_D$, then $\gamma_F > 0$ may only depend on the diameter

$$\operatorname{diam}(\Omega) := \inf\{r > 0 : \exists \boldsymbol{\zeta} \in \mathbb{R}^d, \boldsymbol{\nu} \in \mathbb{R}^d, |\boldsymbol{\nu}| = 1 : \; 0 \leq \langle \boldsymbol{\zeta} - \boldsymbol{\xi}, \boldsymbol{\nu}\rangle \leq r/2 \; \forall \boldsymbol{\xi} \in \Omega\} \,.$$

Proof. We will only tackle the case $\Gamma_D = \Gamma$. Remembering Remark 2.40 we need only establish the estimate for $u \in C_0^\infty(\overline{\Omega})$.

We can choose a Cartesian coordinate system such that

$$\Omega \subset \{\boldsymbol{\xi} \in \mathbb{R}^d, \; 0 < \xi_d < r := \operatorname{diam}(\Omega)\} \,.$$

Moreover, as u is supported inside Ω we can extend all integrations to a tensor product domain $\Omega' \times]0, r[$ that contains Ω. Below, whenever integration over Ω is performed, the reader is well advised to replace it by such a tensor product domain in order to understand the manipulations.

Then, by the fundamental theorem of calculus and setting $\boldsymbol{\xi}' := (\xi_1, \ldots, \xi_{d-1})^T$

$$u(\boldsymbol{\xi}) = \int_0^{\xi_d} \frac{\partial u}{\partial \xi_d}(\boldsymbol{\xi}', \eta_d) \,\mathrm{d}\eta_d \,.$$

Thus, a simple application of the Cauchy-Schwarz inequality accomplishes the proof:

$$\begin{aligned}\int_\Omega |u|^2 \,\mathrm{d}\boldsymbol{\xi} &= \int_\Omega \left| \int_0^{\xi_d} \frac{\partial u}{\partial \xi_d}(\boldsymbol{\xi}', \eta_d) \,\mathrm{d}\eta_d \right|^2 \mathrm{d}\boldsymbol{\xi} \leq \int_\Omega \xi_d \int_0^{\xi_d} \left| \frac{\partial u}{\partial \xi_d}(\boldsymbol{\xi}', \eta_d) \right|^2 \mathrm{d}\eta_d \,\mathrm{d}\boldsymbol{\xi} \\ &\leq r \int_\Omega \int_0^r \left| \frac{\partial u}{\partial \xi_d}(\boldsymbol{\xi}', \eta_d) \right|^2 \mathrm{d}\eta_d \,\mathrm{d}\boldsymbol{\xi} \leq r^2 \int_\Omega \left| \frac{\partial u}{\partial \xi_d}(\boldsymbol{\xi}) \right|^2 \mathrm{d}\boldsymbol{\xi} \,.\end{aligned}$$

For the case $\Gamma_D \neq \Gamma$ the proof has to employ compactness arguments, see [43, § 7]. □

Exercise 2.24. Show that the solution $u + u_g$ of (EVP) is indenpedent of u_g, that is, the concrete extension of the boundary data is immaterial.

The statement of the Poincaré-Friedrichs inequality carries over to higher degree Sobolev spaces.

Lemma 2.62. *With $\gamma_F = \gamma_F(\Omega, \Gamma_0)$ from the previous lemma, we have for $m \in \mathbb{N}$*

$$\|u\|^2_{H^m(\Omega)} \leq \left(1 + d\gamma_F^2 + d^2\gamma_F^4 + \cdots + d^m\gamma_F^{2m}\right) \, |u|_{H^m(\Omega)} \quad \forall u \in H^m_{\Gamma_0}(\Omega) \, .$$

Proof. Applying Lemma 2.61 to derivatives we obtain

$$\begin{aligned} |u|^2_{H^{m-1}(\Omega)} = \sum_{|\boldsymbol{\alpha}|=m-1} \|\partial^{\boldsymbol{\alpha}} u\|^2_{L^2(\Omega)} &\leq \gamma_F \sum_{|\boldsymbol{\alpha}|=m-1} \sum_{k=1}^{d} \left\|\partial^{\boldsymbol{\alpha}+\epsilon_k} u\right\|^2_{L^2(\Omega)} \\ &\leq d\gamma_F^2 \sum_{|\boldsymbol{\alpha}|=m} \|\partial^{\boldsymbol{\alpha}} u\|^2_{L^2(\Omega)} = d\gamma_F^2 \, |u|^2_{H^m(\Omega)} \, . \end{aligned}$$

From this the assertion follows by simple induction. □

Corollary 2.63. *If $|\Gamma_D| > 0$ or c uniformly positive almost everywhere in Ω, then* (EVP) *has a unique solution $u \in H^1_{\Gamma_D}(\Omega)$ that satisfies*

$$\exists \gamma = \gamma(\Omega, \Gamma_D, \mathbf{A}, c) : \quad \|u\|_{H^1(\Omega)} \leq \gamma \left\{ \|f\|_{(H^1_{\Gamma_D}(\Omega))^*} + \|\widetilde{g}\|_{H^{1/2}(\Gamma)} + \left\|\widetilde{h}\right\|_{H^{-1/2}(\Gamma)} \right\} ,$$

where $\widetilde{g} \in H^{1/2}(\Gamma)$ is a suitable extension of g, and $\widetilde{h} \in H^{-1/2}(\Gamma)$ an extension of h by zero.

Next, we assume $c \equiv 0$ and examine the pure Neumann problem, *ie.* $\Gamma = \Gamma_N$. First we point out to a **compatibility conditions** inherent in the equations (E2P): as a consequence of Gauss' theorem Thm. 2.17

$$\left.\begin{array}{ll} \operatorname{div} \mathbf{j} = f & \text{in } \Omega \\ \langle \mathbf{j}, \mathbf{n} \rangle = h & \text{on } \Gamma \end{array}\right\} \quad \Rightarrow \quad \int_\Omega f \, \mathrm{d}\boldsymbol{\xi} = \int_\Gamma h \, \mathrm{d}S \tag{NCC}$$

After the considerations of Sect. 2.5, *cf.* (FWP), one is lead to the weak formulation: seek $u \in H^1(\Omega)$ such that

$$\int_\Omega \langle \mathbf{A} \, \mathbf{grad} \, u, \mathbf{grad} \, v \rangle \, \mathrm{d}\boldsymbol{\xi} = \int_\Omega fv \, \mathrm{d}\boldsymbol{\xi} + \int_\Gamma hv \, \mathrm{d}S \quad \forall v \in H^1(\Omega) \, . \tag{NVP}$$

However, it is easy to see that we cannot expect a unique solutions, because constant functions will belong to the kernel of the bilinear form underlying (NVP). Besides, testing with constants in (NVP), we recover (NCC) as a necessary condition for the existence of solutions.

To weed out the constants we can switch to the subspace

$$H^1_*(\Omega) := \{v \in H^1(\Omega) : \int_\Omega v \,\mathrm{d}\boldsymbol{\xi} = 0\}$$

On this subspace we have an analogue of the Poincaré-Friedrichs inequality of Lemma 2.61, which can be inferred from the next lemma.

Lemma 2.64. *There is a constant $\gamma = \gamma(\Omega)$ such that*

$$\|u\|_{L^2(\Omega)} \leq \gamma \left\{ \left| \int_\Omega u \,\mathrm{d}\boldsymbol{\xi} \right| + |u|_{H^1(\Omega)} \right\} \quad \forall u \in H^1(\Omega) \,.$$

Proof. The proof is based on *compactness arguments* that will be elaborated in Sect. 4.1. Summing up, they guarantee that any bounded sequence in $H^1(\Omega)$ will have a subsequence that converges in $L^2(\Omega)$.

This can be used for an indirect proof: assume that the assertion of the lemma was false. Then, for any $n \in \mathbb{N}$ we can find $u_n \in H^1(\Omega)$, $\|u_n\|_{H^1(\Omega)} = 1$ such that

$$n^{-1} \, \|u_n\|_{L^2(\Omega)} \geq \left\{ \left| \int_\Omega u_n \,\mathrm{d}\boldsymbol{\xi} \right| + |u_n|_{H^1(\Omega)} \right\} \,. \tag{2.38}$$

Owing to the above compactness result we may assume that $\{u_n\}_{n=1}^\infty$ converges in $L^2(\Omega)$ with limit $w \in L^2(\Omega)$. By (2.38), necessarily

$$\int_\Omega w \,\mathrm{d}\boldsymbol{\xi} = 0 \quad \text{and} \quad |w|_{H^1(\Omega)} = 0 \quad \Rightarrow \quad \|w\|_{L^2(\Omega)} = 1 \,,$$

which is an obvious contradiction. □

Thus, the bilinear form from (NVP) will be elliptic on $H^1_*(\Omega)$, which is the proper space for the variational formulation corresponding to the pure Neumann boundary value problem problem. Existence of a unique solution $u \in H^1_*(\Omega)$ follows.

Remark 2.65. One is tempted to ensure uniqueness of solutions of (NVP) by demanding $u(\boldsymbol{\xi}_0) = 0$ for some $\boldsymbol{\xi}_0 \in \Omega$. However, this approach is flawed, because the mapping $u \mapsto u(\boldsymbol{\xi}_0)$ is *unbounded* on $H^1(\Omega)$, see Example 2.44.

Therefore, such a strategy may lead to severely ill-conditioned linear systems of equations when employed in the context of a Galerkin discretization.

> The general rule is that in order to impose constraints one has to resort to functionals/operators/mappings that are continuous on the relevant function spaces.

Exercise 2.25. Show that $H^1_*(\Omega)$ is a Hilbert space.

Exercise 2.26. Give an indirect proof of Lemma 2.61 following the strategy of the proof of Lemma 2.64.

Exercise 2.27. The variational problem (NVP) when posed on $H^1_*(\Omega)$ always possesses a unique solution. How can this be reconciled with the compatibility condition (NCC)?

Different Sobolev spaces have to be employed for a rigorous statement of the dual variational problem (FVD) with $\Gamma_N = \Gamma$: now we employ an extension $\mathbf{j}_h \in H(\mathrm{div};\Omega)$ of $h \in H^{-1/2}(\Gamma)$ and get: find $\mathbf{j} \in H_0(\mathrm{div};\Omega)$ with

$$\int_\Omega c^{-1} \,\mathrm{div}(\mathbf{j}+\mathbf{j}_h)\,\mathrm{div}\,\mathbf{v} + \left\langle \mathbf{A}^{-1}(\mathbf{j}+\mathbf{j}_h), \mathbf{v}\right\rangle \,\mathrm{d}\boldsymbol{\xi} = \int_\Omega c^{-1} f \,\mathrm{div}\,\mathbf{v}\,\mathrm{d}\boldsymbol{\xi} \qquad \text{(VPD)}$$

for all $\mathbf{v} \in H_0(\mathrm{div};\Omega)$. The assumptions on the coefficient functions instantly involve the $H(\mathrm{div};\Omega)$-ellipticity of the underlying bilinear form.

Exercise 2.28. Derive the variational formulation in Sobolev spaces of the pure Dirichlet problem ($\Gamma = \Gamma_D$) for the dual formulation (E2D) of the second-order elliptic boundary value problem. Are the Dirichlet boundary conditions essential or natural in this case?

Existence and uniqueness of solutions of (EVP) and of (VPD) do not necessarily mean that we will recover weak solutions of (FL) and (EL). Fortunately, this is the case:

Proposition 2.66. *Let $u \in H^1_{\Gamma_D}(\Omega)$ be the solution of* (EVP)*. Then $u \in L^2(\Omega)$ and $\mathbf{j} := -\mathbf{A}\,\mathbf{grad}(u+u_g) \in L^2(\Omega)$ are distributional solutions of* (FL) *and* (EL)*. Moreover, $u+u_g$ is a distributional solution of* (E2P) *(with $\Gamma_R = \emptyset$).*

Proof. The equation (FL) is satisfied by virtue of the defintion of $\mathbf{j}$. Next, (FL) can be established by using test functions $v \in C_0^\infty(\Omega)$. This is an immediate consequence of the definition (2.16) of the distributional divergence.

Thus, the functions are regular enought to allow integration by parts according to (FGF). This will give

$$-\int_\Gamma \langle \mathbf{j}, \mathbf{n}\rangle \, v \,\mathrm{d}S = \int_{\Gamma_N} hv \,\mathrm{d}S \quad \forall v \in C^\infty(\overline{\Omega}),\ v_{|\Gamma_D} = 0 \quad \Rightarrow \quad \langle \mathbf{j}, \mathbf{n}\rangle = -h \text{ in } H^{-1/2}(\Gamma)\,.$$

Obviously, this implies the Neumann and Robin boundary conditions, *cf.* the arguments in Sect. 2.6. □

A similar proposition can be formulated for (VPD).

Example 2.67. We consider the homogeneous Dirichlet boundary value problem for the *biharmonic operator*

$$\begin{aligned} \Delta^2 u &= f \quad && \text{in } \Omega\,, \\ u &= 0 && \text{on } \Gamma\,, \\ \langle \mathbf{grad}\, u, \mathbf{n}\rangle &= 0 && \text{on } \Gamma\,, \end{aligned} \qquad \text{(BIH)}$$

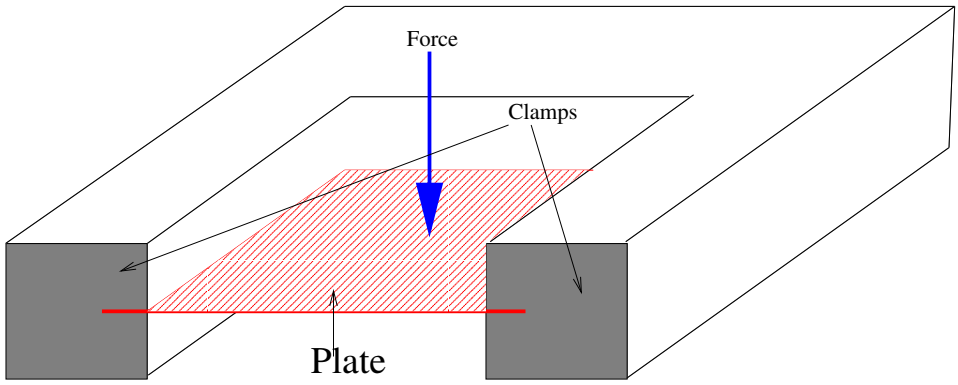


Figure 2.7: Clamped plate (cross section) whose vertical deflection u is described by (BIH)

on a computational domain $\Omega \subset \mathbb{R}^2$. This equation provides a mathematical model for small transversal deflections of a thin clamped plate.

First we derive the formal weak formulation by applying (FGF) twice

$$\begin{aligned}\int_\Omega \Delta^2 u\, v \,\mathrm{d}\boldsymbol{\xi} &= -\int_\Omega \langle \mathbf{grad}(\Delta u), \mathbf{grad}\, v\rangle \,\mathrm{d}\boldsymbol{\xi} + \int_\Gamma \langle \mathbf{grad}\,\Delta u\rangle\, v \,\mathrm{d}S \\ &= \int_\Omega \Delta u\,\Delta v \,\mathrm{d}\boldsymbol{\xi} + \int_\Gamma \langle \mathbf{grad}\,\Delta u\rangle\, v - \langle \mathbf{grad}\, v, \mathbf{n}\rangle\, \Delta u \,\mathrm{d}S\,,\end{aligned}$$

which gives us

$$\int_\Omega \Delta u\,\Delta v \,\mathrm{d}\boldsymbol{\xi} + \int_\Gamma \langle \mathbf{grad}\,\Delta u\rangle\, v - \langle \mathbf{grad}\, v, \mathbf{n}\rangle\, \Delta u \,\mathrm{d}S = \int_\Omega f v \,\mathrm{d}\boldsymbol{\xi} \quad \forall v$$

We see that the boundary conditions are essential (the homogeneous boundary conditions make sense for the test function) and arrive at the variational problem in Sobolev spaces: seek $u \in H_0^2(\Omega)$ such that

$$\int_\Omega \Delta u\,\Delta v \,\mathrm{d}\boldsymbol{\xi} = \int_\Omega f v \,\mathrm{d}\boldsymbol{\xi} \quad \forall v \in H_0^2(\Omega)\,. \tag{2.39}$$

The continuity of the bilinear form and right hand side from (2.39) is immediate. In order to demonstrate the ellipticity of the bilinear form, we first recall the density result of Thm. 2.52. It hints that for proving ellipticity we can restrict ourselves to functions in $C_0^\infty(\Omega)$.

Next, we use the integration by parts formula (IPF) and see that

$$\int_\Omega \frac{\partial^2 u}{\partial \xi_1^2}\cdot\frac{\partial^2 u}{\partial \xi_2^2}\,\mathrm{d}\boldsymbol{\xi} = -\int_\Omega \frac{\partial^3 u}{\partial \xi_1^2 \partial \xi_2}\cdot\frac{\partial u}{\partial \xi_2}\,\mathrm{d}\boldsymbol{\xi} = \int_\Omega \frac{\partial^2 u}{\partial \xi_1 \partial \xi_2}\cdot\frac{\partial^2 u}{\partial \xi_2 \partial \xi_1}\,\mathrm{d}\boldsymbol{\xi}$$

for any smooth compactly supported function. This means that

$$\int_\Omega |\Delta u|^2 \, \mathrm{d}\boldsymbol{\xi} = \int_\Omega \left| \frac{\partial^2 u}{\partial \xi_1^2} \right|^2 + 2 \frac{\partial^2 u}{\partial \xi_1^2} \cdot \frac{\partial^2 u}{\partial \xi_2^2} + \left| \frac{\partial^2 u}{\partial \xi_2^2} \right|^2 \, \mathrm{d}\boldsymbol{\xi} \geq 2 \, |u|^2_{H^2(\Omega)} \ .$$

Appealing to Lemma 2.62 we infer $H^2(\Omega)$-ellipticity, and, in turns, existence and uniqueness of solutions of (2.39).

3 Primal Finite Element Methods

In Chapter 1 we saw that the Galerkin discretization of a linear variational problem (LVP) posed on a Banach space V entails finding suitable finite dimensional trial and test spaces $W_n, V_n \subset V$. In this context, "suitable" means that some discrete inf-sup-conditions have to be satisfied, see Thm. 1.30.

In this chapter we only consider linear variational problems that arise from the primal weak formulation of boundary value problems as discussed in Sect. 2.5, see (EVP), (NVP), and (VPD). These variational problems are set in Sobolev spaces and feature elliptic bilinear forms according to Def. 1.20. Hence, if trial and test space agree, which will be the case throughout this chapter, stability of the discrete variational problem is not an issue, *cf.* Remark 1.32.

In this setting, the construction of V_n has to address two major issues

1. In light of Thm. 1.30 V_n must be able to approximate the solution $u \in V$ of the linear variational problem well in the norm of V.

2. The space V_n must possess a basis $\mathfrak{B}_V$ that allows for efficient assembly of a stiffness matrix with desirable properties (e.g. well conditioned and/or sparse, *cf.* Sect. 1.5).

The finite element methods tries to achieve these goals by employing

- spaces of functions that are piecewise smooth and "simple" and
- locally supported basis function of these spaces.

3.1 Meshes

For the remainder of this section let $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$, stand for a computational domain according to Def. 2.5.

Definition 3.1. *A* ***mesh*** $\mathcal{M}$ *of* Ω *is a collection* $\{K_i\}_{i=1}^{M}$, $M := \sharp\mathcal{M}$[1], *of connected* open *subsets* $K_i \subset \Omega$ *such that*

[1] $\sharp$ denotes the cardinality of a finite set

- *the closure of each K_i is the C^∞-diffeomorphic image of a closed d-dimensional polytope (that is, the convex hull of $d+1$ points in $\mathbb{R}^d$),*
- $\bigcup_i \overline{K_i} = \overline{\Omega}$ *and* $K_i \cap K_j = \emptyset$ *if* $i \neq j$, $i, j \in \{1, \ldots, M\}$.

The K_i are called ***cells*** *of the mesh.*

Remark 3.2. Sometimes the smoothness requirement on the diffeomorphism is relaxed and mapping that are continuous but only piecewise C^∞ are admitted.

Following the terminology of Sect. 2.1, each cell is an interval ($d = 1$), a Lipschitz polygon ($d = 2$), or a Lipschitz polyhedron ($d = 3$). Therefore, we can refer to vertices, edges ($d > 1$), and faces ($d = 3$) of a cell appealing to the geometric meaning of the terms. Meshes are a crucial building block in the design of the finite dimensional trial and test spaces used in the finite element method.

Already contained in the definition of a mesh is the notion of **reference cells**. By them we mean a finite set $\widehat{K}_1, \ldots, \widehat{K}_P$, $P \in \mathbb{N}$, of d-dimensional polytopes such that all cells of the mesh can be obtained from one of the $\widehat{K}_i$ under a suitable diffeomorphism.

We recall that a mapping

$$\boldsymbol{\Phi} : \mathbb{R}^d \mapsto \mathbb{R}^d, \quad \boldsymbol{\xi} \mapsto \mathbf{F}\boldsymbol{\xi} + \boldsymbol{\tau}\ , \quad \mathbf{F} \in \mathbb{R}^{d,d} \text{ regular}\ , \boldsymbol{\tau} \in \mathbb{R}^d \tag{AFF}$$

represents a bijective **affine mapping** of d-dimensional Euklidean space.

Definition 3.3. *A mesh $\mathcal{M}$ of $\Omega \subset \mathbb{R}^d$ is called* ***affine equivalent****, if all its cells arise as affine images of a single d-dimensional (reference) polytope.*

A family of meshes $\{\mathcal{M}_n\}_{n\in\mathbb{N}}$ is affine equivalent, if this is true for each of its members and if the same reference polytope can be chosen for all $\mathcal{M}_n$, $n \in \mathbb{N}$.

Definition 3.4. *For $d = 3$ the* ***set of (topological) faces*** *of a mesh $\mathcal{M}$ is given by*

$$\mathcal{F}(\mathcal{M}) := \left\{\text{interior}(\overline{K_i} \cap \overline{K_j}),\ 1 \leq i < j \leq \sharp\mathcal{M}\right\} \cup \{\textit{(geometric) faces } \subset \partial\Omega\}\ ,$$

the ***set of (topological) edges*** *of $\mathcal{M}$ is defined as*

$$\mathcal{E}(\mathcal{M}) := \left\{\text{interior}(\overline{F} \cap \overline{F'}),\ F, F' \in \mathcal{F}(\mathcal{M}), F \neq F'\right\}\ ,$$

whereas the ***set of (topological) nodes*** *is*

$$\mathcal{N}(\mathcal{M}) := \left\{\overline{E} \cap \overline{E'},\ E, E' \in \mathcal{E}(\mathcal{M}), E \neq E'\right\}\ .$$

Similarly, we can define sets of edges and nodes for $d = 2$, and the set of nodes for $d = 1$. Often the term ***face*** *is used for components of a mesh of dimension $d-1$, that is, for faces in three dimensions, edges in two dimensions, and nodes in one dimension. We still write $\mathcal{F}(\mathcal{M})$ for the set of these (generalized) faces.*

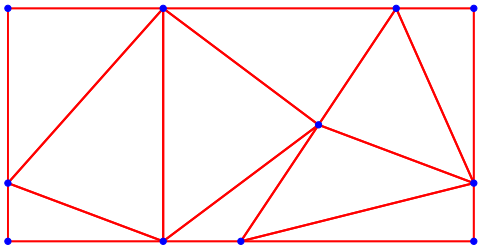


Figure 3.1: Two-dimensional mesh and the sets of edges (red) and nodes (blue)

Remark 3.5. The "is contained in the closure of" **incidence relations** $\mathcal{N}(\mathcal{M}) \times \mathcal{E}(\mathcal{M}) \mapsto \{\text{true}, \text{false}\}$, $\mathcal{E}(\mathcal{M}) \times \mathcal{F}(\mathcal{M}) \mapsto \{\text{true}, \text{false}\}$, etc., describe the **topology** of a mesh. The locations of nodes and shape of cells are features connected wity the **geometry** of the mesh.

Definition 3.6. *A mesh $\mathcal{M} = \{K_i\}_{i=1}^M$ of Ω is called a **triangulation**, if $\overline{K}_i \cap \overline{K}_j$, $1 \leq i < j \leq M$, agrees with a geometric face/edge/vertex of K_i* or *K_j. The cells of a triangulation are sometimes called **elements**.*

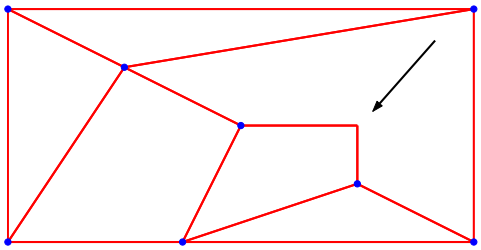


Figure 3.2: A mesh that is not a triangulation. The arrow points at the culprit.

An important concept is that of the **orientation** of the geometric objects of a triangulation.

Definition 3.7. *The (inner) **orientation** of a node is always set to $+1$. Orienting an edge amounts to prescribing a direction. The orientation of a cell for $d = 2$ or a face for $d = 3$ can be fixed by specifying an ordering of the edges along its boundary. Finally, in the case $d = 3$ the orientation of a cell is given by "inside" and "outside".*

We point out that the orientation of a face for $d = 3$ can also be imposed by fixing a crossing direction (outer orientation). If all geoemetric objects of a triangulation are equipped with an orientation, we will call it an **oriented triangulation**.

Definition 3.8. *A triangulation $\mathcal{M} := \{K_i\}_{i=1}^M$ of Ω is called **conforming**, if $\overline{K}_i \cap \overline{K}_j$, $1 \leq i < j \leq M$, is a (geometric) face of both K_i and K_j.*

Unless clearly stated otherwise we will tacitly assume that all meshes that will be used for the construction of finite element spaces in the remainder of this chapter are conforming.

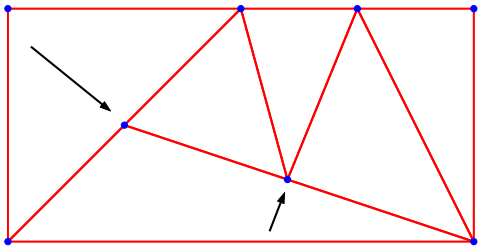


Figure 3.3: A triangulation that is not conforming and possesses two hanging nodes.

Definition 3.9. *A node of a mesh $\mathcal{M}$ that is located in the interior of a geometric face of one of its cells is known as* ***hanging (dangling) node****.*

In the case of triangulations we can distinguish special classes:

- **simplicial triangulations** that entirely consist of triangles ($d = 2$) or tetrahedra ($d = 3$), whose edges/faces might be curved, nevertheless.
- **quadrilateral** ($d = 2$) and **hexahedral** ($d = 3$) triangulations, which only comprise cells of these shapes. Curved edges or faces are admitted, again.

Exercise 3.1. Let $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_4 \in \mathbb{R}^3$ stand for the coordinate vectors of the four vertices of a tetrahedron K. Determine the affine mapping (AFF) that takes the "unit tetrahedron"

$$\widehat{K} := \text{convex}\left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$$

to K. When will the the matrix $\mathbf{F}$ of this affine mapping be regular?

Corollary 3.10. *Any family of simplicial triangulations is affine equivalent.*

Example 3.11. In the case of a conforming simplicial triangulation the orientation of all geometric objects can be fixed by sorting the vertices. This will induce an ordering of the vertices of all cells, edges, and faces, which, in turns, defines their orientation.

Remark 3.12. A quadrilaterial triangulation need not be affine equivalent, because, for instance, there is no affine map taking a square to general trapezoid.

Exercise 3.2. Let $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_4 \in \mathbb{R}^2$ denote the coordinate vectors belonging to the four vertices of a quadrilateral K in the plane. Determine an analytic description of a simple smooth mapping $\boldsymbol{\Phi} : \mathbb{R}^2 \mapsto \mathbb{R}^2$ from the "reference square"

$$\widehat{K} := \text{convex}\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}$$

to K. Compute the Jacobian $D\boldsymbol{\Phi}$ and its determinant.

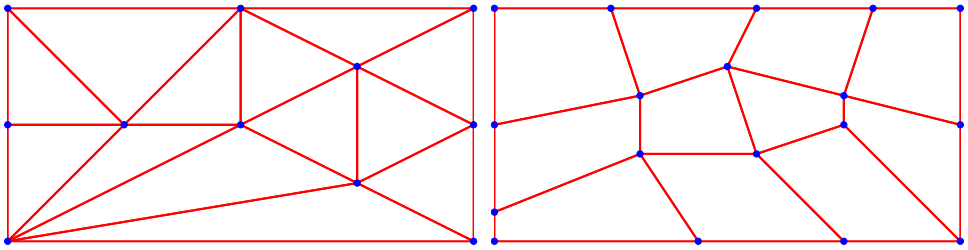


Figure 3.4: Examples of triangular and quadrilateral meshes in two dimensions

The term grid is often used as a synonym for triangulation, but we will reserve it for meshes with a locally **translation invariant** structure. These can be **tensor product grids**, that is meshes whose cells are quadrilaterals ($d = 2$) or hexahedra ($d = 3$) with parallel sides. Images of such meshes under a C^1-diffeomorphism will also be called (parametric) grids. Meshes that lack the regular structure of a grid are often dubbed **unstructured grids**, really an oxymoron.

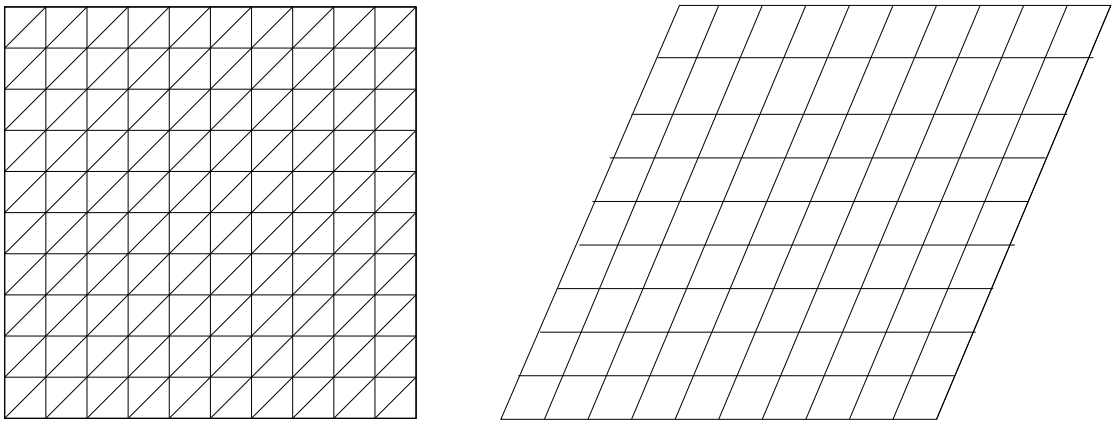


Figure 3.5: Example of triangular and quadrilateral grids in two dimensions

Automatic **mesh generation** is a challenging subject, which deals with the design of algorithms that create a mesh starting from a description of Ω. Such a description can be given

- in terms of geometric primitives (ball, brick, etc.) whose unions or intersections constitute Ω.
- by means of a parameterization of the faces of Ω.
- through a function $f : \mathbb{R}^d \mapsto \mathbb{R}$, whose sign indicates whether a point is located inside Ω or outside.
- by a mesh covering the surface of Ω and a direction of the exterior unit normal.

Various strategies can be employed for automatic grid generation:

- advancing front method that build cells starting from the boundary.
- Delaunay refinement techniques that can create a mesh starting from a mesh for $\partial\Omega$ or a "cloud" of points covering Ω.
- the quadtree ($d = 2$) or octree ($d = 3$) approach, which fills Ω with squares/cubes of different sizes supplemented by special measures for resolving the boundary.
- mapping techniques that split Ω into sub-domains of "simple" shape (curved triangles, parallelograms, bricks), endow those with parametric grids and glue these together.

Remark 3.13. Traditional codes for the solution of boundary value problems based on the finite element method usually read the geometry from a file describing the topolgy and geometry of the underlying mesh. Then an approximate solution is computed and written to file in order to be read by post-processing tools like visualization software, see Fig. 3.6.

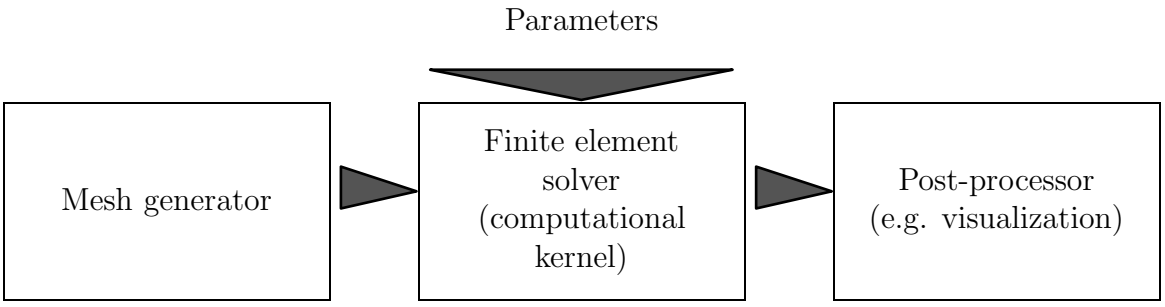


Figure 3.6: Flow of data in traditional finite element simulations

Remark 3.14. A typical file format for a mesh of a simplicial conforming triangulation of a two-dimensional polygonal domain is the following:

$$
\begin{array}{lll}
\# \texttt{Two-dimensional simplicial mesh} & & \\
N \in \mathbb{N} & \# \texttt{Number of nodes} & \\
\xi_1\ \eta_1 & \# \texttt{Coordinates of first node} & \\
\xi_2\ \eta_2 & \# \texttt{Coordinates of second node} & \\
\vdots & & \\
\xi_N \eta_N & \# \texttt{Coordinates of } N\texttt{-th node} & \\
M \in \mathbb{N} & \# \texttt{Number of triangles} & \\
n_1^1\ n_2^1\ n_3^1 \quad X_1 & \# \texttt{Indices of nodes of first triangle} & \\
n_1^2\ n_2^2\ n_3^2 \quad X_2 & \# \texttt{Indices of nodes of second triangle} & \\
\vdots & & \\
n_1^M\ n_2^M\ n_3^M \quad X_M & \# \texttt{Indices of nodes of } M\texttt{-th triangle} &
\end{array}
\tag{3.1}
$$

Here, X_i, $i = 1, \ldots, M$, is an additional piece of information that may, for instance, describe what kind of material properties prevail in triangle $\#i$. In this case X_i may be an integer index into a look-up table of material properties or the actual value of a coefficient function inside the triangle.

Additional information about edges located on $\partial\Omega$ may be provided in the following form:

$$
\begin{array}{lll}
K \in \mathbb{N} & \# \texttt{Number of edges on } \partial\Omega & \\
n_1^1\ n_2^1 \quad Y_1 & \# \texttt{Indices of endpoints of first edge} & \\
n_1^2\ n_2^2 \quad Y_2 & \# \texttt{Indices of endpoints of second edge} & \\
\vdots & & \\
n_1^K\ n_2^K \quad Y_K & \# \texttt{Indices of endpoints of } K\texttt{-th edge} &
\end{array}
\tag{3.2}
$$

where Y_k, $k = 1, \ldots, K$, provides extra information about the type of boundary conditions to be imposed on edge $\#k$. Some file formats even list all edges of the mesh in the format (3.2).

Please note that the ordering of the nodes in the above file formats implies an orientation of triangles and edges.

Exercise 3.3. Write a program (in the programming language of your choice) that reads in the description of a 2D mesh in the format (3.1) (without comments and extra information) and outputs an extended format containing a list of *all* edges according to (3.2). The ordering of the nodes of the edges can be arbitrary.

Exercise 3.4. Consider $\Omega := \{\boldsymbol{\xi} \in \mathbb{R}^2, |\boldsymbol{\xi}| < 1\}$ and the point set

$$\mathcal{P} := (\Omega \cap h\mathbb{Z}^2) \cup (\partial\Omega \cap ((h\mathbb{Z} \times \mathbb{R}) \cup (\mathbb{R} \times h\mathbb{Z}))) ,$$

where $h > 0$. Use the delaunay-Funktion of MATLAB in order to create triangular mesh $\mathcal{M}$ with $\mathcal{N}(\mathcal{M}) = \mathcal{P}$. Plot this mesh. Can it be regarded as a mesh of Ω.

Bibliographical notes. For a comprehensive account on mesh generation see [17]. An interesting algorithm for Delaunay meshing is described in [35, 37]. Also the internet offers plenty of information, see `http://www.andrew.cmu.edu/user/sowen/mesh.html`. Free mesh generation software is also available, NETGEN (`http://www.hpfem.jku.at/netgen/`), Triangle (`http://www-2.cs.cmu.edu/afs/cs/project/quake/public/www/triangle.html`), and GRUMMP (`http://tetra.mech.ubc.ca/GRUMMP/`), to name only a few. However, the most sophisticated mesh generation tools are commercial products and their algorithmic details are classified.

3.2 Polynomials

No other class of functions matches polynomials in terms of "simplicity". Moreover, since truncated Taylor series are polynomials, we expect good *local* approximation properties from them. Thus piecewise polynomials on meshes will be the building blocks of trial and test spaces in the finite element method.

Definition 3.15. *Given a domain $K \subset \mathbb{R}^d$, $d \in \mathbb{N}$, we write*

$$\mathcal{P}_m(K) := \operatorname{span}\{\boldsymbol{\xi} \in K \mapsto \boldsymbol{\xi}^{\boldsymbol{\alpha}} := \xi_1^{\alpha_1} \cdot \cdots \cdot \xi_d^{\alpha_d}, \boldsymbol{\alpha} \in \mathbb{N}_0^d, |\boldsymbol{\alpha}| \le m\}$$

for the vector space of ***d-variate polynomials*** *of (total) degree m, $m \in \mathbb{N}_0$. The space of* ***homogeneous polynomials*** *$\widetilde{\mathcal{P}}_m(K)$ is obtained by demanding $|\boldsymbol{\alpha}| = m$ in the above definition.*

If $\mathbf{m} = (m_1, \ldots, m_d)^T \in \mathbb{N}_0^d$ we designate by

$$\mathcal{Q}_{\mathbf{m}}(K) := \operatorname{span}\{\boldsymbol{\xi} \in K \mapsto \xi_1^{\alpha_1} \cdot \cdots \cdot \xi_d^{\alpha_d}, 0 \le \alpha_k \le m_k, 1 \le k \le d\}$$

the space of ***tensor product polynomials*** *of maximal degree m_k in the k-th coordinate direction.*

If $m < 0$ or $m_j < 0$, we adopt the convention $\mathcal{P}_m(K) = \{0\}$ and $\mathcal{Q}_{\mathbf{m}}(K) = \{0\}$.

It is easy to see that the restriction of $\mathcal{P}_p(K)$ to an affine subset S of K agrees with the space $\mathcal{P}_p(S)$. An analoguous statement does not hold for $\mathcal{Q}_p(K)$.

Lemma 3.16. *If $K \neq \emptyset$ is an open subset of $\mathbb{R}^d$, then the dimensions of the spaces of polynomials are given by*

$$\dim \mathcal{P}_m(K) = \binom{d+m}{m} , \ m \in \mathbb{N} , \quad \dim \mathcal{Q}_{\mathbf{m}}(K) \ = (m_1+1) \cdot \cdots \cdot (m_d+1) , \ \mathbf{m} \in \mathbb{N}_0^d .$$

Proof. Obviously, the monomials used to define the spaces in Def. 3.15 supply bases. Thus, the dimension $\dim \mathcal{P}_m(K)$ is equal to the number of ways in which m can be written as the sum of d non-negative integers. This, in turns, agrees with the number of ways to distribute m indistinguishable objects to $d+1$ containers. The dimension of $\dim \mathcal{Q}_{\mathbf{m}}(K)$ is established by a simple counting argument. □

We recall that an univariate polynomial of degree p, $p \in \mathbb{N}$, is already uniquely determined by prescribing its value for $p+1$ different arguments.

Exercise 3.5. What is the dimension of the space $\mathcal{P}_p(K) \cap H_0^1(K)$ for a triangle $K \subset \mathbb{R}^2$ and $p \in \mathbb{N}$? For the simplest non-trivial case describe a basis of this space. Answer the corresponding question for $(\mathcal{P}_1(K))^2 \cap H_0(\mathrm{div}; K)$.

Exercise 3.6. For a triangle $K \subset \mathbb{R}^2$ find a basis of the space $\{\mathbf{u} \in \mathcal{P}_1(K) : \mathrm{div}\, \mathbf{u} \in \mathcal{P}_0(K)\}$.

Exercise 3.7. If $K \subset \mathbb{R}^2$ is a triangle, determine an $L^2(K)$-orthonormal basis of $\mathcal{P}_1(K)$, that is, find $p_1, p_2, p_3 \in \mathcal{P}_1(K)$ such that

$$(p_j, p_k)_{L^2(K)} = \delta_{j,k} := \begin{cases} 0 & \text{if } j \neq k\,, \\ 1 & \text{if } j = k\,, \end{cases} \quad j,k = 1,2,3\,.$$

3.3 Abstract finite elements

Here, we give a formal recipe for obtaining so-called finite element subspaces of a space V of functions $\Omega \subset \mathbb{R}^d \mapsto \mathbb{R}^l$, $m \in \mathbb{N}$, Ω computational domain, on which a variational problem is posed. Usually, this function space will be one of the Sobolev spaces introduced in Sect. 2.7. These finite element subspace will serve as trial and test spaces V_n in the context of a Galerkin discretization.

The construction of a finite element space starts from two main ingredients:

1. A mesh $\mathcal{M}$ of the computational domain Ω, see Sect. 3.1.
2. A finite-dimensional **local trial space** $\Pi_K \subset (C^\infty(\overline{K}))^l$ for each cell $K \in \mathcal{M}$.

Remark 3.17. Sometimes the smoothness requirement on the local trial spaces is relaxed to allow spaces Π_K of piecewise smooth functions.

We can now state the following *preliminary* definition:

Definition 3.18 (preliminary). *Given a mesh $\mathcal{M}$ of a computational domain $\Omega \subset \mathbb{R}^d$, a family of local trial spaces $\{\Pi_K\}_{K\in\mathcal{M}}$, and a Sobolev space V of functions on Ω we call*

$$V_n := \{v \in V : \ v_{|K} \in \Pi_K \quad \forall K \in \mathcal{M}\} \subset V$$

*a **finite element space** that is V-**conforming**.*

Since $\sharp\mathcal{M} < \infty$, all these finite element spaces have finite dimension. As such they qualify as trial and test spaces for a Galerkin discretization.

What does it mean $v \in V$? To answer this question recall that piecewise smooth functions belong to a Sobolev space if and only if they feature certain continuity properties, see Lemmas 2.27 and 2.30. In other words,

> given a mesh $\mathcal{M}$ and local trial spaces Π_K for each $K \in \mathcal{M}$, a function u on Ω with $u_{|K} \in \Pi_K$ belongs to V, if and only if it satisfies suitable continuity conditions across (intercell) faces.

For concrete spaces these continuity conditions read

Space	Continuity condition	Quantity
$V = H^1(\Omega)$	$\rightarrow$ global continuity of the function	$>$ potential type
$V = H^2(\Omega)$	$\rightarrow$ continuity of the function and its gradient	$>$ deflection type
$V = H(\mathrm{div};\Omega)$	$\rightarrow$ continuity of the normal component	$>$ flux type
$V = H(\mathbf{curl};\Omega)$	$\rightarrow$ continuity of tangential components	$>$ gradient type
$V = L^2(\Omega)$	$\rightarrow$ no continuity required	$>$ density type

It is no coincidence that these continuity conditions are closely linked to the natural trace operators for the various Sobolev spaces that we studied in Sect. 2.7.4. We recall that these natural trace operators are the following:

Space	Natural trace	see
$V = H^1(\Omega)$	$\rightarrow$ pointwise restriction R_1	Thm. 2.49
$V = H^2(\Omega)$	$\rightarrow$ pointwise restrictions $(\mathsf{R}_1, \mathsf{R}_1 \circ \mathbf{grad})$	Cor. 2.29
$V = H(\mathrm{div};\Omega)$	$\rightarrow$ normal components trace $\mathsf{R}_\mathbf{n}$	Lemma 2.60
$V = H(\mathbf{curl};\Omega)$	$\rightarrow$ tangential components trace $\mathsf{R}_\times$	Exercise 2.23
$V = L^2(\Omega)$	$\rightarrow$ no trace $\mathsf{R} = 0$	

Notation: Below, we will write R for the natural trace belonging to the Sobolev space V.

Thus we can rephrase the above requirement.

> Given a mesh $\mathcal{M}$ of Ω and local trial spaces Π_K for each $K \in \mathcal{M}$, a function u on Ω with $u_{|K} \in \Pi_K$ belongs to V, if and only if the natural traces from both sides of an (intercell) face agree.

In short, the preceding considerations provide a recipe for gluing together the functions in local trial spaces. After having fixed the Π_K we can "easily" determine the resulting finite element space. However, the result may not be particularly useful.

Example 3.19. Let $\Omega :=]0;1[^2$ be equipped with the grid

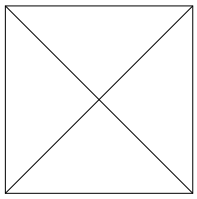
$$\mathcal{M} := \{]ih, (i+1)h[\times]jh, (j+1)h[,\ 0 \le i, j < N,\ h := 1/N\}$$

with $N \in \mathbb{N}$. For each square $K \in \mathcal{M}$ we use the local space $\mathcal{P}_1(K)$. However, if we aim for a $H_0^1(\Omega)$-conforming finite element space, we will end up with a trivial functions space only: Lemma 2.27 tells us that the functions have to be globally continuous. Besides, on $\partial\Omega$ the functions have to vanish and, since any $v \in \mathcal{P}_1(K)$ is already fixed by prescribing three values in points that are not collinear, the finite element function turns out to be zero in the corner square. By induction it is finally seen to vanish erverywhere in Ω.

Exercise 3.8. Consider a quadrilateral triangulation of $\Omega :=]0;1[^2$ that consists of nine congruent squares. Pick the local trial space $\mathcal{P}_2(K)$ for each square K. Compute a basis of the $H_0^1(\Omega)$-conforming finite element space that emerges from this choice.

Exercise 3.9.

For $\Omega =]0;1[^2$ consider the mesh sketched beside. For all cells we choose $\Pi_K = (\mathcal{P}_1(K))^2$. Find bases of the resulting $(H_0^1(\Omega))^2$-conforming and $H_0(\mathrm{div};\Omega)$-conforming finite element spaces.



In fact, Def. 3.18 covers all standard finite element spaces, but we have seen that actually finding V_n may be difficult or even impossible, in particular on large unstructured meshes. We badly need a guideline for choosing and glueing the local spaces. The tool for doing this are so-called degrees of freedom (d.o.f.).

Definition 3.20. *Given a cell K of some mesh with associated local trial space Π_K, we call linear functionals $(C^\infty(\overline{K}))^l \mapsto \mathbb{R}$ **local degrees of freedom** if they provide a basis for the dual space $(\Pi_K)^*$. We write Σ_K for a generic set of local degrees of freedom for Π_K.*

Remark 3.21. The property that Σ_K is a dual basis is also known as the **unisolvence** of the functionals in Σ_K. It is equivalent to

$$\sharp\Sigma_K = \dim \Pi_K \quad \text{and} \quad \forall v \in \Pi_K : \quad l(v) = 0\ \forall l \in \Sigma_K \quad \Rightarrow \quad v = 0\ .$$

Prescribing the values of all d.o.f. in Σ_K will single out a unique $v \in \Pi_K$ that produces exactly those values $l(v)$, $l \in \Sigma_K$. The common parlance is that the degrees of freedom *fix the function v*.

Definition 3.22. *Given a local d.o.f. $\Sigma_K = \{l_1^K, \ldots, l_k^K\}$ for Π_K, K a cell of a mesh, there are $k := \dim \Pi_K$ **local shape functions** $b_1^K, \ldots, b_k^K$ such that*

$$l_m^K(b_j^K) = \delta_{mj}\ , \quad m, j \in \{1, \ldots, k\}\ .$$

As Π_K is a basis of the dual space, the local shape functions are uniquely defined.

Remark 3.23. There is a perfect duality linking local degrees of freedom and the local shape functions: precribing the latter will also uniquely define the former.

Exercise 3.10. For a triangle $K \subset \mathbb{R}^2$ use $\Pi_K := \mathcal{P}_1(K)$ and

$$\Sigma_K := \{u \mapsto u(\boldsymbol{\mu}),\, u \mapsto \tfrac{\partial}{\partial \xi_1} u(\boldsymbol{\mu}),\, u \mapsto \tfrac{\partial}{\partial \xi_2} u(\boldsymbol{\mu})\}\ ,$$

where $\boldsymbol{\mu}$ is the center of gravity of K. Is Σ_K a set of degrees of freedom ofr Π_K? If so, compute the shape functions.

The local degrees of freedom must be carefully chosen, lest they fail to enforce the crucial continuity conditions.

Definition 3.24. *Let K be a cell of a mesh $\mathcal{M}$ and F be a face/edge/node of the mesh that is contained in $\overline{K}$. Given a local trial space Π_K, $\Pi_K \subset (C^\infty(\overline{K}))^l$, and a set Σ_K of local degrees of freedom, a linear functional $l \in \Sigma_K$ is called* ***localized/supported on*** *F or* ***associated with*** *F, if*

$$l(v) = 0 \quad \forall v \in (C^\infty(\overline{K}))^l,\ \mathrm{supp}(v) \cap F = \emptyset\ .$$

Notation: The d.o.f. localized on a face F of K form the set $\Sigma_K(F)$.
By duality, localized degrees of freedom permit us to talk about "local shape functions associated with faces/edges/nodes".

Definition 3.25. *A* ***finite element*** *is a triple (K, Π_K, Σ_K) such that*

(i) K is a cell of a mesh $\mathcal{M}$ of the computational domain $\Omega \subset \mathbb{R}^d$.

(ii) $\Pi_K \subset (C^\infty(\overline{K}))^l$, and $\dim \Pi_K < \infty$.

(iii) Σ_K is a set of local degrees of freedom.

A finite element is called V ***-conforming****, if*

(iv) for any face F of K the degrees of freedom localized on F uniquely determine the natural trace $\mathsf{R}\, u_{|F}$ of a $u \in \Pi_K$ onto F.

Corollary 3.26. $\sharp\Sigma_K(F) = \dim(\mathsf{R}(\Pi_K)_{|F})$

Example 3.27. Let K be a triangle and consider a H^1-conforming finite element $(K, \mathcal{P}_m(K), \Sigma_K)$ that is to serve as the building block for a H^1-conforming finite element space. This means that the degrees of freedom associated with an edge E of K have to fix the values of local trial functions in every point of $\overline{E}$, because the pointwise restriction is the natural trace for $H^1(\Omega)$.

The restriction of $\Pi_K = \mathcal{P}_m(K)$ to an edge E of K yields $\mathcal{P}_m(E)$, which is a space of dimension $m+1$. Consequently, exactly $m+1$ degrees of freedom must be associated with E and they form the set $\Sigma_K(E) \subset \Sigma_K$.

Pick an endpoint $\boldsymbol{\nu}$ of E and write E' for another edge sharing it. We aim to show that $\Sigma_K(E) \cap \Sigma_K(E') \neq \emptyset$. Assume that the intersection was empty. Then, in light of requirement (iv) from Def. 3.25 we can choose the values of d.o.f. $\in \Sigma_K(E)$ such that this fixes $\mathsf{R}_1\, v_E \equiv 1$. On the other hand, demanding that all d.o.f. in $\Sigma_K(E')$ vanish will involve $\mathsf{R}_1\, v_{E'} \equiv 0$. This is impossible, because polynomials cannot jump at $\boldsymbol{\nu}$.

The degree(s) of freedom contained in $\Sigma_K(E) \cap \Sigma_K(E')$ will be localized on the vertex $\boldsymbol{\nu}$. The simplest example is the point evaluation $v \mapsto v(\boldsymbol{\nu})$.

Bibliographical notes. This abstract approach to finite elements can be found in [12, Ch. 2] and [9, Ch. 3].

3.4 Finite element spaces

The localization of degrees of freedom paves the way for converting them into global degrees of freedom. The starting point is the following additional constraint:

> Given a conforming triangulation $\mathcal{M}$ and a family $\{(K, \Pi_K, \Sigma_K)\}_{K\in\mathcal{M}}$ of V-conforming finite elements built on it, we demand that for each face F adjacent to two cells $K, K' \in \mathcal{M}$ there is a bijection ("matching") $\kappa : \Sigma_K(F) \mapsto \Sigma_{K'}(F)$ such that the requirement
>
> $$l(v) = \kappa(l)(v') \quad \forall l_K \in \Sigma_K(F) \quad \text{for } v \in \Pi_K, v' \in \Pi_{K'}$$
>
> ensures that the natural traces $\mathsf{R}\, v_{|F}$ and $\mathsf{R}\, v'_{|F}$ from K and K', respectively, coincide. Degrees of freedom linked by κ are called **matching d.o.f.**. An **interior d.o.f.** that is not localized on any face, is regarded as matching itself.

Remark 3.28. Matching local degrees of freedom have the same support.

Now we are in a position to give the *practical definition* of a finite element space. To keep it simple we assume that no boundary conditions are imposed in the definition of V. This is true for the Sobolev spaces introduced in Sect. 2.7.2. However, spaces like $H^1_0(\Omega), H_0(\mathrm{div};\Omega)$ are (temporarily) excluded.

Definition 3.29 (practical). *Given a mesh $\mathcal{M}$ of a computational domain $\Omega \subset \mathbb{R}^d$, a Sobolev space V of functions on Ω, and a family of V-conforming finite elements $\{(K, \Pi_K, \Sigma_K)\}_{K\in\mathcal{M}}$ with matching degrees of freedom we call*

$$V_n := \{v \in L^2(\Omega) :\ v_{|K} \in \Pi_K\ \forall K \in \mathcal{M} \text{ and matching d.o.f. agree}\}$$

*a **finite element space** that is V-**conforming**.*

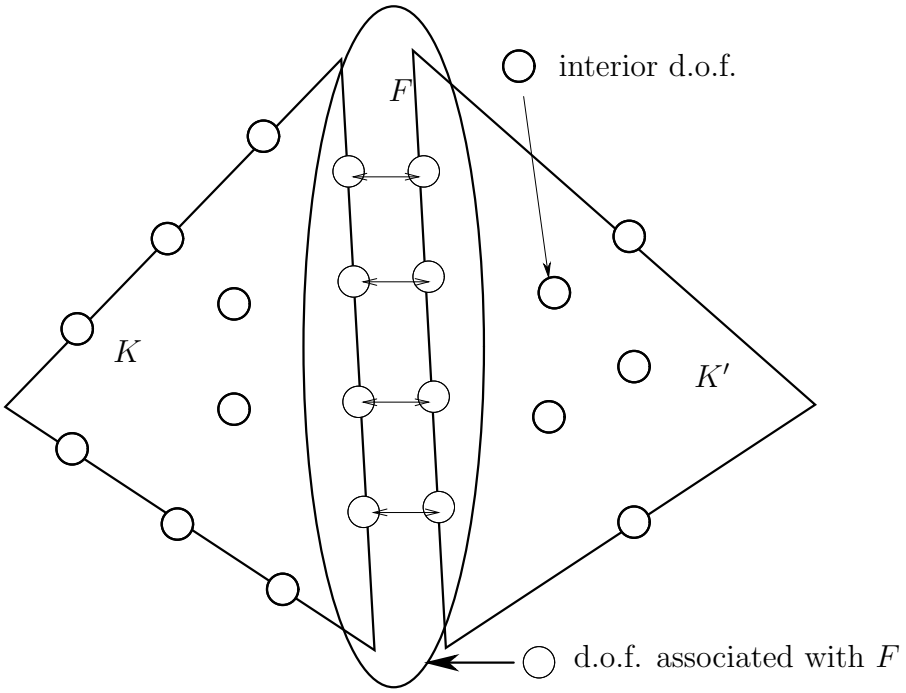


Figure 3.7: Matching d.o.f. across an intercell face. Double arrows indicate the bijection κ

The equality of matching degrees of freedom should be understood as follows: a function u on Ω that is $\mathcal{M}$-piecewise in Π_K for every $K \in \mathcal{M}$ will be restricted to two adjacent cells K, K' with shared face F. If $\kappa : \Sigma_K(F) \mapsto \Sigma_{K'}(F)$ denotes the matching, we expect

$$l(u_{|K}) = \kappa(l)(u_{|K'}) \quad \forall l \in \Sigma_K(F) \; . \tag{3.3}$$

By definition of "matching" the functions in V_n feature continuity of their natural traces across intercell faces:

Corollary 3.30. *A finite element space V_n from Def 3.29 satisfies $V_n \subset V$.*

> Finding local trial spaces that allow for degrees of freedoms that match across intercell faces is challenging aspect of the construction of finite element spaces.

In Sect. 2.8 we have learned that essential boundary conditions have to be enforced on trial and test functions in the case of a weak formulation of a boundary value problem. Let us assume that we want to get a finite element subspace of

$$V_0 := \{v \in V : \; \mathsf{R}\, v = 0 \;\; \text{on } \Gamma_0 \subset \Gamma\} \; ,$$

where $\overline{\Gamma}_0$ is the union of closed faces of the underlying triangulation $\mathcal{M}$. Thanks to the localization of global/local d.o.f. we easily find the proper extension of Def. 3.29:

$$V_{n,0} := \{v \in V_n : \ l(v) = 0 \quad \forall \text{ d.o.f. } l \text{ supported on nodes/edges/faces } \subset \Gamma_0\} \ .$$

Homogeneous essential boundary conditions on a part of Γ are enforced in the finite element context by setting all d.o.f. localized on that part to zero.

Remark 3.31. Given the setting of Def. 3.29 and assuming the absence of essential boundary conditions in V, in many cases both definitions Def. 3.29 and Def. 3.18 provide exactly the same finite element space, but *not necessarily so*.

Remark 3.32. Following the recipe of Def. 3.29 different families of finite elements can lead to the same finite element space, provided they differ only in their local degrees of freedom. Examples will be given in Sect. 3.10.

Exercise 3.11. Let $K \subset \mathbb{R}^2$ be a triangle with vertices $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \boldsymbol{\nu}_3$ and oriented edges E_1, E_2, E_3 (with midpoints $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3$). Write $\mathbf{n}_j$ for the exterior unit normal vector at edge E_j, $j = 1, 2, 3$, and $\boldsymbol{\tau}_j$ for unit vectors in the direction of E_j.

We aim for a $H(\mathrm{div})$-conforming finite element and choose $\Pi_K := (\mathcal{P}_1(K))^2$. The following sets of linear functionals on Π_K should be considered:

(A) $\Sigma_K := \{\mathbf{v} \mapsto v_j(\boldsymbol{\nu}_i),\ j = 1, 2,\ i = 1, 2, 3\}$,

(B) $\Sigma_K := \{\mathbf{v} \mapsto \langle \mathbf{v}, \mathbf{n}_i \rangle (\boldsymbol{\mu}_i),\ i = 1, 2, 3,\ \mathbf{v} \mapsto \boldsymbol{\tau}_i^T D\mathbf{v}(\boldsymbol{\mu}_i)\mathbf{n}_i,\ i = 1, 2, 3\}$, $D\mathbf{v}$ Jacobi matrix.

(i) Show that both sets provide valid local degrees of freedom.

(ii) Determine the local shape functions in each case.

(iii) Show that only the choice (B) leads to a $H(\mathrm{div})$-conforming finite element.

3.5 Global shape functions

We take for granted a V-conforming finite element space V_n according to Def. 3.29. The concept of "matching" permits us to convert local degrees of freedom into **global degrees of freedom** by lumping together all matching local degrees of freedom: a global degree of freedom will be an equivalence class of matching local degrees of freedom. It can be viewed as linear form on the finite element space V_n in the following sense.

Given a global degree of freedom $\underline{l}$ and $u_n \in V_n$ the evaluation of $\underline{l}(u)$ proceeds as follows:

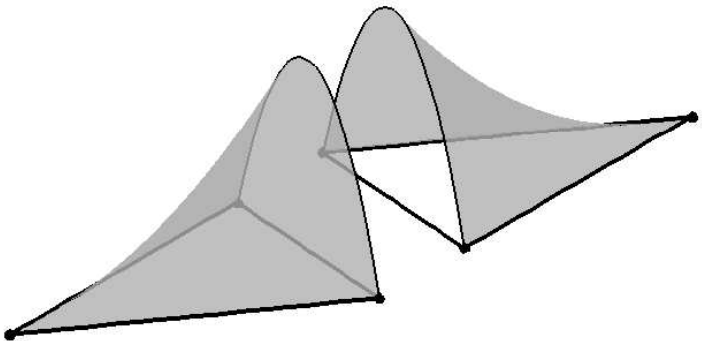


Figure 3.8: Matching of local shape functions associated with an edge for a Lagrangian finite element of degree 2

1. Pick a cell $K \in \mathcal{M}$ and a local d.o.f. $l \in \Sigma_K$ that is a representative of $\underline{l}$.
2. Define $\underline{l}(u_n) := l(u_{n|K})$.

Thanks to the notion of "matching", $\underline{l}$ is well defined and its linearity is clear.

Notation: The set of global degrees of freedom belonging to a finite element space V_n will be denoted by $\mathrm{gdof}(V_n)$.

Remark 3.33. Interior local degrees of freedom directly qualify as global degrees of freedom. Moreover, similar to local degrees of freedom their global counterparts are *localized on cells/faces/edges/nodes* of the mesh.

Owing to locality global degrees of freedom can be evaluated for functions that do not belong to the finite element space, for instance, for $w \in (C^\infty(\overline{\Omega}))$: for $\underline{l} \in \mathrm{gdof}(V_n)$ we pick a $K \in \mathcal{M}$ and $l \in \Sigma_K$ that is a representative of the equivalence class $\underline{l}$ and has the same support as $\underline{l}$. Then one defines $\underline{l}(w) := l(w_{|K})$.

Theorem 3.34. *For a finite element space V_n the set* $\mathrm{gdof}(V_n)$ *forms a basis of the dual space* $(V_n)^*$.

Proof. Prescribing the values of all global d.o.f. means that all local d.o.f. are prescribed as well. This will fix function $v_K \in \Pi_K$ for each cell $K \in \mathcal{M}$, which can be combined to a function $v \in V_n$. Linear independence of the global d.o.f. can now be demonstrated using the function that arises from setting exactly one global d.o.f. to one and the rest of them to zero.

Further, unisolvence of the local d.o.f. implies the same for $\mathrm{gdof}(V_n)$. □

Definition 3.35. *Given a finite element space V_n the set of* ***global shape functions*** *or* ***nodal basis functions*** $\mathfrak{B}(V_n) := \{b^1, \ldots, b^N\}$, $N := \dim V_n$, *is the basis of V_n that is dual to* $\mathrm{gdof}(V_n)$.

By definition there will be a one-to-one correspondence between global d.o.f. and global shape functions. The global degrees of freedom are localized and this property carries over the local shape functions.

Theorem 3.36. *Given a V-conforming finite element space V_n, let $\underline{l} \in \mathrm{gdof}(V_n)$ be localized on F, which may be a node/edge/face/cell of the mesh $\mathcal{M}$. Then the global shape function b_F belonging to $\underline{l}$ ($\underline{l}(b_F) = 1$) satisfies*

$$\mathrm{supp}(b_F) = \bigcup \{\overline{K} : \ K \in \mathcal{M}, \ F \subset \overline{K}\} \ .$$

Proof. If $K \in \mathcal{M}$ is not contained in $\bigcup\{\overline{K} : \ K \in \mathcal{M}, \ F \subset \overline{K}\}$, then no local d.o.f. in the equivalence class of $\underline{l}$ will belong to Σ_K. Hence, they will all vanish for $b_{F|K}$, which, by unisolvence, implies $b_{F|K} \equiv 0$. □

Global finite element shape functions have local supports

Moreover, the matching of local d.o.f. ensures that

the restriction of a global shape function onto a cell
agrees with a local shape function on this cell.

Remark 3.37. The "matching condition" for local d.o.f. is equivalent to demanding that the related local shape function can be "sewn together" across intercell faces to yield a function in V.

Remark 3.38. The statement of Remark 3.23 carries over to global shape functions. In particular, the construction of finite element spaces can also start from local/global shape functions. This perspective is completely equivalent to the approach via degrees of freedom ("dual view", see Fig. 3.9).

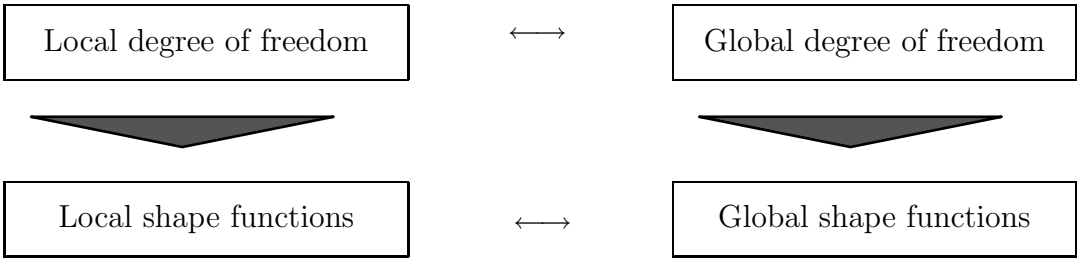


Figure 3.9: Duality of degrees of freedom and shape functions

Numerous example of finite elements and corresponding global shape functions for different Sobolev spaces V will be given in Sect. 3.8.

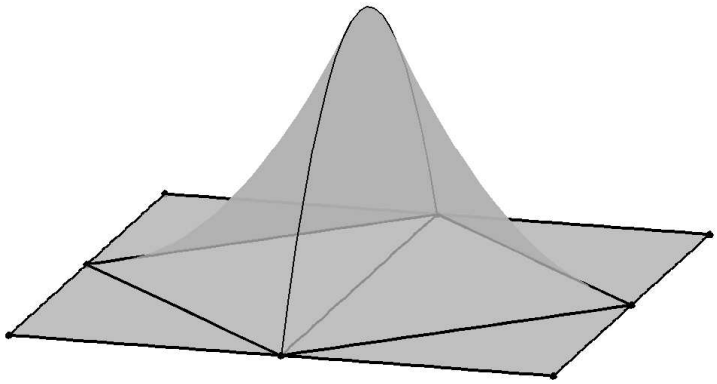


Figure 3.10: Global shape function localized on an intercell edge for a finite element space based on Lagrangian finite elements of degree 2

3.6 Finite element interpolation operators

Again, we assume that the triangulation $\mathcal{M}$ is equipped with a family of V-conforming finite elements $(K, \Pi_K, \Sigma_K)_{K\in\mathcal{M}}$ according to Def. 3.25 and that they possess mathing degrees of freedom. Then Def. 3.29 will give us the V-conforming finite element space V_n.

As explained in Remark 3.21 a local trial function $v \in \Pi_K$, $K \in \mathcal{M}$, is known once we know the values $l(v)$, $l \in \Sigma_K$. However, any degree of freedom $l \in \Sigma_K$ can be evaluated for any function $\in (C^\infty(\overline{K}))^l$. This defines a mapping $(C^\infty(\overline{K}))^l \mapsto \Pi_K$.

Definition 3.39. *Given a finite element $(K, \Pi_K, \Sigma_K = \{l_1, \ldots, l_k\})$ with local shape functions $b_1^K, \ldots, b_k^K$, $k := \dim \Pi_K$, we define the* ***local interpolation operator*** $\mathsf{I}_K : (C^\infty(\overline{K}))^l \mapsto \Pi_K$ *by*

$$\mathsf{I}_K(w) = \sum\nolimits_{j=1}^{k} l_j(w)\, b_j^K \quad \forall w \in (C^\infty(\overline{K}))^l \ .$$

Analoguously, writing X for the space of functions for which all global degrees of freedom are well defined, we can introduce the (global) ***finite element interpolation operator*** $\mathsf{I}(V_n) : X \mapsto V_n$ *by*

$$\mathsf{I}(V_n)(w) = \sum\nolimits_{j=1}^{N} \underline{l}_j(w)\, b_j \quad \forall w \in X \ ,$$

where $N := \dim V_n$ and $\{b_1, \ldots, b_N\}$ is the set of global shape functions with b_j belonging to the global d.o.f. $\underline{l}_j$.

Notation: If the underlying finite element space is clear, we will simply write I for the (global) finite element interpolation operator.

Remark 3.40. Obviously $(C^\infty(\overline{\Omega}))^l \subset X$ and $V_n \subset X$.

Lemma 3.41. *Both interpolation operators I_K, $K \in \mathcal{M}$, and $\mathsf{I}(V_n)$ are linear projections.*

Proof. Linearity is inherited from the linear functional in the definition. The relationships $\mathsf{I}_K^2 = \mathsf{I}_K$ and $\mathsf{I}^2 = \mathsf{I}$ immediately follow from the fact that the local/global shape functions have been introduced as basis functions dual to the local/global d.o.f. □

Theorem 3.42. *The global finite element interpolation operator introduced in Def. 3.39 is local in the sense that*

$$\forall K \in \mathcal{M},\, w \in X \cap V: \quad w_{|K} \equiv 0 \quad \Rightarrow \quad \mathsf{I}(w)_{|K} \equiv 0\,.$$

Proof. Pick a $K \in \mathcal{M}$. If $w_{|K} \equiv 0$ and $w \in V$, then $\mathsf{R}_{\partial K}\, w \equiv 0$. Then

$$\forall l \in \Sigma_K: \quad l(\mathsf{I}(w)_{|K}) = l(w_{|K}) = 0 \quad \Rightarrow \quad \mathsf{I}(w)_{|K} \equiv 0\,,$$

due to the unisolvence of Σ_K and the fact that $\mathsf{I}(w)_{|K} \in \Pi_K$. □

Putting it shortly, the values of w in K completely determine its (global) finite element interpolant on K.

Remark 3.43. The finite element interpolation operator for a V-conforming finite element space need not be bounded on V.

3.7 Parametric finite elements

By definition, *cf.* Def. 3.1, the cells of a mesh $\mathcal{M}$ of a computational domain $\Omega \subset \mathbb{R}^d$ are diffeomorphic images of some d-polytope. Reasonable meshes can usually described by specifying a finite set of **reference cells** $\widehat{K}_1, \ldots, \widehat{K}_P$, $P \in \mathbb{N}$, and suitable diffeomorphism.

This way of constructing a mesh can be adopted for finite elements, as well.

Definition 3.44. *Two V-conforming finite elements (K, Π_K, Σ_K) and $(\widehat{K}, \Pi_{\widehat{K}}, \Sigma_{\widehat{K}})$ are called **parametric equivalent**, if*

(i) there is a diffeomorphism $\mathbf{\Phi} : \overline{\widehat{K}}_p \mapsto \overline{K}$.

(ii) the local trial spaces $\Pi_K, \Pi_{\widehat{K}}$ satisfy

$$\Pi_{\widehat{K}} = \mathsf{XT}_{\mathbf{\Phi}}(\Pi_K)\,, \tag{3.4}$$

where XT stands for the natural (pullback) transformation of functions in V, see Sect. 2.2.

Sobolev space	quantity	transformation
$H^1(\Omega)$	potential type	Pullback $\mathsf{FT}_{\mathbf{\Phi}}$, see (FT)
$H(\mathrm{div};\Omega)$	flux type	Piola transformation $\mathsf{PT}_{)}\,\mathbf{\Phi}$, sse (PT)
$L^2(\Omega)$	density type	$\mathsf{DT}_{\mathbf{\Phi}}$, see (DT)
$H(\mathbf{curl};\Omega)$	gradient type	$\mathsf{GT}_{\mathbf{\Phi}}$, see (GT)
$H^2(\Omega)$	—	Pullback $\mathsf{FT}_{\mathbf{\Phi}}$, see (FT)

Table 3.1: Natural transformations associated with certain Sobolev spaces

(iii) the degrees of freedom are connected by

$$\forall\, l \in \Sigma_K : \quad \exists\, \widehat{l} \in \Sigma_{\widehat{K}} : \quad l(v) = \widehat{l}(\mathsf{XT}_{\mathbf{\Phi}}(v)) \quad \forall v \in \Pi_K\ . \tag{3.5}$$

If $\mathbf{\Phi}$ is an affine mapping (AFF)*, then the finite elements are called* ***affine equivalent****.*

Remark 3.45. Parametric equivalence defines an equivalence relation on finite elements.

Remark 3.46. If (K, Π_K, Σ_K) and $(\widehat{K}, \Pi_{\widehat{K}}, \Sigma_{\widehat{K}})$ are parametric equivalent, then the transformation $\mathsf{XT}_{\mathbf{\Phi}}$ establishes a bijection between the shape functions on K and $\widehat{K}$, respectively.

Remark 3.47. Given two cells $K, \widehat{K}$ of a mesh and a V-conforming finite element on $\widehat{K}$, we can use Def. 3.44 to *define* a V-conforming finite element on K. Observe that, if $\mathbf{\Phi}$ is a diffeomorphism, this also holds true for the the transformation $\mathsf{XT}_{\mathbf{\Phi}}$. Hence $\Pi_{\widehat{K}}$ and $\Sigma_{\widehat{K}}$ defined by (3.4) and (3.5), respectively, will satisfy all the requirements of Def. 3.25. Hence, it is enough to specify finite elements for the reference cells and use this *parametric construction of finite elements* to obtain finite elements for the actual cells.

The notion of parametric equivalence can be extended to families of finite elements.

Definition 3.48. *A family of finite element is called* ***parametric (affine) equivalent*** *if there are only a finite number of parametric (affine) equivalence classes.*

Given a family of meshes $\mathcal{M}_{nn\in\mathbb{N}}$, an associated family of finite element spaces is called ***parametric (affine) equivalent****, if the set $(K, \Pi_K, \Sigma_K)_{K\in\mathcal{M}_n, n\in\mathbb{N}}$ of all underlying finite elements is parametric (affine) equivalent.*

Parametric and, in particular, affine equivalent families of finite element spaces, will play a key role in estimating the best approximation errors incurred in the case of finite element spaces.

Exercise 3.12. Given a $H(\mathrm{div};\Omega)$-conforming finite element (K, Π_K, Σ_K), K a simplex in $\mathbb{R}^d$, show that the parametric construction of a finite element $(\widehat{K}, \Pi_{\widehat{K}}, \Sigma_{\widehat{K}})$ as explained in Remark 3.47 yields another $H(\mathrm{div};\Omega)$-conforming finite element. In particular, show that requirement (iv) from Def. 3.25 is satisfied.

3.8 Particular finite elements

Now, we give a survey of important types of conforming finite elements that pervade the Galerkin discretization of boundary value problems for partial differential equations.

3.8.1 H^1-conforming Lagrangian finite elements

H^1-Conformity entails global continuity of functions in a finite element space. Lagrangian finite elements use degrees of freedom based on point evaluations in order to guarantee this necessary condition: they belong to the class of **C^0-elements**.

We first consider the case of simplicial meshes. They are naturally affine equivalent in the sense of Def. 3.3 and lend themselves to an affine equivalent construction of finite elements, see Sect. 3.7. Thus, we only need to specify a finite element for a (generic) **reference simplex** $\widehat{K}$.

Definition 3.49. *Given $d+1$ points $\boldsymbol{\nu}^1, \ldots, \boldsymbol{\nu}^{d+1} \in \mathbb{R}^d$ that do not lie in a hyperplane the **barycentric coordinates** $\lambda_1 = \lambda_1(\boldsymbol{\xi}), \ldots, \lambda_{d+1} = \lambda_{d+1}(\boldsymbol{\xi}) \in \mathbb{R}$ of $\boldsymbol{\xi} \in \mathbb{R}^d$ are uniquely defined by*

$$\lambda_1 + \cdots + \lambda_{d+1} = 1 \quad , \quad \lambda_1 \boldsymbol{\nu}^1 + \cdots + \lambda_{d+1} \boldsymbol{\nu}^{d+1} = \boldsymbol{\xi} \, .$$

The barycentric coordinates can be obtained by solving

$$\begin{pmatrix} \nu_1^1 & \cdots & \nu_1^{d+1} \\ \vdots & & \vdots \\ \nu_d^1 & \cdots & \nu_d^{d+1} \\ 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_d \\ \lambda_{d+1} \end{pmatrix} = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_d \\ 1 \end{pmatrix} \, , \tag{3.6}$$

which shows their uniqueness and existence, if the points $\boldsymbol{\nu}^j$ are not complanar. The convex hull of $\boldsymbol{\nu}^1, \ldots, \boldsymbol{\nu}^d$ can be described by

$$\operatorname{convex}\{\boldsymbol{\nu}^1, \ldots, \boldsymbol{\nu}^d\} = \{\boldsymbol{\xi} \in \mathbb{R}^d, \ 0 \le \lambda_i(\boldsymbol{\xi}) \le 1, \ 1 \le i \le d+1\} \, .$$

Corollary 3.50. *Given $d+1$ points $\boldsymbol{\nu}^1, \ldots, \boldsymbol{\nu}^{d+1} \in \mathbb{R}^d$ as in Def. 3.49, the barycentric coordinates are affine linear functions on $\mathbb{R}^d$, which satisfy*

$$\lambda_j(\boldsymbol{\nu}^i) = \delta_{ij} \quad 1 \le i, j \le d+1 \, .$$

Lemma 3.51. *Given $d+1$ points $\boldsymbol{\nu}^1, \ldots, \boldsymbol{\nu}^{d+1} \in \mathbb{R}^d$ as in Def. 3.49 and writing $\lambda_1, \ldots, \lambda_{d+1}$ for the associated barycentric coordinate functions, we have for all $m \in \mathbb{N}$*

$$\mathcal{P}_m(\mathbb{R}^d) = \operatorname{span}\{\lambda_1^{\alpha_1} \cdot \cdots \cdot \lambda_{d+1}^{\alpha_{d+1}}, \ \alpha_i \in \mathbb{N}_0, \ \sum\nolimits_{i=1}^{d+1} \alpha_i = m\} \, .$$

Proof. Identifying $\xi_{d+1} \equiv 1$ we can write

$$\mathcal{P}_m(\mathbb{R}^d) = \operatorname{span}\{\xi_1^{\alpha_1} \cdot \dots \cdot \xi_{d+1}^{\alpha_{d+1}},\ \alpha_i \in \mathbb{N}_0,\ \sum_{i=1}^{d+1} \alpha_i = m\}\ .$$

The identity (3.6) immediately confirms that all ξ_j, $j = 1, \dots, d+1$ can be written as a linear combination of the barycentric coordinate functions. This finishes the proof. □

Lemma 3.52. *Let K be a non-degenerate d-simplex with vertices $\boldsymbol{\nu}^1, \dots, \boldsymbol{\nu}^{d+1}$. Then*

$$\mathcal{P}_m(K) \cap \{v \in C^\infty(\overline{K}) :\ v_{|\partial K} = 0\} = \operatorname{span}\{p \cdot \lambda_1 \cdot \dots \cdot \lambda_{d+1},\ p \in \mathcal{P}_{m-d-1}(K)\}\ .$$

The functions in the latter set furnish a basis of $\mathcal{P}_m(K)$.

Proof. Let the face F of the simplex K be spanned by the vertices $\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_d$. Assume that $p \in \mathcal{P}_m(K)$ has zero restriction to F.

By Lemma 3.51 the "barycentric monomials" $\boldsymbol{\xi} \mapsto \lambda_1^{\alpha_1} \cdot \dots \cdot \lambda_{d+1}^{\alpha_{d+1}}$, $\boldsymbol{\alpha} := (\alpha_1, \dots, \alpha_d) \in \mathbb{N}^d$, $|\boldsymbol{\alpha}| = m$, form a basis of $\mathcal{P}_m(F)$. Hence, they must not contribute to the representation of p according to Lemma 3.51. The remaining terms all contain at least one power of λ_{d+1}.

Applying this argument to all faces gives the desired representation. □

Aware of Lemma 3.16, we instantly infer the dimension of the space of polynomials that vanish on all faces of a simplex.

Corollary 3.53. $$\dim \mathcal{P}_m(K) \cap \{v \in C^\infty(\overline{K}) :\ v_{|\partial K} = 0\} = \binom{m-1}{d}$$

Definition 3.54. *Given a non-degenerate simplex $K \in \mathbb{R}^d$, we define the H^1-conforming* ***simplicial Lagrangian finite element of degree*** *$m \in \mathbb{N}$ by*

(i) $\Pi_K := \mathcal{P}_m(K)$.

(ii) $\Sigma_K := \{v \in C^\infty(\overline{K}) \mapsto v(\boldsymbol{\xi}),\ \boldsymbol{\xi} \in \mathcal{N}_K\}$, *where*

$$\mathcal{N}_K := \{\boldsymbol{\xi} \in \overline{K} :\ \lambda_j(\boldsymbol{\xi}) \in \{0, \tfrac{1}{m}, \dots, \tfrac{m-1}{m}, 1\},\ j = 1, \dots, d+1\}\ . \tag{3.7}$$

Theorem 3.55. *Def. 3.54 describes a valid H^1-conforming finite element.*

Proof. Let us start with the observation, that the degrees of freedom associated with a sub-simplex S (vertex, edge, etc.) of K are exactly those evaluations on points in the closure of S.

To prove unisolvence we first note that Lemma 3.16 and a simple counting argument show

$$\dim \mathcal{P}_m(K) = \sharp \mathcal{N}_K\ .$$

We still have to show that

$$v \in \Pi_K :\quad l(v) = 0 \quad \forall l \in \Sigma_K \quad \Rightarrow \quad v = 0\ .$$

To do so we proceed by induction: assume that $p \in \mathcal{P}_m(K)$ is zero at all $\boldsymbol{\xi} \in \mathcal{N}_K$.

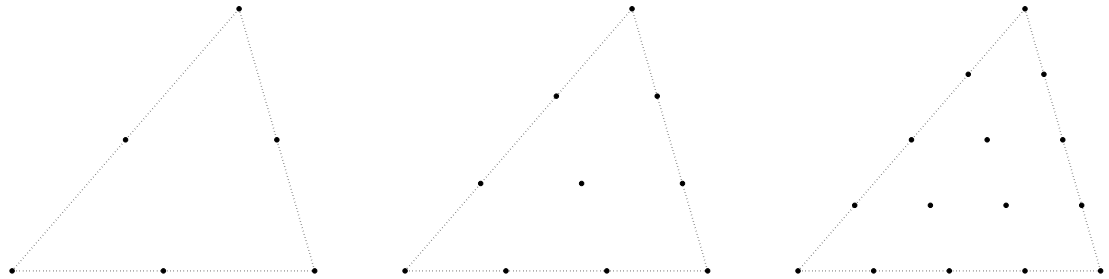


Figure 3.11: Location of local sampling points for triangular Lagrangian finite elements of degree 2 (left), degree 3 (middle), and degree 4 (right)

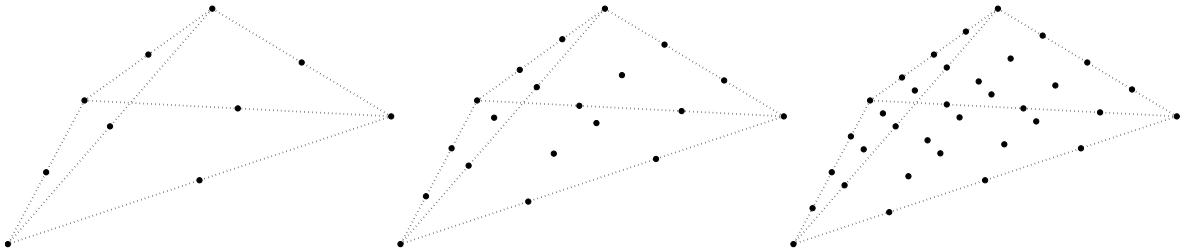


Figure 3.12: Point sets $\mathcal{N}$ for tetrahedral Lagrangian finite elements of degree 1 (left), degree 2 (middle), and degree 3 (right)

- For $d = 1$ the assertion of the theorem is evident, because a polynomial of degree m is uniquely determined by its values for $d + 1$ different arguments.
- For $m = 1$ Lemma 3.51 and Corollary 3.50 settle the issue.

For general m, d note that the faces of K are simplices of dimension $d - 1$. For any face S it is clear that $\mathcal{N}_K \cap \overline{S}$ equals $\mathcal{N}_S$, where $\mathcal{N}_S$ is an analogue of the set $\mathcal{N}_K$ from (3.7). By the induction assumption for dimension $d - 1$ we find $p_{|\partial K} \equiv 0$.

Next, we appeal to Lemma 3.52 to conclude that

$$p = \lambda_1 \cdot \dots \cdot \lambda_{d+1}\, q \quad \text{for some } q \in \mathcal{P}_{m-d-1}(K)\ ,$$

where

$$q(\boldsymbol{\xi}) = 0 \quad \text{for} \quad \boldsymbol{\xi} \in \mathcal{N}' := \{\boldsymbol{\xi} \in \overline{K} :\ \lambda_j(\boldsymbol{\xi}) \in \{\tfrac{1}{m}, \dots, \tfrac{m-1}{m}\},\ j = 1, \dots, d+1\}\ .$$

For $m \leq d$ the set $\mathcal{N}'$ is empty, but in this case q has to be trivial. For $m = d+1$ the set $\mathcal{N}'$ reduces to a single point (the barycenter of K) and q will be constant. For $m > d+1$ the convex hull of $\mathcal{N}'$ will be another non-degenerate d-simplex, for which $\mathcal{N}'$ provides the set of evaluation points according to (3.7) for a Lagrangian finite element of degree $m - d - 1$. This matches the degree of q and the induction hypothesis bears out $q \equiv 0$.

In this case unisolvence of the point evaluations already guarantees the requirement (iv) of Def. 3.25. Repeating the above arguments and appealing to our initial observation,

we see that a trial function will vanish on a face S of the simplex if it evaluates to zero in all points $\mathcal{N}_K \cap \overline{S}$. □

Using the barycentric coordinate functions $\lambda_1, \ldots, \lambda_{d+1}$ on a simplex K the local shape functions for Lagrangian finite elements of degree m can be expressed conveniently, see Table 3.2 for triangles.

	Shape functions supported on		
	vertices $\boldsymbol{\nu}_i$	edges $[\boldsymbol{\nu}_i, \boldsymbol{\nu}_j]$	triangle
$m = 1$	λ_i $i = 1, 2, 3$	—	—
$m = 2$	$-\lambda_i(1 - 2\lambda_i)$ $i = 1, 2, 3$	$4\lambda_i\lambda_j$, $1 \le i < j \le 3$	—
$m = 3$	$1/2\lambda_i(1 - 3\lambda_i)(2 - 3\lambda_i)$, $i = 1, 2, 3$	$-9/2\lambda_i\lambda_j(1 - 3\lambda_i)$, $-9/2\lambda_i\lambda_j(1 - 3\lambda_j)$, $1 \le i < j \le 3$	$9\lambda_1\lambda_2\lambda_3$ "Bubble function"

Table 3.2: Local shape functions for triangular Lagrangian finite elements of degree m

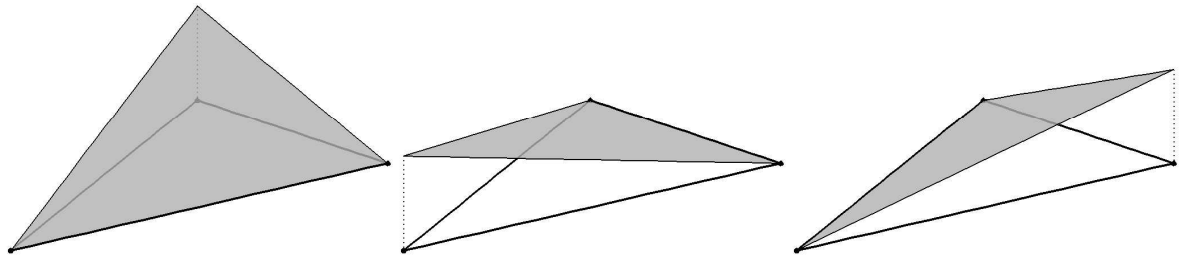


Figure 3.13: Graphical representation of shape functions for triangular Lagrangian finite element of degree 1

There are variants of Lagrangian finite elements for cells of tensor product geometry (rectangles and bricks).

Definition 3.56. *For $K =]0,1[^d$ the Lagrangian finite element (K, Π_K, Σ_K) of degree $m \in \mathbb{N}$ is is defined by*

(i) $\Pi_K := \mathcal{Q}_m(K)$.

(ii) $\Sigma_K := \{v \mapsto v(\boldsymbol{\xi}),\ \boldsymbol{\xi} \in \mathcal{N}_K\}$, *where*

$$\mathcal{N}_K := \{\boldsymbol{\alpha}/m,\ \boldsymbol{\alpha} \in \{0, \ldots, m\}^d\}\ .$$

Theorem 3.57. *The triple (K, Π_K, Σ_K) from Def. 3.56 is a finite element.*

	Shape functions supported on		
	vertices $\boldsymbol{\nu}_i$	edges $[\boldsymbol{\nu}_i, \boldsymbol{\nu}_j]$	triangle
$m = 1$	$\lambda_i \mu_j$ $i, j = 1, 2$	—	—
$m = 2$	$\lambda_i(1 - 2\lambda_i)\mu_j(1 - 2\mu_j)$ $i, j = 1, 2$	$-4\lambda_i(1 - 2\lambda_i)\mu_1\mu_2$, $-4\mu_i(1 - 2\mu_i)\lambda_1\lambda_2$, $i = 1, 2$	$16\lambda_1\lambda_2\mu_1\mu_2$
$m = 3$	$1/4\lambda_i(1 - 3\lambda_i)(2 - 3\lambda_i)$ $\mu_j(1 - 3\mu_j)(2 - 3\mu_j)$ $i, j = 1, 2$	$9/4\mu_1\mu_2\lambda_i(1 - 3\lambda_i)$ $(2 - 3\lambda_i)(1 - 3\mu_j)$, $9/4\lambda_1\lambda_2\mu_i(1 - 3\mu_i)$ $(2 - 3\mu_i)(1 - 3\lambda_j)$, $i, j = 1, 2$	$-81/4\lambda_1\lambda_2\mu_1\mu_2$ $(1 - 3\lambda_i)(1 - 3\mu_j)$, $i, j = 1, 2$

Table 3.3: Shape functions for Lagrangian finite elements on unit square $\{\boldsymbol{\xi} \in \mathbb{R}^2 : 0 < \xi_i < 1\}$. Here, we have used the abbreviations $\lambda_1 = \xi_1$, $\lambda_2 = 1 - \xi_1$, $\mu_1 = \xi_2$, $\mu_2 = 1 - \xi_2$.

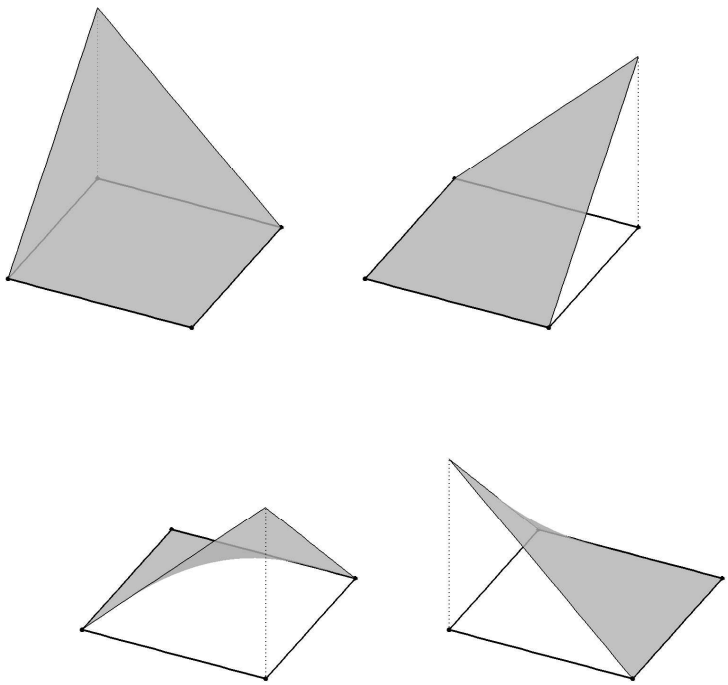


Figure 3.14: Sketches of shape functions for Lagrangian finite element of degree 1 on the unit square

Proof. Exactly the same induction thechnique as in the proof of Thm. 3.55 can be used, much facilitated by the tensor product structure. □

Remark 3.58. The principle behind the construction of Lagrangian finite elements on rectangles/bricks is a **tensor product approach**. If the cell K arises by forming the tensor products of two lower-dimensional "cells" K_α, K_β, that is

$$K := \{\boldsymbol{\xi} = (\boldsymbol{\xi}_\alpha, \boldsymbol{\xi}_\beta) \in \mathbb{R}^d : \ \boldsymbol{\xi}_\alpha \in K_\alpha \subset \mathbb{R}^{d_\alpha}, \boldsymbol{\xi}_\beta \in K_\beta \subset \mathbb{R}^{d_\beta}, d_\alpha + d_\beta = d\} ,$$

and $(K_\alpha, \Pi_\alpha, \Sigma_\alpha)$, $(K_\alpha, \Pi_\alpha, \Sigma_\alpha)$ are Lagrangian finite elements on K_α, K_β, whose sets of local degrees of freedom comprise point evaluations in $\mathcal{N}_\alpha \subset \mathbb{R}^{d_\alpha}$, $\mathcal{N}_\beta \subset \mathbb{R}^{d_\beta}$, respectively, then

$$\begin{aligned}\Pi_K :=& \{v(\boldsymbol{\xi}) = v_\alpha(\boldsymbol{\xi}_\alpha) \cdot v(\boldsymbol{\xi}_\beta) : \ v_\alpha \in \Pi_\alpha, \ v_\beta \in \Pi_\beta\} ,\\ \Sigma_K :=& \{v \mapsto v((\boldsymbol{\eta}_\alpha, \boldsymbol{\eta}_\beta)) : \ \boldsymbol{\eta}_\alpha \in \mathcal{N}_\alpha, \ \boldsymbol{\eta}_\beta \in \mathcal{N}_\beta\}\end{aligned}$$

supplies a Lagrangian finite element for K.

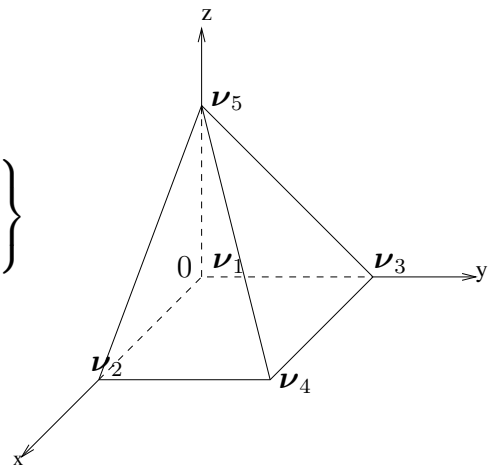
Exercise 3.13. Describe the Lagrangian finite element of degree 1 on a prism

$$K := \{\boldsymbol{\xi} \in \mathbb{R}^3 : \ 0 < \xi_1, \xi_2, \xi_3 < 1, \ \xi_1 + \xi_2 < 1\} .$$

Remark 3.59.
Pyramids are important for creating conforming triangulations that can contain both tetrahedra and hexahedra. For the pyramid

$$K := \text{convex}\left\{ \begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}0\\1\\0\end{pmatrix}, \begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}1\\1\\0\end{pmatrix}, \begin{pmatrix}0\\0\\1\end{pmatrix} \right\}$$

there exists no local trial space $\Pi_K \subset C^\infty(\overline{K})$ such that $\Pi_{K|F} \in \mathcal{P}_1(F)$ for every triangular face F of K and $\Pi_{K|Q} \in \mathcal{Q}_1(Q)$ for the quadrilateral face.



Assume that there is such a trial space and let $v \in \Pi_K$ assume the value 1 in $(0,0,0)^T$ and zero in all the other vertices. Then restricted to the faces we have

(a)	$v(x,y,z) = (1-x)(1-y)$	in convex$\{(0,0,0),(1,0,0),(1,1,0),(0,1,0)\}$
(b)	$v(x,y,z) = (1-x-z)$	in convex$\{(0,0,0),(1,0,0),(0,0,1)\}$
(c)	$v(x,y,z) = (1-y-z)$	in convex$\{(0,0,0),(0,1,0),(0,0,1)\}$
(d)	$v(x,y,z) = 0$	in convex$\{(1,0,0),(1,1,0),(0,0,1)\}$
(e)	$v(x,y,z) = 0$	in convex$\{(1,1,0),(0,1,0),(0,0,1)\}$

Using (d) and (e), we get

$$\mathbf{grad}\, v(0,0,1) \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix} = \mathbf{grad}\, v(0,0,1) \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} = \mathbf{grad}\, v(0,0,1) \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix} = 0.$$

If $\mathbf{grad}\, v$ is continuous at $(0,0,1)$, this implies $\mathbf{grad}\, v(0,0,1) = 0$, but (b) and (c) give

$$\mathbf{grad}\, v(0,0,1) \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = -1.$$

This is a contradiction.

However, when we relax the requirement $\Pi_K \subset C^1(\overline{K})$ suitable local trial spaces can be found. The result is known as a **singular finite element** characterized by a lack of smoothness of the local trial space. For instance, we can use the space spanned by the following functions

$$\begin{aligned} b_1 &= \frac{(1-\xi_3-\xi_1)(1-\xi_3-\xi_2)}{1-\xi_3}\,, & b_2 &= \frac{\xi_1(1-\xi_3-\xi_2)}{1-\xi_3}\,, \\ b_3 &= \frac{(1-\xi_3-\xi_1)\xi_2}{1-\xi_3}\,, & b_4 &= \frac{\xi_1\xi_2}{1-\xi_3}\,, \\ b_5 &= \xi_3\,. \end{aligned}$$

These functions satisfy $b_i(\boldsymbol{\nu}_j) = \delta_{ij}$, $i,j = 1,\ldots,5$, which renders them the shape functions belonging to the point evaluations at vertices of K a set of degrees of freedom. In particular these shape functions will match Lagrangian finite elements in adjacent tetrahedra and hexahedra.

Summing up, for all kinds of Lagrangian finite elements on a conforming triangulation the local degrees of freedom supported on vertices, edges (and faces) will be matching by construction. Hence,

> for Lagrangian H^1-conforming finite element spaces the resulting global degrees of freedom will be point evaluations at vertices, on edges/faces, and in the interior of cells.

Example 3.60. On a 2D simplicial triangulation the global degrees of freedom for Lagrangian finite elements of uniform degree 2 are given by point evaluations in vertices and midpoints of edges, *cf.* Fig. 3.11.

Example 3.61. Consider a conforming tetrahedral mesh in three dimensions and Lagrangian finite elements of uniform degree 3 on it. The global degrees of freedom boil down to point evaluations, one in vertices, two for each edge and one in the barycenter of each face, *cf.* Fig. 3.12.

Assuming matching local d.o.f. Def. 3.29 can be used to introduce the $H^1(\Omega)$-conforming **Lagrangian finite element spaces** of uniform degree m, $m \in \mathbb{N}$ on simplicial/tensor product triangulations. They will agree with the spaces, *cf.* Remark 3.31,

$$\mathcal{S}_m(\mathcal{M}) := \begin{cases} \{v \in H^1(\Omega) : \ v_{|K} \in \mathcal{P}_m(K)\} & \text{if } K \text{ is a simplex,} \\ \{v \in H^1(\Omega) : \ v_{|K} \in \mathcal{Q}_{(m,\ldots,m)}(K)\} & \text{if } K \text{ is a tensor product cell} \end{cases}$$

In light of Remarks 3.58, 3.59 it is clear how introduce $\mathcal{S}_m(\mathcal{M})$, if prism and pyramids occur in a three-dimensional triangulation.

Exercise 3.14. Let $\mathcal{M}$ be a conforming simplicial triangulation of a two-dimensional polygonal domain. For each $K \in \mathcal{M}$ with vertices $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \boldsymbol{\nu}_3$ and center of gravity $\boldsymbol{\gamma} := 1/3(\boldsymbol{\nu}_1 + \boldsymbol{\nu}_2 + \boldsymbol{\nu}_3)$ we choose the following finite-dimensional local trial space

$$\Pi_K \ := \ \mathcal{P}_3(K)$$

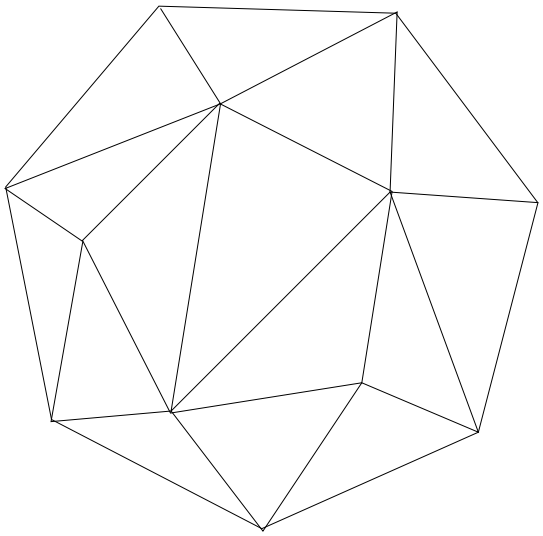
and the following sets of local degrees of freedom

$$\begin{aligned} \Sigma_K \ := \ & \{v \mapsto v(\boldsymbol{\nu}_i),\ i = 1,2,3, \\ & v \mapsto \langle \mathbf{grad}\, v(\boldsymbol{\nu}_i), \boldsymbol{\nu}_j - \boldsymbol{\nu}_i \rangle\ ,\ i = 1,2,3,\ j \in \{1,2,3\} \setminus \{i\}, \\ & v \mapsto v(\boldsymbol{\gamma})\}, \\ \Sigma'_K \ := \ & \{v \mapsto v(\boldsymbol{\nu}_i),\ i = 1,2,3, \\ & v \mapsto \frac{\partial v}{\partial \xi_i}(\boldsymbol{\nu}_j),\ i = 1,2,\ j = 1,2,3, \\ & v \mapsto v(\boldsymbol{\gamma})\} \end{aligned}$$

(i) Show that (K, Π_K, Σ_K), $K \in \mathcal{M}$, is a H^1-conforming finite element and that all of them are affine equivalent. This family of finite elements is called the **cubic triangular Hermite element**.

(ii) Show that (K, Π_K, Σ'_K), $K \in \mathcal{M}$, is a H^1-conforming finite element but fails to be affine equivalent.

(iii) Describe global degrees of freedom arising from the choice of Σ_K or Σ'_K.



(iv) For the triangulation sketched to the right compute the dimensions of the cubic Lagrangian finite element space $\mathcal{S}_3(\mathcal{M})$, of the space obtained from cubic triangular Hermite elements, and of the finite element space corresponding to subtask (ii).

Bibliographical notes. A comprehensive discussion of Lagrangian (and Hermitian) finite elements in given in [12, Ch, 2].

3.8.2 Whitney finite elements

In Sect. 2.2 we distinguished between different types of quantities and presented the associated transformations. In the previous section we merely discussed finite elements for potential type quantities. Now we aim to introduce suitable finite elements for the other types. We emphasize that all the finite elements introduced in this section can serve as the basis for parametric construction *using the right transformation*, see Sect. 2.2.

We first consider flux type quantities. To motivate the construction, we recall that the fundamental evaluation for these quantities consists of calculating the total flux through an oriented surface. Hence, the construction of finite element spaces has to rely on oriented meshes $\mathcal{M}$, see Def. 3.7. In particular, every face F will be endowed with a unit normal $\mathbf{n}_F$ pointing into the "crossing direction" for F. This has to be distiguished from the exterior unit normal vectorfield $\mathbf{n}_{\partial K}$ on the surface of a cell $K \in \mathcal{M}$: in general $\mathbf{n}_{\partial K|F} = \pm \mathbf{n}_F$.

Remark 3.62. One has to be careful when using a parametric construction for finite elements, whose degrees of freedom rely on orientation. It this case it might happen that even for a conforming simplicial triangulation more than one reference simplex is needed: the reference simplices differ in the orientation of their faces/edges. The three oriented reference triangles for $d = 2$ are depicted in Fig. 3.15.

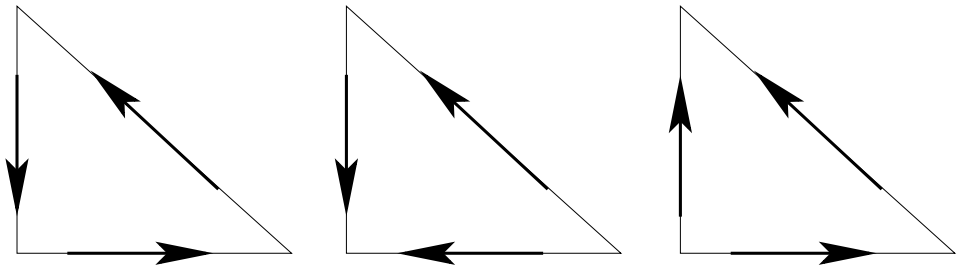


Figure 3.15: Three reference triangles with different edge orientations

To begin with, we discuss the case of an oriented conforming simplicial triangulation $\mathcal{M}$ of the computational domain $\Omega \subset \mathbb{R}^d$.

Definition 3.63. *Let $K \subset \mathbb{R}^d$ be a non-degenerate simplex $\in \mathcal{M}$. Then we define the* ***face element*** (K, Π_K, Σ_K) *by*

(i) $\Pi_K := \{\boldsymbol{\xi} \in K \mapsto \boldsymbol{\alpha} + \beta\, \boldsymbol{\xi},\ \boldsymbol{\alpha} \in \mathbb{R}^d,\ \beta \in \mathbb{R}\}$,

(ii) $\Sigma_K := \left\{ \mathbf{v} \mapsto \int_F \langle \mathbf{v}, \mathbf{n}_F \rangle \, \mathrm{d}S,\ F \text{ face of } K \right\}$

Theorem 3.64. *The face element constitutes a $H(\mathrm{div})$-conforming finite element.*

Proof. Clearly, $\Pi_K \subset (C^\infty(\overline{K}))^d$ and $\dim \Pi_K = \sharp\Sigma_K = d+1$. Following Remark 3.21, for the unisolvence of Σ_K it remains to show that

$$\mathbf{v} \in \Pi_K \quad \text{and} \quad l(\mathbf{v}) = 0 \quad \forall l \in \Sigma \quad \Rightarrow \quad \mathbf{v} = 0 \,.$$

First, we note that the normal components of functions in Π_K on any face F of K are constant, because for a fixed $\boldsymbol{\xi} \in F$

$$\langle \boldsymbol{\alpha} + \beta(\boldsymbol{\xi} + \boldsymbol{\tau}), \mathbf{n}_F \rangle = \langle \boldsymbol{\alpha} + \beta\boldsymbol{\xi}, \mathbf{n}_F \rangle$$

for all vectors $\boldsymbol{\tau} \in \mathbb{R}^d$ tangential (complanar) to F. Thus vanishing d.o.f. for $\mathbf{v} \in \Pi_K$ mean that $\langle \mathbf{v}, \mathbf{n}_F \rangle \equiv 0$ for every face of K. This implies that $\mathbf{v}$ is constant:

$$0 = \int_{\partial K} \langle \mathbf{v}, \mathbf{n} \rangle \, \mathrm{d}S = \int_K \operatorname{div} \mathbf{v} \, \mathrm{d}\boldsymbol{\xi} = d|K|\,\beta \quad \Rightarrow \quad \beta = 0 \,.$$

Finally, a constant that has vanishing inner product with all $d+1$ face normals must be zero. These considerations also show that the d.o.f. on a face F already fixes the normal component $\langle \mathbf{v}, \mathbf{n}_F \rangle$ for a $\mathbf{v} \in \Sigma$. Since this is the natural trace for $H(\mathrm{div})$, the element turns out to be $H(\mathrm{div})$-conforming. □

It goes without saying that the local d.o.f. of these $H(\mathrm{div})$-conforming finite elements match at interelement faces. Traces onto edges and vertices do not make any sense here. Hence, the conversion of local d.o.f. into global d.o.f. does not encounter any difficulties.

Lemma 3.65. *The finite elements defined in Thm. 3.64 are affine equivalent after a suitably flipping of the orientations of some faces.*

Proof. In the case of $H(\mathrm{div})$-conforming finite elements we have to use the Piola transform $\mathsf{PT}_{\boldsymbol{\Phi}}$, see (PT), where $\boldsymbol{\Phi}$ is an affine mapping of the form (AFF). Then $D\boldsymbol{\Phi} = \mathbf{F}$ and $\det D\boldsymbol{\Phi} \equiv \text{const.}$. One easily computes

$$\mathsf{PT}_{\boldsymbol{\Phi}}(\boldsymbol{\xi} \mapsto \boldsymbol{\alpha} + \beta\boldsymbol{\xi})(\widetilde{\boldsymbol{\xi}}) = |\det D\boldsymbol{\Phi}|\, \mathbf{F}^{-1}(\boldsymbol{\alpha} + \beta(\mathbf{F}\widetilde{\boldsymbol{\xi}} + \boldsymbol{\tau})) \,.$$

Hence, the Piola transformation will leave the local trial spaces invariant, *ie.* the requirement (3.4) of Def. 3.44 is fulfilled. The relationship (3.5) follows from Lemma 2.8. □

Lemma 3.66. *Let $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_{d+1}$ the vertices of a non-degenerate simplex $K \in \mathcal{M}$. Then the local shape functions $\{\boldsymbol{b}_F\}_F$, F face of K, for the $H(\mathrm{div})$-conforming finite element from Thm. 3.64 are given by*

$$\boldsymbol{b}_F(\boldsymbol{\xi}) := \pm \frac{1}{d\,|K|} \, (\boldsymbol{\xi} - \boldsymbol{\nu}_F) \,,$$

where $\boldsymbol{\nu}_F$ is the unique vertex of K that is not located on $\overline{F}$, and $|F|, |K|$ stand for the volume of F and K, respectively. The $+$-sign applies, if $\mathbf{n}_F$ points into the exterior of K, otherwise the $-$-sign has to be used.

Proof. It is clear that for any other face $F' \neq F$ we have $\langle \boldsymbol{b}_F(\boldsymbol{\xi}), \mathbf{n}_{F'} \rangle = 0$ for all $\boldsymbol{\xi} \in F'$. Hence, $\boldsymbol{b}_F$ will be in the kernel of the d.o.f. that is associated with F'.

Moreoever, the Hessian normal form of a hyperplane in $\mathbb{R}^d$ tells us that

$$\operatorname{dist}(\boldsymbol{\nu}_F, F) = \langle \boldsymbol{\xi} - \boldsymbol{\nu}, \mathbf{n}_F \rangle \ .$$

Using the formula

$$|K| = \operatorname{dist}(\boldsymbol{\nu}_F, F) \cdot |F|/d$$

we conclude that the local d.o.f. belonging to F will yield ± 1 when applied to $\boldsymbol{b}_F$. □

For $\boldsymbol{\eta} \in \mathbb{R}^2$ we can define the counterclockweise rotation by $\pi/2$

$$\begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix}^{\perp} = \begin{pmatrix} -\eta_2 \\ \eta_1 \end{pmatrix} .$$

Then, writing $\lambda_1, \ldots, \lambda_{d+1}$ for the barycentric coordinate functions associated with the vertices of the simplex K, we find an alternative representation for the local shape functions from Lemma 3.66 and $d = 2$

$$\boldsymbol{b}_F = \pm \ (\lambda_i \operatorname{\mathbf{grad}} \lambda_j - \lambda_j \operatorname{\mathbf{grad}} \lambda_i)^{\perp} \ , \tag{3.8}$$

where $\boldsymbol{\nu}_i$ and $\boldsymbol{\nu}_j$ are the endpoints of the edge F. For $d = 3$ a similar formula reads

$$\boldsymbol{b}_F = \pm \ (\lambda_i \operatorname{\mathbf{grad}} \lambda_j \times \operatorname{\mathbf{grad}} \lambda_k + \lambda_k \operatorname{\mathbf{grad}} \lambda_i \times \operatorname{\mathbf{grad}} \lambda_j + \lambda_j \operatorname{\mathbf{grad}} \lambda_k \times \operatorname{\mathbf{grad}} \lambda_i) \ , \tag{3.9}$$

when $F = \operatorname{convex}\{\boldsymbol{\nu}_i, \boldsymbol{\nu}_j, \boldsymbol{\nu}_k\}$.

Exercise 3.15. Confirm the formulas (3.8) and (3.9).

The local shape functions in two dimensions are plotted in Figure 3.16.

Second, we examine the case of square/brick-shaped cells. Thanks to the tool of parametric construction, we merely need to look at $K =]0;1[^d$ (However, remember Remark 3.62).

Definition 3.67. *For $K =]0,1[^d$ the **face element** (K, Π_K, Σ_K) is defined by*

(i) $\Pi_K := \mathcal{Q}_{\boldsymbol{\epsilon}_1}(K) \times \cdots \times \mathcal{Q}_{\boldsymbol{\epsilon}_d}(K)$, $\boldsymbol{\epsilon}_j$ *= j-th unit vector,*

(ii) $\Sigma_K := \left\{ \mathbf{v} \mapsto \int_F \langle \mathbf{v}, \mathbf{n}_F \rangle \, \mathrm{d}S, \ F \text{ face of } K \right\}$

Theorem 3.68. *The face element on $K =]0,1[^d$ provides a $H(\operatorname{div})$-conforming finite element.*

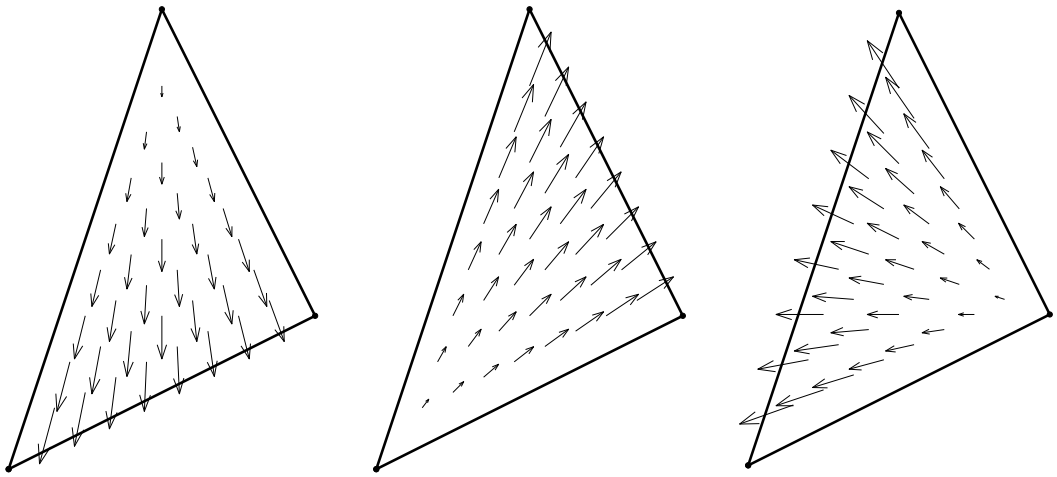


Figure 3.16: Local shape functions for lowest degree $H(\mathrm{div})$-conforming triangular finite elements. The crossing directions of the faces are outward w.r.t. K, throughout.

Proof. Since $\dim \Pi_K = \sharp\Sigma_K$, we only have to establish unisolvence of the local degreees of freedom. Assume that $\mathbf{v} \in \Pi_K$ belongs to the kernels of all local d.o.f. Its first component reads $v_1 = \alpha\xi_1 + \beta$, $\alpha, \beta \in \mathbb{R}$. Consider the faces $F := \{0\}\times]0,1[^{d-1}$ and $F := \{1\}\times]0,1[^{d-1}$ with $\mathbf{n}_F = (1,0,\ldots,0)^T$. Obviously, $\langle \mathbf{v}, \mathbf{n}_F\rangle_{|F}$ is constant and zero integral over F will force it to be zero, which means $\beta = 0$ and $\alpha + \beta = 0$: the first component of $\mathbf{v}$ must vanish. The same argument can be applied in all other coordinate directions. $H(\mathrm{div})$-Conformity of the finite element is seen as in the proof of 3.64. □

Figure 3.17 displays the four local shape functions for the $H(\mathrm{div})$-conforming element presented in Def. 3.67 in two dimensions.

Next, we examine $H(\mathbf{curl})$-conforming finite elements for gradient type quantities in $d = 3$. The natural restriction is the tangential components trace, see Exercise 2.23. We recall that the typical evaluations for gradient type quantities are integrals along directed paths. This provides a hint about how to choose degrees of freedom. We also conclude that we will need an orientation of the edges of the underlying triangulation.

Initially, we focus on an oriented conforming tetrahedral triangulation $\mathcal{M}$.

Definition 3.69. *For a tetrahedron $K \in \mathcal{M}$ with oriented edges we define the* ***edge element*** (K, Π_K, Σ_K) *by*

(i) $\Pi_K := \{\boldsymbol{\xi} \in K \mapsto \boldsymbol{\alpha} + \boldsymbol{\beta} \times \boldsymbol{\xi},\ \boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^3\}$,

(ii) $\Sigma_K := \left\{ \mathbf{v} \mapsto \int_E \mathbf{v} \cdot \mathrm{d}\boldsymbol{s},\ E \text{ edge of } K \right\}$.

Theorem 3.70. *The edge element from Def. 3.69 is a $H(\mathbf{curl})$-conforming finite element.*

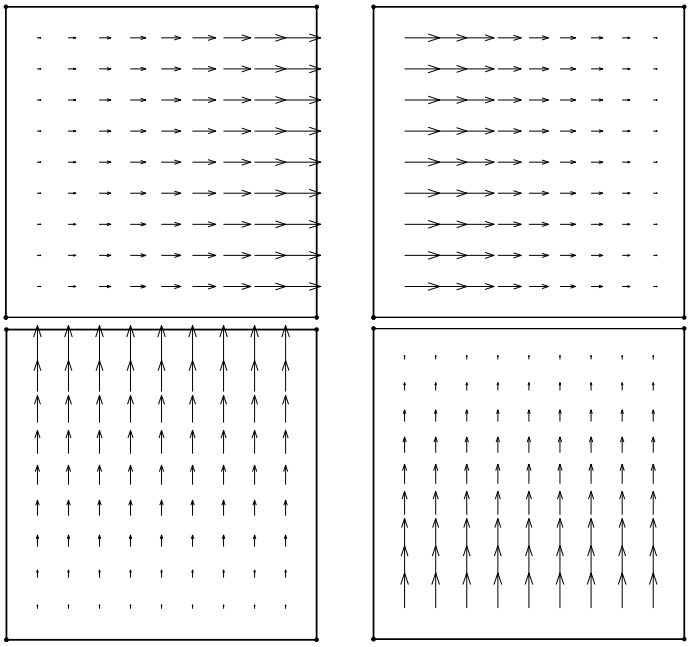


Figure 3.17: Local shape functions for lowest degree $H(\mathrm{div})$-conforming square finite element. Faces bear orientation of ∂K.

Proof. We find that the tangential component of $\mathbf{v} \in \Pi_K$ on a straight line $\boldsymbol{\xi} = \boldsymbol{\zeta} + \rho\,\boldsymbol{\tau}$, $\boldsymbol{\zeta}, \boldsymbol{\tau} \in \mathbb{R}^3$, $\rho \in \mathbb{R}$, is constant:

$$\langle \boldsymbol{\alpha} + \boldsymbol{\beta} \times (\boldsymbol{\zeta} + \rho\,\boldsymbol{\tau}), \boldsymbol{\tau} \rangle = \langle \boldsymbol{\alpha} + \boldsymbol{\beta} \times \boldsymbol{\zeta}, \boldsymbol{\tau} \rangle \quad \forall \rho \in \mathbb{R}\,.$$

Thus, for $\mathbf{v} \in \Pi_K$ vanishing local d.o.f. immediately imply vanishing tangential components along all edges of K. Necessarily, all components of $\mathbf{v}$ will be zero in the vertices of K. Since they are affine linear, this can only be the case, if they are zero everywhere.

Pick a face F of K. If the local d.o.f. associated with edges of F vanish, then the components of $\mathbf{v}$ tangential to F will be zero in all vertices of F. This means that $\mathbf{v}$ will have zero tangential components everywhere on F and $H(\mathbf{curl})$-conformity is established. □

Lemma 3.71. *The finite elements defined in Def. 3.69 are affine equivalent, possibly after flipping the directions of some edges of K.*

Proof. The transformation fitting $H(\mathbf{curl})$ is (GT). If $\boldsymbol{\Phi}(\widetilde{\boldsymbol{\xi}}) = \mathbf{F}\widetilde{\boldsymbol{\xi}} + \boldsymbol{\tau}$ according to (AFF), we find

$$(\mathsf{GT}_{\boldsymbol{\Phi}}\,\mathbf{u})(\widetilde{\boldsymbol{\xi}}) = \mathbf{F}^T(\mathbf{u}(\mathbf{F}\widetilde{\boldsymbol{\xi}} + \boldsymbol{\tau}))\,,$$

which means

$$\mathsf{GT}_{\boldsymbol{\Phi}}(\boldsymbol{\xi} \mapsto \boldsymbol{\alpha} + \boldsymbol{\beta} \times \boldsymbol{\xi})(\widetilde{\boldsymbol{\xi}}) = \mathbf{F}^T(\boldsymbol{\alpha} + \boldsymbol{\beta} \times (\mathbf{F}\widetilde{\boldsymbol{\xi}} + \boldsymbol{\tau}))\,.$$

Note that in the expression

$$\mathbf{F}^T\big(\boldsymbol{\beta} \times \mathbf{F}\widetilde{\boldsymbol{\xi}}\big) = \left\{ \mathbf{F}^T \begin{pmatrix} 0 & -\beta_3 & \beta_2 \\ \beta_3 & 0 & -\beta_1 \\ -\beta_2 & \beta_1 & 0 \end{pmatrix} \mathbf{F} \right\} \widetilde{\boldsymbol{\xi}}$$

the matrix in braces is skew-symmetric again. Thus, its multiplication onto $\widetilde{\boldsymbol{\xi}}$ is equivalent to a vector product with a constant vector. This shows invariance of the local spaces, see (3.4). The invariance of the local d.o.f., *cf.* (3.5) is a consequence of Lemma 2.7. □

Remark 3.72. Assuming matching orientations of edges, it is clear that the edge based degrees of freedom from Thm. 3.70 will match. The set of global degrees of freedom will comprise all path integrals along edges of $\mathcal{M}$.

The local shape functions corresponding to the $H(\mathbf{curl})$-conforming finite element from Thm. 3.70 have a convenient representation in terms of barycentric coordinate functions: if we write $\boldsymbol{b}_{i,j}$ for the local shape functions associated with the edge connecting vertex i and vertex j of K, then

$$\boldsymbol{b}_{i,j} = \pm \left(\lambda_i \, \mathbf{grad}\, \lambda_j - \lambda_j \, \mathbf{grad}\, \lambda_i\right) . \tag{3.10}$$

Exercise 3.16. Verify the formula (3.10).

On a cube $K =]0,1[^3$, which can serve as reference cell for a conforming hexahedral triangulation, a similar $H(\mathbf{curl})$-conforming finite element is available.

Definition 3.73. *For $K =]0,1[^3$ the* ***edge element*** (K, Π_K, Σ_K)*, $K =]0,1[^3$ is given by*

(i) $\Pi_K = \mathcal{Q}_{(0,1,1)}(K) \times \mathcal{Q}_{(1,0,1)}(K) \times \mathcal{Q}_{(1,1,0)}(K)$,

(ii) $\Sigma := \left\{ \mathbf{v} \mapsto \int_E \mathbf{v} \cdot \mathrm{d}\boldsymbol{s},\ E \text{ edge of } K \right\}$.

Proof. A counting argument confirms $\dim \Pi_K = \sharp\Sigma_K = 6$. Further, by the very definition of Π_K tangential components of local trial functions along edges of K are constant. Hence, for $\mathbf{v} \in \Pi_K$ vanishing local degrees of freedom will involve $\mathbf{v} = 0$ for all vertices of K. The bilinear components of $\mathbf{v}$ must vanish everywhere. We conclude as in the proof of Thm. 3.70. □

Theorem 3.74. *The edge element on $K =]0,1[^3$ is a $H(\mathbf{curl})$-conforming finite element.*

Remark 3.75. Edge and face elements introduced in Definitions 3.69, 3.73, 3.63, and 3.67 owe their names to the location of their degrees of freedom.

Notation: For the global face element finite element space on a mesh $\mathcal{M}$ of $\Omega \subset \mathbb{R}^d$ we adopt the notation $\mathcal{W}_F(\mathcal{M})$. The global space arising from edge elements built on a conforming triangulation $\mathcal{M}$ of $\Omega \subset \mathbb{R}^3$ is denoted by $\mathcal{W}_E(\mathcal{M})$. By conformity we have

$$\mathcal{W}_F(\mathcal{M}) \subset H(\mathrm{div};\Omega) \quad , \quad \mathcal{W}_E(\mathcal{M}) \subset H(\mathbf{curl};\Omega) \, .$$

Remark 3.76. Variants of face and edge finite elmenents of higher polynomial degrees are available, but rather technical, see [31] and [25].

Remark 3.77. $H(\mathrm{div})$-conforming finite elements have to be used for the Galerkin discretization of the variational problems (VPD) and (2.9) arising from second-order elliptic boundary value problems. $H(\mathbf{curl})$-conforming finite elements play a key role in computational electromagnetism, *cf.* Remarks 2.15, 2.16.

We have not yet looked at quantities of density type with related Sobolev space $L^2(\Omega)$. In this case there is *no continuity condition* and the construction of finite elements becomes simple. We can skip the formal developments and jump right to the definition of the space

$$\mathcal{Q}_0(\mathcal{M}) := \{v \in L^2(\Omega) : \ v_{|K} \equiv \text{const.} \ \forall K \in \mathcal{M}\}$$

of piecewise constants on a mesh $\mathcal{M}$ of a computational domain Ω. This space can easily be put into a finite element framework: the local = global degrees of freedom are given by

$$\mathrm{gdof}(\mathcal{Q}_0(\mathcal{M})) := \left\{ v \mapsto \int_K v \, \mathrm{d}\boldsymbol{\xi}, \ K \in \mathcal{M} \right\} \, .$$

We remark that the related finite element interpolation operator $\mathsf{I}(\mathcal{Q}_0(\mathcal{M}))$, see Sect. 3.6, coincides with the standard $L^2(\Omega)$-orthogonal projection onto $\mathcal{Q}_0(\mathcal{M})$.

Now we study special properties of face and edge element spaces and, in particular, of their related finite element interpolation operators, see Sect. 3.6. To that end we fix a conforming triangulation $\mathcal{M}$ of a computational domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, for which a single d-simplex and/or d-dimensional cube can serve as reference cells (apart from orientation, *cf.* Remark 3.15). On this mesh we consider the space $\mathcal{W}_F(\mathcal{M})$ of face elements, and, for $d = 3$, the space $\mathcal{W}_E(\mathcal{M})$ of edge elements introduced above. For the sake of brevity, we write I_F and I_E for the finite element interpolation operators $\mathsf{I}(\mathcal{W}_F(\mathcal{M}))$ and $\mathsf{I}(\mathcal{W}_E(\mathcal{M}))$, respectively, see Def. 3.39. Further, let $\mathsf{Q}_0 : L^2(\Omega) \mapsto \mathcal{Q}_0(\mathcal{M})$ stand for the $L^2(\Omega)$-orthogonal projection.

Theorem 3.78. *For any simplicial conforming triangulation holds true*

$$\mathrm{div} \circ \mathsf{I}_F = \mathsf{Q}_0 \circ \mathrm{div} \quad \textit{on } (C^\infty(\overline{\Omega}))^d \, .$$

Proof. First, we have to ensure that $\operatorname{div} \mathcal{W}_F(\mathcal{M}) \subset \mathcal{Q}_0(\mathcal{M})$. This is clear from Def. 3.63, because, obviously, $\operatorname{div} \Pi_K = \mathcal{P}_0(K)$ for any $K \in \mathcal{M}$.

Next, appealing to the definition of the orthogonal projection Q_0, we have to show

$$\int_K \operatorname{div}(\mathbf{v} - \mathsf{I}_F \mathbf{v}) \, \mathrm{d}\boldsymbol{\xi} = 0 \quad \forall K \in \mathcal{M},\ \mathbf{v} \in (C^\infty(\overline{\Omega}))^d \, .$$

Recalling the definition of the local/global d.o.f. for $\mathcal{W}_F(\mathcal{M})$, see Thm. 3.64 and Thm. 3.68, this is straightforward from Gauss theorem Thm. 2.17. □

Theorem 3.79. *In the case of a tetrahedral mesh we have*

$$\mathbf{curl} \circ \mathsf{I}_E = \mathsf{I}_F \circ \mathbf{curl} \quad \textit{on} \quad (C^\infty(\overline{\Omega}))^3 \, .$$

Proof. Pick a tetrahedron $K \in \mathcal{M}$ and write Π_K^F / Π_K^E for the local trial spaces from Def. 3.63 and Def. 3.69, respectively. Then elementary vector analysis shows $\mathbf{curl}\, \Pi_K^E \subset \Pi_K^F$. Moreover, from $\operatorname{div} \circ \mathbf{curl} = 0$ we learn that $\mathbf{curl}\, \mathcal{W}_E(\mathcal{M}) \subset H(\operatorname{div}; \Omega)$, which implies normal continuity. Necessarily, then $\mathbf{curl}\, \mathcal{W}_E(\mathcal{M}) \subset \mathcal{W}_F(\mathcal{M})$.

By definition of the various global degrees of freedom for the finite element spaces we still have to show

$$\int_F \mathbf{curl}(\mathbf{v} - \mathsf{I}_E \mathbf{v}) \cdot \mathbf{n}_F \, \mathrm{d}S = 0 \quad \forall F \in \mathcal{F}(\mathcal{M}),\ \mathbf{v} \in (C^\infty(\overline{\Omega}))^3 \, .$$

Thanks to the definition of the d.o.f. for edge elements this can be instantly deduced from Stokes' theorem applied to the faces of the mesh. □

Remark 3.80. Combining Thm. 3.78 and Thm. 3.79 for $d = 3$ we see that the diagram

$$\begin{array}{ccccc} C^\infty(\overline{\Omega}))^3 & \xrightarrow{\mathbf{curl}} & C^\infty(\overline{\Omega}))^3 & \xrightarrow{\operatorname{div}} & C^\infty(\overline{\Omega}) \\ \downarrow{\scriptstyle \mathsf{I}_E} & & \downarrow{\scriptstyle \mathsf{I}_F} & & \downarrow{\scriptstyle \mathcal{Q}_0} \\ \mathcal{W}_E(\mathcal{M}) & \xrightarrow{\mathbf{curl}} & \mathcal{W}_F(\mathcal{M}) & \xrightarrow{\operatorname{div}} & \mathcal{Q}_0(\mathcal{M}) \end{array}$$

commutes. Hence, the assertions of Thm. 3.78 and Thm. 3.79 are often called **commuting diagram properties**.

Exercise 3.17. On a triangle K the following vectorfields are considered

$$\boldsymbol{b}_{ij}^1 := (\lambda_i \, \mathbf{grad}\, \lambda_j - \lambda_j \, \mathbf{grad}\, \lambda_i)^\perp \, , \boldsymbol{b}_{ij}^2 := (\lambda_i \, \mathbf{grad}\, \lambda_j + \lambda_j \, \mathbf{grad}\, \lambda_i)^\perp \, , \quad 1 \le i < j \le 3 \, .$$

Characterize the local space Π_K spanned by these functions. Find the associated set Σ_K of local degrees of freedom. Prove that (K, Π_K, Σ_K) is a $H(\operatorname{div})$-conforming finite element. How have the local d.o.f. be changed to ensure matching local d.o.f. within a conforming triangulation. What global degrees of freedom will result?

Remark 3.81. The families of finite elements we have used presented can be regarded as **discrete differential forms**. Their lowest degree simplicial variants were first discovered by Whitney [42] and used for theoretical purposes in differential geometry.

Bibliographical notes. A comprehensive survey of $H(\operatorname{div})$-conforming finite elements in given in [10, Ch. 3]. A similar treatment of edge elements is given in [25, Ch. 3].

3.8.3 $H^2(\Omega)$-conforming finite elements

Whereas for H^1-conforming elements the global degrees of freedom only need to ensure the continuity of the functions in the finite element space, the additional continuity of all first derivatives is required for H^2-conformity, see Sect. 2.4. This can only be achieved by fairly intricate constructions of finite elements. Having the plate problem Example 2.67 in mind, we restrict ourselv to $d = 2$ and conforming simplicial triangulations. Proofs will be skipped and for details the reader is referred to [12, Ch. 2,§2.2.1].

The "simplest" H^2-conforming finite element involving a full polynomial space is the **Argyris triangle** (K, Π_K, Σ_K) with

(i) K is a non-degenerate triangle with vertices $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \boldsymbol{\nu}_3$

(ii) $\Pi_K = \mathcal{P}_5(K)$,

(iii)

$$\begin{aligned}\Sigma_K := \{&v \mapsto v(\boldsymbol{\nu}_i),\ i = 1,2,3,\\ &v \mapsto \frac{\partial v}{\partial \xi_j}(\boldsymbol{\nu}_i),\ i = 1,2,3, j = 1,2,\\ &v \mapsto \frac{\partial^2 v}{\partial \xi_j \xi_k}(\boldsymbol{\nu}_i)),\ i = 1,2,3, j,k = 1,2,\\ &v \mapsto \langle \mathbf{grad}\, v(1/2(\boldsymbol{\nu}_i + \boldsymbol{\nu}_j)), \mathbf{n}_{ij}\rangle\ ,\ 1 \le i < j <\le 3\}\ ,\end{aligned}$$

where $\mathbf{n}_{ij}$ is a normal to the edge connecting $\boldsymbol{\nu}_i$ and $\boldsymbol{\nu}_j$.

Another example that uses fewer local degrees of freedom is **Bell's triangle**, which is defined by

(i) K is a non-degenerate triangle with vertices $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \boldsymbol{\nu}_3$

(ii) $\Pi_K = \{v \in \mathcal{P}_5(K) :\ \langle (\mathbf{grad}\, v)(\cdot), \mathbf{n}_F\rangle \in \mathcal{P}_3(F)\ \forall\, F \text{ edge of } K\}$,

(iii)

$$\begin{aligned}\Sigma_K := \{&v \mapsto v(\boldsymbol{\nu}_i),\ i = 1,2,3,\\ &v \mapsto \frac{\partial v}{\partial \xi_j}(\boldsymbol{\nu}_i),\ i = 1,2,3, j = 1,2,\\ &v \mapsto \frac{\partial^2 v}{\partial \xi_j \xi_k}(\boldsymbol{\nu}_i)),\ i = 1,2,3, j,k = 1,2\}\ .\end{aligned}$$

It can be shown that 18 is the minimal dimension of a polynomial local trial space on a triangle in order to accommodate global C^1-continuity. However, we may dispense with polynomial local trial spaces and resort to so-called **composite finite elements**

that rely on piecewise polynomial local trial spaces. This sacrifices $\Pi_K \subset C^\infty(\overline{K})$, *cf.* Remark 3.17.

An example is the **Xie-Clough-Tocher triangle** with a local piecewise cubic trial space of dimension 12:

(i) K is a non-degenerate triangle with vertices $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \boldsymbol{\nu}_3$

(ii) K is split into three sub-triangles K_1, K_2, K_3 by connecting its vertices with the barycenter $\frac{1}{3}(\boldsymbol{\nu}_1 + \boldsymbol{\nu}_2 + \boldsymbol{\nu}_3)$. Then

$$\Pi_K := \{v \in C^1(\overline{K}) : \ v_{|K_i} \in \mathcal{P}_3(K_i),\ i = 1,2,3\} \ .$$

(iii)

$$\begin{aligned}\Sigma_K := \{&v \mapsto v(\boldsymbol{\nu}_i),\ i = 1,2,3,\\ &v \mapsto \frac{\partial v}{\partial \xi_j}(\boldsymbol{\nu}_i),\ i = 1,2,3, j = 1,2,\\ &v \mapsto \langle \mathbf{grad}\, v(1/2(\boldsymbol{\nu}_i + \boldsymbol{\nu}_j)), \mathbf{n}_{ij}\rangle \ , \ 1 \le i < j <\leq 3\} \ .\end{aligned}$$

The bottom line is that H^2-conforming finite elements require relatively large local trial spaces and "complicated" degrees of freedom.

Remark 3.82. For none of the above families of H^2-conforming finite elements on triangles it is possible to choose local degrees of freedom that convert them into affine equivalent families. The culprit is he presence of a face normal vector in the definitions of Σ_K.

Bibliographical notes. The details of finite elements for variational problems in $H^2(\Omega)$ are discussed in [12, Ch. 6].

3.9 Algorithmic issues

A finite element scheme for the discretization of a boundary value problem has to be implemented carefully in order to achieve maximum efficiency.

3.9.1 Assembly

Assembly is a common term for computing and storing the matrix and right hand side vector of the linear system of equations (LSE) arising from the finite element discretization of a linear variational problem, *cf.* (LVP),

$$u \in V : \quad \mathsf{b}(u,v) = \langle f, v\rangle_{V^* \times V} \quad \forall v \in V$$

posed on some function space V, see Sect. 1.5.

We write $V_n \subset V$ for the finite element space built on a triangulation $\mathcal{M}$ of the computational domain Ω, see Sect. 3.3. It must posses a basis $\mathfrak{B} = \{b^1, \dots, b^N\}$, $N := \dim V_n$, of locally supported global shape functions as explained in Sect. 3.5. By (1.24), the crucial matrix $\mathbf{B}$ and right hand side vector $\boldsymbol{\varphi}$ are given by

$$\mathbf{B} = \big(\mathsf{b}(b^j, b^i)\big)_{i,j=1}^N \in \mathbb{R}^{N,N} \quad , \quad \boldsymbol{\varphi} = \Big(\big\langle f, b^i \big\rangle_{V^* \times V}\Big)_{i=1}^N \in \mathbb{R}^N \, . \tag{3.11}$$

In the finite element context $\mathbf{B}$ is called **stiffness matrix** and $\boldsymbol{\varphi}$ is known as **load vector**.

In Ch. 2 we have learned that for relevant boundary value problems the bilinear form b and right hand side functional will always involve a single integration over the computational domain Ω. In particular they can be written in terms of **cell contributions**:

$$\mathsf{b}(u, v) = \sum_{K \in \mathcal{M}} \mathsf{b}_K(u_{|K}, v_{|K}) \quad , \quad \langle f, v \rangle_{V^* \times V} = \sum_{K \in \mathcal{M}} f_K(v_{|K}) \, . \tag{3.12}$$

where the b_K are bilinear forms on $V_{|K}$, and f_K are local source term functionals supported on K.

Example 3.83. In the case of the primal variational formulation of a second-order elliptic boundary value problem with source function $f \in L^2(\Omega)$ we trivially have

$$\mathsf{b}(u, v) := \int_\Omega \langle \mathbf{A}\,\mathbf{grad}\, u, \mathbf{grad}\, v \rangle \, \mathrm{d}\boldsymbol{\xi} = \sum_K \underbrace{\int_K \langle \mathbf{A}\,\mathbf{grad}\, u, \mathbf{grad}\, v \rangle \, \mathrm{d}\boldsymbol{\xi}}_{=: \mathsf{b}_K(u_{|K}, v_{|K})} \, ,$$

$$\langle f, v \rangle_{V^* \times V} := \int_\Omega f v \, \mathrm{d}\boldsymbol{\xi} = \sum_K \underbrace{\int_K f v \, \mathrm{d}\boldsymbol{\xi}}_{=: f_K(v_{|K})} \, .$$

In Sect. 3.5 we have found that the restrictions of global shape functions to individual cells $K \in \mathcal{M}$ will coincide with local shape functions from the local trial spaces Π_K. Combined with (3.12) this will be key to assembling the

Definition 3.84. *Given the set $\{b_1^K, \dots, b_k^K\}$, $k = \dim \Pi_K$, of local shape functions for the finite element (K, Π_K, Σ_K), $K \in \mathcal{M}$, we call*

$$\mathbf{B}_K := \big(\mathsf{b}_K(b_i^K, b_j^K)\big)_{j,i=1}^k \in \mathbb{R}^{k,k}$$

*the **element stiffness matrix** for K. The **element load vector** $\boldsymbol{\varphi}_K \in \mathbb{R}^k$ is given by*

$$\boldsymbol{\varphi}_K := \big(f_K(b_i^K)\big)_{i=1}^k \, .$$

Theorem 3.85. *The stiffness matrix and load vector can be obtained from their cell counterparts by*

$$\mathbf{B} = \sum_K \mathbf{T}_K^T \mathbf{B}_K \mathbf{T} \quad , \quad \boldsymbol{\varphi} = \sum_K \mathbf{T}_K \boldsymbol{\varphi}_K \ , \tag{ASS}$$

with the ***T-matrices*** $\mathbf{T}_K \in \mathbb{R}^{k_K, N}$, $k_K := \dim \Pi_K$, *defined by*

$$(\mathbf{T}_K)_{ij} := \begin{cases} 1 & , \text{ if } \quad b^j_{|K} = b^K_i \ , \\ 0 & , \text{ if } \quad \operatorname{supp}(b^j) \cap K = \emptyset \ , \end{cases} \qquad 1 \le i \le k_K,\ 1 \le j \le N \ .$$

Proof.

$$\begin{aligned} (\mathbf{B})_{ij} = \mathsf{b}(b^j, b^i) &= \sum_{K \in \mathcal{M}} \mathsf{b}_K(b^j{}_{|K}, b^i{}_{|K}) \\ &= \sum_{K \in \mathcal{M},\, \operatorname{supp}(b^j) \cap K \neq \emptyset,\, \operatorname{supp}(b^i) \cap K \neq \emptyset} \mathsf{b}_K(b^K_{l(j)}, b^K_{l(i)}) \\ &= \sum_{K \in \mathcal{M},\, \operatorname{supp}(b^j) \cap K \neq \emptyset,\, \operatorname{supp}(b^i) \cap K \neq \emptyset} (\mathbf{B}_K)_{l(i), l(j)} \ , \end{aligned}$$

where $l(i) \in \{1, \ldots, k_K\}$, $1 \le i \le N$, is the index of the local shape function corresponding to the global shape function b^i on K. Hence,

$$(\mathbf{B})_{ij} = \sum_{K \in \mathcal{M},\, \operatorname{supp}(b^j) \cap K \neq \emptyset,\, \operatorname{supp}(b^i) \cap K \neq \emptyset} \ \sum_{l=1}^{k} \sum_{n=1}^{k} (\mathbf{T}_K)_{li} \, (\mathbf{B}_K)_{ln} \, (\mathbf{T}_K)_{nj} \ .$$

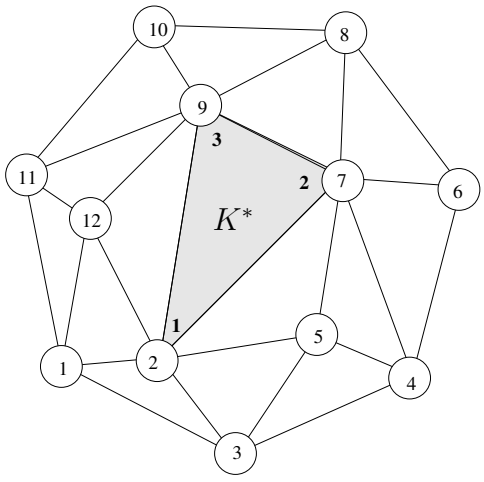
□

Example 3.86.

We consider Lagrangian finite elements (see Sect. 3.8.1) of degree 1 on the mesh sketches beside. The T-matrix for the marked triangle and the triangulation sketched beside reads

$$\mathbf{T}_{K^*} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \ ,$$

if the indicated numbering of global and local shape functions is used.



Remark 3.87. In Sect. 3.8.2 we encountered the situation where the use of only one reference element could mean that a global shape function restricted to a cell is equal to the negative of a local shape function. In this case the concept of the T-matrices can be generalized to allow for entries $\in \{-1, 0, 1\}$.

We observe that the bilinear forms b underlying the weak formulation of the boundary value problems discussed in Ch. 2 are *local* in the sense that

$$u, v \in V : \quad |\operatorname{supp}(v) \cap \operatorname{supp}(w)| = 0 \quad \Rightarrow \quad \mathsf{b}(u, v) = 0 \,. \tag{LOC}$$

Hence, if there is no overlap of the supports of two global shape functions, then the corresponding entry of the stiffness matrix will be zero.

Lemma 3.88. *Let the global shape function $b_F \in V_n$ be associated with the node/edge/face F of a conforming triangulation $\mathcal{M}$. Then the row and column of the finite element stiffness matrix $\mathbf{B}$ corresponding to b_F have at most*

$$\sum_{K \in \mathcal{M},\, F \subset \overline{K}} \dim \Pi_K$$

non-zero entries.

Proof. This estimate is immediate from Thm. 3.36 and (LOC). □

As a consequence, if

- there is an upper bound $n_{\mathcal{M}}$ for the number of closed intersecting cells of a conforming triangulation $\mathcal{M}$,
- the dimensions of the trial spaces are uniformly below a small constant $D \in \mathbb{N}$ for all cells,

then

> the finite element stiffness matrix will have at most $n_{\mathcal{M}} \cdot D \cdot N$, $N := \dim V_n$, non-zero entries. This means that for $N \gg 1$, *ie.* meshes with many elements, only a small fraction of the entries of the stiffness matrix will be non-zero, it will be **sparse**.

Remark 3.89. As a consequence of (LOC), finite element stiffness matrices often turn out to be **structurally symmetric**, even if $\mathbf{B} = \mathbf{B}^T$ fails to hold. This means

$$b_{ij} \neq 0 \quad \Leftrightarrow \quad b_{ji} \neq 0 \,, \quad 1 \leq i, j \leq N \,.$$

Example 3.90. If the smallest angles of the triangles of a two-dimensional conforming simplicial triangulation are bounded from below by $\alpha > 0$, then at most $\lfloor 2\pi/\alpha \rfloor$ (closed) triangles can intersect.

Example 3.91. For H^1-conforming linear Lagrangian finite elements on the triangulation of Example 3.86 and assuming the numbering of global shape functions given there, we obtain the following **sparsity pattern** for the stiffness matrix.

$$\mathbf{B} = \begin{pmatrix} * & * & * & 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & * \\ * & * & * & 0 & * & 0 & * & 0 & * & 0 & 0 & * \\ * & * & * & * & * & * & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & * & * & * & * & 0 & 0 & 0 & 0 & 0 \\ 0 & * & * & * & * & 0 & * & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & * & 0 & * & * & * & 0 & 0 & 0 & 0 \\ 0 & * & 0 & * & * & * & * & * & * & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & * & * & * & * & * & 0 & 0 \\ 0 & * & 0 & 0 & 0 & 0 & * & * & * & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & * & * & * & 0 \\ * & 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & * & * & * \\ * & * & 0 & 0 & 0 & 0 & 0 & 0 & * & 0 & * & * \end{pmatrix},$$

where $*$ stands for a potential non-zero entry. If we replaced $*$ with 1, the matrix would describe the **graph of the mesh**.

Remark 3.92. When dealing with sparse matrices in MATLAB, one must use special *sparse matrix functions* like the sparse function for creating matrices. Gross inefficiency results, if one inadvertently carries out conversions to dense matrices. In MATLAB the sparsity pattern of a matrix can be visualized by means of the spy command.

Exercise 3.18. Compute a tight upper bound for the number of non-zero entries of the stiffness matrix, if cubic H^1-conforming Lagrangian elements according to Def. 3.54 are used on the triangulation of 3.86 in order to discretize the variational problem (NVP).

Answer the same question for the lowest order $H(\mathrm{div})$-conforming space from Def. 3.63 and a variational problem of the type (FWD).

3.9.2 Local computations

Two main options exist for the computation of the entries of element stiffness matrices.

As an example for **analytic evaluation**, which is the first option, we present the case of degree m Lagrangian finite elements on a triangle (see Def. 3.54 in Sect. 3.8.1) applied for the Galerkin discretization of the bilinear form

$$\mathsf{b}(u,v) := \int_\Omega \langle \mathbf{grad}\, u, \mathbf{grad}\, v \rangle \, \mathrm{d}\boldsymbol{\xi} \,. \tag{3.13}$$

Pick a generic triangle K. All local shape functions b_i on K, will possess a representation in terms of the barycentric coordinate functions $\lambda_1, \lambda_2, \lambda_3$ of K, see Def. 3.49 and

Lemma 3.51:

$$b_i = \sum_{\boldsymbol{\alpha}\in\mathbb{N}_0^3,|\boldsymbol{\alpha}|=m} \kappa_{\boldsymbol{\alpha}}\, \lambda_1^{\alpha_1}\lambda_2^{\alpha_2}\lambda_3^{\alpha_3}\ , \quad \kappa_{\boldsymbol{\alpha}} \in \mathbb{R}\ , \tag{3.14}$$

which involves

$$\begin{aligned}\mathbf{grad}\, b_i = \sum_{\boldsymbol{\alpha}\in\mathbb{N}_0^3,|\boldsymbol{\alpha}|=m} \kappa_{\boldsymbol{\alpha}}\Big(&\alpha_1\lambda_1^{\alpha_1-1}\lambda_2^{\alpha_2}\lambda_3^{\alpha_3}\,\mathbf{grad}\,\lambda_1 + \alpha_2\lambda_1^{\alpha_1}\lambda_2^{\alpha_2-1}\lambda_3^{\alpha_3}\,\mathbf{grad}\,\lambda_2 + \\ &\alpha_3\lambda_1^{\alpha_1-1}\lambda_2^{\alpha_2}\lambda_3^{\alpha_3-1}\,\mathbf{grad}\,\lambda_3\Big)\ .\end{aligned} \tag{3.15}$$

The bottom line is that for the evaluation of the localised bilinear form b_K for local shape functions we have to compute integrals of the form

$$\int_K \lambda_1^{\beta_1}\lambda_2^{\beta_2}\lambda_3^{\beta_3}\,\langle \mathbf{grad}\,\lambda_i, \mathbf{grad}\,\lambda_j\rangle\ \mathrm{d}\boldsymbol{\xi}\ , \quad i,j\in\{1,2,3\},\ \beta_k\in\mathbb{N}\ . \tag{3.16}$$

Let the vertices $\boldsymbol{\nu}^1,\boldsymbol{\nu}^2,\boldsymbol{\nu}^3$ of the triangle K be arranged in counterclockwise fashion. Then we find the formulas

$$\begin{aligned}\lambda_1(\boldsymbol{\xi}) &= \frac{1}{2|K|}\left(\boldsymbol{\xi} - \begin{pmatrix}\nu_1^2\\ \nu_2^2\end{pmatrix}\right)\cdot\begin{pmatrix}\nu_2^2-\nu_2^3\\ \nu_1^3-\nu_1^2\end{pmatrix}\ ,\\ \lambda_2(\boldsymbol{\xi}) &= \frac{1}{2|K|}\left(\boldsymbol{\xi} - \begin{pmatrix}\nu_1^3\\ \nu_2^3\end{pmatrix}\right)\cdot\begin{pmatrix}\nu_2^3-\nu_2^1\\ \nu_1^1-\nu_1^3\end{pmatrix}\ ,\\ \lambda_3(\boldsymbol{\xi}) &= \frac{1}{2|K|}\left(\boldsymbol{\xi} - \begin{pmatrix}\nu_1^1\\ \nu_2^1\end{pmatrix}\right)\cdot\begin{pmatrix}\nu_2^1-\nu_2^2\\ \nu_1^2-\nu_1^1\end{pmatrix}\ .\end{aligned}$$

The gradients of the barycentric coordinate functions are constant and read

$$\mathbf{grad}\,\lambda_1 = \frac{1}{2|K|}\begin{pmatrix}\nu_2^2-\nu_2^3\\ \nu_1^3-\nu_1^2\end{pmatrix}\ ,\ \mathbf{grad}\,\lambda_2 = \frac{1}{2|K|}\begin{pmatrix}\nu_2^3-\nu_2^1\\ \nu_1^1-\nu_1^3\end{pmatrix}\ ,\ \mathbf{grad}\,\lambda_3 = \frac{1}{2|K|}\begin{pmatrix}\nu_2^1-\nu_2^2\\ \nu_1^2-\nu_1^1\end{pmatrix}\ .$$

Geometric computations show

$$\begin{aligned}&\left(\langle \mathbf{grad}\,\lambda_i, \mathbf{grad}\,\lambda_j\rangle\right)_{i,j=1}^3 \\ &\quad = \frac{1}{2|K|}\begin{pmatrix}1-\cot\omega_3-\cot\omega_2 & -\cot\omega_3 & -\cot\omega_2\\ -\cot\omega_3 & 1-\cot\omega_3-\cot\omega_1 & -\cot\omega_1\\ -\cot\omega_2 & -\cot\omega_1 & 1-\cot\omega_2-\cot\omega_1\end{pmatrix}\ ,\end{aligned} \tag{3.17}$$

where ω_i is the angle at vertex $\boldsymbol{\nu}^i$.

Exercise 3.19. Prove the formula (3.17).

Apart from (3.17), the next formula is the second ingredient for successfully tackling (3.16).

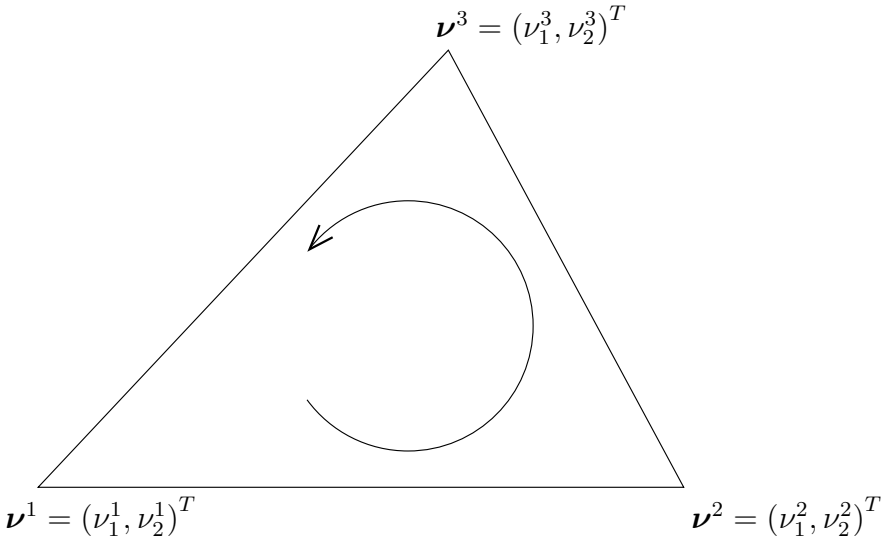


Figure 3.18: Generic triangular element

Lemma 3.93. *For any non-degenerate triangle K and $\beta_1, \beta_2, \beta_3 \in \mathbb{N}$,*

$$\int_K \lambda_1^{\beta_1} \lambda_2^{\beta_2} \lambda_3^{\beta_3} \, \mathrm{d}\boldsymbol{\xi} = 2|K| \cdot \frac{\beta_1!\, \beta_2!\, \beta_3!}{(\beta_1 + \beta_2 + \beta_3 + 2)!} \, .$$

Proof. The first step amounts to a transformation of K to the "unit triangle" $\widehat{K} :=$ convex $\left\{ \binom{0}{0}, \binom{1}{0}, \binom{0}{1} \right\}$, which leads to

$$\begin{aligned}
\int_K \lambda_1^{\beta_1} \lambda_2^{\beta_2} \lambda_3^{\beta_3} \, \mathrm{d}\boldsymbol{\xi} &= 2|K| \int_0^1 \int_0^{1-\xi_1} \xi_1^{\beta_1} \xi_2^{\beta_2} (1 - \xi_1 - \xi_2)^{\beta_3} \, \mathrm{d}\xi_2 \mathrm{d}\xi_1 \\
&= 2|K| \int_0^1 \xi_1^{\beta_1} \int_0^1 (1 - \xi_1)^{\beta_2 + \beta_3 + 1} s^{\beta_2} (1 - s)^{\beta_3} \, \mathrm{d}s \, \mathrm{d}\xi_1 \\
&= 2|K| \int_0^1 \xi_1^{\beta_1} (1 - \xi_1)^{\beta_2 + \beta_3 + 1} \, \mathrm{d}\xi_1 \cdot B(\beta_2 + 1, \beta_3 + 1) \\
&= 2|K| \; B(\beta_1 + 1, \beta_2 + \beta_3 + 2) \cdot B(\beta_2 + 1, \beta_3 + 1) \, ,
\end{aligned}$$

where we used the substitution $s(1 - \xi_1) = \xi_2$ to arrive at Euler's beta functions

$$B(\alpha, \beta) := \int_0^1 t^{\alpha - 1} (1 - t)^{\beta - 1} \, \mathrm{d}t \, , \quad 0 < \alpha, \beta < \infty \, .$$

Using the known formula $\Gamma(\alpha+\beta)\,B(\alpha,\beta)=\Gamma(\alpha)\Gamma(\beta)$, Γ the Gamma function, we end up with

$$\int_K \lambda_1^{\beta_1}\lambda_2^{\beta_2}\lambda_3^{\beta_3}\,\mathrm{d}\boldsymbol{\xi} = 2|K|\cdot\frac{\Gamma(\beta_1+1)\Gamma(\beta_2+1)\Gamma(\beta_3+1)}{\Gamma(\beta_1+\beta_2+\beta_3+3)}\,.$$

Then, $\Gamma(n)=(n-1)!$ finishes the proof. □

Remark 3.94. Lemma 3.93 can be extended to simplices in dimension d. If λ_i $i=1,\ldots,d+1$ stand for barycentric coordinate functions of a d-simplex K, we have

$$\int_K \lambda_1^{\alpha_1}\cdot\cdots\cdot\lambda_{d+1}^{\alpha_{d+1}}\,\mathrm{d}\boldsymbol{\xi} = d!|K|\,\frac{\alpha_1!\alpha_2!\cdot\cdots\cdot\alpha_{d+1}!}{(\alpha_1+\alpha_2+\cdots+\alpha_{d+1}+d)!}\quad\forall\boldsymbol{\alpha}\in\mathbb{N}_0^{d+1}\,.$$

Exercise 3.20. Compute the element stiffness matrix related to the $L^2(\Omega)$ inner product (which plays the role of the bilinear form) and quadratic Lagrangian finite elements on a triangle.

We remark that the global matrix related to the $L^2(\Omega)$ inner product is often called **mass matrix**.

Exercise 3.21. For $\tau>0$ consider the bilinear form

$$\mathsf{b}(\mathbf{u},\mathbf{v}) := \int_\Omega \operatorname{div}\mathbf{u}\cdot\operatorname{div}\mathbf{v}\,\mathrm{d}\boldsymbol{\xi} + \tau\int_\Omega \langle\mathbf{u},\mathbf{v}\rangle\,\mathrm{d}\boldsymbol{\xi}\,,\quad \mathbf{u},\mathbf{v}\in H(\operatorname{div};\Omega)\,.$$

Compute the element stiffness matrix for this bilinear form, when lowest degree $H(\mathrm{div})$-conforming finite elements introduced in Thm. 3.64 are used on a triangular cell in two dimensions.

3.9.3 Numerical quadrature

Another approach to local evaluation is the use of **numerical quadrature**. It amounts to the following approximation of the integrals occuring in the definition of the bilinear form and of the source term:

$$\int_\Omega f(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} \approx \sum_{K\in\mathcal{M}}|K|\sum_{l=1}^{P_K}\omega_l^K\,f(\boldsymbol{\pi}_l^K)\,. \tag{NUQ}$$

This formula makes sense for $\mathcal{M}$-piecewiese continuous functions f that allow continuous extensions from K onto ∂K for all $K\in\mathcal{M}$. The ω_l^K are called the **local weights**, the points $\boldsymbol{\pi}_l^K\in K$ are the **local nodes**. Together, they constitute a **local quadrature rule** on K. We point out that

only quadrature rules with positive weights are numerically stable.

Local quadrature formulas are usually defined on reference cells: if $K \in \mathcal{M}$, $\widehat{K}$ a corresponding reference cell, and $\boldsymbol{\Phi} : \widehat{K} \mapsto K$ the diffeomorphism connecting both, then the local quadrature rule

$$\int_{\widehat{K}} f(\widehat{\boldsymbol{\xi}})\,\mathrm{d}\widehat{\boldsymbol{\xi}} \approx |\widehat{K}| \sum_{l=1}^{P} \widehat{\omega}_l\, f(\widehat{\boldsymbol{\pi}}_l)$$

on $\widehat{K}$ will spawn a quadrature rule on K by setting $\omega_l^K := \widehat{\omega}_l$ and $\boldsymbol{\pi}_l^K := \boldsymbol{\Phi}(\widehat{\boldsymbol{\pi}}_l)$.

Remark 3.95. If K is a simplex, then the use of a reference cell can be avoided, when the quadrature nodes are described through their barycentric coordinates.

The quality of a local quadrature rule on K is gauged by determining the largest space of polynomials on $\widehat{K}$ that the corresponding quadrature rule on the reference cell manages to integrate exactly.

Example 3.96. If K is a triangle, the we can use the "unit triangle" $\widehat{K} := \text{convex}\left\{\binom{0}{0}, \binom{1}{0}, \binom{0}{1}\right\}$ as reference cell. Writing the quadrature rules on $\widehat{K}$ as finite sets of pairs $(\widehat{\omega}_1, \widehat{\boldsymbol{\pi}}_1), \ldots, (\widehat{\omega}_P, \widehat{\boldsymbol{\pi}}_P)$, $P \in \mathbb{N}$, we find that

$$\left\{\left(\frac{1}{3}, \binom{0}{0}\right), \left(\frac{1}{3}, \binom{0}{1}\right), \left(\frac{1}{3}, \binom{1}{0}\right)\right\} \tag{3.18}$$

is exact for $\mathcal{P}_1(\widehat{K})$. The similar rule

$$\left\{\left(\frac{1}{3}, \binom{1/2}{0}\right), \left(\frac{1}{3}, \binom{0}{1/2}\right), \left(\frac{1}{3}, \binom{1/2}{1/2}\right)\right\} \tag{3.19}$$

is exact even for $\mathcal{P}_2(\widehat{K})$! If we use inner points of the triangle, we can achieve exactness for surprisingly high polynomial degrees. The very simple rule

$$\left\{\left(1, \binom{1/3}{1/3}\right)\right\} \tag{3.20}$$

is already exact for $\mathcal{P}_1(\widehat{K})$, whereas the 7-point rule

$$\begin{aligned}\left\{\left(\frac{9}{80}, \binom{1/3}{1/3}\right), \left(\frac{155+\sqrt{15}}{2400}, \binom{6+\sqrt{15}/21}{6+\sqrt{15}/21}\right), \left(\frac{155+\sqrt{15}}{2400}, \binom{9-2\sqrt{15}/21}{6+\sqrt{15}/21}\right),\right.\\ \left(\frac{155+\sqrt{15}}{2400}, \binom{6+\sqrt{15}/21}{9-2\sqrt{15}/21}\right), \left(\frac{155-\sqrt{15}}{2400}, \binom{6-\sqrt{15}/21}{9+2\sqrt{15}/21}\right),\\ \left.\left(\frac{155-\sqrt{15}}{2400}, \binom{9+2\sqrt{15}/21}{6-\sqrt{15}/21}\right), \left(\frac{155-\sqrt{15}}{2400}, \binom{6-\sqrt{15}/21}{6-\sqrt{15}/21}\right)\right\}\end{aligned} \tag{3.21}$$

is exact even for quintic polynomials $\in \mathcal{P}_5(\widehat{K})$. Many more quadrature rules on triangles and tetrahedra are known, see [39, Ch. 3].

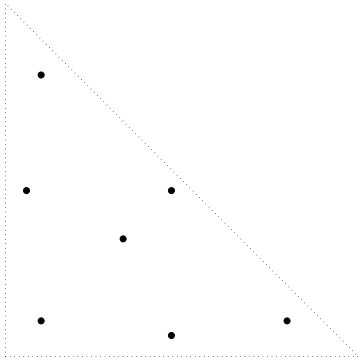


Figure 3.19: Evaluation points for the 7-point rule on the reference triangle.

Remark 3.97. Usually, in a finite element code only a few local quadrature rules will be needed. Their nodes and weights should be stored in look-up tables.

Example 3.98. For a quadrilateral cell K the natural reference cell is the unit square $\widehat{K} := \text{convex}\left\{\binom{0}{0}, \binom{1}{0}, \binom{0}{1}, \binom{1}{1}\right\}$. On $\widehat{K}$ a straightforward tensor product construction yields suitable quadrature rules: if $\{(\omega_1, \pi_1), \ldots, (\omega_P, \pi_P)\}$, $P \in \mathbb{N}$, is a quadrature rule on the interval $]0,1[$ that is exact for $\mathcal{P}_m(]0,1[)$, then

$$\left\{ \begin{array}{ccc} (\omega_1^2, \binom{\pi_1}{\pi_1}) & \cdots & (\omega_1\omega_P, \binom{\pi_1}{\pi_P}) \\ \vdots & & \vdots \\ (\omega_1\omega_P, \binom{\pi_P}{\pi_1}) & \cdots & (\omega_P^2, \binom{\pi_P}{\pi_P}) \end{array} \right\}$$

will yield a quadrature rule on $\widehat{K}$ that is exact for $\mathcal{Q}_{(m,m)}(\widehat{K})$.

Numerical analysis has provided plenty of quadrature rules on $]0,1[$:

- classical Newton-Cotes formulas that rely on an equidistant distribution of quadrature nodes.
- Gauss-Legendre quadrature rules that are exact for $\mathcal{P}_{2P+1}(]0,1[)$ using only P nodes.
- Gauss-Lobatto quadrature rules that rely on P nodes including the endpoints of the interval and remain exact for polynomials up to degree $2P$.

Remark 3.99. Quadrature rules for triangles can be obtained from quadrature rules on a square using the **Duffy transformation**, see Fig. 3.20. When $K = \text{convex}\{\binom{0}{0}, \binom{1}{0}, \binom{1}{1}\}$ and $\widehat{K} =]0,1[^2$, then

$$\int_K f(\xi_1, \xi_2)\, \mathrm{d}\boldsymbol{\xi} = \int_{\widehat{K}} f(\widehat{\xi}_1(1-\widehat{\xi}_2), \widehat{\xi}_2)\,(1-\widehat{\xi}_2)\, \mathrm{d}\widehat{\boldsymbol{\xi}}\ .$$

If $f \in \mathcal{P}_m(K)$, then the integrand on the right hand side will belong to $\mathcal{Q}_{m,m+1}(\widehat{K})$.

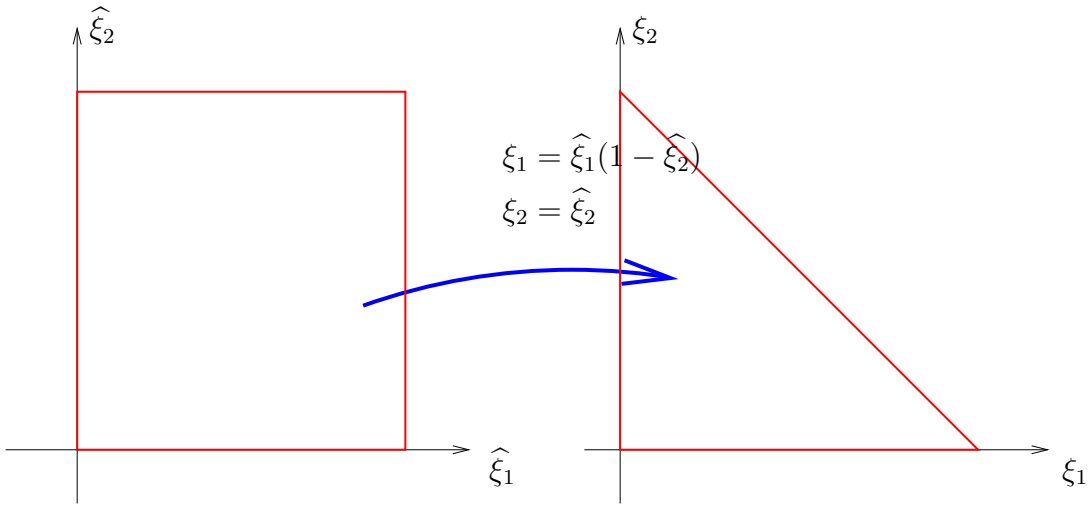


Figure 3.20: "Duffy transformation" of a square into a triangle

Exercise 3.22. Solve Exercise 3.20 using the quadrature rule (3.18). What will be the properties of the resulting (approximate) mass matrix.

Exercise 3.23. The Poisson equation, that is, an elliptic boundary value problem with bilinear form (3.13) has been discretized on a mesh composed of square cells using Lagrangian finite elements of degree m, $m \in \mathbb{N}$. The computation of element stiffness matrices employs a tensor product Gauss-Legendre quadrature rule. How many quadrature nodes per cells are required in order to avoid any quadrature error?

There are cases where we cannot help using numerical quadrature. For instance, we may want to solve the variational problem

$$u \in H^1(\Omega): \quad \int_\Omega \langle \mathbf{A}\,\mathbf{grad}\,u, \mathbf{grad}\,v\rangle + c\,uv\,\mathrm{d}\boldsymbol{\xi} = \int_\Omega fv\,\mathrm{d}\boldsymbol{\xi} \quad \forall v \in H^1(\Omega)\,. \tag{3.22}$$

Usually, the coefficients and source functions will

- be supplied as complicated analytical expressions, or
- they are only accessible via point evaluations, because they are provided by a sub-routine of the finite element code.

The latter case is very common, if the coefficient functions themselves are the result of numerical approximation. In these two cases an analytical evaluation of the element stiffness matrix and load vector is no longer feasible and numerical quadrature has to be used.

The use of numerical quadrature inevitably introduces another approximation, which will contribute to the overall discretization error. The general rule is that

The error due to numerical quadrature must not dominate the total discretization error in the relevant norms.

In Sect. 4.7 we will see that this can be ensured by selecting quadrature rules that are exact for polynomials of sufficiently high degree.

Remark 3.100. An alternative to numerical quadrature for (3.22) is polynomial interpolation of coefficients and source functions followed by analytical evaluation of the localised integrals.

3.9.4 Boundary approximation

According to our concept of computational domain Ω, see Def. 2.5, we may encounter smooth, but curved parts of $\partial\Omega$. Parametric finite elements, introduced in Sect. 3.7, offer a way to deal with them.

To illustrate the idea we consider a computational domain $\Omega \subset \mathbb{R}^2$ that has been approximately triangulated by a conforming simplicial triangulation. We pick a triangle $\widehat{K}$, for which two vertices $\boldsymbol{\nu}^1, \boldsymbol{\nu}^2$ are located on Γ, see Figure 3.21.

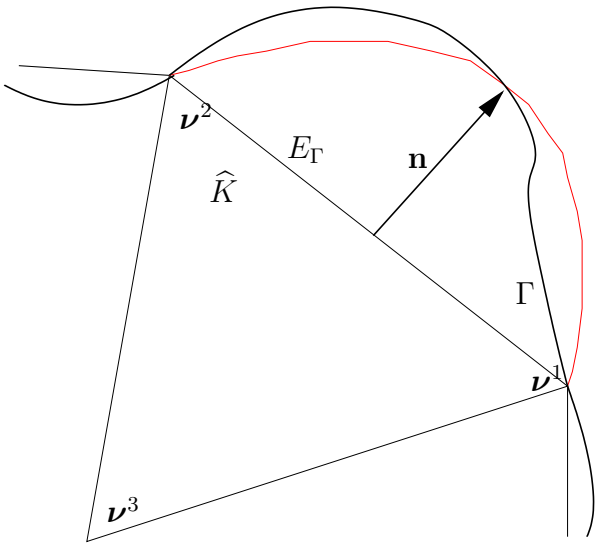


Figure 3.21: Triangle at curved boundary and quadratic approximation of the boundary

Using a suitable mapping we aim to fit the shape of $\widehat{K}$ to the boundary by replacing the straigt edge E_Γ by a piece of a parabola: The part of Γ between $\boldsymbol{\nu}^1$ and $\boldsymbol{\nu}^2$ is viewed as the graph of a function over E_Γ. Then quadratic polynomial interpolation with nodes in $\boldsymbol{\nu}^1, \boldsymbol{\nu}^2$ and the midpoint of E_Γ is carried out.

To convert this idea into a transformation for $\widehat{K}$, let δ denote the distance of the intersection point of the perpendicular bisector of E_Γ with Γ. Then, writing λ_i for the

barycentric coordinate functions on $\widehat{K}$, the mapping

$$\mathbf{\Phi}(\widehat{\boldsymbol{\xi}}) := \widehat{\boldsymbol{\xi}} + 4\delta\, \lambda_1(\widehat{\boldsymbol{\xi}})\lambda_2(\widehat{\boldsymbol{\xi}}) \cdot \mathbf{n}\ ,$$

$\mathbf{n}$ standing for the exterior unit normal to E_Γ, will take $\widehat{K}$ to a altered cell K with curved boundary, which will become a cell of the final **boundary fitted mesh**. We note that

$$D\mathbf{\Phi} = Id + 4\delta\, \mathbf{n} \cdot \mathbf{grad}(\lambda_1\lambda_2)^T \quad , \quad \det(D\mathbf{\Phi}) = 4\delta\ \langle \mathbf{n}, \mathbf{grad}(\lambda_1\lambda_2)\rangle\ ,$$

that is, both the Jacobi matrix and its determinant are polynomial of degree 1. This kind of polynomial boundary approximation can be generalized to higher polynomial degree and three dimensions. This is an example for the use of parametric finite elements, see Sect. 3.7.

Summing up, polynomial boundary approximation relies on *polynomial interpolation* of parts of the boundary that are considered as graphs of a function over straight edges/flat faces of a preliminary triangulation.

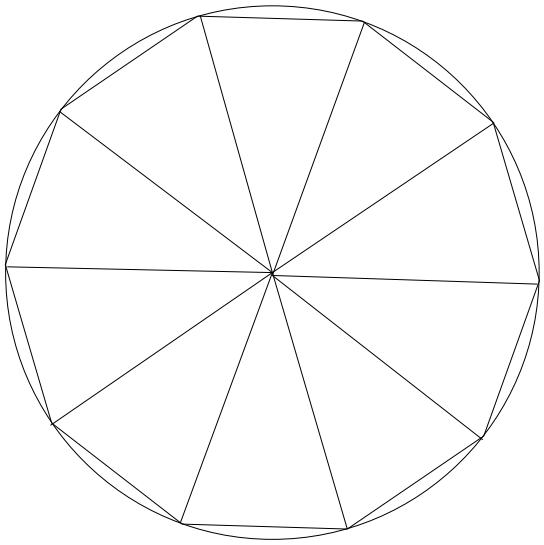
Example 3.101. Assume that we want to solve the Poisson equation on the domain Ω with curved boundary. The related bilinear form is (3.13). We employ a *transformation approach* to the computation of the element stiffness matrix for the cell $K = \mathbf{\Phi}(\widehat{K})$. Using the transformation (GT) of gradients, we get

$$\int_K \langle \mathbf{grad}\, u, \mathbf{grad}\, v\rangle\ \mathrm{d}\boldsymbol{\xi} = \int_{\widehat{K}} \left\langle D\mathbf{\Phi}^T\, \mathbf{grad}\, \widehat{u}, D\mathbf{\Phi}^T\, \mathbf{grad}\, \widehat{v}\right\rangle |\det D\mathbf{\Phi}|\, \mathrm{d}\widehat{\boldsymbol{\xi}}\ .$$

If parametric Lagrangian finite elements of degree 1 are used on K, we merely have to integrate cubic polynomials over $\widehat{K}$. This can be done analytically using the techniques outlined in Sect. 3.9.2.

Exercise 3.24.

A raw mesh of the unit disk $\Omega := \{\boldsymbol{\xi} \in \mathbb{R}^2 : |\boldsymbol{\xi}| < 1\}$ is obtained by inscribing a regular dekagon into the unit circle, see sketch beside. A better mesh results from piecewise quadratic boundary approximation as explained above. Compute the stiffness matrix related to the bilinear form (3.13) for triangular Lagrangian finite elements of degree 1 on the boundary fitted mesh.
Specify a parametric mapping for the cells of the raw mesh that will yield a perfect boundary approximation.



3.9.5 Data structures

The issue of data structures mainly concerns the internal representation of the mesh and of the stiffness matrix in a finite element code. For the sake of simplicity, we will only consider conforming triangulations consisting of simplices or tensor product cells.

The bare minimum of information the mesh data structure has to provide is

1. a unique identification and the geometric location of all nodes,
2. a possibility to traverse the set of nodes of every cell in a fixed order,
3. a way to run through the edges/faces of a cell in predifined order,
4. and a method for iterating through all cells of the mesh.

We will take this minimal information for granted but little else. Of course, often a wealth of problem dependent information (e.g. about coefficients and boundary conditions) has to be available on the cell level.

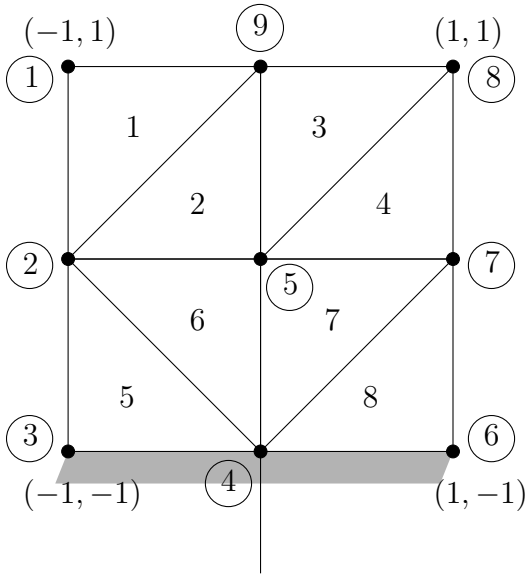
Two different schemes are widely used for storing mesh information.

1. **FORTRAN/MATLAB-style array oriented data layout.** For a conforming d-dimensional simplicial triangulation $\mathcal{M}$ the coordinates of the nodes (set $\mathcal{N}(\mathcal{M})$, see Def. 3.4) are stored in an $\sharp\mathcal{N}(\mathcal{M}) \times d$ array `coords` of real numbers. Another $\sharp\mathcal{M} \times (d+1)$-array `cells` of integers provides the numbers of the nodes that form the vertices of each simplex. These numbers refer to the indices of the nodes in the `coord`-array. This data layout is very similar to the file format discussed in Remark 3.14.

   Example 3.102. For a simple two-dimensional simplicial triangulation with 9 nodes the array oriented description is given in Fig. 3.22. This is only rudimentary information. For instance, one may want to supplement it with a $\sharp\mathcal{E}_\Gamma(\mathcal{M}) \times 2$ integer array `bdedges` that tells the edges on the boundary Γ along with a flag specifying the relevant boundary conditions.

   Of course, also "maximal" information about the mesh can be kept in an array oriented fashion. For instance, in the case of a conforming three-dimensional simplicial triangulation, the `cells` array may be of dimension $\sharp\mathcal{M} \times 14$, where each line contains four vertex indices, six edge indices, and four face indices.

2. **C++/JAVA-style object oriented data layout.** Each node and cell of the mesh $\mathcal{M}$ corresponds to a dynamically allocated instance of class `Node` and `Cell`, respectively. A special object of class `Mesh` takes care of global mesh management.



i	Coordinates	
1	-1	1
2	-1	0
3	-1	-1
4	0	-1
5	0	0
6	1	-1
7	1	0
8	1	1
9	0	1

Array `coord`

K_j	Vertex indices		
1	1	2	9
2	2	5	9
3	5	8	9
4	5	7	8
5	3	4	2
6	4	5	2
7	4	7	5
8	4	6	7

Array `cells`

Figure 3.22: Array oriented storage of mesh information for a two-dimensional simplicial triangulation.

Faces on the boundary may be stored separately in order to impose boundary conditions. A possible basic C++ implementation in depicted in Fig. 3.23, 3.24. The `id` data member of `Node` is optional and supplies a unique "identifier" for a node.

Of course, more information can easily be included into this data layout by introducing classes for edges/faces, too. Possibilities for extension are almost unlimited.

```
class Node {
  private:
    double x,y;
    ID id;
  public:
    Node(double x,double y,ID id=0);
    Point getCoords(void) const;
    ID getId(void) const;
};
```

```
class Cell {
  private:
    const vector<Node*> vertices;
  public:
    Cell(const vector<Node*>
         &vertices);
    int NoNodes(void) const;
    const Node &getNode(int) const;
};
```

Figure 3.23: Class skeletons for nodes and cells of a two-dimensional triangulation

The finite element stiffness matrix $\mathbf{B}$ will usually be sparse, see Sect. 3.9.1, and therefore it would be grossly wasteful to store it in a regular $N \times N$-array of real numbers, N the dimension of the finite element space V_n.

```
class BdFace {
  private:
    const vector<Node*> vertices;
    BdCond bdcond;
  public:
    BdFace(const vector<Node*>
           &vertices);
    int NoNodes(void) const;
    const Node &getNode(int) const;
    BdCond getBdCond(void) const;
};
```

```
class Mesh {
  private:
    list<Node> nodes;
    list<Cell> cells;
    list<BdFace> bdfaces;
  public:
    Mesh(istream &file);
    virtual ~Mesh(void);
    const list<Node> &
          Nodes(void) const;
    const list<Cell> &
          Cells(void) const;
    const list<BdFace> &
          BdFaces(void) const;
}
```

Figure 3.24: Classes for boundary faces and mesh management

A much more efficient data layout is offered by the wide used **compressed row storage (CRS)**, an array based storage scheme. This matrix representation is based on three arrays: one for real numbers (`val`), and the other two for integers (`col_ind`, `row_ptr`). The `val` array stores the values of the non-zero elements of the matrix $\mathbf{B} \in \mathbb{R}^{N,N}$ as they are traversed in a row-wise fashion. It contains $\mathrm{nnz}(\mathbf{B})$ elements, where

$$\mathrm{nnz}(\mathbf{B}) := \sharp\{(i,j) \in \{1,\ldots,N\}^2 : \ (\mathbf{B})_{ij} \neq 0\}$$

is the total number of non-zero entries of $\mathbf{B}$. The `col_ind` array has length $\mathrm{nnz}(\mathbf{B})$ and contains the column indices of the elements in `val`. That is, if $\texttt{val}[k] = b_{ij}$, then $\texttt{col\_ind}[k] = j$, $1 \leq k \leq \mathrm{nnz}(\mathbf{B})$, $1 \leq i,j \leq N$. The `row_ptr` array provides the locations in the `val` array that start a row, that is, if $\texttt{val}[k] = b_{ij}$, then $\texttt{row\_ptr}[i] \leq k < \texttt{row\_ptr}[i{+}1]$, $1 \leq k \leq \mathrm{nnz}(\mathbf{B})$, $1 \leq i \leq N$. By convention, $\texttt{row\_ptr}[N{+}1] = \mathrm{nnz}(\mathbf{B}){+}1$.

A matrix $\mathbf{B}$ stored in the CRS format requires nnz real numbers and $\mathrm{nnz}+N+1$ integers. For sparse matrices this means a considerable saving in memory, compared to the N^2 real numbers for dense storage, *cf.* Lemma 3.88.

For symmetric finite element stiffness matrices the **diagonal CRS format** is popular: the strictly lower triangular part of the matrix is stored in the usual CRS format, whereas the diagonal resides in an extra N-array `diag`. If the matrix is merely structurally symmetric, see Remark 3.89, then strictly lower and upper triangular parts can be stored in CRS fashion using two `val` arrays `val_lower` and `val_upper`.

Example 3.103. As an example for the standard CRS format we consider the non-

symmetric matrix $\mathbf{A}$ defined by

$$A = \begin{pmatrix} 10 & 0 & 0 & 0 & -2 & 0 \\ 3 & 9 & 0 & 0 & 0 & 3 \\ 0 & 7 & 8 & 7 & 0 & 0 \\ 3 & 0 & 8 & 7 & 5 & 0 \\ 0 & 8 & 0 & 9 & 9 & 13 \\ 0 & 4 & 0 & 0 & 2 & -1 \end{pmatrix}.$$

The CRS format for this matrix is then specified by the following arrays

`val`	10	-2	3	9	3	7	8	7	3 ... 9	13	4	2	-1
`col_ind`	1	5	1	2	6	2	3	4	1 ... 5	6	2	5	6

`row_ptr`	1	3	6	9	13	17	20

.

Exercise 3.25. A *structurally symmetric* sparse matrix $\mathbf{B} \in \mathbb{R}^{N,N}$ is stored in diagonal CRS format involving the arrays

`val_diag` : entries of diagonal of $\mathbf{B}$ (length N)
`val_upper` : non-zero entries of strictly upper triangular part of $\mathbf{B}$
`val_lower` : non-zero entries of strictly lower triangular part of $\mathbf{B}$
`col_ind` : column indices for strictly lower triangular part
`row_ptr` : row starts for strictly lower triangular part

Write a MATLAB function

```
y = matvec(x,val_diag,val_upper,val_lower,col_ind,row_ptr)
```

that performs the multipilcation of $\mathbf{B}$ with $x \in \mathbb{R}^N$ and returns the result in `y`.

Remark 3.104. Sometimes one can completely dispense with storing the stiffness matrix $\mathbf{B}$ This is the case, when the only operation needed is the multiplication of $\mathbf{B}$ with a vector. In many cases this can be sufficient for computing an approximate solution of the linear system (LSE). Then only the element stiffness matrices need be computed and formula (ASS) from Thm. 3.85 is used to carry out a matrix×vector-multiplication. No matrix entries will be kept in memory permanently.

Bibliographical notes. Storage formats for sparse matrices are reviewed in [5, Sect. 4.3]. Finite element data structures are mainly discussed in the manuals of finite element codes like UG `http://cox.iwr.uni-heidelberg.de/~ug/intro.html`, NETGEN `http://www.hpfem.jku.at/`, KASKADE and PLTMG [3]. Articles dealing with these issues are, among others, [29, 6, 24, 16]. A MATLAB implementation is discussed in [1].

3.9.6 Algorithms

We mainly discuss algorithmic issues connected with assembly of the stiffness matrix and load vector, see Sect. 3.9.1. An important guideline is that

> assembly should be conducted in a cell oriented fashion.

We illustrate the main principles of an efficient assembly in the case of quadratic Lagrangian finite elements used for the discretization of a second-order elliptic boundary value problem in primal form. The basic data structures for mesh representation are taken for granted, see Sect. 3.9.5.

The core assembly procedure in C++-syntax is listed as Algorithm 3.1. It is designed as a method of class `Mesh`, see Fig. 3.24, because access to global information is indispensable. Further, we assume that a method

```
int getEdgeID(const Cell &,unsigned int) const          (3.23)
```

of class `Mesh` is provided that computes the matrix/vector index corresponding to the degree of freedom located on the edge opposite to vertex K. Assuming constant complexity of all elementary function calls, we notice that

> the total computational effort is of the order $O(\sharp\mathcal{M}) = O(N)$, $N := \dim V_n$.

Remark 3.105. The method `getEdgeID` can be implemented using an associative storage scheme based on *hash tables*. The hash key can be computed from the index numbers of the endoints of the edge. After initialization the hash map is traversed once in order to set the global edge numbers.

In the above example we had to rely on a numbering of the edges in order to handle the global d.o.f. associated with them. This reflects the fact that

> assembly entails a numbering of the global shape functions/global d.o.f.

This can be achieved in several different ways:

1. Cells can be provided with information about the index numbers of their edges and faces. However their efficient initialization poses a challenge.

2. The index numbers of d.o.f. can be stored in separate data structures as the hash table of Remark 3.105.

```
void Mesh::assemble(StiffnesMatrix &B,LoadVector &f) {
  B = 0; f = 0;
  for(K=Cells().begin();K!=Cells().end();K++) {
    // Computation of local stiffness matrix and load vector
    FullMatrix BK(compLocalStiffnessMatrix(*K));
    vector<double> fK(compLocalLoadVector(*K));

    // Retrieval of row/column indices corresponding to local d.o.f.
    int nodeid[3];
    for(i=1;i<=3;i++) { nodeid[i] = (K->getNode(i)).getId(); }
    int edgeid[3];
    for(i=1;i<=3;i++) { edgeid[i] = getEdgeID(K,i); }

    // Update of contributions to global stiffness matrix and load vector
    for(i=1;i<=3;i++) {
      for(j=1;j<=3;j++) {
        B(nodeid[i],nodeid[j]) += BK(i,j);
        B(nodeid[i],edgeid[j]) += BK(i,3+j);
        B(edgeid[i],nodeid[j]) += BK(i+3,j);
        B(edgeid[i],edgeid[j]) += BK(i+3,j+3);
      }
      f(nodeid[i]) = fK[i];
      f(edgeid[i]) = fK[i+3];
    }
}
```

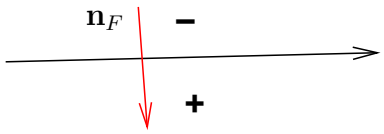
Algorithm 3.1: Assembly of finite element stiffness matrix for quadratic Lagrangian finite elements on a simplicial triangulation in two dimensions.

Exercise 3.26. On a two-dimensional simplicial mesh $\mathcal{M}$ we have computed the solution of a second-order elliptic boundary value problems by means of quadratic finite elements. The coefficients of the solution, that is, the values of the global degrees of freedom according to Def. 3.54, are stored in the vector $\boldsymbol{\mu} \in \mathbb{R}^N$.

Design a method `double Mesh::calcL2(vector<double> mu)` that computes the L^2-norm of the finite element solution. The same data structures as in Algorithm 3.1 should be used.

Exercise 3.27. We are given a two-dimensional oriented simplicial triangulation $\mathcal{M}$, stored in object oriented fashion, see Sect. 3.9.5 and, in particular, Figures 3.23 and 3.24. Each node is tagged with a unique integer indentifier (`id`), see Fig. 3.23, and an edge F is pointing from the endpoint with smaller `id` to that with larger `id`.

Let $\mathcal{F}_\Omega(\mathcal{M})$ stand for the set of interior edges of $\mathcal{M}$. For each oriented edge $F \in \mathcal{F}_\Omega(\mathcal{M})$ we distinguish its "+-side" and "−-side" as indicated in the drawing beside. The crossing direction of F is from "−" to "+". The unit vector $\mathbf{n}_F$ is supposed to point in this direction.



We consider the mesh-dependent bilinear form

$$\mathsf{b}(u,v) := \sum_{F\in\mathcal{F}(\mathcal{M})} \int_F [\langle \mathbf{grad}\, u, \mathbf{n}_F\rangle]_F \, [\langle \mathbf{grad}\, v, \mathbf{n}_F\rangle]_F \, \mathrm{d}S \;, \quad u,v \in \mathcal{S}_1(\mathcal{M}) \;,$$

where $[v]_F$ stands for the jump of a $\mathcal{M}$-piecewise continuous function v across the edge F: $[v]_F = v^+ - v^-$, v^+ being the value of v on the "+-side", v^- the value on the "−-side". If $F \subset \Gamma$, we set $[v]_F := v_{|F}$.

Devise a method

```
void jumpassemble(StiffnessMatrix &B) const
```

of the class `Mesh` that computes the stiffness matrix arising from the biliner form `b` for linear Lagrangian finite elements according to Def. 3.54. You may use information about the total number of nodes, edges, and cells of $\mathcal{M}$, provided by methods `int Mesh::getNoNodes()`, `int Mesh::getNoEdges()`, `int Mesh::getNoCells()`, and a method `int Mesh::getEdgeID(const Cell &,int)` as in (3.23) that returns a unique integer identifier $\in \{1,\ldots,\sharp\mathcal{E}(\mathcal{M})\}$ when invoked with the index numbers of the two endpoints of an edge.

3.9.7 Treatment of essential boundary conditions

For $g \in H^{1/2}(\Gamma)$ we examine the Dirichlet problem: seek $u \in H^1(\Omega)$, $\mathsf{R}_1\, u = g$ on Γ, such that

$$\int_\Omega \langle \mathbf{grad}\, u, \mathbf{grad}\, v\rangle \, \mathrm{d}\boldsymbol{\xi} = \int_\Omega f v \, \mathrm{d}\boldsymbol{\xi} \quad \forall v \in H^1_0(\Omega) \;. \tag{3.24}$$

Here, the Dirichlet boundary conditions play the role of essential boundary conditions. From Sect. 2.8 we recall that it took an extension u_g of the Dirichlet data g to arrive at the proper linear variational problem (EVP).

Let us study the Galerkin finite element discretization of (3.24) by means of a H^1-coforming finite element space V_n. Imagine V_n as Lagrangian finite element space $\mathcal{S}_m(\mathcal{M})$ built on a conforming triangulation of $\mathcal{M}$. The treatment of essential boundary conditions comprises two steps:

1. The first complication is that, probably, $g \notin \mathsf{R}_1(V_n)$. Thus, in a first step we have to replace g by $G_n \in \mathsf{R}_1(V_n)$ obtained by some *projection* onto $\mathsf{R}(V_n)$. Often, the finite element interpolation operator $\mathsf{I}(V_n)$ restricted to Γ is used. Here, "restricted" means that only global d.o.f. supported on Γ are considered in the interpolation. Of course, sufficient smoothness of g has to be assumed.

Example 3.106. If $V_n = \mathcal{S}_1(\mathcal{M})$ and $g \in C^0(\Gamma)$, we can simply perform linear interpolation based on the values of g in $\mathcal{N}(\mathcal{M}) \cap \Gamma$ in order to get g_n.

2. Given g_n we can resort to **trivial discrete extension**: we define $u_{g,n} \in V_n$ by

$$\underline{l}(u_{g,n}) := \begin{cases} \underline{l}(g_n) & \text{if } \underline{l} \text{ is supported on } \Gamma, \\ 0 & \text{otherwise.} \end{cases}$$

Example 3.107. Continuing Example 3.106, in this case $u_{g,n} \in \mathcal{S}_1(\mathcal{M})$ is just that function that agrees with G_n on Γ and vanishes on all nodes of $\mathcal{M}$ in the interior of Ω.

After these two steps the discrete variational problem corresponding to (3.24) reads: seek $u_n \in V_{n,0}$ such that

$$\int_\Omega \langle \mathbf{grad}(u_n + u_{g,n}), \mathbf{grad}\, v_n \rangle \, \mathrm{d}\boldsymbol{\xi} = \int_\Omega f v_n \, \mathrm{d}\boldsymbol{\xi} \quad \forall v \in V_{n,0} \,, \tag{3.25}$$

where $V_{n,0} := V_n \cap H_0^1(\Omega)$.

On the algebraic level, things look slightly different. The set of global shape functions is partitioned into those associated with nodes/edges/faces in Ω and those associated with nodes/edges/faces on Γ. This induces a partitioning of the final linear systems of equations

$$\begin{pmatrix} \mathbf{B}_{\Omega\Omega} & \mathbf{B}_{\Gamma\Omega} \\ \mathbf{B}_{\Omega\Gamma} & \mathbf{B}_{\Gamma\Gamma} \end{pmatrix} \begin{pmatrix} \boldsymbol{\mu}_\Omega \\ \boldsymbol{\mu}_\Gamma \end{pmatrix} = \begin{pmatrix} \boldsymbol{\varphi}_\Omega \\ \boldsymbol{\varphi}_\Gamma \end{pmatrix} \,. \tag{3.26}$$

Actually, the coefficients in $\boldsymbol{\mu}_\Gamma$ are no unknowns, because u_n is already fixed on Γ, that is, we know

$$\left(\boldsymbol{\mu}_\Gamma\right)_i = \underline{l}(g_n) \,, \tag{3.27}$$

if the i-th component of $\boldsymbol{\mu}_\Gamma$ belongs to the global degree of freedom $\underline{l}$ supported on Γ. Hence, use (3.27) to determine $\boldsymbol{\mu}_\Gamma$, which yields the reduced linear system

$$\mathbf{B}_{\Omega\Omega}\boldsymbol{\mu}_\Omega = \boldsymbol{\varphi}_\Omega - \mathbf{B}_{\Gamma\Omega}\boldsymbol{\mu}_\Gamma \,. \tag{3.28}$$

Exercise 3.28. The coefficient vector $\boldsymbol{\mu}_\Omega$ computed via (3.28) fits the solution $u_n + u_{g,n} \in V_n$ of (3.25).

3.9.8 Non-conforming triangulations

Now we consider that situation that a (closed) face of some cell of the triangulation $\mathcal{M}$ coincides with the union of several faces of other cells: the mesh $\mathcal{M}$ is no longer

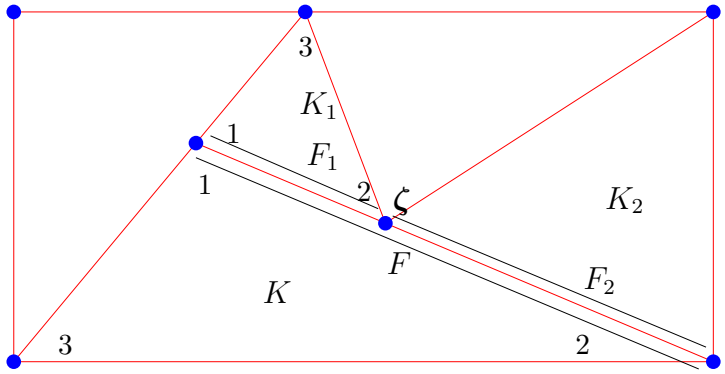


Figure 3.25: Model of a non-conforming simplicial triangulation in 2D

conforming, see Def. 3.8. As a two dimensional model case we will study the simplicial mesh of Fig. 3.25.

In Fig. 3.25 we focus on the edge F that is composed of the sub-edges F_1 and F_2. The common parlance calls F the **master edge**, whereas F_1 and F_2 are the **slave edges**. As a general rule, the global shape functions/degrees of freedom are defined with respect to the master edge.

Here we only discuss the example of second degree H^1-conforming Lagrangian finite elements on the triangulation of Fig. 3.25. This will sufficiently convey the abstract principles. We remember that for triangular Lagrangian finite elements of degree 2, see Def. 3.54, three local shape functions have non-zero restriction to an edge. If K is a triangle with vertices $\boldsymbol{\nu}^1, \boldsymbol{\nu}^2, \boldsymbol{\nu}^3$ the shape functions associated with the edge connecting $\boldsymbol{\nu}^1$ and $\boldsymbol{\nu}^2$ have the barycentric representations

$$b_1 = \lambda_1(1-2\lambda_2) \quad , \quad b_{12} = 4\lambda_1\lambda_2 \quad , \quad b_2 = \lambda_2(1-2\lambda_1) \; .$$

Here, b_1 is associated with $\boldsymbol{\nu}^1$, b_{12} with the midpoint of the edge, and b_2 belongs to $\boldsymbol{\nu}^2$. In Fig. 3.26 the traces of these shape function on an edge are depicted.

The restriction of the global shape functions to the master edge F should either be zero or agree with one of the three three local shape functions from K. This rules out using the standard shape functions on the cells K_1 and K_2, which are adjacent to F from the "slave side". The mismatch is illustrated in Fig. 3.27.

The local shape functions on the cells K_1 and K_2 have to be modified. Let $\{b_1^1, b_{12}^1, b_2^1\}$ be the set of standard local shape functions on K_1 associated with the endpoints and midpoint of F_1. The local shape functions on K that belong to $\overline{F}$ are b_1, b_{12}, b_2 If we assume that the point $\boldsymbol{\zeta}$ splits F with a ratio of $1:2$ ($|F_1| = 1/3|F|$), then, on F_1,

$$\begin{aligned} b_{1|F_1} &= 1 \cdot b_1^1 + 5/9 \cdot b_{12} + 2/9 \cdot b_{12}; , \\ b_{12|F_1} &= 0 \cdot b_1^1 + 5/9 \cdot b_{12} + 8/9 \cdot b_2 \; , \\ b_{2|F_1} &= 0 \cdot b_1^1 - 1/9 \cdot b_{12} - 1/9 \cdot b_2^1 \; . \end{aligned}$$

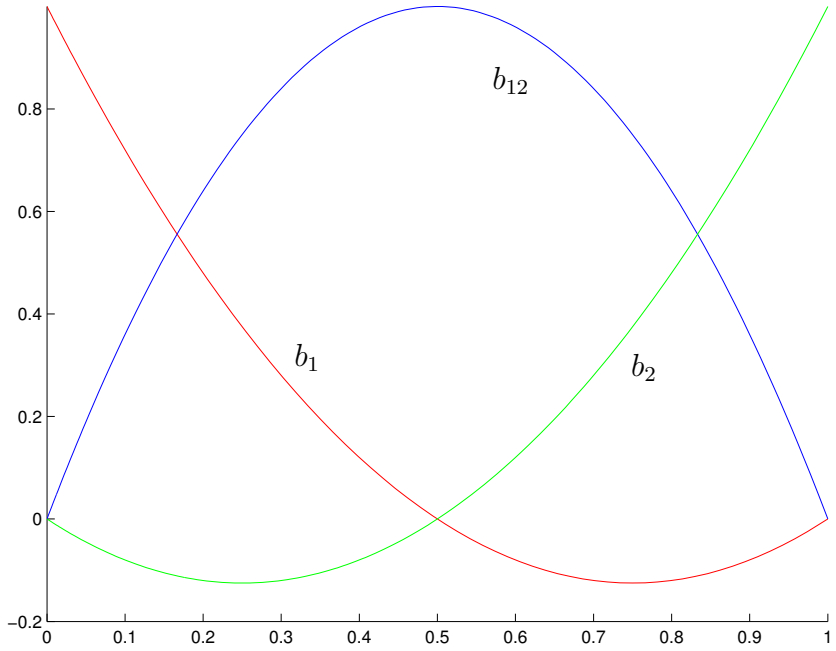


Figure 3.26: Traces of second degree Lagrangian shape functions on an edge, which is parameterized over $[0, 1]$.

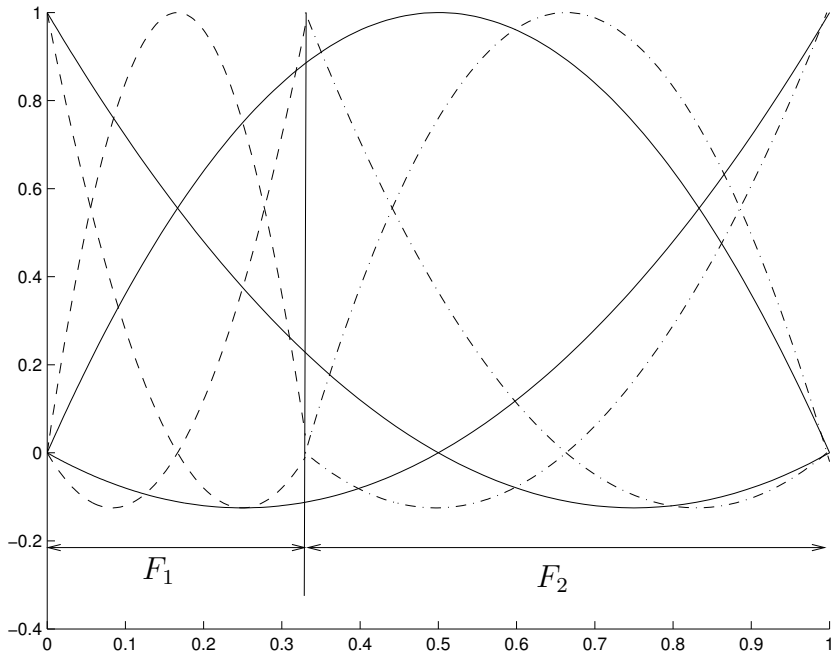


Figure 3.27: Local shape functions on edge F both from master cell (solid line) and slave cells (dashed / dotted line)

That is, to achieve a matching of local shape functions across F_1, the standard local shape functions on K_1 have to be replaced by suitable linear combinations. The coefficients are given by the above relationships.

We assume the numbering of local shape functions for quadratice Lagrangian finite elements that has already been used in Algorithm 3.1: the basis functions associated with the vertices will come first, then those associated with the midpoints of edges. Moreover, we use the numbering of vertices of K_1 indicated in Fig. 3.25. Then the re-combination of local shape functions can formally be expressed through the equation

$$\begin{pmatrix} \widetilde{b}_1 \\ \widetilde{b}_2 \\ \widetilde{b}_3 \\ \widetilde{b}_{23} \\ \widetilde{b}_{31} \\ \widetilde{b}_{12} \end{pmatrix} = \begin{pmatrix} 1 & 2/9 & 0 & 0 & 0 & 5/9 \\ 0 & -1/9 & 0 & 0 & 0 & -1/9 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 8/9 & 0 & 0 & 0 & 5/9 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_{23} \\ b_{31} \\ b_{12} \end{pmatrix} . \tag{3.29}$$

The new local shape functions are $\{\widetilde{b}_1, \widetilde{b}_2, \widetilde{b}_3, \widetilde{b}_{23}, \widetilde{b}_{31}, \widetilde{b}_{12}\}$.

The matrix in (3.29) is called **S-matrix** of K_1. The S-matrix $\mathbf{S}_K$ relates the restrictions of global shape functions on a cell $K \in \mathcal{M}$ to the standard local shape functions that are used to compute the element stiffness matrix $\mathbf{B}_K$ and element load vector $\boldsymbol{\varphi}_K$. In the general case $\mathbf{S}_K$ has dimension $k_K \times k_K$, $k_K := \dim \Pi_K$. The S-matrix differs from the identity only for those cells that are adjacent to the "slave side" of a face/edge.

Lemma 3.108. *For a non-conforming triangulation the formulas of Thm. 3.85 have to be altered into*

$$\mathbf{B} = \sum_{K \in \mathcal{M}} \mathbf{T}_K^T \mathbf{S}_K \mathbf{B} \mathbf{S}_K^T \mathbf{T}_K \quad , \quad \boldsymbol{\varphi} = \sum_{K \in \mathcal{M}} \mathbf{T}_K^T \mathbf{S}_K \boldsymbol{\varphi}_K \, . \tag{3.30}$$

Proof. Tagging the restrictions of global shape functions to $K \in \mathcal{M}$ by a tilde we have

$$\widetilde{b}_i^K = \sum_{l=1}^{k_K} s_{il} \, b_l^K \, .$$

Write $\widetilde{\mathbf{B}}_K$ for the element stiffness matrix computed with respect to the new local shape functions $\widetilde{b}_i^K$. Then

$$\big(\widetilde{\mathbf{B}}\big)_{ij} = \mathsf{b}(\widetilde{b}_j^K, \widetilde{b}_i^K) = \sum_{l=1}^{k_K} \sum_{m=1}^{k_K} s_{il} s_{jm} \mathsf{b}(b_m, b_l) = \sum_{l=1}^{k_K} \sum_{m=1}^{k_K} s_{il} s_{jm} \big(\mathbf{B}_K\big)_{lm} \, ,$$

which means

$$\widetilde{\mathbf{B}}_K = \mathbf{S}_K \mathbf{B}_K \mathbf{S}_K^T \, .$$

□

Exercise 3.29. Consider degree 2 Lagrangian finite elements on the triangulation depicted in Fig. 3.25. Compute the S-Matrix for cell K_2.

Exercise 3.30. Face elements according to Def. 3.63 are to be used on the non-conforming triangulation from Fig. 3.25. Determine the S-Matrix belonging to cell K_1 in this case.

3.9.9 Static condensation

There may be global shape functions whose supports coincide with the closure of a cell of the mesh. For instance, Lagrangian finite elements of degree m, $m \in \mathbb{N}$, on a two-dimensional simplicial triangulation feature those if $m \geq 3$. Sometimes, these global shape functions are called **interior basis functions**.

If the bilinear form of the variational problem has property (LOC), then all entries of the stiffness matrix corresponding to pairs of interior basis functions belonging to different cells will vanish. If we sort the global shape functions in way that counts the interior basis functions last, then the linear system of equations resulting from finite element Galerkin discretization will have the following block structure

$$\mathbf{B}\boldsymbol{\mu} = \begin{pmatrix} \mathbf{B}_{oo} & \mathbf{B}_{oi} \\ \mathbf{B}_{io} & \mathbf{B}_{ii} \end{pmatrix} \begin{pmatrix} \boldsymbol{\mu}_o \\ \boldsymbol{\mu}_i \end{pmatrix} = \begin{pmatrix} \boldsymbol{\varphi}_o \\ \boldsymbol{\varphi}_i \end{pmatrix} = \boldsymbol{\varphi} \,, \tag{3.31}$$

where $\mathbf{B}_{ii}$ is due to the coupling among interior basis functions. The important observation is that $\mathbf{B}_{ii}$ is *block-diagonal* with blocks corresponding to the cells of the mesh. Each block will be of a size equal to the number of interior basis functions of the related cell. For instance, for Lagrangian finite elements with fixed degree m this number will be relatively small.

Thus, the block-diagonal matrix $\mathbf{B}_{ii}$ can be inverted with a numerical effort proportional to $\sharp\mathcal{M}$. Thus (3.31) is converted into the smaller **Schur-complement** system

$$\left(\mathbf{B}_{oo} - \mathbf{B}_{oi}\mathbf{B}_{ii}^{-1}\mathbf{B}_{io}\right)\boldsymbol{\mu}_o = \boldsymbol{\varphi}_o - \mathbf{B}_{oi}\mathbf{B}_{ii}^{-1}\boldsymbol{\varphi}_i \,. \tag{3.32}$$

Due to its reduced size, this system may be easier to solve than the full system (3.31).

Exercise 3.31. Consider Lagrangian finite elements of degree 4 on a very fine two-dimensional quadrilateral triangulation $\mathcal{M}$ of the kind depicted in Fig. 3.5 (right). "Very fine" means that cells adjacent to Γ need not be taken into account in the considerations. By what percentage can static condensation reduce the number of unknowns?

3.10 Spectral H^1-conforming elements

According to Thm. 1.30 enlarging the finite element space promises to yield a better Galerkin solution w.r.t. the V-norm. There are two basic ways to "enlarge" a Lagrangian finite element space:

1. the **h-version** of finite elements uses Lagrangian finite elements of the same degree on a mesh with more cells ("finer mesh")

2. the **p-version** of finite elements keeps the mesh, but raises the degree m of the Lagrangian finite elements. This leads to **spectral approximations**.

The **hp-version** of finite elements blends both ideas.

Whenever higher order Lagrangian finite elements are used it turns out to be crucial to pick appropriate local ($\rightarrow$ global) shape functions in order to obtain a linear system of equations (LSE) with a reasonably conditioned matrix (see Sect. 1.5). Actually, Def. 3.54 determines the local shape functions, but for $m > 1$ alternative local degrees of freedom/local shape functions are possible that will yield the same space $\mathcal{S}_m(\mathcal{M})$, *cf.* Remark 3.32.

Example 3.109. Consider the Lagrangian finite element of degree two on a triangle K. Instead of the shape functions from Table 3.2 we can also use the **p-hierarchical local shape functions**. Using barycentric coordinates in K, see Def. 3.49, they can be written as

$$\{b_i = \lambda_i,\ i = 1,2,3,\ 4\lambda_i\lambda_j,\ 1 \le i < j \le 3\}\ . \tag{3.33}$$

This basis is called hierarchical because a subset is a basis of $\mathcal{P}_1(K)$: the entire basis emerges by augmenting the set of shape functions for the first degree Lagrangian finite element on K. Obviously the shape functions match across interelement faces (dual view according to Remark 3.38) and will yield $\mathcal{S}_m(\mathcal{M})$.

Exercise 3.32. Determine local d.o.f. on $\mathcal{P}_1(K)$ in the form of combinations of point evaluations that fit the local shape functions from (3.33).

Example 3.109 teaches us how the set of local shape functions of a Lagrangian finite element (K, Π_K, Σ_K) can be altered without affecting the matching property across interelement faces:

> A local shape function supported on a face/edge/vertex F of K can be modified by adding contributions from other shape functions that are supported on a face/edge F' such that $F \subset \overline{F}'$.

However, the number of local d.o.f./local shape functions associated with a vertex/edge/face/cell must always remain the same.

Example 3.110. Consider cubic Lagrangian finite elements on a triangle K, that is, $\Pi_K = \mathcal{P}_3(K)$, see Def. 3.54. Denote by b_i^K the local shape function associated with the vertex i, $i = 1,2,3$. By b_{iji}^K, b_{ijj}^K, $1 \le i < j \le 3$, we mean the local shape function related to the point in $\mathcal{N}$ on the edge connecting vertices i and j that is closer to vertex i. Finally, b_{123}^K will stand for the interior local shape function.

As in (3.29) we can use a matrix notation to describe a change of local shape functions. The above rule will mean that any new set $\{\widetilde{b}_i^K, \widetilde{b}_{iji}^K, \widetilde{b}_{ijj}^K, \widetilde{b}_{123}^K\}$ will be generated by

$$\begin{pmatrix} \widetilde{b}_1^K \\ \widetilde{b}_2^K \\ \widetilde{b}_3^K \\ \widetilde{b}_{121}^K \\ \widetilde{b}_{122}^K \\ \widetilde{b}_{131}^K \\ \widetilde{b}_{133}^K \\ \widetilde{b}_{232}^K \\ \widetilde{b}_{233}^K \\ \widetilde{b}_{123}^K \end{pmatrix} = \begin{pmatrix} * & 0 & 0 & * & * & * & * & * & * & * \\ 0 & * & 0 & * & * & * & * & * & * & * \\ 0 & 0 & * & * & * & * & * & * & * & * \\ 0 & 0 & 0 & * & * & 0 & 0 & 0 & 0 & * \\ 0 & 0 & 0 & * & * & 0 & 0 & 0 & 0 & * \\ 0 & 0 & 0 & 0 & 0 & * & * & 0 & 0 & * \\ 0 & 0 & 0 & 0 & 0 & * & * & 0 & 0 & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & * & * & * \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & * \end{pmatrix} \begin{pmatrix} b_1^K \\ b_2^K \\ b_3^K \\ b_{121}^K \\ b_{122}^K \\ b_{131}^K \\ b_{133}^K \\ b_{232}^K \\ b_{233}^K \\ b_{123}^K \end{pmatrix},$$

where $*$ indicates potentially non-zero entries of the matrix. Besides, the transformation matrix has to be regular.

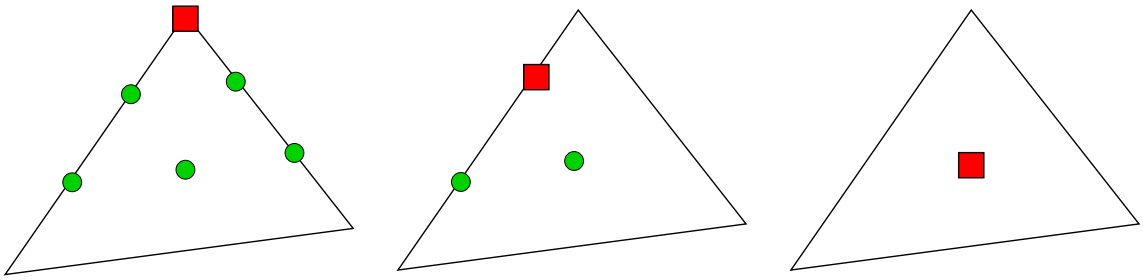


Figure 3.28: Possible modifications of local shape functions for Lagrangian finite elements of degree 3. The disks indicate local shape function that can contribute to a modification of the local shape functions marked by a square.

We have seen that for higher degree Lagrangian finite elements we have ample freedom to choose local and global shape functions. In light of the insights gained in Sect. 1.5, we cannot hope to influence the discretization error by changing the basis, but we can hope to "improve" the stiffness matrix. The usual objective is to

achieve a small condition number of the stiffness matrix.

Thus, the "optimal" set of global shape functions will depend on the variational problem that we want to tackle.

Usually the task of finding an optimal set of global shape functions is impossible to solve. Instead one tries to reduce the condition number of the element stiffness matrices.

Remark 3.111. We examine the case of a symmetric positive definite bilinear form b that is composed of local contributions according to (3.12). We apply a Galerkin finite element discretization and obtain the symmetric positive definite stiffness matrix $\mathbf{B}$, which, by Thm. 3.85 has a representation

$$\mathbf{B} = \sum_{K\in\mathcal{M}} \mathbf{T}_K^T \mathbf{B}_K \mathbf{T}_K \ .$$

Further, assume that at most D, $D \in \mathbb{N}$, cells of a mesh $\mathcal{M}$ share a node. This implies

$$|\boldsymbol{\eta}|^2 \le \sum_{K\in\mathcal{M}} |\mathbf{T}_K\boldsymbol{\eta}|^2 \le D|\boldsymbol{\eta}|^2 \quad \forall \boldsymbol{\eta} \in \mathbb{R}^N \ ,$$

because each coefficient of $\boldsymbol{\eta}$ will occur at most D times as an entry of one of the vectors $\mathbf{T}_K\boldsymbol{\eta}$. Hence,

$$\begin{aligned}
\boldsymbol{\eta}^T\mathbf{B}\boldsymbol{\eta} &= \sum_{K\in\mathcal{M}} \big(\mathbf{T}_K\boldsymbol{\eta}\big)^T \mathbf{B}_K \big(\mathbf{T}_K\boldsymbol{\eta}\big) \\
&\le \max_K \ \lambda_{\max}(\mathbf{B}_K) \sum_{K\in\mathcal{M}} |\mathbf{T}_K\boldsymbol{\eta}|^2 \le D \ \max_K \ \lambda_{\max}(\mathbf{B}_K)|\boldsymbol{\eta}|^2 \ , \\
&\ge \min_K \ \lambda_{\min}(\mathbf{B}_K) \sum_{K\in\mathcal{M}} |\mathbf{T}_K\boldsymbol{\eta}|^2 \le \min_K \ \lambda_{\min}(\mathbf{B}_K)|\boldsymbol{\eta}|^2 \ .
\end{aligned}$$

As a consequence, the spectral condition number (see Sect. 1.5) satisfies

$$\kappa(\mathbf{B}) \le \frac{D \ \max_K \lambda_{\max}(\mathbf{B}_K)}{\min_K \lambda_{\min}(\mathbf{B}_K)} \ .$$

This shows the close relationship between the condition number of the stiffness matrix and the extremal eigenvalues of the element stiffness matrices.

Remark 3.112. For the p-version of Lagrangian finite elements static condensation (see Sect. 3.9.9) becomes an essential tool. Besides significantly reducing the size of the linear system of equations, static condensation also brings about a considerable improvement of the conditioning of the linear system.

Exercise 3.33. Consider the second degree Lagrangian finite element on an equilateral triangle K. Determine a modified admissible set of local shape functions that achieves optimal spectral conditioning of the element stiffness matrix with respect to the bilinear form

$$(u,v) \mapsto \int_K \langle \mathbf{grad}\, u, \mathbf{grad}\, v\rangle \ \mathrm{d}\boldsymbol{\xi} \ .$$

Of course, the local shape functions must remain compliant with H^1-conformity. Moreover, one may always assume that the new shape functions remain invariant with respect to permutations of the vertices of the triangle.

For tensor product cells like $\widehat{K} =]-1,1[^2$ local shape functions for Lagrangian finite elements of degree m, $m \in \mathbb{N}$, can be obtained by tensor product construction from bases of $\mathcal{P}_m(]-1,1[)$, *cf.* Remark 3.58.

Example 3.113. A **nodal basis** of $\mathcal{P}_m(]-1,1[)$ is fixed by specifying $m+1$ "nodes"

$$-1 =: \widehat{\xi}_0 < \widehat{\xi}_1 < \cdots < \widehat{\xi}_m := 1 \,.$$

Then the nodal (or Lagrange) shape functions are given by

$$\widehat{b}_j^m(\xi) := \frac{\prod_{i=0}^{m} (\xi - \widehat{\xi}_i)}{\prod_{\substack{i=0 \\ i\neq j}}^{p} (\widehat{\xi}_i - \widehat{\xi}_j)} \,. \tag{3.34}$$

Evidently, $\widehat{b}_j^m \in \mathcal{P}_m(]-1,1[)$ and $\widehat{b}_j^m(\widehat{\xi}_i) = \delta_{ij}$ i.e. $\widehat{b}_j^m(\xi)$ vanishes in all nodes except node number j.

Many choices of nodes are possible:

- **equidistant nodes:**

$$\widehat{\xi}_j^m := -1 + \frac{2j}{m}, \quad j = 0, \ldots, m \,. \tag{3.35}$$

  These are not recommended for high m because the element stiffness matrices for bilinear forms pertaining to $H^1(\Omega)$-elliptic problems will become extremely ill-conditioned.

- **Chebysev nodes:**

$$\widehat{\xi}_j^m := -\cos\Big(j\,\frac{\pi}{m}\Big) \quad j = 0, 1, \ldots, m \,. \tag{3.36}$$

  This option leads to reasonable well-conditioned element stiffness matrices for variational problems in $H^1(\Omega)$.

- **Lobatto nodes:** $\widehat{\xi}_j^m$ are the $m+1$ zeros of

$$(1-\xi^2)\, L'_m(\xi) \,, \tag{3.37}$$

  where $L_p(\xi)$ is the p-th Legendre polynomial. These points (which are used in spectral methods) are as good as (3.36), but, in addition, they are suitable for numerical integration.

Please note that all these local shape functions are not p-hierarchical, because when m is increased they all change. This necessitates a complete re-computation of all element matrices, when the polynomial degree of the Lagrange finite elements is raised by only one.

Example 3.114. The shape functions for $\mathcal{P}_m(]-1,1[)$ can also be chosen to yield a p-hierarchical basis, see Example 3.109. For instance, they can be constructed from the Legendre polynomials $\{L_p\}_{p=0}^{\infty}$. They are obtained by applying the Gram-Schmidt process in $L^2(-1,1)$ to the nomials $\{\xi^p\}_{p=0}^{\infty}$. For classical reasons, the $\{L_p(\xi)\}$ are normalized by the condition

$$L_p(1) = 1\,. \tag{3.38}$$

Legendre polynomials satisfy the **Legendre differential equation**

$$\big((1-\xi^2)L_p'(\xi)\big)' + p(p+1)L_p(\xi) = 0 \text{ in }]-1,1[\tag{3.39}$$

and the **orthogonality**

$$\int_{-1}^{1} L_n(\xi)L_m(\xi)d\xi = \begin{cases} \frac{2}{2n+1} & \text{if } n = m \\ 0 & \text{else}\,. \end{cases} \tag{3.40}$$

Also

$$L_p(\xi) = \frac{L_{p+1}'(\xi) - L_{p-1}'(\xi)}{2p+1}\,,\quad p \geq 1\,. \tag{3.41}$$

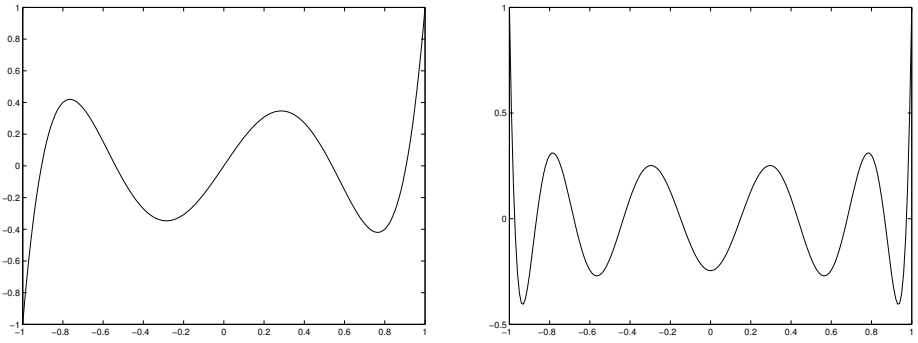


Figure 3.29: Legendre polynomials of degree 5 and 10.

A p-hierarchic sequence of shape functions $\widehat{b}_j(\xi)$ on $]-1,1[$ is then given by

$$\widehat{b}_0(\xi) = \frac{1+\xi}{2} \quad,\quad \widehat{b}_1(\xi) = \frac{1-\xi}{2} \tag{3.42}$$

$$\widehat{b}_j(\xi) = \sqrt{\frac{2j-1}{2}} \int_{-1}^{\xi} L_{j-1}(\eta)d\eta, \quad j \geq 2\,. \tag{3.43}$$

This means that

$$\widehat{b}_0' = 1\,,\ \widehat{b}_1' = -1\,,\ \widehat{b}_j' = \sqrt{\frac{2j-1}{2}}\, L_{j-1},\ j > 1\,.$$

Moreover, from (3.40) we conclude

$$\int_{-11}^{1} p(\xi) L_j(\xi)\,\mathrm{d}\xi = 0 \quad \forall p \in \mathcal{P}_{j-1}(]-1,1[)\ .$$

This means that the element matrix arising from the local shape functions (3.42), (3.43) and the bilinear form

$$(u,v) \mapsto \int_{-1}^{1} u'(\xi)\cdot v'(\xi)\,\mathrm{d}\xi$$

reads

$$\mathbf{B}_{]-1,1[} = \begin{pmatrix} 1/2 & -1/2 & 0 & 0 & \cdots & 0 \\ -1/2 & 1/2 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & & & \ddots & & \vdots \\ \vdots & & & & \ddots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$$

Bibliographical notes. The gist and foundations of spectral methods are discussed in [7]. A complete treatment of the hp-version of Lagrangian finite elements is given in [36].

4 Basic Finite Element Theory

In this chapter we focus on V-elliptic linear variational problems (see Sect. 1.2 and Def. 1.20) and their finite element Galerkin discretization.

> The main objective of finite element theory is to predict the dependence of certain norms of the finite element discretization error on discretization parameters like features of the mesh and the type of finite elements (**a-priori error estimate**, see Sect. 1.4).

We will mainly discuss these issues for the following $H_0^1(\Omega)$-elliptic model problem that arises as the primal variational formulation of the pure homogeneous Dirichlet problem for a second-order elliptic boundary value problem, see Sect. 2.8: seek $u \in H_0^1(\Omega)$ such that

$$\mathsf{b}(u,v) := \int_\Omega \langle \mathbf{A}\,\mathbf{grad}\,u, \mathbf{grad}\,v\rangle \,\mathrm{d}\boldsymbol{\xi} = \int_\Omega fv\,\mathrm{d}\boldsymbol{\xi} \quad \forall v \in H_0^1(\Omega)\,. \tag{4.1}$$

Here, $\Omega \subset \mathbb{R}^d$, $d = 1,2,3$, is a computational domain according to Def. 2.5. The coefficient function $\mathbf{A} \in (L^\infty(\Omega))^{d,d}$ is supposed to be uniformly positive definite, see (UPD), whereas $f \in L^2(\Omega)$. $H^1(\Omega)$-Ellipticity of b is a consequence of Lemma 2.61.

4.1 The Bramble-Hilbert lemma

We start with a fundamental concept of functional analysis:

Definition 4.1. *An linear operator $\mathsf{K} : V \mapsto W$, V, W normed vector spaces, is called **compact**, if the image of any bounded sequence in V contains a sub-sequence that converges in W. The set of compact operators $V \mapsto W$ is denoted by $K(V,W)$.*

Corollary 4.2. *For Banach spaces U, V, W we have*

(i) $K(V,W) \subset L(V,W)$,

(ii) If $\mathsf{K} \in K(U,V)$, $\mathsf{S} \in L(V,W)$, *then* $\mathsf{S} \circ \mathsf{K} \in K(U,W)$.

Compact operators are a generalization of the simple operators that have a finite-dimensional image space (range).

Theorem 4.3. *For Banach spaces V, W the space of compact operators is the closure of the set of linear operators $V \mapsto W$ with finite dimensional range in $L(V, W)$.*

In infinite-dimensional spaces compact operators can never be invertible.

Theorem 4.4. *An isomorphism $\mathsf{T} : V \mapsto W$ of Banach spaces V, W is compact, if and only if $\dim V = \dim W < \infty$.*

Theorem 4.5 (Peetre-Tartar lemma). *Let U, V, W Banach spaces, $\mathsf{K} \in K(V, W)$, $\mathsf{S} \in L(V, U)$. If*

$$\exists \gamma > 0 : \quad \|v\|_V \leq \gamma \big(\|\mathsf{K}\, v\|_W + \|\mathsf{S}\, v\|_U \big) \quad \forall v \in V \ , \tag{4.2}$$

then

(i) $\dim \mathrm{Ker}(\mathsf{S}) < \infty$,

(ii) $\exists \gamma' > 0 : \quad \inf_{p \in \mathrm{Ker}(\mathsf{S})} \|v - p\|_V \leq \gamma' \, \|\mathsf{S}\, v\|_U \quad \forall v \in V.$

Proof. (i): On $N := \mathrm{Ker}(\mathsf{S})$ we have

$$\|v\|_V \leq \gamma \ \|\mathsf{K}\, v\|_W \quad \forall v \in N \ ,$$

that is, $\mathsf{K}_{|N} : N \mapsto \mathsf{K}(N)$ is an isomorphism. On the other hand, $\mathsf{K}_{|N} \in K(N, \mathsf{K}(N))$, which, by Thm. 4.4, can only be satisfied, if $\dim N < \infty$.

(ii): Assume that the assertion was false. Then there would exist a sequence $\{v_l'\}_{l=1}^{\infty} \subset V/N$ such that

$$\|v_l'\|_{V/N} \geq l \ \|\mathsf{S}\, v_l'\|_U \quad \forall l \in \mathbb{N} \, .$$

By rescaling we conclude the existence of a sequence $\{v_l'\}_{l=1}^{\infty} \subset V/N$ — for convenience notations have not been changed — such that

$$\|v_l'\|_{V/N} = 1 \quad \text{and} \quad \lim_{l \to \infty} \mathsf{S}\, v_l' = 0 \, .$$

We have

$$\|v\|_{V/N} = \inf_{p \in N} \|v + p\|_V \ ,$$

and since $\dim N < \infty$ the infimum is attained. This means that we can find a sequence $\{v_l\}_{l=1}^{\infty} \subset V$ such that

$$\|v_l\|_V = 1 \quad \text{and} \quad \lim_{l \to \infty} \|\mathsf{S}\, v_l\|_U = 0 \, .$$

Therefore, we can extract a subsequence $\{v_k\}_{k=1}^{\infty}$ such that $\{\mathsf{K}\, v_k\}_{k=1}^{\infty}$ converges in W. By (4.2) this will be a Cauchy sequence in V. As a consequence

$$u := \lim_{k\to\infty} v_k \in V \quad \text{and} \quad u \in N \;\Rightarrow\; \|u\|_{V/N} = 0 \,.$$

This is a contradiction, because we assumed $\|v_k\|_{V/N} = 1$. □

This sophisticated theorem is extremely useful for Sobolev spaces, because their imbeddings offer a wealth of compact operators.

Theorem 4.6 (Rellich's theorem). *Let $\Omega \subset \mathbb{R}^d$ a bounded domain with Lipschitz-continuous boundary. If $m, m' \in \mathbb{N}$, $m > m'$, then the continuous embedding $H^m(\Omega) \subset H^{m'}(\Omega)$ is also compact.*

One consequence of this is that under the assumptions on Ω made in the theorem, any bounded sequence in $H^1(\Omega)$ has a sub-sequence that converges in $L^2(\Omega)$. We have used this in the proof of Lemma 2.64.

The combination of Thm. 4.5 and Thm. 4.6 yields an extremely useful tool for finite element analysis. Given a bounded Lipshitz-domain Ω ,we apply Thm. 4.5 with

- $V = H^m(\Omega)$, $m \in \mathbb{N}$, $U = (L^2(\Omega))^l$, $l = \binom{m+d}{d}$, $W = H^{m-1}(\Omega)$,
- K = the embedding $H^m(\Omega) \mapsto H^{m-1}(\Omega)$, which is compact according to Thm. 4.6,
- and $\mathsf{S} := (\partial^{\boldsymbol{\alpha}})_{|\boldsymbol{\alpha}|=m} : V \mapsto U$,

Then, the concrete form of (4.2) is straightforward

$$\|v\|_{H^m(\Omega)} = \left(\|v\|^2_{H^{m-1}(\Omega)} + \|\mathsf{S}\, v\|^2_{L^2(\Omega)}\right)^{1/2} \le \left(\|v\|_{H^{m-1}(\Omega)} + \|\mathsf{S}\, v\|_{L^2(\Omega)}\right) \quad \forall v \in H^m(\Omega) \,.$$

The assertion (i) is clear because

$$\mathrm{Ker}(\mathsf{S}) = \mathcal{P}_{m-1}(\Omega) \,.$$

In this setting Thm. 4.5 has the following consequence:

Lemma 4.7 (Bramble-Hilbert lemma). *If $\Omega \subset \mathbb{R}^d$ is a bounded Lipshitz-domain and $m \in \mathbb{N}$, then*

$$\exists \gamma = \gamma(m, \Omega) > 0 : \qquad \inf_{p \in \mathcal{P}_{m-1}(\Omega)} \|v - p\|_{H^m(\Omega)} \le \; \gamma\, |v|_{H^m(\Omega)} \quad \forall v \in H^m(\Omega) \,.$$

In other words, the norm on the quotient space $H^m(\Omega)/\mathcal{P}_{m-1}(\Omega)$ is equivalent to the seminorm $|\cdot|_{H^m(\Omega)}$.

Exercise 4.1. Demonstrate how Lemma 2.64 can be obtained by a straightforward application of Lemma 4.7.

There is also a version of the Bramble-Hilbert lemma for polynomials of total degree $< m$. It is based on the estimate

$$\exists \gamma = \gamma(m, \Omega) > 0 : \quad \|v\|_{H^m(\Omega)} \leq \gamma \left(\|v\|_{H^{m-1}(\Omega)} + \left(\sum_{i=1}^{d} \left\| \frac{\partial^m v}{\partial \xi_i^m} \right\|_{L^2(\Omega)}^2 \right)^{1/2} \right) ,$$

which can be found in [38]. Thus, we can apply Thm. 4.5 with

- $V = H^m(\Omega)$, $U = (L^2(\Omega))^d$, $W = H^{m-1}(\Omega)$,
- K as the embedding $H^m(\Omega) \mapsto H^{m-1}(\Omega)$, which is compact according to Thm. 4.6,
- and $\mathsf{S} = (\partial^{(m,0,\ldots,0)}, \partial^{(0,m,0,\ldots,0)}, \ldots, \partial^{(0,\ldots,0,m)})^T$.

Obviously, we now have

$$\mathrm{Ker}(\mathsf{S}) = \mathcal{Q}_{m-1}(\Omega) ,$$

which makes Thm. 4.5 yield the following result:

Lemma 4.8. *If $\Omega \subset \mathbb{R}^d$ is a bounded Lipshitz-domain and $m \in \mathbb{N}$, then*

$$\exists \gamma = \gamma(m, \Omega) > 0 : \quad \inf_{p \in \mathcal{Q}_{m-1}(\Omega)} \|v - p\|_{H^m(\Omega)} \leq \gamma \left(\sum_{i=1}^{d} \left\| \frac{\partial^m v}{\partial \xi_i^m} \right\|_{L^2(\Omega)}^2 \right)^{1/2} \quad \forall v \in H^m(\Omega) .$$

Remark 4.9. The above reasonings can immediately be adopted for Sobolev spaces based on $L^p(\Omega)$, $p \geq 1$, that are no longer Hilbert spaces. For instance, as an analogue to Lemma 4.7 we get that on a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$

$$\exists \gamma = \gamma(m, \Omega) > 0 : \quad \inf_{p \in \mathcal{P}_{m-1}(\Omega)} \|v - p\|_{W^{m,\infty}(\Omega)} \leq \gamma \, |v|_{W^{m,\infty}(\Omega)} \quad \forall v \in W^{m,\infty}(\Omega) ,$$

where

$$\|v\|_{W^{m,\infty}(\Omega)} := \max_{|\boldsymbol{\alpha}| \leq m} \|\partial^{\boldsymbol{\alpha}} v\|_{L^\infty(\Omega)} ,$$
$$|v|_{W^{m,\infty}(\Omega)} := \max_{|\boldsymbol{\alpha}| = m} \|\partial^{\boldsymbol{\alpha}} v\|_{L^\infty(\Omega)} .$$

For $W^{m,\infty}(\Omega)$ a version of Lemma 4.8 exists, too.

4.2 Transformation techniques

The transformation approach has already proved useful for the construction of finite elements, see Sect. 3.7, and for the evaluation of element stiffness matrices, see Example 3.101. It will also play a pivotal role in finite element theory for parametric families of finite elements.

Our goal is to prove *asymptotic estimates for localized (quasi-)interpolation operators in the h-version of finite elements*. These are estimates of the following form

$$\|u - \mathsf{I}\, u\|_X \leq \gamma\, \epsilon(h)\, \|u\|_Y \ , \tag{4.3}$$

where $\|\cdot\|_X, \|\cdot\|_Y$ are suitable norms, $\gamma > 0$ is a constant that may only depend on the (continuous) boundary value problem and controllable properties of the underlying mesh, ϵ is a "simple" function tending to zero as its argument $\to 0$, and h stands for the *meshwidth*, see below. For the sake of simplicity we will assume that all cells of the underlying finite element mesh arise from a single reference cell $\widehat{K}$.

Usually, transformation techniques aiming at (4.3) involve the following steps:

(I) Localization: express $\|u - \mathsf{I}\, u\|_X$ by means of contributions of cells $K \in \mathcal{M}$ or, at least, neighborhoods of cells.

(II) Perform natural pullback $\mathsf{XT}_{\mathbf{\Phi}_K}$ of $u - \mathsf{I}\, u_{|K}$, $K \in \mathcal{M}$, to the reference cell $\widehat{K}$, $K = \mathbf{\Phi}_K(\widehat{K})$, and establish a relationship between $\|\mathsf{XT}_{\mathbf{\Phi}_K}(u - \mathsf{I}\, u)\|_{X,\widehat{K}}$ and $\|u - \mathsf{I}\, u\|_{X,K}$.

(III) Derive an estimate of the form

$$\exists \widehat{\gamma} > 0 : \quad \left\| \widehat{u} - \widehat{\mathsf{I}}\widehat{u} \right\|_{X,\widehat{K}} \leq \widehat{\gamma}\, \|\widehat{u}\|_{Y,\widehat{K}} \quad \forall u \ ,$$

where $\widehat{\mathsf{I}} := \mathsf{XT}_{\mathbf{\Phi}_K} \circ \mathsf{I}_{|K}$.

(IV) Determine the impact of the transformation on the Y-norm.

These steps can be summarized by the chain of estimates

$$\begin{aligned} \|u - \mathsf{I}\, u\|_X^2 &\overset{\text{(I)}}{=} \sum_{K \in \mathcal{M}} \left\| (u - \mathsf{I}\, u)_{|K} \right\|_{X,K}^2 \\ &\overset{\text{(II)}}{\leq} \sum_{K \in \mathcal{M}} \gamma_K\, \|\mathsf{XT}_{\mathbf{\Phi}_K}(u - \mathsf{I}\, u)\|_{X,\widehat{K}}^2 \\ &\overset{\text{(III)}}{\leq} \widehat{\gamma} \sum_{K \in \mathcal{M}} \gamma_K\, \|\mathsf{XT}_{\mathbf{\Phi}_K}(u)\|_{Y,\widehat{K}}^2 \\ &\overset{\text{(IV)}}{\leq} \widehat{\gamma} \sum_{K \in \mathcal{M}} \gamma_K \gamma_K'\, \|u\|_{Y,K}^2 \ . \end{aligned}$$

Here the "constants" γ_K, γ'_K will depend on the norms and the shape of K.

First we investigate the behavior of standard Sobolev seminorms $|\cdot|_{H^m(K)}$, $m \in \mathbb{N}_0$, under the standard pullback $\mathsf{FT}_{\mathbf{\Phi}}$, where, for the sake of simplicity, $\mathbf{\Phi} : \widehat{K} \mapsto K$ is an affine mapping $\mathbb{R}^d \mapsto \mathbb{R}^d$, $\mathbf{\Phi}(\widehat{\boldsymbol{\xi}}) := \mathbf{F}\widehat{\boldsymbol{\xi}} + \boldsymbol{\tau}$ according to (AFF).

Lemma 4.10. *If $\mathbf{\Phi} : \widehat{K} \mapsto K$ is an affine mapping $\widehat{\boldsymbol{\xi}} \mapsto \mathbf{F}\widehat{\boldsymbol{\xi}} + \boldsymbol{\tau}$, then, for all $m \in \mathbb{N}_0$,*

$$|\widehat{u}|_{H^m(\widehat{K})} \le \binom{m+d}{d} d^m \, \|\mathbf{F}\|^m \, |\det(\mathbf{F})|^{-1/2} \, |u|_{H^m(K)} \quad \forall u \in H^m(K) \, ,$$
$$|u|_{H^m(K)} \le \binom{m+d}{d} d^m \, \|\mathbf{F}^{-1}\|^m \, |\det(\mathbf{F})|^{1/2} \, |\widehat{u}|_{H^m(\widehat{K})} \quad \forall u \in H^m(\widehat{K}) \, .$$

with $\|\mathbf{F}\|$ denoting the matrix norm of $\mathbf{F}$ associated with the Euclidean vector norm.

Proof. Without loss of generality we can assume that $u \in C^\infty(\overline{K})$. Let $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $|\boldsymbol{\alpha}| = m$, $m \in \mathbb{N}_0$. Then, the m-th Gateaux-derivative $D^m : \mathbb{R}^d \times \cdots \times \mathbb{R}^d \mapsto \mathbb{R}$ allows to express

$$\partial^{\boldsymbol{\alpha}} \widehat{u} = D^m \widehat{u}(\widehat{\boldsymbol{\xi}})(\boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^m) \, ,$$

with $\boldsymbol{\delta}^1 = \cdots = \boldsymbol{\delta}^{\alpha_1} = \boldsymbol{\epsilon}_1$, $\boldsymbol{\delta}^{\alpha_1+1} = \cdots = \boldsymbol{\delta}^{\alpha_1+\alpha_2} = \boldsymbol{\epsilon}_2$, etc. Remember that $\boldsymbol{\epsilon}_k$ designates the k-th unit vector in $\mathbb{R}^d$. We deduce that

$$|\partial^{\boldsymbol{\alpha}} \widehat{u}(\widehat{\boldsymbol{\xi}})| \le \|D^m \widehat{u}(\widehat{\boldsymbol{\xi}})\| := \sup\{D^m \widehat{u}(\widehat{\boldsymbol{\xi}})(\boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^m), \, \boldsymbol{\delta}^k \in \mathbb{R}^d, \, |\boldsymbol{\delta}^k| = 1\} \, ,$$

which implies

$$|\widehat{u}|^2_{H^m(\widehat{K})} = \sum_{|\boldsymbol{\alpha}|=m} \int_{\widehat{K}} \left|\partial^{\boldsymbol{\alpha}} \widehat{u}(\widehat{\boldsymbol{\xi}})\right|^2 \, \mathrm{d}\widehat{\boldsymbol{\xi}} \le \binom{m+d}{m} \int_{\widehat{K}} \left\|D^m \widehat{u}(\widehat{\boldsymbol{\xi}})\right\|^2 \, \mathrm{d}\widehat{\boldsymbol{\xi}} \, . \tag{4.4}$$

The chain rule gives

$$D^m \widehat{u}(\widehat{\boldsymbol{\xi}})(\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_m) = D^m u(\mathbf{\Phi}(\widehat{\boldsymbol{\xi}}))(\mathbf{F}\boldsymbol{\delta}_1, \ldots, \mathbf{F}\boldsymbol{\delta}_m) \, . \tag{4.5}$$

This means that

$$\|D^m \widehat{u}(\widehat{\boldsymbol{\xi}})\| \le \|\mathbf{F}\|^m \, \|D^m u(\mathbf{\Phi}(\widehat{\boldsymbol{\xi}}))\| \, .$$

Next, we use the transformation formula for multidimensional integrals and apply it to (4.4):

$$|\widehat{u}|^2_{H^m(\widehat{K})} \le \binom{m+d}{m} \int_K \|\mathbf{F}\|^{2m} \|D^m u(\boldsymbol{\xi})\|^2 \, |\det(\mathbf{F})|^{-1} \, \mathrm{d}\boldsymbol{\xi} \, . \tag{4.6}$$

Finally, observe that

$$\begin{aligned}\|D^m u(\boldsymbol{\xi})\| &= \sup\{D^m u(\boldsymbol{\xi})(\boldsymbol{\delta}^1,\ldots,\boldsymbol{\delta}^m),\, \boldsymbol{\delta}^k \in \mathbb{R}^d, \left|\boldsymbol{\delta}^k\right| = 1\} \\ &\leq \sup\{\sum_{\alpha_1=1}^{d}\cdots\sum_{\alpha_m=1}^{d} |D^m u(\delta^1_{\alpha_1}\boldsymbol{\epsilon}_{\alpha_1},\ldots,\delta^m_{\alpha_m}\boldsymbol{\epsilon}_{\alpha_m})|,\, \boldsymbol{\delta}^k \in \mathbb{R}^d, \left|\boldsymbol{\delta}^k\right| = 1\} \\ &\leq \sum_{\alpha_1=1}^{d}\cdots\sum_{\alpha_m=1}^{d} |D^m u(\boldsymbol{\epsilon}_{\alpha_1},\ldots,\boldsymbol{\epsilon}_{\alpha_m})| \\ &\leq d^m \ \max\{|\partial^{\boldsymbol{\alpha}} u(\boldsymbol{\xi})|,\ |\boldsymbol{\alpha}| = m\}\ .\end{aligned}$$

□

Remark 4.11. If $\boldsymbol{\Phi}$ is a general C^∞-diffeomorphism $\widehat{K} \mapsto K$, then the analogue of (4.5) will involve derivatives of u from Du up to $D^m u$ and derivatives $D\boldsymbol{\Phi}$ up to $D^m\boldsymbol{\Phi}$. Thus, in order to estimate the Sobolev-seminorm $|\widehat{u}|_{H^m(\widehat{K})}$, we have to resort to the full Sobolev norm $\|u\|_{H^m(K)}$ and vice versa.

Let us consider an affine equivalent simplicial triangulation $\mathcal{M}$, see Def. 3.3. We fix a reference simplex $\widehat{K}$ and find affine mappings $\boldsymbol{\Phi}_K : \widehat{K} \mapsto K$, $\boldsymbol{\Phi}_K(\widehat{\boldsymbol{\xi}}) := \mathbf{F}_K\widehat{\boldsymbol{\xi}} + \boldsymbol{\tau}_K$ for each $K \in \mathcal{M}$. In light of the general strategy outlined above, we have to establish bounds for $\|\mathbf{F}\|$, $\|\mathbf{F}^{-1}\|$, $|\det(\mathbf{F})|$, and $|\det(\mathbf{F})|^{-1}$ that depend on *controllable* geometric features of $\mathcal{M}$.

Definition 4.12. *Given a cell K of a mesh $\mathcal{M}$ we define its* ***diameter***

$$h_K := \sup\{|\boldsymbol{\xi} - \boldsymbol{\eta}|\,,\ \boldsymbol{\xi},\boldsymbol{\eta} \in K\}\ ,$$

and the maximum radius of an inscribed ball

$$r_K := \sup\{r > 0 :\ \exists \boldsymbol{\xi} \in K :\ |\boldsymbol{\xi} - \boldsymbol{\eta}| < r\ \Rightarrow\ \boldsymbol{\eta} \in K\}\ .$$

The ratio h_K/r_K is called the ***shape regularity measure*** *ρ_K of K.*

Lemma 4.13. *If $\widehat{K}, K \subset \mathbb{R}^d$, $d = 2, 3$, are a generic non-degenerate simplices and $\boldsymbol{\Phi}_K : \widehat{K} \mapsto K$, $\boldsymbol{\Phi}_K(\widehat{\boldsymbol{\xi}}) := \mathbf{F}\widehat{\boldsymbol{\xi}} + \boldsymbol{\tau}$, the associated bijective affine mapping, then*

$$\left(\frac{h_K}{h_{\widehat{K}}}\right)^d \rho_K^{1-d} = \frac{h_K r_K^{d-1}}{h_{\widehat{K}}^d} \leq |\det(\mathbf{F})| = \frac{|K|}{|\widehat{K}|} \leq \frac{h_K^d}{h_{\widehat{K}} r_{\widehat{K}}^{d-1}} = \left(\frac{h_K}{h_{\widehat{K}}}\right)^d \rho_{\widehat{K}}^{d-1}\ , \tag{4.7}$$

$$\|\mathbf{F}\| \leq \frac{h_K}{2r_{\widehat{K}}} = \tfrac{1}{2}\rho_{\widehat{K}}\frac{h_K}{h_{\widehat{K}}} \quad , \quad \left\|\mathbf{F}^{-1}\right\| \leq \frac{h_{\widehat{K}}}{2r_K} = \tfrac{1}{2}\rho_K\frac{h_{\widehat{K}}}{h_K}\ . \tag{4.8}$$

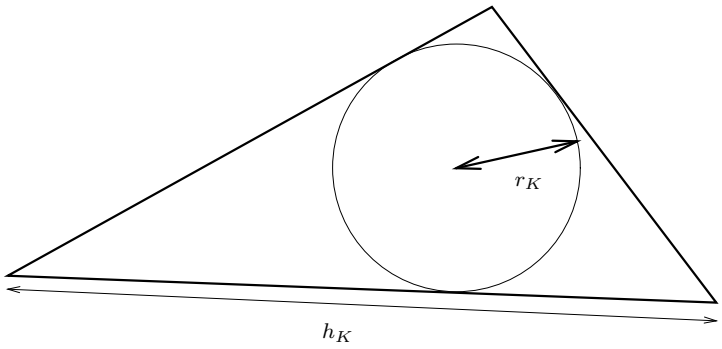


Figure 4.1: Diameter h_K and r_K for a triangular cell

Proof. The inequalities (4.7) can be concluded from the volume formula for simplices by elementary geometric considerations.

Write $\widehat{\boldsymbol{\zeta}} \in \widehat{K}$ for the center of the largest inscribed ball of $\widehat{K}$. Then estimates (4.8) follow from

$$\|\mathbf{F}\| = \sup\{|\mathbf{F}\widehat{\boldsymbol{\xi}}|,\ |\widehat{\boldsymbol{\xi}}| = 1\} = \tfrac{1}{2} r_{\widehat{K}}^{-1} \sup\{|\boldsymbol{\Phi}(\widehat{\boldsymbol{\xi}}) - \boldsymbol{\Phi}(\widehat{\boldsymbol{\zeta}})|,\ |\widehat{\boldsymbol{\xi}} - \widehat{\boldsymbol{\zeta}}| = 2r_{\widehat{K}}\} \le h_K / 2r_{\widehat{K}} \ ,$$

because both $\boldsymbol{\Phi}(\widehat{\boldsymbol{\xi}})$ and $\boldsymbol{\Phi}(\widehat{\boldsymbol{\zeta}})$ lie inside K. A role reversal of $\widehat{K}$ and K establishes the other estimate. □

The shape regularity measure of a simplex can be calculated from bounds for the smallest and largest angles enclosed by edge/face normals. We give the result for two dimensions:

Lemma 4.14. *If the smallest angle of a triangle K is bounded from below by $\alpha > 0$, then*

$$\sin(\alpha/2)^{-1} \le \rho_K \le 2\sin(\alpha/2)^{-1} \ .$$

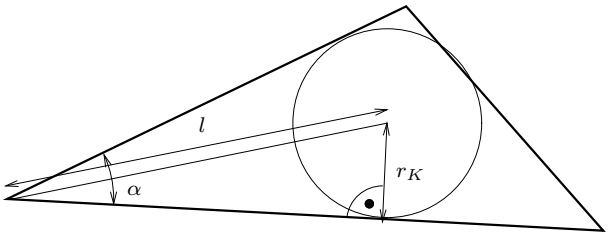


Figure 4.2: Angle condition for shape regularity of a triangle

Proof. It is immediate from Fig. 4.2 that

$$1/2 h_K \sin(\alpha/2) \le l \sin(\alpha/2) = r_K \le h_K \sin(\alpha/2) \ .$$

□

Lemma 4.13 clearly shows that *uniform shape-regularity* of the cells is key to achieving a uniform behavior of the Sobolev seminorms under transformation to a reference element.

Definition 4.15. *Given a mesh $\mathcal{M}$ its* ***meshwidth*** *can be computed by*

$$h_{\mathcal{M}} := \max\{h_K,\ K \in \mathcal{M}\}\ ,$$

whereas its ***shape regularity measure*** *is defined as*

$$\rho_{\mathcal{M}} := \max\{\rho_K,\ K \in \mathcal{M}\}\ .$$

Here, the notations from Def. 4.12 have been used. Moreover, the ***quasi-uniformity measure*** *of $\mathcal{M}$ is the quantity*

$$\mu_{\mathcal{M}} := \max\{h_K/h_{K'},\ K, K' \in \mathcal{M}\}\ .$$

Remark 4.16. Usually software for simplicial mesh generation employs elaborate algorithms to ensure that the angles of the triangles/tetrahedra do not become very small or close to π. Hence, it is not unreasonable to assume good shape regularity of simplicial meshes that are used for finite element computations.

The choice of reference simplices is arbitrary. So we may just opt for

$$\widehat{K} := \text{convex}\left\{\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}\right\} \quad \text{for } d = 2\ , \tag{4.9}$$

$$\widehat{K} := \text{convex}\left\{\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0\\1\\0\end{pmatrix}, \begin{pmatrix}0\\0\\1\end{pmatrix}\right\} \quad \text{for } d = 3\ . \tag{4.10}$$

Corollary 4.17. *Let $\mathcal{M}$ be a simplicial triangulation and choose the reference simplex according to* (4.9) *and* (4.10)*, respectively. Then the affine mappings $\boldsymbol{\Phi}_K : \widehat{K} \mapsto K$, $\boldsymbol{\Phi}_K(\widehat{\boldsymbol{\xi}}) := \mathbf{F}_K\widehat{\boldsymbol{\xi}} + \boldsymbol{\tau}_K$, $K \in \mathcal{M}$, satisfy*

$$\frac{\rho_{\mathcal{M}}^{1-d}}{\mu_{\mathcal{M}}^{d}}\, h_{\mathcal{M}}^{d} \le |\det(\mathbf{F}_K)| \le h_{\mathcal{M}}^{d}\ , \quad \|\mathbf{F}_K\| \le h_{\mathcal{M}}\ , \quad \left\|\mathbf{F}_K^{-1}\right\| \le \rho_{\mathcal{M}}\mu_{\mathcal{M}}\, h_{\mathcal{M}}^{-1}\ .$$

Lemma 4.18. *The number of cells of a conforming simplicial triangulation $\mathcal{M}$ that share a vertex can be bounded by means of the shape-regularity measure of $\mathcal{M}$.*

Proof. For $d = 2$ the assertion is immediate from Lemma 4.14. Similar elementary geometric considerations settle the case $d = 3$. □

4.3 Fundamental estimates

In this section we consider H^1-conforming Lagrangian finite elements of fixed degree $m \in \mathbb{N}$ on simplicial meshes, *cf.* Sect. 3.8.1 and, in particular, 3.54. We recall that they form affine equivalent families of finite elements in the sense of Def. 3.44. This makes them amenable to the application of transformation techniques.

Besides the Bramble-Hilbert lemma 4.7 the following result will be used frequently.

Lemma 4.19. *If $\|\cdot\|_1$ and $\|\cdot\|_2$ are two norms on a finite dimensional vector space X, then they are equivalent in the sense that*

$$\exists 0 < \underline{\gamma} \le \overline{\gamma} : \quad \underline{\gamma}\,\|v\|_1 \le \|v\|_2 \le \overline{\gamma}\,\|v\|_1 \quad \forall v \in X\ .$$

It is the main idea behind the proof of **L^2-stability** of the bases provided by the global shape functions.

Lemma 4.20. *Let $\mathcal{M}$ be a simplicial mesh of the computational domain $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$, and $\mathcal{S}_m(\mathcal{M}) \subset H^1(\Omega)$ the Lagrangian finite element space on $\mathcal{M}$ of uniform polynomial degree $m \in \mathbb{N}$ equipped with the basis $\mathfrak{B} := \{b^1, \ldots, b^N\}$, $N := \dim \mathcal{S}_m(\mathcal{M})$, of global shape functions. This basis is L^2-**stable** in the sense that*

$$\underline{\gamma}\ \left\|\sum\nolimits_{l=1}^{N} \mu_l b^l\right\|^2_{L^2(\Omega)} \le \sum\nolimits_{l=1}^{N} \mu_l^2 \left\|b^l\right\|^2_{L^2(\Omega)} \le \overline{\gamma}\left\|\sum\nolimits_{l=1}^{N} \mu_l b^l\right\|^2_{L^2(\Omega)}\ ,$$

with constants $0 < \underline{\gamma} \le \overline{\gamma}$ that only depend on d.

Proof. We start with localization: consider $v_n \in \mathcal{S}_m(\mathcal{M})$ and its restriction $v_K := v_{n|K}$ to a cell $K \in \mathcal{M}$. We can rely on a representation in terms of local shape functions

$$v_K = \sum\nolimits_{l=1}^{N_K} \mu_l b_l\ .$$

Let $\widehat{K}$ denote the single reference element for $\mathcal{M}$ and $\widehat{v}_K$ the affine pullback of v_K to $\widehat{K}$. We note that $\widehat{v}_K$ belongs to the fixed finite dimensional space $\mathcal{P}_m(\widehat{K})$ and that

$$\widehat{v}_K = \sum\nolimits_{l=1}^{N_K} \mu_l \widehat{b}_l\ ,$$

where $\widehat{b}_l$ are the local shape functions on $\widehat{K}$. Obviously,

$$\widehat{v}_K \mapsto \sum\nolimits_{l=1}^{N_K} \mu_l^2 \left\|\widehat{b}_l\right\|^2_{L^2(\widehat{K})}$$

is the square of a norm on $\mathcal{P}_m(\widehat{K})$. Thus, combining twice Lemma 4.10 (for $m = 0$) and Lemma 4.19 we obtain

$$\|v_K\|^2_{L^2(K)} = |K|/|\widehat{K}|\ \ \|\widehat{v}_K\|^2_{L^2(\widehat{K})} \approx |K|/|\widehat{K}|\ \sum\nolimits_{l=1}^{N_K} \mu_l^2 \left\|\widehat{b}_l\right\|^2_{L^2(\widehat{K})} = \sum\nolimits_{l=1}^{N_K} \mu_l^2 \left\|b_l\right\|^2_{L^2(K)}\ ,$$

where $\approx$ stands for a two-sided estimate involving constants that merely depend on d. Summing over all cells finishes the proof. □

The assertion of the lemma is not affected by a rescaling of the global shape functions. Hence, we may assume $\left\|b^l\right\|_{L^2(\Omega)} = 1$, $l = 1, \ldots, N$. The coefficient isomorphism $\mathsf{C}_n : \mathbb{R}^N \mapsto \mathcal{S}_m(\mathcal{M})$ related to the resulting basis will satisfy

$$\|\mathsf{C}_n\|^2_{\mathbb{R}^N \mapsto L^2(\Omega)} \leq \underline{\gamma}^{-1} \quad , \quad \left\|\mathsf{C}_n^{-1}\right\|^2_{L^2(\Omega) \mapsto \mathbb{R}^N} \leq \overline{\gamma} \, , \tag{4.11}$$

where the constants $\underline{\gamma}$, $\overline{\gamma}$ are those from Lemma 4.20. It is not difficult to derive (4.11) from Def. 1.8 of the operator norm.

Assume that the rescaled global shape functions are used in the context of a Galerkin finite element discretization of the $L^2(\Omega)$ inner product

$$\mathsf{b}(u, v) := \int_\Omega u \cdot v \, \mathrm{d}\boldsymbol{\xi} \, , \quad u, v \in L^2(\Omega) \, .$$

This will give us the **mass matrix** $\mathbf{M} \in \mathbb{R}^{N,N}$, *cf.* Exercise 3.20. Now, we can use (4.11) and apply Lemma 1.46:

$$\kappa(\mathbf{M}) \leq \underline{\gamma}^{-1} \overline{\gamma} \, .$$

We conclude that

> The spectral condition number of the diagonally scaled mass matrix arising from Lagrangian finite elements of fixed polynomial degree does not depend on the finite element mesh.

The Lagrangian finite element space $\mathcal{S}_m(\mathcal{M})$ has finite dimension, which, by Lemma 4.19, implies the equivalence of all norms, in particular of all Sobolev norms. However, such an equivalence cannot hold true for the infinite-dimensional function spaces. We expect that the constants in the norm equivalence will blow up or tend to zero when we choose larger and larger finite element spaces. The next lemma gives a quantitative estimate, a so-called **inverse estimate**.

Lemma 4.21 (Inverse estimate). *If $\mathcal{M}$ is a simplicial mesh of $\Omega \subset \mathbb{R}^d$, then for $0 \leq k < l$, $m \in \mathbb{N}$,*

$$\exists \gamma = \gamma(d, l, k, m, \rho_{\mathcal{M}}, \mu_{\mathcal{M}}) > 0 : \quad |v_n|_{H^l(\Omega)} \leq \gamma \, h_{\mathcal{M}}^{l-k} \, |v_n|_{H^k(\Omega)} \quad \forall v_n \in \mathcal{S}_m(\mathcal{M}) \, .$$

Proof. Using the transformation technique, Lemma 4.10, Corollary 4.17, and Lemma 4.19 the proof can be accomplished easily. □

Example 4.22. The second-order elliptic boundary value problem

$$-\Delta u + u = f \quad \text{in } \Omega \subset \mathbb{R}^d \quad , \quad \langle \mathbf{grad}\, u, \mathbf{n} \rangle = 0 \quad \text{on } \Gamma \tag{4.12}$$

is discretized by means of linear Lagrangian finite elements on a simplicial mesh $\mathcal{M}$. The bilinear form associated with (4.12) is

$$\mathsf{b}(u,v) := \int_\Omega \langle \mathbf{grad}\, u, \mathbf{grad}\, v \rangle + uv \,\mathrm{d}\boldsymbol{\xi} \;, \quad u, v \in H^1(\Omega) \;, \tag{4.13}$$

which agrees with the $H^1(\Omega)$-inner product.

We use (4.9) as reference triangle and invoke Lemma 4.10 for $m = 0, 1$:

$$\begin{array}{rcl} |v_n|^2_{H^1(K)} & \approx & h_{\mathcal{M}}^{d-2}\, |\widehat{v}_n|^2_{H^1(\widehat{K})} \;, \\ \|v_n\|^2_{L^2(K)} & \approx & h_{\mathcal{M}}^{d}\, \|\widehat{v}_n\|^2_{L^2(\widehat{K})} \end{array} \quad \forall v_n \in \mathcal{S}_1(\mathcal{M}) \;, \tag{4.14}$$

where $\approx$ designates a two-sided estimate with constants depending on $\rho_{\mathcal{M}}$ and $\mu_{\mathcal{M}}$. On $\mathcal{P}_1(\widehat{K})$ all norms are equivalent (Lemma 4.19) so that

$$\|\widehat{v}\|_{H^1(\widehat{K})} \approx \left(\sum_{j=1}^{d+1} v(\widehat{\boldsymbol{\nu}}_j)^2 \right)^{1/2} \qquad \|\widehat{v}\|_{L^2(\widehat{K})} \approx \left(\sum_{j=1}^{d+1} v(\widehat{\boldsymbol{\nu}}_j)^2 \right)^{1/2} \quad \forall v \in \mathcal{P}_1(\widehat{K}) \;. \tag{4.15}$$

Here, $\approx$ stands for equivalence involving only universal constants, and $\widehat{\boldsymbol{\nu}}_1, \ldots, \widehat{\boldsymbol{\nu}}_{d+1}$ are the vertices of $\widehat{K}$. Merging (4.14) and (4.15) yields

$$\begin{array}{l} \|v_n\|^2_{H^1(K)} \le \gamma\, h_{\mathcal{M}}^{d-2}\, \|\widehat{v}_n\|^2_{H^1(\widehat{K})} \le \gamma\, h_{\mathcal{M}}^{d-2} \sum_{j=1}^{d+1} \widehat{v}_n(\widehat{\boldsymbol{\nu}}_j)^2 \;, \\ \|v_n\|^2_{H^1(K)} \ge \gamma\, h_{\mathcal{M}}^{d}\, \|\widehat{v}_n\|^2_{L^2(\widehat{K})} \ge \gamma\, h_{\mathcal{M}}^{d} \sum_{j=1}^{d+1} \widehat{v}_n(\widehat{\boldsymbol{\nu}}_j)^2 \end{array} \qquad \forall v_n \in \mathcal{S}_1(\mathcal{M}),\ K \in \mathcal{M} \;.$$

Here, $\gamma > 0$ are *generic constants* that only depend on $\rho_{\mathcal{M}}$ and $\mu_{\mathcal{M}}$. Moreover, we have tacitly assumed $h_{\mathcal{M}}^d \ll h_{\mathcal{M}}^{d-2}$, which makes sense for reasonable fine meshes. Hence, in the first estimate only the H^1-seminorm has been taken into account, whereas in the second estimate the L^2-norm was supposed to be dominant.

Summing up and appealing to Lemma 4.18, we end up with

$$\underline{\gamma}\, h_{\mathcal{M}}^d \sum\nolimits_{\boldsymbol{\xi} \in \mathcal{N}(\mathcal{M})} v_n(\boldsymbol{\xi})^2 \le \|v_n\|^2_{H^1(\Omega)} \le \overline{\gamma}\, h_{\mathcal{M}}^{d-2} \sum\nolimits_{\boldsymbol{\xi} \in \mathcal{N}(\mathcal{M})} v_n(\boldsymbol{\xi})^2 \;, \tag{4.16}$$

with other constants $0 < \underline{\gamma} = \underline{\gamma}(\rho_{\mathcal{M}}, \mu_{\mathcal{M}})$, $0 < \overline{\gamma} = \overline{\gamma}(\rho_{\mathcal{M}}, \mu_{\mathcal{M}})$. Repeating the analysis for the mass matrix, from Lemma 1.46 we get the bound

$$\kappa(\mathbf{B}) \le \gamma h_{\mathcal{M}}^{-2}$$

for the spectral condition number of the stiffness matrix $\mathbf{B}$ arising from $\mathcal{S}_1(\mathcal{M})$ and the bilinear form (4.13). The constant γ will only depend on $\rho_{\mathcal{M}}$ and $\mu_{\mathcal{M}}$. This demonstrates that the spectral condition number of $\mathbf{B}$ may deteriorate like $O(h_{\mathcal{M}}^{-2})$ for a sequence of meshes that are uniformly shape-regular and quasi-uniform.

Exercise 4.2. Show that the estimate (4.16) is sharp by plugging in suitable functions $v_n \in \mathcal{S}_1(\mathcal{M})$.

4.4 Interpolation error estimates

In Sect. 1.4 we have learned that it takes knowledge about the best approximation error of the exact solution in the trial space in order to gauge the discretization error of a Galerkin scheme for a linear variational problem, see Thm. 1.30.

Usually, in the case of finite elements the best approximation error remains elusive, but thanks to the *locality* statement of Thm. 3.42 the interpolation errors for the finite element interpolation operators (see Def. 3.39) can be estimated by means of transformation techniques. At least, this gives us an upper bound for the best approximation error.

Here, we focus on H^1-conforming Lagrangian finite elements of uniform polynomial degree $m \in \mathbb{N}$ on a simplicial triangulation $\mathcal{M}$ of a computational domain $\Omega \subset \mathbb{R}^d$, $d = 1, 2, 3$. Eventually we want to examine the dependence of the interpolation error on the meshwidth $h_{\mathcal{M}}$ of $\mathcal{M}$. Thus, the results will be relevant for the h-version of finite elements.

Remark 4.23. The considerations of this section carry over to the other affine equivalent families of finite elements presented in Sect. 3.8.2.

Theorem 4.24. *Let* I *stand for the finite element interpolation operator belonging to the Lagrangian finite element space* $\mathcal{S}_m(\mathcal{M})$ *on a simplicial mesh* $\mathcal{M}$*. Then, for* $2 \le t \le m+1$, $0 \le r \le t$

$$\exists \gamma = \gamma(t, r, m, \rho_{\mathcal{M}}) : \quad \|u - \mathsf{I}\,u\|_{H^r(\Omega)} \le \gamma\, h_{\mathcal{M}}^{t-r}\, |u|_{H^t(\Omega)} \quad \forall u \in H^t(\Omega)\ .$$

Proof. Single out $K \in \mathcal{M}$ and write $\widehat{K}$ for the assocoiated reference element according to (4.9) or (4.10), respectively. They are linked by the affine mapping $\boldsymbol{\Phi}_K : \widehat{K} \mapsto K$.

The crucial observation is that the local finite element interpolation operators commute with the pullback $\mathsf{FT}_{\boldsymbol{\Phi}_K}$:

$$\mathsf{FT}_{\boldsymbol{\Phi}_K}(\mathsf{I}_K\, u) = \widehat{\mathsf{I}}(\mathsf{FT}_{\boldsymbol{\Phi}_K}\, u) \quad \forall u \in H^t(K)\ .$$

This is an immediate consequence of the parametric (affine) equivalence of the finite elements on K and $\widehat{K}$, *cf.* Def. 3.44.

Next, we appeal to Lemma 4.10 and Cor. 4.17 and estimate

$$\|u - \mathsf{I}_K\, u\|_{H^r(K)} \le \gamma\, h_K^{d/2-r} \left\|\widehat{u} - \widehat{\mathsf{I}}\widehat{u}\right\|_{H^r(\widehat{K})} \le \gamma\, h_K^{d/2-r} \inf_{p \in \mathcal{P}_m(\widehat{K})} \left\|(\widehat{u} - p) - \widehat{\mathsf{I}}(\widehat{u} - p)\right\|_{H^r(\widehat{K})}\ ,$$

because, by Lemma 3.41, $\widehat{\mathsf{I}}$ preserves polynomials in $\Pi_{\widehat{K}} = \mathcal{P}_m(\widehat{K})$. Note that, thanks to Thm. 2.46, functions in $H^t(\widehat{K})$ are uniformly bounded and continuous for $t \ge 2$ so

that $\widehat{\mathsf{I}} : H^t(\widehat{K}) \mapsto \mathcal{P}_m(\widehat{K})$ is continuous. This enables us to apply the Bramble Hilbert Lemma

$$\|u - \mathsf{I}_K\, u\|_{H^r(K)} \leq \gamma\ \ h_K^{d/2-r} \inf_{p\in\mathcal{P}_m(\widehat{K})} \|\widehat{u} - p\|_{H^t(\widehat{K})} \leq \gamma\ \ h_K^{d/2-r}\ \ |\widehat{u}|_{H^t(\widehat{K})} \leq \gamma\ \ h_K^{t-r}\, |u|_{H^t(K)}\ .$$

All the constants only depend on the norms and ρ_K. Morover, in the final step Lemma 4.10 was invoked once more. Summing over all elements yields the result. □

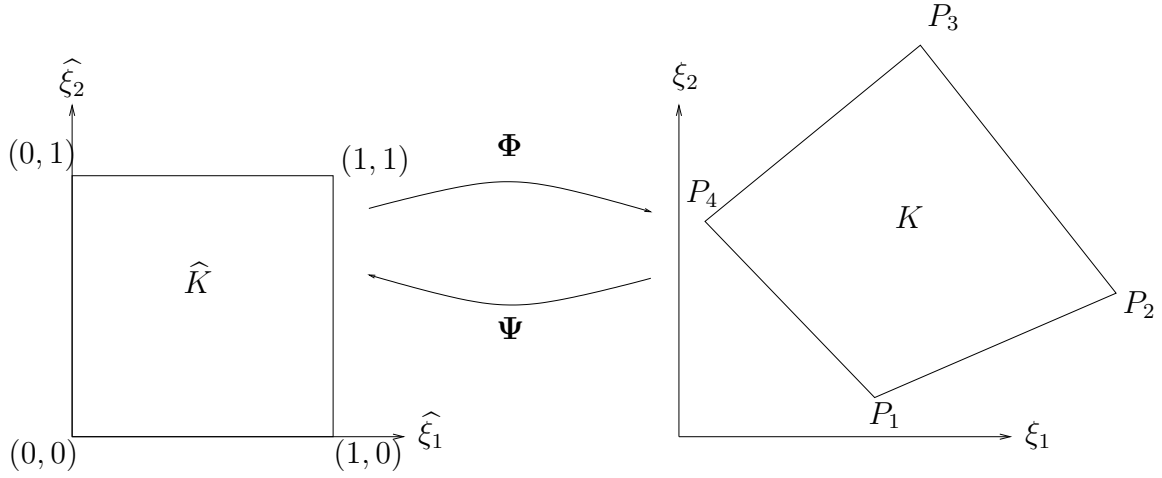
Remark 4.25. The proof essentially hinges on the continuity of the local finite element interpolation operator in $H^t(K)$. Of course, we cannot expect any interpolation error estimate in Sobolev norms for which the finite element interpolation is not bounded.

Remark 4.26. The statement of Thm. 4.24 remains true for tensor product Lagrangian finite elements according to Def. 3.56 on affine equivalent meshes.

Exercise 4.3. Consider the following coordinate transformation $\mathbf{\Phi} : \widehat{K} \to K$ which matches the unit square $]0;1[^2$ to an arbitrary quadrangle with vertices P_1, P_2, P_3 and P_4

$$\mathbf{\Phi}(\widehat{\xi_1},\widehat{\xi_2}) \ \ = \ \ P_1\widehat{\xi_1}\widehat{\xi_2} + P_2\widehat{\xi_1}(1-\widehat{\xi_2}) + P_3(1-\widehat{\xi_1})\widehat{\xi_2} + P_4(1-\widehat{\xi_1})(1-\widehat{\xi_2})$$

and the inverse transformation $\mathbf{\Psi} := \mathbf{\Phi}^{-1}$.



(i) Compute the Jacobi matrix $D\mathbf{\Phi}(\boldsymbol{\xi})$ and show that for an arbitrary matrix $\mathbf{A} \in \mathbb{R}^{N,N}$, $N \in \mathbb{N}$ there holds

$$\|\mathbf{A}\| \ \ \leq \ \ \|\mathbf{A}\|_F$$

where $\|.\|_F$ denotes the Frobenius norm on $\mathbb{R}^{N,N}$ which is defined by

$$\|\mathbf{A}\|_F = \left(\sum_{i,j=1}^{N} a_{ij}^2\right)^{1/2}, \qquad \mathbf{A} = (a_{ij})_{i,j=1}^{N}\,.$$

(ii) Compute the determinant $|\det D\boldsymbol{\Phi}(\boldsymbol{\xi})|$.

(iii) Show that for an arbitrary $u \in H^1(\widehat{K})$ there holds

$$|u|_{H^1(K)} \leq \max_{\boldsymbol{\xi}\in K} \|D\boldsymbol{\Psi}(\boldsymbol{\xi})\|_F |\det D\boldsymbol{\Psi}(\boldsymbol{\xi})|^{-1/2} |\widehat{u}|_{H^1(\widehat{K})}.$$

where $\widehat{u} \in H^1(\widehat{K})$ is defined by $\widehat{u}(\widehat{\boldsymbol{\xi}}) := u(\boldsymbol{\Phi}(\widehat{\boldsymbol{\xi}}))$.

(iv) Show that for an arbitrary $u \in H^2(\widehat{K})$ there holds

$$\left(\sum_{k=1}^{2} \left\| \frac{\partial^2 \widehat{u}}{\partial \widehat{\xi}_k^2} \right\|_{L^2(\widehat{K})}^2 \right)^{1/2} \leq \max_{\widehat{\boldsymbol{\xi}}\in\widehat{K}} \|D\boldsymbol{\Phi}(\widehat{\boldsymbol{\xi}})\|_F^2 |\det D\boldsymbol{\Phi}(\widehat{\boldsymbol{\xi}})|^{-1/2} \|\widehat{u}\|_{H^2(\widehat{K})}$$

where $\widehat{u} \in H^2(\widehat{K})$ is defined by $\widehat{u}(\widehat{\boldsymbol{\xi}}) := u(\boldsymbol{\Phi}(\widehat{\boldsymbol{\xi}}))$.

(v) Show that

$$|u - \mathsf{I}\, u|_{H^1(K)} \leq \gamma \max_{\boldsymbol{\xi}\in K} \|D\boldsymbol{\Psi}(\boldsymbol{\xi})\|_F |\det D\boldsymbol{\Psi}(\boldsymbol{\xi})|^{-1/2} \max_{\widehat{\boldsymbol{\xi}}\in\widehat{K}} \|D\boldsymbol{\Phi}(\widehat{\boldsymbol{\xi}})\|_F^2 |\det D\boldsymbol{\Phi}(\widehat{\boldsymbol{\xi}})|^{-1/2} \|u\|_{H^2(K)}$$

holds for all $u \in H^2(K)$, where γ is a constant.

Bibliographical notes. Basic interpolation estimates for Lagrangian finite elements are elaborated in [8, §6] and [12, Sect. 3.1].

4.5 A priori error estimates for Lagrangian finite elements

We consider a second order elliptic boundary value problem on a polygon/polyhedron $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, with homogeneous Dirichlet boundary conditions

$$-\operatorname{div}(\mathbf{A}\,\mathbf{grad}\, u) = f \ \text{ in } \Omega \quad , \quad u = 0 \ \text{ on } \Gamma \,, \tag{4.17}$$

whose variational formulation reads (see Sect. 2.8 and (4.1)): seek $u \in H_0^1(\Omega)$ such that

$$\mathsf{b}(u, v) := \int_\Omega \langle \mathbf{A}\,\mathbf{grad}\, u, \mathbf{grad}\, v \rangle \, \mathrm{d}\boldsymbol{\xi} = \int_\Omega f v \, \mathrm{d}\boldsymbol{\xi} \quad \forall v \in H_0^1(\Omega) \,. \tag{4.18}$$

To discretize (4.18) we can employ Lagrangian finite elements of fixed polynomial degree $m \in \mathbb{N}$ on an affine equivalent mesh $\mathcal{M}$. This will give us a discrete solution $u_n \in \mathcal{S}_m(\mathcal{M})$.

According to the results of Sect. 1.4 and 2.8, a Galerkin finite element discretization of (4.18) will be quasi-optimal. Hence bounds for the H^1-norm of the discretization error

can be derived from the interpolation error estimates: if the exact solution u belongs to $H^2(\Omega)$ we obtain with $\gamma = \gamma(\Omega, \mathbf{A}) > 0$

$$\|u - u_n\|_{H^1(\Omega)} \leq \gamma \inf_{v_n \in \mathcal{S}_m(\mathcal{M})} \|u - v_n\|_{H^1(\Omega)} \leq \gamma \|u - \mathsf{I}\, u\|_{H^1(\Omega)} \ ,$$

where I is the finite element interpolation operator onto $\mathcal{S}_m(\mathcal{M})$. We continue by applying Thm. 4.24 and end up with

$$\|u - u_n\|_{H^1(\Omega)} \leq \gamma\, h_{\mathcal{M}}^{t-1}\, |u|_{H^t(\Omega)} \quad \text{for } 2 \leq t \leq m+1 \ \text{ and } \ u \in H^t(\Omega) \ , \tag{4.19}$$

where $\gamma = \gamma(\Omega, \mathbf{A}, \rho_{\mathcal{M}}, \mu_{\mathcal{M}})$. Let us discuss this a priori finite element discretization error estimate:

1. The estimate (4.19) hinges on the fact that the exact solution u is "smoother" (in terms of Sobolev norms) than merely belonging to $H^1(\Omega)$. For general $f \in H^{-1}(\Omega)$ this must never be taken for granted. However, for a restricted class of problems (4.18) with extra smoothness of the right hand side, e.g. $f \in H^r(\Omega)$, **elliptic shift theorems** may guarantee that $u \in H^t(\Omega)$ for $t > r$. For instance, for smooth Ω we can expect $u \in H^{r+2}(\Omega)$.

   Example 4.27. For $d = 1$ we have $u \in H^{r+2}(\Omega)$, if $f \in H^r(\Omega)$.

2. The bound from (4.19) can be converted into an **asymptotic a priori error estimate** by considering a sequence $\mathcal{M}_n$, $n \in \mathbb{N}$, of simplicial meshes of Ω. They are assumed to be *uniformly* shape-regular, that is,

   $$\exists \gamma > 0 : \quad \rho_{\mathcal{M}_n} < \gamma \quad \forall n \in \mathbb{N} \ .$$

   Moreover, the meshes are to become infinitely fine

   $$h_{\mathcal{M}_n} \to 0 \quad \text{as} \quad n \to \infty \ .$$

   Then the statement of (4.19) can be expressed by

   $$\|u - u_n\|_{H^1(\Omega)} = O(h_{\mathcal{M}_n}^{t-1}) \quad \text{for } n \to \infty \ . \tag{4.20}$$

   If (4.20) holds, common parlance says that the h-version finite element solutions enjoy convergence of the order $t - 1$ as the meshwidth tends to zero.

Remark 4.28. With considerable extra effort, more sophisticated best approximation estimates can be derived: for $m, t \geq 1$ we have

$$\inf_{v_n \in \mathcal{S}_m(\mathcal{M})} \|u - v_n\|_{H^1(\Omega)} \leq \gamma(\rho_{\mathcal{M}}, \mu_{\mathcal{M}}) \left(\frac{h_{\mathcal{M}}}{m}\right)^{\min\{m+1,t\}-1} \|u\|_{H^t(\Omega)} \ . \tag{4.21}$$

This paves the way for a-priori error estimates for the p-version of H^1-conforming elements.

4.6 Duality techniques

We still deal with the variational problem (4.1) and its Galerkin discretization based on the Lagrangian finite element space $\mathcal{S}_m(\mathcal{M})$ on a simplicial mesh $\mathcal{M}$. Now, we aim to establish an estimate of the discretization error in the $L^2(\Omega)$-norm.

This is beyond the scope of the theory presented in Ch. 1 and will rely on particular techniques for elliptic boundary value problems.

Assumption 4.29. *We assume that* (4.17) *is* ***2-regular****, that is, all* $u \in H^1_0(\Omega)$ *with* $-\operatorname{div}(\mathbf{A}\,\mathbf{grad}\,u) \in L^2(\Omega)$ *satisfy*

$$u \in H^2(\Omega) \quad \textit{and} \quad \|u\|_{H^2(\Omega)} \leq \gamma \,\|\operatorname{div}(\mathbf{A}\,\mathbf{grad}\,u)\|_{L^2(\Omega)} \ ,$$

with a constant $\gamma = \gamma(\mathbf{A},\Omega) > 0$ *independent of* u.

We write u_n for the unique solution of the discrete variational problem

$$u_n \in \mathcal{S}_m(\mathcal{M}) \cap H^1_0(\Omega): \quad \mathsf{b}(u_n, v_n) = \int_\Omega f v_n \,\mathrm{d}\boldsymbol{\xi} \quad \forall v_n \in \mathcal{S}_m(\mathcal{M}) \ .$$

Write $u \in H^1_0(\Omega)$ for the exact solution of (4.18) and $e := u - u_n \in H^1_0(\Omega)$ for the discretization error. From Sect. 1.4 we recall the *Galerkin orthogonality* (1.16)

$$\mathsf{b}(e, v_n) = 0 \quad \forall v_n \in \mathcal{S}_m(\mathcal{M}) \ .$$

The solution $w \in H^1_0(\Omega)$ of the *dual linear variational problem*

$$w \in H^1_0(\Omega): \quad \mathsf{b}(w, v) = \int_\Omega e v \,\mathrm{d}\boldsymbol{\xi} \quad \forall v \in H^1_0(\Omega) \ , \tag{4.22}$$

will be a solution of the the elliptic boundary value problem

$$-\operatorname{div}(\mathbf{A}\,\mathbf{grad}\,w) = e \text{ in } \Omega \quad , \quad w = 0 \ \text{ on } \Gamma \ .$$

Since $e \in L^2(\Omega)$, by Assumption 4.29 we know

$$w \in H^2(\Omega) \quad , \quad \|w\|_{H^2(\Omega)} \leq \gamma \,\|e\|_{L^2(\Omega)} \ , \tag{4.23}$$

with $\gamma = \gamma(\Omega, \mathbf{A}) > 0$.

Next, we plug $v = e$ into (4.22) and arrive at

$$\|e\|^2_{L^2(\Omega)} = \mathsf{b}(w, e) = \inf_{v_n \in \mathcal{S}_m(\mathcal{M})} \mathsf{b}(w - v_n, e) \ ,$$

where Galerkin orthogonality came into play. We may now plug in $v_n := \mathsf{I}\,w$, where I is the finite element interpolation operator for $\mathcal{S}_m(\mathcal{M})$. Then we can use the the continuity of b in $H^1(\Omega)$ and the interpolation error estimate of Thm. 4.24 for $r = 1$ and $t = 2$:

$$\|e\|^2_{L^2(\Omega)} \leq \mathsf{b}(w - \mathsf{I}\,w, e) \leq \gamma \ \|w - \mathsf{I}\,w\|_{H^1(\Omega)} \cdot \|e\|_{H^1(\Omega)} \leq \gamma \, h_{\mathcal{M}} \ |w|_{H^2(\Omega)} \cdot \|e\|_{H^1(\Omega)} \ .$$

Here, the final constant γ will depend on $\mathbf{A}$, m, $\rho_{\mathcal{M}}$, and $\mu_{\mathcal{M}}$, but not on u or u_n. Eventually, we resort to the 2-regularity in the form of estimate (4.23) and cancel one power of $\|e\|_{L^2(\Omega)}$.

This technique is known as **duality technique**, because it relies on the dual variational problem (4.22). Sometimes the term "Aubin-Nitsche trick" can be found. Summing up we have proved the following result:

Theorem 4.30. *Assuming 2-regularity according to Assumption 4.29, we obtain*

$$\|u - u_n\|_{L^2(\Omega)} \leq \gamma\, h_{\mathcal{M}}\, \|u - u_n\|_{H^1(\Omega)} \ ,$$

where the constant $\gamma > 0$ depends on $\Omega, \mathbf{A}, m, \rho_{\mathcal{M}}, \mu_{\mathcal{M}}$.

Remark 4.31. Thm. 4.30 tells us that under suitable assumptions in the h-version of finite elements we can gain another power of $h_{\mathcal{M}}$ when measuring the discretization error in the $L^2(\Omega)$-norm. More generally, often we can expect that, sloppily speaking,

> the weaker the norm of the discretization error that we consider the faster it will converge to zero as $h_{\mathcal{M}} \to 0$.

What remains to be settled is whether Assumption 4.29 is reasonable. This is part of **elliptic regularity theory**. In particular, we have the following result [19]

Theorem 4.32. *If the computational domain $\Omega \subset \mathbb{R}^d$ is convex or has C^1-boundary and $\mathbf{A} \in C^1(\overline{\Omega})$, then the elliptic boundary value problem* (4.17) *is 2-regular.*

4.7 Estimates for quadrature errors

As explained in Sect. 3.9.3, usually the finite element discretization of (4.18) will rely on local numerical quadrature for the computation of the stiffness matrix and of the load vector.

The use of numerical quadrature will inevitably perturb the finite element Galerkin solution and introduce another contribution to the total discretization error, which is called **consistency error**. We have already stressed that the choice of the local quadrature rule is guided by the principle that

> the error due to numerical quadrature must not dominate the total discretization error (in the relevant norms).

As far as the h-version of finite elements in concerned this guideline can be rephrased as follows:

> the impact of numerical quadrature must not affect the order of convergence in terms of the meshwidth.

4.7.1 Abstract estimates

We consider a linear variational problem (LVP) on a Banach space V

$$u \in V : \quad \mathsf{b}(u,v) = \langle f, v\rangle_{V^*\times V} \quad \forall v \in V \ ,$$

with V-elliptic bilinear form $\mathsf{b} \in L(V \times V, \mathbb{R})$ (Def. 1.20) and $f \in V^*$, see Sect. 1.2. Existence and uniqueness of a solution $u \in V$ are guaranteed by Thm. 1.17.

Based on $V_n \subset V$, $\dim(V_n) < \infty$, we arrive at the discrete variational problem (DVP), see Sect. 1.4,

$$u_n \in V_n : \quad \mathsf{b}(u_n,v_n) = \langle f, v_n\rangle_{V^*\times V} \quad \forall v_n \in V_n \ .$$

>From an abstract point of view the application of numerical quadrature in a finite element context means that the discrete variational problem will suffer a perturbation

$$u_n \in V_n : \quad \widetilde{\mathsf{b}}(u_n,v_n) = \left\langle \widetilde{f}, v_n\right\rangle_{V^*\times V} \quad \forall v_n \in V_n \ , \tag{4.24}$$

with a bilinear form $\widetilde{\mathsf{b}} \in L(V_n \times V_n, \mathbb{R})$ and $\widetilde{f} \in V_n^*$. The perturbation destroys Galerkin orthogonality and leads to extra terms in the discretization error estimate of Cor. 1.38.

Theorem 4.33. *Beside the assumptions on* b *and* $\widetilde{\mathsf{b}}$ *stated above we demand that*

$$\exists \gamma_1 > 0 : \quad \widetilde{\mathsf{b}}(v_n,v_n) \geq \gamma_1 \left\|v_n\right\|_V^2 \quad \forall v_n \in V_n \ . \tag{4.25}$$

Then (4.24) *will have a unique solution* $u_n \in V_n$*, which satisfies the a-priori error estimate*

$$\begin{aligned} \left\|u-u_n\right\|_V \ \leq \ \gamma\Big(& \inf_{v_n\in V_n} \big(\left\|u-v_n\right\|_V + \sup_{w_n\in V_n} \frac{|\mathsf{b}(v_n,w_n) - \widetilde{\mathsf{b}}(v_n,w_n)|}{\left\|w_n\right\|_V}\big) \\ & + \sup_{w_n\in V_n} \frac{|\left\langle f, w_n\right\rangle_{V\times V^*} - \left\langle \widetilde{f}, w_n\right\rangle_{V\times V^*}|}{\left\|w_n\right\|_V}\Big) \ , \end{aligned}$$

with $\gamma = \gamma(\|\mathsf{b}\|, \gamma_e, \gamma_1) > 0$

Proof. The assumption (4.25) means that for any $v_n \in V_n$

$$\begin{aligned} \gamma_1 \left\|u_n - v_n\right\|_V^2 &\leq \widetilde{\mathsf{b}}(u_n - v_n, u_n - v_n) \\ &= \mathsf{b}(u - v_n, u_n - v_n) + \big(\mathsf{b}(v_n, u_n - v_n) - \widetilde{\mathsf{b}}(v_n, u_n - v_n)\big) \\ &\quad + \big(\widetilde{\mathsf{b}}(u_n, u_n - v_n) - \mathsf{b}(u, u_n - v_n)\big) \\ &= \mathsf{b}(u - v_n, u_n - v_n) + \big(\mathsf{b}(v_n, u_n - v_n) - \widetilde{\mathsf{b}}(v_n, u_n - v_n)\big) \\ &\quad - \langle f, u_n - v_n\rangle_{V\times V^*} + \left\langle \widetilde{f}, u_n - v_n\right\rangle_{V\times V^*} \end{aligned}$$

Next, we exploit the continuity of b and divide by $\|u_n - v_n\|_V$:

$$\begin{aligned}\gamma_1 \|u_n - v_n\|_V \leq \|\mathsf{b}\| \|u - v_n\|_V &+ \frac{|\mathsf{b}(v_n, u_n - v_n) - \widetilde{\mathsf{b}}(v_n, u_n - v_n)|}{\|u_n - v_n\|_V} \\ &+ \frac{|\langle f, u_n - v_n\rangle_{V\times V^*} - \left\langle \widetilde{f}, u_n - v_n \right\rangle_{V\times V^*}|}{\|u_n - v_n\|_V} \\ \leq \|\mathsf{b}\| \|u - v_n\|_V &+ \sup_{w_n \in V_n} \frac{|\mathsf{b}(v_n, w_n) - \widetilde{\mathsf{b}}(v_n, w_n)|}{\|w_n\|_V} \\ &+ \sup_{w_n \in V_n} \frac{|\langle f, w_n\rangle_{V\times V^*} - \left\langle \widetilde{f}, w_n \right\rangle_{V\times V^*}|}{\|w_n\|_V}\end{aligned}$$

As $v_n \in V_n$ has been arbitrary, the triangle inequality

$$\|u - u_n\|_V \leq \|u - v_n\|_V + \|u_n - v_n\|_V$$

finishes the proof. □

The assumption (4.25) is called **h-ellipticity**. In the h-version of finite elements we want γ_1 to be indepedent of the meshwidth ("uniform h-ellipticity"). The two terms

$$\sup_{w_n \in V_n} \frac{|\mathsf{b}(v_n, w_n) - \widetilde{\mathsf{b}}(v_n, w_n)|}{\|w_n\|_V},$$
$$\sup_{w_n \in V_n} \frac{|\langle f, w_n\rangle_{V\times V^*} - \left\langle \widetilde{f}, w_n \right\rangle_{V\times V^*}|}{\|w_n\|_V},$$

are called **consistency (error) terms**. They have to be tackled, when we aim to gauge the impact of numerical quadrature quantitatively.

4.7.2 Uniform h-ellipticity

In the sequel, our investigations will focus on (4.18) discretized by means of Lagrangian finite elements of uniform polynomial degree m on a simplicial triangulation $\mathcal{M}$ of a polygonal/polyhedral computational domain Ω.

Applying local quadrature rules of the form (NUQ), the perturbed bilinear form reads

$$\widetilde{\mathsf{b}}(u_n, v_n) := \sum_{K\in\mathcal{M}} |K| \sum_{l=1}^{P_K} \omega_l^K \left\langle \mathbf{A}(\boldsymbol{\pi}_l^K)\, \mathbf{grad}\, u_n(\boldsymbol{\pi}_l^K), \mathbf{grad}\, v_n(\boldsymbol{\pi}_l^K)\right\rangle \quad u_n, v_n \in \mathcal{S}_m(\mathcal{M})\,. \tag{4.26}$$

For the analysis we must rely on a certain smoothness of the coefficient function $\mathbf{A}$:

Assumption 4.34. *The restriction of the coefficient function* $\mathbf{A} : \Omega \mapsto \mathbb{R}^{d,d}$ *to any cell* $K \in \mathcal{M}$ *belongs to* $C^m(K)^{d,d}$ *and can be extended to a function* $\in C^m(\overline{K})^{d,d}$.

Lemma 4.35. *Let* $\mathbf{A}$ *satisfy* (UPD) *and Assumption 4.34, and the local quadrature weights* ω_l^K *be positive. If the local quadrature rules are exact for polynomials up to degree* $2m-2$, *then*

$$\widetilde{\mathsf{b}}(v_n, v_n) \geq \underline{\gamma}\, |v_n|^2_{H^1(\Omega)} \quad \forall v_n \in \mathcal{S}_m(\mathcal{M}) \; .$$

where the constant $\underline{\gamma}$ *is that from* (UPD).

Proof. Since $\mathbf{A}$ is uniformly positive definite and the quadrature weights are positive

$$\begin{aligned}\widetilde{\mathsf{b}}(v_n, v_n) &\geq \sum_{K\in\mathcal{M}} |K| \sum_{l=1}^{P_K} \omega_l^K \underline{\gamma} |\operatorname{\mathbf{grad}} v_n(\boldsymbol{\pi}_l^K)|^2 \geq \underline{\gamma} \sum_{K\in\mathcal{M}} |K| \sum_{l=1}^{P_K} \omega_l^K |\operatorname{\mathbf{grad}} v_n(\boldsymbol{\pi}_l^K)|^2 \\ &= \underline{\gamma}\, |v_n|^2_{H^1(\Omega)} \; ,\end{aligned}$$

because on each $K \in \mathcal{M}$ we know $\operatorname{\mathbf{grad}} v_n \in \mathcal{P}_{m-1}(K)^d$ so that the numerical quadrature of $|\operatorname{\mathbf{grad}} v_n|^2$ is exact. □

4.7.3 Consistency

We first examine the consistency error term arising from numerical quadrature applied to the right hand side. The analysis hinges on the smoothness of the source function f:

Assumption 4.36. *The restriction of the source function* $f : \Omega \mapsto \mathbb{R}$ *to any cell* $K \in \mathcal{M}$ *belongs to* $C^m(K)$ *and can be extended to a function* $\in C^m(\overline{K})$.

The analysis of the consistency errors for the h-version of finite elements is based on transformation techniques, see Sect. 4.2. They can be applied, if the local quadrature rules stem from a single quadrature formula on the reference simplex, *cf.* Sect. 3.9.3.

Assumption 4.37. *We assume that the local quadrature formulas arise from a single quadrature formula on the reference simplex* $\widehat{K}$.

In particular, this means that on each cell the same number of quadrature nodes and the same quadrature weights are used, which amounts to

$$\left\langle \widetilde{f}, w_n \right\rangle_{V\times V^*} = \sum_{K\in\mathcal{M}} |K| \sum_{l=1}^{P} \omega_l \cdot (f w_n)(\boldsymbol{\pi}_l^K) \; .$$

Let us zero in on the case $m = 1$, i.e. linear Lagrangian finite elements, and assume that the quadrature rule is exact for polynomials of degree 0 (constants). As usual, localization will be the first crucial step

$$\langle f, w_n \rangle_{V\times V^*} - \left\langle \widetilde{f}, w_n \right\rangle_{V\times V^*} = \sum_{K\in\mathcal{M}} \int_K f w_n \,\mathrm{d}\boldsymbol{\xi} - |K| \sum_{l=1}^{P} \omega_l \cdot (f w_n)(\boldsymbol{\pi}_l^K) \; .$$

Let us single out a $K \in \mathcal{M}$. Then, we use the transformation formula and carry out transformation to the reference element

$$\int_K f w_n \,\mathrm{d}\boldsymbol{\xi} - |K| \sum_{l=1}^{P} \omega_l (f w_n)(\boldsymbol{\pi}_l^K) = |\det \mathbf{F}_K| \, \widehat{\mathcal{E}}_Q(\widehat{f}\widehat{w}_n) \,, \tag{4.27}$$

where $\widehat{\boldsymbol{\xi}} \mapsto \mathbf{F}_K \widehat{\boldsymbol{\xi}} + \boldsymbol{\tau}_K$ is the affine mapping taking $\widehat{K}$ to K and

$$\widehat{\mathcal{E}}_Q(\widehat{f}\widehat{w}_n) := \int_{\widehat{K}} \widehat{f}\widehat{w}_n \,\mathrm{d}\widehat{\boldsymbol{\xi}} - |K| \sum_{l=1}^{P} \omega_l \cdot (\widehat{f}\widehat{w}_n)(\widehat{\boldsymbol{\pi}}_l) \,. \tag{4.28}$$

The linear continuous quadrature error functional $\widehat{\mathcal{E}}_Q : L^\infty(\widehat{K}) \mapsto \mathbb{R}$ satisfies

$$\widehat{\mathcal{E}}_Q(p) = 0 \quad \forall p \in \mathcal{P}_0(\widehat{K}) \,.$$

Hence, we can appeal to the generalization of the Bramble-Hilbert lemma presented in Remark 4.9.

$$|\widehat{\mathcal{E}}_Q(w)| = \inf_{p \in \mathcal{P}_0(\widehat{K})} |\widehat{\mathcal{E}}_Q(w-p)| \leq \gamma \inf_{p \in \mathcal{P}_0(\widehat{K})} \|w-p\|_{W^{1,\infty}(\widehat{K})} \leq \gamma \, |w|_{W^{1,\infty}(\widehat{K})} \,, \tag{4.29}$$

with universal constants $\gamma > 0$. We want to apply this estimate to $w = \widehat{f}\widehat{w}_n$. The product rule gives

$$\left|\widehat{f}\widehat{w}_n\right|_{W^{1,\infty}(\widehat{K})} \leq \left|\widehat{f}\right|_{W^{1,\infty}(\widehat{K})} \|\widehat{w}_n\|_{L^\infty(\widehat{K})} + |\widehat{w}_n|_{W^{1,\infty}(\widehat{K})} \left\|\widehat{f}\right\|_{L^\infty(\widehat{K})} \,. \tag{4.30}$$

In light of the fact that $\widehat{w}_n \in \mathcal{P}_1(\widehat{K})$ we can use Lemma 4.19 and conclude

$$\left|\widehat{f}\widehat{w}_n\right|_{W^{1,\infty}(\widehat{K})} \leq \left|\widehat{f}\right|_{W^{1,\infty}(\widehat{K})} \|\widehat{w}_n\|_{L^2(\widehat{K})} + |\widehat{w}_n|_{H^1(\widehat{K})} \left\|\widehat{f}\right\|_{L^\infty(\widehat{K})} \,. \tag{4.31}$$

Now, recall from Lemma 4.10 that

$$\|\widehat{w}_n\|_{L^2(\widehat{K})} \leq \gamma h_{\mathcal{M}}^{-d/2} \|w_n\|_{L^2(K)} \quad , \quad |\widehat{w}_n|_{H^1(\widehat{K})} \leq \gamma h_{\mathcal{M}}^{1-d/2} |w_n|_{H^1(K)} \,. \tag{4.32}$$

Further, a simple application of the chain rule reveals

$$\left|\widehat{f}\right|_{W^{1,\infty}(\widehat{K})} \leq \gamma h_{\mathcal{M}} |f|_{W^{1,\infty}(K)} \quad , \quad \|\widehat{f}\|_{L^\infty(\widehat{K})} = \|f\|_{L^\infty(K)} \,. \tag{4.33}$$

In all these estimates the constants only depend on $\rho_{\mathcal{M}}$ and $\mu_{\mathcal{M}}$. Combining (4.29), (4.31), (4.32), and (4.33) we get

$$|\widehat{\mathcal{E}}_Q(\widehat{f}\widehat{w}_n)| \leq \gamma h_{\mathcal{M}}^{1-d/2} \Big(|f|_{W^{1,\infty}(K)} \|v_n\|_{L^2(K)} + \|f\|_{L^\infty(K)} |v_n|_{H^1(K)} \Big) \,.$$

By Cor. 4.17 $|\det \mathbf{F}_K| \leq h_{\mathcal{M}}^d$, which, when combined with (4.27) leads to

$$\int_K f w_n \,\mathrm{d}\boldsymbol{\xi} - |K| \sum_{l=1}^{P} \omega_l(f w_n)(\boldsymbol{\pi}_l^K) \leq \gamma h_{\mathcal{M}}^{1+d/2} \|f\|_{W^{1,\infty}(K)} \|w_n\|_{H^1(K)} \ ,$$

with $\gamma = \gamma(\rho_{Mesh}, \mu_{\mathcal{M}})$. Summation over all cells yields

$$\langle f, w_n \rangle_{V \times V^*} - \left\langle \widetilde{f}, w_n \right\rangle_{V \times V^*} \leq \gamma h_{\mathcal{M}}^{1+d/2} \|f\|_{W^{1,\infty}(\Omega)} \sum_{K \in \mathcal{M}} \|w_n\|_{H^1(K)} \ . \tag{4.34}$$

As a consequence of the Cauchy-Schwarz inequality

$$\sum_{K \in \mathcal{M}} \|w_n\|_{H^1(K)} \leq (\sharp \mathcal{M})^{1/2} \cdot \|w_n\|_{H^1(\Omega)} \ , \tag{4.35}$$

whereas, with a constant $\gamma > 0$ depending on $\rho_{\mathcal{M}}$ and $\mu_{\mathcal{M}}$,

$$\sharp \mathcal{M} \leq \gamma \, |\Omega| \, h_{\mathcal{M}}^{-d} \ . \tag{4.36}$$

Merging (4.34), (4.35), and (4.36) we obtain

$$\langle f, w_n \rangle_{V \times V^*} - \left\langle \widetilde{f}, w_n \right\rangle_{V \times V^*} \leq \gamma h_{\mathcal{M}} \|f\|_{W^{1,\infty}(\Omega)} \|w_n\|_{H^1(\Omega)} \ .$$

This implies a behavior like $O(h_{\mathcal{M}})$ of the consistency error term due to numerical quadrature for the load vector. This estimate matches the a-priori error estimate (4.19) of the H^1-norm of the discretization error. In sense, using a quadrature rule that is exact merely for constants complies with the fundamental guideline about proper use of numerical quadrature.

In [12, Ch. 4, §4.1] a more general theorem is proved:

Theorem 4.38. *If simplicial Lagrangian finite elements of uniform degree $m \in \mathbb{N}$ are used to discretize* (4.18) *and the load vector is computed by a numerical quadrature scheme that complies with Assumption 4.37 and is exact for polynomials up to degree $2m-2$, then*

$$\langle f, w_n \rangle_{V \times V^*} - \left\langle \widetilde{f}, w_n \right\rangle_{V \times V^*} \leq \gamma h_{\mathcal{M}}^m \ \|f\|_{W^{m,\infty}(\Omega)} \|w_n\|_{H^1(\Omega)} \quad \forall w_n \in \mathcal{S}_m(\mathcal{M}) \ ,$$

where $\gamma = \gamma(m, \rho_{\mathcal{M}}, \mu_{\mathcal{M}}) > 0$.

We still have to bound the consistency error term introduced by numerical quadrature applied to the bilinear form b of (4.18). Assumptions 4.34 and 4.37 will remain in effect. Again, we only study the case of linear Lagrangian finite elements $m = 1$. In addition, we will only consider the case of scalar coefficient $\mathbf{A}(\boldsymbol{\xi}) = a(\boldsymbol{\xi})\mathbf{I}_d$.

>From (4.26) we get for $v_n, w_n \in \mathcal{S}_1(\mathcal{M})$, whose gradients are locally constant,

$$\begin{aligned}
&\mathsf{b}(v_n, w_n) - \widetilde{\mathsf{b}}(v_n, w_n) \\
&\quad = \sum_{K\in\mathcal{M}} \int_K \langle \mathbf{A}\,\mathbf{grad}\, v_n, \mathbf{grad}\, w_n\rangle \,\mathrm{d}\boldsymbol{\xi} - |K| \sum_{l=1}^{P} \omega_l (\langle \mathbf{A}\,\mathbf{grad}\, v_n, \mathbf{grad}\, w_n\rangle)(\boldsymbol{\pi}_l^K) \\
&\quad = \sum_{K\in\mathcal{M}} \langle \mathbf{grad}\, v_n, \mathbf{grad}\, w_n\rangle_{|K} \left(\int_K a(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi} - |K| \sum_{l=1}^{P} \omega_l a(\boldsymbol{\pi}_l^K) \right) \\
&\quad = \sum_{K\in\mathcal{M}} \langle \mathbf{grad}\, v_n, \mathbf{grad}\, w_n\rangle_{|K} \,|\det(\mathbf{F}_K)|\, \widehat{\mathcal{E}}_Q(\widehat{a}) \;,
\end{aligned}$$

where we used the quadrature error term (4.28) on the reference simplex. Note that $\widehat{a} = \mathsf{FT}_{\mathbf{\Phi}_K}\, a$ is a pullback that depends on K. We single out a cell $K \in \mathcal{M}$. Then we can reuse (4.29), (4.33) and get

$$\widehat{\mathcal{E}}_Q(\widehat{a}) \le \gamma\, |\widehat{a}|_{W^{1,\infty}(\widehat{K})} \le \gamma h_{\mathcal{M}}\, |a|_{W^{1,\infty}(K)} \;,$$

where $\gamma = \gamma(\rho_{\mathcal{M}}, \mu_{\mathcal{M}}) > 0$. Eventually,

$$\begin{aligned}
\mathsf{b}(v_n, w_n) - \widetilde{\mathsf{b}}(v_n, w_n) &\le \gamma h_{\mathcal{M}} \sum_{K\in\mathcal{M}} |\det(\mathbf{F}_K)|\, \langle \mathbf{grad}\, v_n, \mathbf{grad}\, w_n\rangle_{|K}\, |a|_{W^{1,\infty}(K)} \\
&\le \gamma h_{\mathcal{M}}\; |a|_{W^{1,\infty}(\Omega)}\; (v_n, w_n)_{H^1(\Omega)} \\
&\le \gamma h_{\mathcal{M}}\; |a|_{W^{1,\infty}(\Omega)}\, \|v_n\|_{H^1(\Omega)}\, \|w_n\|_{H^1(\Omega)} \;.
\end{aligned}$$

This means that the corresponding consistency error term displays a dependence like $O(h_{\mathcal{M}})$ (with constants depending on shape-regularity and quasi-uniformity). Again, the weakest quadrature rule turns out to be sufficient to preserve first order convergence in the meshwidth.

A more general result is proved in [12, Ch. 4,§4.1]:

Theorem 4.39. *Assume that the exact solution u of (4.18) belongs to $H^{m+1}(\Omega)$, $m \in \mathbb{N}$. If simplicial Lagrangian finite elements of uniform degree m are used to discretize (4.18) and $\widetilde{\mathsf{b}}$ arises from a numerical quadrature scheme that fits Assumption 4.37 and is exact for polynomials up to degree $2m-2$, then*

$$\mathsf{b}(\mathsf{I}\, u, w_n) - \widetilde{\mathsf{b}}(\mathsf{I}\, u, w_n) \le \gamma h_{\mathcal{M}}^m\; \|\mathbf{A}\|_{W^{m,\infty}(\Omega)}\, \|u\|_{H^{m+1}(\Omega)}\; \|w_n\|_{H^1(\Omega)} \quad \forall w_n \in \mathcal{S}_m(\mathcal{M}) \;,$$

where $\gamma = \gamma(\mathbf{A}, m, \rho_{\mathcal{M}}, \mu_{\mathcal{M}}) > 0$.

5 Special Finite Element Methods

In this chapter we discuss the theoretical aspect of finite element methods that do not belong to the standard finite element schemes introduced in Ch. 3 and analyzed in the previous chapter.

5.1 Non-conforming finite element schemes

Definition 5.1. *A finite element method is called* ***non-conforming****, if the finite element trial or test space is not a subspace of the function spaces that occur in the definition of the continuous variational problem.*

A rationale for using such a peculiar approach will be given in Sect. 5.3.4.

Remark 5.2. Another rationale for using non-conforming finite elements is that conforming elements might be dismissed as too complicated. This has been the main reason for using non-conforming schemes for variational problems posed on $H^2(\Omega)$, see Sect. 3.8.3.

5.1.1 Abstract theory

We consider the linear variational problem (LVP)

$$u \in V : \quad \mathsf{b}(u,v) = \langle f, v\rangle_{V^*\times V} \quad \forall v \in V \ ,$$

where $\mathsf{b} \in L(V \times V, \mathbb{R})$ is a V-elliptic bilinear form, and $f \in V^*$.

In the case of a non-conforming Galerkin discretization based on the vector space V_n the discrete variational problem reads

$$u_n \in V_n : \quad \widetilde{\mathsf{b}}(u_n, v_n) = \left\langle \widetilde{f}, v_n \right\rangle_{V_n^*\times V_n} \quad \forall v_n \in V_n \ . \tag{5.1}$$

Since $V_n \not\subset V$ is admitted, the original bilinear form b might not make sense for elements of V_n. Therefore it has to be replaced by $\widetilde{\mathsf{b}} : V_n \times V_n \mapsto \mathbb{R}$. The same is true of the right hand side functional.

Strictly speaking, (5.1) has nothing to do with (LVP), we cannot even compare the respective solutions $u \in V$ and $u_n \in V_n$ in the V-norm. In order to state discretization

error estimates we have to introduce a special norm $\|\cdot\|_{h,V}$ for functions in $V_n \cup V$ that agrees with $\|\cdot\|_V$ on V.

Remark 5.3. In most practical cases $\|\cdot\|_{h,V}$ will be a **split norm**, adding contributions from mesh cells. In a sense, it is a mesh dependent norm.

We make the following assumptions on $\widetilde{\mathsf{b}}$:

$$\exists \gamma_C > 0: \quad \widetilde{\mathsf{b}}(v, w_n) \le \gamma_C \|v_n\|_{h,V} \|w_n\|_{h,V} \qquad \forall v \in V \cup V_n,\ w_n \in V_n\ , \tag{5.2}$$

$$\exists \gamma_1 > 0: \quad \widetilde{\mathsf{b}}(v_n, v_n) \ge \gamma_1 \|v_n\|_{h,V}^2 \qquad \forall v_n \in V_n \tag{5.3}$$

Here, (5.2) asserts the *continuity* of $\widetilde{\mathsf{b}}$, and (5.3) is known as h**-ellipticity**. Under these assumptions, Thm. 1.17 ensures existence and uniqueness of a solution $u_n \in V_n$ of (5.1). Its relationship to the solution $u \in V$ of (LVP) is disclosed by the next theorem.

Theorem 5.4 (Strang's second lemma). *If $\mathsf{b} \in L(V \times V, \mathbb{R})$ is V-elliptic according to Def. 1.20 and $\widetilde{\mathsf{b}}$ satisfies* (5.2) *and* (5.3)*, then*

$$\|u - u_n\|_{h,V} \le \gamma \left(\inf_{v_n \in V_n} \|u - v_n\|_{h,V} + \sup_{w_n \in V_n} \frac{\left| \widetilde{\mathsf{b}}(u, w_n) - \left\langle \widetilde{f}, w_n \right\rangle_{V_n^* \times V_n} \right|}{\|w_n\|_{h,V}} \right) ,$$

with $\gamma = \gamma(\gamma_C, \gamma_1)$.

Proof. We pick an arbitrary $v_n \in V_n$ and get from (5.3) and (5.2)

$$\begin{aligned} \gamma_1 \|u_n - v_n\|_{h,V}^2 &\le \widetilde{\mathsf{b}}(u_n - v_n, u_n - v_n) \\ &= \widetilde{\mathsf{b}}(u - v_n, u_n - v_n) + \left(\left\langle \widetilde{f}, u_n - v_n \right\rangle_{V_n^* \times V_n} - \widetilde{\mathsf{b}}(u, u_n - v_n) \right) \\ &\le \gamma_C \|u - v_n\|_{h,V} \|u_n - v_n\|_{h,V} + \sup_{w_n \in V_n} \left(\frac{\left| \widetilde{\mathsf{b}}(u, w_n) - \left\langle \widetilde{f}, w_n \right\rangle_{V_n^* \times V_n} \right|}{\|w_n\|_{h,V}} \right) \end{aligned}$$

As $v_n \in V_n$ was arbitrary, the assertion of the theorem follows from the triangle inequality as in the proof of Thm. 4.33. □

The extra term in the a-priori error estimate is another **consistency error term**.

5.1.2 The Crouzeix-Raviart element

Definition 5.5. *For a triangle $K \in \mathbb{R}^2$ the* ***Crouzeix-Raviart finite element*** *is defined by*

(i) $\Pi_K = \mathcal{P}_1(K)$,

(ii) $\Sigma_K = \{v \mapsto |F_i|^{-1} \int_{F_i} v(\boldsymbol{\xi}) \,\mathrm{d}S(\boldsymbol{\xi}),\ i = 1,2,3\}$, *where* F_i *is the edge opposite to vertex* i, $i = 1,2,3$.

Exercise 5.1. Compute the local shape functions for the Crouzeix-Raviart element in terms of barycentric coordinate functions.

Remark 5.6. On Π_K indentical local degrees of freedom can be defined by $\Sigma_K = \{v \mapsto v(\boldsymbol{\mu}_i),\ i = 1,2,3\}$, where $\boldsymbol{\mu}_i$ is the midpoint of the edge opposite to vertex i. However, when evaluated for more general functions, these d.o.f. are clearly different.

Corollary 5.7. *The Crouzeix-Raviart element is affine equivalent.*

Obviously, this finite element is not H^1-conforming, because there is only one local degree of freedom associated with an edge, which must fail to fix the restriction of a local trial function onto that edge, *cf.* Example 3.27.

The global degrees of freedom of the Crouzeix-Raviart finite element coincide with averages over the edges of a simplicial triangulation. For a functions that are linear on an edge the same average means that they coincide at the midpoint of the edge. Hence, the Crouzeix-Raviart finite element space reads

$$\begin{aligned}\mathcal{CR}(\mathcal{M}) := \{\, v \in L^2(\Omega) :\ & v_{|K} \in \mathcal{P}_1(K)\, \forall K \in \mathcal{M}, \\ & v \text{ continuous at midpoints of interior edges of } \mathcal{M}\} \,.\end{aligned}$$

A global shape function is sketched in Fig. 5.1.

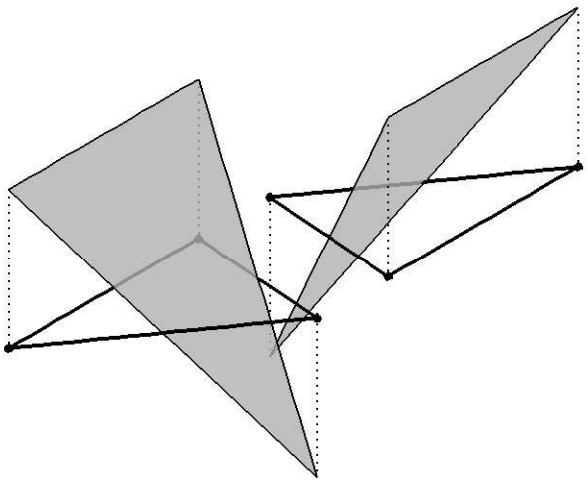


Figure 5.1: Global shape function for Crouzeix-Raviart element

Lemma 5.8. *If $\mathsf{I}_{\mathcal{CR}}$ denotes the finite element interpolation operator associated with $\mathcal{CR}(\mathcal{M})$ on a two-dimensional simplicial triangulation $\mathcal{M}$ of $\Omega \subset \mathbb{R}^2$, then for $0 \le r \le t \le 2$, $t \ge 1$,*

$$\sum_{K\in\mathcal{M}} \|u - \mathsf{I}_{\mathcal{CR}}\, u\|^2_{H^r(K)} \le \gamma h_{\mathcal{M}}^{2(t-r)} \, |u|^2_{H^t(\Omega)} \quad \forall u \in H^t(\Omega) \; ,$$

with $\gamma = \gamma(t, r, \rho_{\mathcal{M}}) > 0$.

Proof. Local transformation techniques in conjunction with the Bramble-Hilbert Lemma 4.7 and Lemma 4.10 accomplish the proof as in the case of Thm. 4.24. □

In contrast to the finite element interpolation operators for Lagrangian finite elements, $\mathsf{I}_{\mathcal{CR}}$ is bounded on $H^1(\Omega)$:

Lemma 5.9. $\quad \exists \gamma = \gamma(\rho_{\mathcal{M}}) : \quad \sum\nolimits_{K\in\mathcal{M}} |\mathsf{I}_{\mathcal{CR}}\, v|_{H^1(K)} \le \gamma \, |v|_{H^1(\Omega)} \quad \forall v \in H^1(\Omega)$

Proof. Pick any triangle $K \in \mathcal{M}$ and remember that the Crouzeix-Raviart element is affine equivalent. Thus, by the transformation formulas of Lemma 4.10, the equivalence of norms on finite dimensional spaces (Lemma 4.19), and the trace theorem Thm. 2.49,

$$|\mathsf{I}_{\mathcal{CR}}\, v|_{H^1(K)} \le \gamma \left|\widehat{\mathsf{I}}_{\mathcal{CR}} \widehat{v}\right|_{H^1(\widehat{K})} \le \gamma \sum_{j=1}^{3} |\widehat{F}_j|^{-1} \left| \int_{\widehat{F}_j} v \, \mathrm{d}S \right| \le \gamma \, |\widehat{v}|_{H^1(\widehat{K})} \le \gamma \, |v|_{H^1(K)} \; ,$$

where all constants only depend on ρ_K. □

The Crouzeix-Raviart finite element on a triangulation of a polygonal computational domain $\Omega \subset \mathbb{R}^2$ shall be used to solve

$$-\Delta u = f \in L^2(\Omega) \;\text{ in } \Omega \quad , \quad u = 0 \;\text{ on } \Gamma \; ,$$

which amounts to the variational problem

$$u \in H^1_0(\Omega) : \quad \mathsf{b}(u, v) := \int_\Omega \langle \mathbf{grad}\, u, \mathbf{grad}\, v \rangle \, \mathrm{d}\boldsymbol{\xi} = \int_\Omega f v \, \mathrm{d}\boldsymbol{\xi} \quad \forall v \in H^1_0(\Omega) \; . \tag{5.4}$$

The usual way to impose the essential homogeneous Dirichlet boundary condition is to set global degrees of freedom located on Γ to zero. We also do this for the Crouzeix-Raviart element, so that we get the trial/test space

$$V_n := \{ v \in \mathcal{CR}(\mathcal{M}) : \; v = 0 \text{ at midpoints of edges } \in \mathcal{E}(\mathcal{M}) \cap \Gamma \} \; .$$

In this setting the natural candidate for the mesh-dependent norm is

$$\|v\|_{h,V} := \left(\sum\nolimits_{K\in\mathcal{M}} |v|^2_{H^1(K)} \right)^{1/2} .$$

For $v \in H^1(\Omega)$ this agrees with $|\cdot|_{H^1(\Omega)}$, which, by the Poincaré-Friedrichs inequality Lemma 2.61, is a norm on $H^1_0(\Omega)$. That $\|\cdot\|_{h,V}$ is a norm on V_n as well can be seen by the following argument: if $\|v_n\|_{h,V} = 0$ the function v_n must be locally constant. Continuity at midpoints of edges then means that $v_n \in V_n$ is constant on Ω. As it vanishes in some points on Γ, we conclude $v_n = 0$.

The modified bilinear form will read

$$\widetilde{\mathsf{b}}(u,v) := \sum_{K\in\mathcal{M}} \int_K \langle \mathbf{grad}\, u, \mathbf{grad}\, v\rangle \,\mathrm{d}\boldsymbol{\xi} \,.$$

For $u, v \in H^1(\Omega)$ is agrees with b. We observe that (5.2) and (5.3) are trivially satisfied with $\gamma_C = \gamma_1 = 1$. Besides, we notice that the right hand side functional need not be altered.

In order to estimate the consistency error term, we have to *assume* that the solution u of (5.4) belongs to $H^2(\Omega)$. Then we can use local integration by parts in order to bound the consistency error term:

$$\begin{aligned}
\widetilde{\mathsf{b}}(u,w_n) - \langle f, w_n\rangle_{H^{-1}(\Omega)\times H^1_0(\Omega)} &= \sum_{K\in\mathcal{M}} \int_K \langle \mathbf{grad}\, u, \mathbf{grad}\, w_n\rangle - f w_n \,\mathrm{d}\boldsymbol{\xi} \\
&= \sum_{K\in\mathcal{M}} \int_{\partial K} \langle \mathbf{grad}\, u, \mathbf{n}_{\partial K}\rangle \; w_n \,\mathrm{d}S - \int_K (\Delta u + f)\, w_n \,\mathrm{d}\boldsymbol{\xi} \\
&= \sum_{K\in\mathcal{M}} \int_{\partial K} \langle \mathbf{grad}\, u, \mathbf{n}_{\partial K}\rangle \; w_n \,\mathrm{d}S \,,
\end{aligned}$$

because $-\Delta u = f$.

$$= \sum_{K\in\mathcal{M}} \sum_{j=1}^{3} \int_{F_j} \langle \mathbf{grad}\, u, \mathbf{n}_{\partial K}\rangle \; (w_n - w_n(F_j)) \,\mathrm{d}S \,,$$

where $w_n(F_j) \in \mathbb{R}$ is the unique value of $w_n \in V_n$ at the midpoint of the edge F_j. We can subtract it, because if

1. F_j is an interior edge, it will occur twice in the sum, with opposite sign, however, due to the different directions of the unit normal vectors, and if
2. $F_j \subset \Gamma$, then, by virtue of the definition of V_n, $w_n(F_j) = 0$.

Now, $w_n - w_n(F_j)$ has average zero on F_j:

$$\int_F w_n - w_n(F_j) \,\mathrm{d}S = 0 \quad \forall F \in \mathcal{E}(\mathcal{M}) \,.$$

Since $\langle \mathbf{grad}\, \mathsf{I}\, u, \mathbf{n}_{\partial K} \rangle_{|F} \in \mathcal{P}_0(F)$, F an edge of K, it is now possible to replace $\langle \mathbf{grad}\, u, \mathbf{n}_{\partial K} \rangle_{F_j}$ with $\langle \mathbf{grad}(u - \mathsf{I}\, u), \mathbf{n}_{\partial K} \rangle_{F_j}$, where I is the finite element interpolation operator for the Lagrangian finite element space $\mathcal{S}_1(\mathcal{M})$. Eventually, we get by the Cauchy-Schwarz inequality

$$\begin{aligned} \widetilde{\mathsf{b}}(u, w_n) - \langle f, w_n \rangle_{H^{-1}(\Omega)\times H^1_0(\Omega)} &= \sum_{K\in\mathcal{M}} \sum_{j=1}^{3} \int_{F_j} \langle \mathbf{grad}(u - \mathsf{I}\, u), \mathbf{n}_{\partial K} \rangle \, (w_n - w_n(F_j)) \, \mathrm{d}S \\ &\leq \sum_{K\in\mathcal{M}} \sum_{j=1}^{3} |u - \mathsf{I}\, u|_{H^1(F_j)} \, \|w_n - w_n(F_j)\|_{L^2(F_j)} \ . \end{aligned} \tag{5.5}$$

On the reference element $\widehat{K}$ the multiplicative trace theorem Thm. 2.49 and the Bramble-Hilbert lemma Lemma 4.7 permit us to estimate

$$\left|\widehat{u} - \widehat{\mathsf{I}}\widehat{u}\right|^2_{H^1(\partial\widehat{K})} \leq \gamma \left|\widehat{u} - \widehat{\mathsf{I}}\widehat{u}\right|_{H^1(\widehat{K})} \left(\left|\widehat{u} - \widehat{\mathsf{I}}\widehat{u}\right|_{H^1(\widehat{K})} + |\widehat{u}|_{H^2(\widehat{K})} \right) \leq \gamma \, |\widehat{u}|^2_{H^2(\widehat{K})} \quad \widehat{u} \in H^2(\widehat{K}) \ .$$

By means of transformation techniques we infer

$$|u - \mathsf{I}\, u|^2_{H^1(\partial K)} \leq \gamma h_K \ |u|^2_{H^2(K)} \ . \tag{5.6}$$

The constant will depend on ρ_K. Moreover, elementary computations show

$$\|w_n - w_n(F_j)\|^2_{L^2(F_j)} \leq \gamma(\rho_K) h_K \ |w_n|^2_{H^1(K)} \ . \tag{5.7}$$

We plug (5.6) and (5.6) into (5.5) and get

$$\begin{aligned} \widetilde{\mathsf{b}}(u, w_n) - \langle f, w_n \rangle_{H^{-1}(\Omega)\times H^1_0(\Omega)} &\leq \gamma \sum_{K\in\mathcal{M}} h_K \, |u|_{H^2(K)} \ |w_n|_{H^1(K)} \\ &\leq \gamma \, h_{\mathcal{M}} \, |u|_{H^2(\Omega)} \ \|w_n\|_{h,V} \ , \end{aligned}$$

where $\gamma = \gamma(\rho_{\mathcal{M}})$. This demonstrates that the consistency error term is of order $O(h_{\mathcal{M}})$ as $h_{\mathcal{M}} \to 0$ for a uniformly shape-regular sequence of simplicial meshes.

The best approximation error for V_n can be estimated easily, because $\mathcal{S}_1(\mathcal{M}) \subset \mathcal{CR}(\mathcal{M})$: from Thm. 4.24 we get

$$\inf_{v_n\in V_n} \|u - v_n\|_{h,V} \leq \inf_{w_n\in\mathcal{S}_1(\mathcal{M})} |u - w_n|_{H^1(\Omega)} \leq \gamma h_{\mathcal{M}} \ |u|_{H^2(\Omega)} \ ,$$

with $\gamma = \gamma(\rho_{\mathcal{M}}, \mu_{\mathcal{M}})$. The final result for the discretization of (5.4) by means of non-conforming Crouzeix-Raviart elements is

$$\|u - u_n\|_{h,V} \leq \gamma(\rho_{\mathcal{M}}, \mu_{\mathcal{M}}) \, h_{\mathcal{M}} \ |u|_{H^2(\Omega)} \ .$$

Exercise 5.2. The elliptic boundary value problem

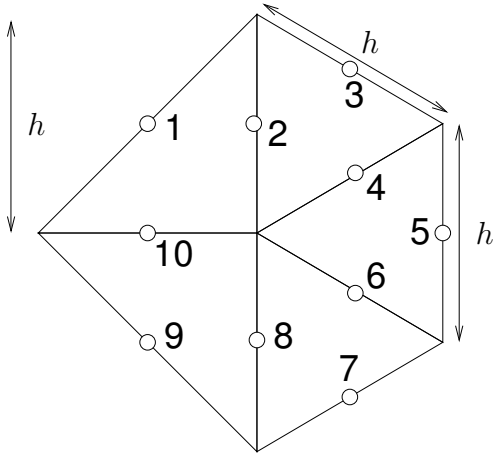
$$-\Delta u + \tau u = f \quad \text{in } \Omega \quad , \quad \langle \mathbf{grad}\, u, \mathbf{n} \rangle = 0 \quad \text{on } \Gamma \,, \tag{5.8}$$

$\tau > 0$ shall be discretized by means of Crouzeiz-Raviart elements on a simplicial triangulation in two dimensions.

(i) Compute the element matrix for an arbitrary triangle with vertices $\boldsymbol{\nu}^1, \boldsymbol{\nu}^2, \boldsymbol{\nu}^3 \in \mathbb{R}^2$.

(ii) Let the triangular grid of Fig. 3.5 have M, $M \in \mathbb{N}$, cells in each coordinate directions. How many unknowns will we encounter in the case of the Crouzeix-Raviart finite element scheme for (5.8)? How many will arise when using piecewise linear H^1-conforming Lagrangian finite elements? How many non-zero entries will the stiffness matrix have in either case?

(iii) Compute the stiffness matrix that arises on the triangular mesh sketched beside. Use the numbering of the global degrees of freedom, which are associated with the edges of the mesh, given in the figure.



Exercise 5.3. The Crouzeix-Raviart element can be generalized to quadrilaterals: the **Rannacher-Turek** finite element on a straight-sided quadrilateral $K \subset \mathbb{R}^2$ is defined by

- $\Pi_K = \operatorname{span}\{1, \xi_1, \xi_2, \xi_1^2 - \xi_2^2\}$, and
- $\Sigma_K = \{v \mapsto |F_i|^{-1} \int_{F_i} v(\xi)\, \mathrm{d}S(\xi),\ i = 1, 2, 3, 4\}$, where $F_1, \ldots, F_4$ are the edges of K.

(i) Verify that (K, Π_K, Σ_K) is a finite element according to Def. 3.25, if K is an axiparallel rectangle.

(ii) Show that this finite element is H^1-nonconforming.

(iii) Describe the local shape functions for the Rannacher-Turek element on an axiparallel rectangle K.

5.2 Mixed finite elements for second-order elliptic boundary value problems

5.2.1 Dual variational problem

In Sect. 2.5 we derived the dual variational formulation of the second-order elliptic boundary value problem

$$\begin{aligned} \mathbf{A}^{-1}\mathbf{j} &= \mathbf{grad}\, u \quad \text{in } \Omega\,, && (5.9)\\ \operatorname{div}\mathbf{j} &= f \quad \text{in } \Omega\,, && (5.10)\\ u &= g \in H^{1/2}(\Gamma) \quad \text{on } \Gamma\,. \end{aligned}$$

by multiplying both equations with test functions and using integration by parts for (5.9). This results in the linear variational problem: seek $u \in L^2(\Omega)$, $\mathbf{j} \in H(\operatorname{div};\Omega)$, such that

$$\begin{aligned} \int_\Omega \langle \mathbf{A}^{-1}\mathbf{j}, \mathbf{q}\rangle \,\mathrm{d}\boldsymbol{\xi} \;+\; \int_\Omega \operatorname{div}\mathbf{q}\cdot u \,\mathrm{d}\boldsymbol{\xi} &= \int_\Gamma g\,\langle \mathbf{q}, \mathbf{n}\rangle \,\mathrm{d}S \qquad \forall \mathbf{q} \in H(\operatorname{div};\Omega)\,,\\ \int_\Omega \operatorname{div}\mathbf{j}\cdot v \,\mathrm{d}\boldsymbol{\xi} &= \int_\Omega f v \,\mathrm{d}\boldsymbol{\xi} \qquad \forall v \in L^2(\Omega)\,. \end{aligned} \tag{5.11}$$

Following the guideline stated in Sect. 2.7 we picked Sobolev spaces that barely render the bilinear forms continuous.

5.2.2 Abstract variational saddle point problems

Let V and Q represent two Hilbert spaces. Given two bilinear forms $\mathsf{a} \in L(V\times V, \mathbb{R})$ and $\mathsf{b} \in L(V\times Q, \mathbb{R})$ and linear forms $g \in V^*, f \in Q^*$, the **abstract saddle point problem** reads: seek $\mathbf{j} \in V$, $u \in Q$ such that

$$\begin{aligned} \mathsf{a}(\mathbf{j},\mathbf{v}) \;+\; \mathsf{b}(\mathbf{v},u) &= \langle g, \mathbf{v}\rangle_{V^*\times V} \qquad \forall \mathbf{v} \in V\,,\\ \mathsf{b}(\mathbf{j},q) &= \langle f, q\rangle_{Q^*\times Q} \qquad \forall q \in Q\,. \end{aligned} \tag{SPP}$$

The meaning of a, b, f, and g for the concrete saddle point problem (5.11) should be clear.

Similar to the approach in Sect. 1.2, *cf.* (1.13), an operator notation of (SPP) can promote understanding: introduce the continuous operators $\mathsf{A} : V \mapsto V^*$ and $\mathsf{B} : V \mapsto Q^*$ associated with the bilinear forms a, b, respectively. Then, (SPP) can be converted into

$$\begin{pmatrix} \mathsf{A} & \mathsf{B}^* \\ \mathsf{B} & 0 \end{pmatrix} \begin{pmatrix} \mathbf{j} \\ u \end{pmatrix} = \begin{pmatrix} g \\ f \end{pmatrix}\,. \tag{5.12}$$

Existence and uniqueness of solutions of (5.12) for any $(g, f) \in V^* \times Q^*$ is guaranteed if

(A) $\mathsf{B} : V \mapsto Q^*$ is surjective and $B^* : Q \mapsto V^*$ is injective.

(B) $\mathsf{A} : \mathcal{N}(\mathsf{B}) \mapsto \mathcal{N}(\mathsf{B})^*$ is bijective, where

$$\mathcal{N}(\mathsf{B}) := \{\mathbf{v} \in V : \ \mathsf{b}(\mathbf{v}, q) = 0 \, \forall q \in Q\} \subset V \ .$$

is the null space of B.

A sufficient condition for (A) is given in the next lemma. It is closely related to Thm. 1.17.

Lemma 5.10. *If and only if* $\mathsf{b} \in L(V \times Q, \mathbb{R})$ *satisfies the* ***inf-sup condition***

$$\exists \beta > 0 : \quad \sup_{\mathbf{v} \in V \setminus \{0\}} \frac{\mathsf{b}(\mathbf{v}, q)}{\|\mathbf{v}\|_V} \geq \beta \, \|q\|_Q \quad \forall q \in Q \ , \tag{5.13}$$

then $\mathsf{B} : V \mapsto Q^*$ *is surjective and* $B^* : Q \mapsto V^*$ *is injective.*

Proof. "⇒": Assume that (5.13) holds and consider the restriction $\mathsf{B}^\perp$ of B to the orthogonal complement $\mathcal{N}(\mathsf{B})^\perp$ of $\mathcal{N}(\mathsf{B})$ in V. This is a Hilbert space, too, and $\mathsf{B}^\perp$ meets all assumptions of Thm. 1.17. Hence,

$$\left(\mathsf{B}^\perp\right)^* : Q \mapsto \left(\mathcal{N}(\mathsf{B})^\perp\right)^*$$

is an isomorphism, which makes

$$\mathsf{B}^\perp : \mathcal{N}(\mathsf{B})^\perp \mapsto Q^*$$

an isomorphism as well as in Exercise 1.2.

"⇒": As above, apply Thm. 1.17 to $\mathsf{B}^\perp : \mathcal{N}(\mathsf{B})^\perp \mapsto Q^*$. □

A sufficient condition for (B) is the V-ellipticity of a on $\mathcal{N}(\mathsf{B})$:

$$\exists \alpha > 0 : \quad \mathsf{a}(\mathbf{v}, \mathbf{v}) \geq \alpha \, \|\mathbf{v}\|_V^2 \quad \forall \mathbf{v} \in V \ . \tag{5.14}$$

Theorem 5.11. *If* (5.13) *and* (5.14) *hold, then* (SPP) *has a unique solution* $(\mathbf{j}, u) \in V \times Q$ *for all* $(g, f) \in V^* \times Q^*$*, which satisfies*

$$\begin{aligned} \|\mathbf{j}\|_V &\leq \frac{1}{\alpha} \, \|g\|_{V^*} + \frac{1}{\beta} \left(1 + \frac{\|\mathsf{a}\|}{\alpha}\right) \|f\|_{Q^*} \ , \\ \|u\|_Q &\leq \frac{1}{\beta} \left(\|g\|_{V^*} + \|\mathsf{a}\| \, \|\mathbf{j}\|_V\right) \end{aligned}$$

Proof. Existence and uniqueness of solution of (SPP) is an immediate consequence of Lemma 5.10: thanks to this lemma we can find $\mathbf{w} \in V$ such that

$$\mathsf{B} \, \mathbf{w} = f \quad \text{and} \quad \|\mathbf{w}\|_V \leq \beta^{-1} \, \|f\|_{Q^*} \ .$$

Testing the first equation of (SPP) with $\mathbf{v}^0 \in \mathcal{N}(\mathsf{B})$ we get

$$\mathsf{a}(\mathbf{j} - \mathbf{w}, \mathbf{v}^0) = g(\mathbf{v}^0) - \mathsf{a}(\mathbf{w}, \mathbf{v}^0) \ .$$

This gives the solution component $\mathbf{j}$. From (5.14) and the triangle inequality we conclude the estimate for $\|\mathbf{j}\|_V$. Then u can be defined by

$$u \in Q : \quad \mathsf{b}(\mathbf{v}, u) = g(\mathbf{v}) - \mathsf{a}(\mathbf{j}, \mathbf{v}) \quad \forall \mathbf{v} \in V \ .$$

The estimate for $\|u\|_Q$ follows straight from (5.13). □

Let us verify the assumptions (5.13) and (5.14) for the dual variational problem (5.11). Here we have

$$V = H(\mathrm{div}; \Omega) \quad , \quad Q = L^2(\Omega) \ ,$$
$$\mathsf{a}(\mathbf{j}, \mathbf{v}) := \int_\Omega \left\langle \mathbf{A}^{-1}\mathbf{j}, \mathbf{v} \right\rangle \, \mathrm{d}\boldsymbol{\xi} \ , \quad \mathsf{b}(\mathbf{v}, q) := \int_\Omega \mathrm{div}\, \mathbf{v} \cdot q \, \mathrm{d}\boldsymbol{\xi} \ .$$

It is clear that in this case $\mathcal{N}(\mathsf{B}) := H(\mathrm{div}\, 0; \Omega)$ and (5.14) is immediate from the property (UPD) of the coefficient function $\mathbf{A}$ and the definition of the $H(\mathrm{div})$-norm.

In order to show (5.13) we pick $q \in L^2(\Omega)$ and define $\widetilde{\mathbf{v}} := -\,\mathbf{grad}\, w$, where

$$w \in H^1_0(\Omega) : \quad \int_\Omega \langle \mathbf{grad}\, w, \mathbf{grad}\, v \rangle \, \mathrm{d}\boldsymbol{\xi} = \int_\Omega qv \, \mathrm{d}\boldsymbol{\xi} \quad \forall v \in H^1_0(\Omega) \ . \tag{5.15}$$

Obviously, by the Poincaré-Friedrichs inequality Lemma 2.61,

$$\|\widetilde{\mathbf{v}}\|_{L^2(\Omega)} \leq \gamma(\Omega) \ \|q\|_{L^2(\Omega)} \ . \tag{5.16}$$

Then, since by definition of w in 5.15

$$-\Delta w = q \quad \Rightarrow \quad \mathrm{div}\, \widetilde{\mathbf{v}} = q \ ,$$

(5.13) is readily established with $\beta = (1 + \gamma(\Omega)^2)^{-1/2}$:

$$\sup_{\mathbf{v} \in H^1_0(\Omega)\setminus\{0\}} \frac{\mathsf{b}(\mathbf{v}, q)}{\|\mathbf{v}\|_V} \geq \frac{\mathsf{b}(\widetilde{\mathbf{v}}, q)}{\|\widetilde{\mathbf{v}}\|_{H(\mathrm{div};\Omega)}} = \frac{\|q\|^2_{L^2(\Omega)}}{\sqrt{\|q\|^2_{L^2(\Omega)} + \|\widetilde{\mathbf{v}}\|^2_{L^2(\Omega)}}} \geq \frac{1}{\sqrt{1 + \gamma(\Omega)^2}} \ \|q\|_{L^2(\Omega)} \ .$$

Exercise 5.4. Show that the abstract variational saddle point problem (SPP) is equivalent to the linear variational problem

$$\mathfrak{u} \in X : \quad \mathsf{c}(\mathfrak{u}, \mathfrak{v}) = \langle h, \mathfrak{v} \rangle_{X^* \times X} \quad \forall \mathfrak{v} \in X \ ,$$

where $X = V \times Q$, $\mathfrak{u} = (u_V, u_Q)$, $\mathfrak{v} = (v_V, v_Q)$,

$$\mathsf{c}(\mathfrak{u}, \mathfrak{v}) = (u_V, v_V) + \mathsf{b}(u_V, v_Q) + \mathsf{b}(v_V, u_Q) \ ,$$
$$\langle h, \mathfrak{v} \rangle_{X^* \times X} = \langle g, v_V \rangle_{V^* \times V} + \langle g, v_Q \rangle_{Q^* \times Q} \ .$$

5.2.3 Discrete variational saddle point problems

The canonical Galerkin discretization of (SPP) relies on finite dimensional trial and test spaces $V_n \subset V$, $Q_n \subset Q$: seek $(\mathbf{j}_n, u_n) \in V_n \times Q_n$ such that

$$\begin{array}{rcll} \mathsf{a}(\mathbf{j}_n, \mathbf{v}_n) + \mathsf{b}(\mathbf{v}_n, u_n) &=& \langle g, \mathbf{v}_n \rangle_{V^* \times V} & \forall \mathbf{v}_n \in V_n \ , \\ \mathsf{b}(\mathbf{j}_n, q_n) &=& \langle f, q_n \rangle_{Q^* \times Q} & \forall q_n \in Q_n \ . \end{array} \tag{DSPP}$$

According to the theory developed in the previous subsection, the variational problem (DSPP) has a uniqe solution $(\mathbf{j}_n, u_n)$, if the **Babuška-Brezzi conditions** (sometimes called the Ladyshenskaya-Babuška-Brezzi conditions, hence LBB-conditions)

$$\exists \beta_n > 0 : \quad \sup_{\mathbf{v}_n \in V_n} \frac{b(\mathbf{v}_n, q_n)}{\|\mathbf{v}_n\|_V} \geq \beta_n \|q_n\|_Q \quad \forall q_n \in Q_n \ , \tag{LBB1}$$

$$\exists \alpha_n > 0 : \quad a(\mathbf{v}_n, \mathbf{v}_n) \geq \alpha_n \|\mathbf{v}_n\|_V^2 \quad \forall \mathbf{v}_n \in \mathcal{N}(\mathsf{B}_n) \ , \tag{LBB2}$$

are satisfied, where

$$\mathcal{N}(\mathsf{B}_n) := \{\mathbf{v}_n \in V_n : \ \mathsf{b}(\mathbf{v}_n, q_n) = 0 \ \forall q_n \in Q_n\} \ . \tag{5.17}$$

Remark 5.12. Of course, the relation

$$\dim V_n \geq \dim Q_n \tag{5.18}$$

is necessary for (LBB1).

The following result is the basis for a-priori error estimates of the Galerkin discretization error.

Theorem 5.13. *Assume* (5.13), (5.14), *and the Babuška-Brezzi conditions* (LBB1) *and* (LBB2). *Let* $(\mathbf{j}, u) \in V \times Q$ *and* $(\mathbf{j}_n, u_n) \in V_n \times Q_n$ *stand for the unique solutions of* (SPP) *and* (DSPP), *respectively. Then*

$$\begin{aligned} \|\mathbf{j} - \mathbf{j}_n\|_V &\leq \gamma_1 \inf_{\mathbf{v}_n \in V_n} \|\mathbf{j} - \mathbf{v}_n\|_V + \frac{\|\mathsf{b}\|}{\alpha_n} \inf_{q_n \in Q_n} \|u - q_n\|_Q \ , \\ \|u - u_n\|_Q &\leq \left(1 + \frac{\|\mathsf{b}\|}{\beta_n}\right) \inf_{q_n \in Q_n} \|u - q_n\|_Q + \frac{\|\mathsf{a}\|}{\beta_n} \|\mathbf{j} - \mathbf{j}_n\|_V \ , \end{aligned}$$

with

$$\gamma_1 := \left(1 + \frac{\|\mathsf{a}\|}{\alpha_n}\right)\left(1 + \frac{\|\mathsf{b}\|}{\beta_n}\right) \ .$$

Moreover, if $\mathcal{N}(\mathsf{B}_n) \subset \mathcal{N}(\mathsf{B})$, *then*

$$\|\mathbf{j} - \mathbf{j}_n\|_V \leq \gamma_1 \inf_{\mathbf{v}_n \in V_n} \|\mathbf{j} - \mathbf{v}_n\|_V \ . \tag{5.19}$$

Proof. Subtracting the first equations of (SPP) and (DSPP), we obtain

$$\mathsf{a}(\mathbf{j}-\mathbf{j}_n,\mathbf{v}_n)+\mathsf{b}(\mathbf{v}_n,u-q_n)=0 \quad \forall \mathbf{v}_n\in V_n\ , \tag{5.20}$$

so that for $q_n\in Q_n$ we find

$$\mathsf{b}(\mathbf{v}_n,q_n-u_n)=-\mathsf{a}(\mathbf{j}-\mathbf{j}_n,\mathbf{v}_n)-\mathsf{b}(\mathbf{v}_n,u-q_n)\ .$$

Using this and (LBB1) we have

$$\beta_n\left\|q_n-u_n\right\|_Q\le\sup_{\mathbf{v}_n\in V\setminus\{0\}}\frac{\mathsf{b}(\mathbf{v}_n,q_n-u_n)}{\left\|\mathbf{v}_n\right\|_V}=\sup_{\mathbf{v}_n\in V\setminus\{0\}}\frac{\mathsf{a}(\mathbf{j}-\mathbf{j}_n,\mathbf{v}_n)+\mathsf{b}(\mathbf{v}_n,u-u_n)}{\left\|\mathbf{v}_n\right\|_V}\ .$$

By the continuity of the bilinear forms, this leads to

$$\beta_n\left\|q_n-u_n\right\|_Q\le\left\|\mathsf{b}\right\|\left\|u-q_n\right\|_Q+\left\|\mathsf{a}\right\|\left\|\mathbf{j}-\mathbf{j}_n\right\|_V\ ,$$

which implies the estimate for $\left\|u-u_n\right\|_Q$ by the triangle-inquality.

Let $\mathbf{w}_n\in V_n$ satisfy $\mathbf{j}_n-\mathbf{w}_n\in\mathcal{N}(\mathsf{B}_n)$. By (LBB2) this implies

$$\begin{aligned}\alpha\left\|\mathbf{j}_n-\mathbf{w}_n\right\|_V&\le\sup_{\mathbf{v}_n\in\mathcal{N}(\mathsf{B}_n)\setminus\{0\}}\frac{\mathsf{a}(\mathbf{j}_n-\mathbf{w}_n,\mathbf{v}_n)}{\left\|\mathbf{v}_n\right\|_V}\\&=\sup_{\mathbf{v}_n\in\mathcal{N}(\mathsf{B}_n)\setminus\{0\}}\frac{\mathsf{a}(\mathbf{j}_n-\mathbf{j},\mathbf{v}_n)+\mathsf{a}(\mathbf{j}-\mathbf{w}_n,\mathbf{v}_n)}{\left\|\mathbf{v}_n\right\|_V}\\&\overset{(5.20)}{=}\sup_{\mathbf{v}_n\in\mathcal{N}(\mathsf{B}_n)\setminus\{0\}}\frac{\mathsf{a}(\mathbf{j}-\mathbf{w}_n,\mathbf{v}_n)-\mathsf{b}(\mathbf{v}_n,u-u_n)}{\left\|\mathbf{v}_n\right\|_V}\end{aligned}$$

If $\mathcal{N}(\mathsf{B}_n)\subset\mathcal{N}(\mathsf{B})$, condition $\mathbf{v}_n\in\mathcal{N}(\mathsf{B}_n)$ involves $\mathbf{v}_n\in\mathcal{N}(\mathsf{B})$ and the continuity of a gives

$$\alpha\left\|\mathbf{j}_n-\mathbf{w}_n\right\|_V\le\sup_{\mathbf{v}_n\in\mathcal{N}(\mathsf{B}_n)\setminus\{0\}}\frac{\mathsf{a}(\mathbf{w}_n-\mathbf{j},\mathbf{v}_n)}{\left\|\mathbf{v}_n\right\|_V}\le\left\|\mathsf{a}\right\|\left\|\mathbf{j}-\mathbf{w}_n\right\|_V\ . \tag{5.21}$$

Next, we aim to estimate $\left\|\mathbf{j}-\mathbf{w}_n\right\|_V$. To that end, for some $\mathbf{v}_n\in V_n$ seek the minimum norm solution $\mathbf{r}_n\in V_n$ of

$$\mathsf{b}(\mathbf{r}_n,q_n)=b(\mathbf{j}-\mathbf{v}_n,q_n)\quad\forall q_n\in Q_n\ .$$

Since, V is a Hilbert space and reflexive, it can be estimated by

$$\left\|\mathbf{r}_n\right\|_V\le\beta_n^{-1}\|\mathsf{b}\|\ \left\|\mathbf{j}-\mathbf{v}_n\right\|_V$$

Observe that $\mathbf{j}-\mathbf{v}_n-\mathbf{r}_n\in\mathcal{N}(\mathsf{B}_n)$ and use the triangle inequality:

$$\left\|\mathbf{j}-\mathbf{v}_n-\mathbf{r}_n\right\|_V\le\left(1+\frac{\|\mathsf{b}\|}{\beta_n}\right)\left\|\mathbf{j}-\mathbf{v}_n\right\|_V\ . \tag{5.22}$$

Then use $\mathbf{w}_n := \mathbf{v}_n + \mathbf{r}_n$ in (5.21) and (5.19) follows from the triangle-inequality.

In the general case $\mathcal{N}(\mathsf{B}_n) \not\subset \mathcal{N}(\mathsf{B})$ we still have $\mathbf{v}_n \in \mathcal{N}(\mathsf{B}_n)$ and for any $q_n \in Q_n$

$$\mathsf{b}(\mathbf{v}_n, u - u_n) = \mathsf{b}(\mathbf{v}_n, u - q_n) \leq \|\mathsf{b}\| \, \|\mathbf{v}_n\|_V \, \|u - q_n\|_Q \ .$$

Comnbining this with the above estimate yields

$$\alpha \, \|\mathbf{j}_n - \mathbf{w}_n\|_V \leq \|\mathsf{a}\| \, \|\mathbf{j} - \mathbf{w}_n\|_V + \|\mathsf{b}\| \, \|u - q_n\|_Q \ ,$$

and, again, (5.22) along with the triangle inequality finishes the proof. □

Anologuous to Thm. 1.30, we conclude that the Galerkin solutions are *quasi-optimal*. Of course, the main task will be to investigate the dependence of α_n and β_n on discretization parameters.

There is a standard way to prove (LBB1). Again, consider the variational saddle point problem (SPP) and its discretized version DSPP.

Definition 5.14. *An operator $\mathsf{F}_n \in L(V, V_n)$ is called a **Fortin projector**, if*

$$\mathsf{b}(\mathbf{v} - \mathsf{F}_n \, \mathbf{v}, q_n) = 0 \quad \forall \mathbf{v} \in V, \ q_n \in Q_n \ . \tag{5.23}$$

Lemma 5.15. *Provided that the inf-sup condition* (5.13) *is satisfied, the existence of a Fortin projector F_n ensures that first Babuška-Brezzi condition* (LBB1) *holds with $\beta_n = \beta \, \|\mathsf{F}_n\|^{-1}_{V \mapsto V}$.*

Proof. For an arbitrary $q_n \in Q_n$ we have

$$\begin{aligned} \beta \, \|q\|_Q &\leq \sup_{\mathbf{v} \in V \setminus \{0\}} \frac{\mathsf{b}(\mathbf{v}, q_n)}{\|\mathbf{v}\|_V} \overset{(5.23)}{=} \sup_{\mathbf{v} \in V \setminus \{0\}} \frac{\mathsf{b}(\mathsf{F}_n \, \mathbf{v}, q_n)}{\|\mathbf{v}\|_V} \leq \sup_{\mathbf{v} \in V \setminus \{0\}} \|\mathsf{F}_n\|_{V \mapsto V} \frac{\mathsf{b}(\mathsf{F}_n \, \mathbf{v}, q_n)}{\|\mathsf{F}_n \, \mathbf{v}\|_V} \\ &\leq \|\mathsf{F}_n\|_{V \mapsto V} \sup_{\mathbf{v}_n \in V_n \setminus \{0\}} \frac{\mathsf{b}(\mathbf{v}_n, q_n)}{\|\mathbf{v}_n\|_V} \ . \end{aligned}$$

□

Exercise 5.5. Consider the discrete variational saddle point problem (SPP) and assume that the first Babuška-Brezzi condition (LBB1) is satisfied. Show by means of an auxiliary saddle point problem that this implies the existence of a Fortin projector whose norm can be bounded in terms of β_n and $\|\mathsf{b}\|$.

Exercise 5.6. Let $\mathsf{a} \in L(V \times V, \mathbb{R})$ and $\mathsf{d} \in L(Q \times Q, \mathbb{R})$ be continuous bilinear forms that are V-elliptic and Q-elliptic, respectively. We examine the variational problem: seek $\mathbf{j} \in V$, $u \in Q$ such that

$$\begin{array}{rcrcll} \mathsf{a}(\mathbf{j}, \mathbf{v}) & + & \mathsf{b}(\mathbf{v}, u) & = & \langle g, \mathbf{v} \rangle_{V^* \times V} & \forall \mathbf{v} \in V \ , \\ \mathsf{b}(\mathbf{j}, q) & - & \mathsf{d}(u, q) & = & \langle f, q \rangle_{Q^* \times Q} & \forall q \in Q \ . \end{array} \tag{5.24}$$

(i) Show that (5.24) is related to an elliptic linear variational problem, see Exercise 5.4.

(ii) Prove the quasi-optimality of any conforming finite element Galerkin solution of (5.24).

5.2.4 A priori error analysis of lowest order finite element scheme

The finite element Galerkin discretization of the variational saddle point problem (5.11) can rely on the finite element spaces introduced in Sect. 3.8.2. On a conforming triangulation $\mathcal{M}$ we approximate functions in

- $H(\mathrm{div};\Omega)$ by face elements according to Def. 3.63, 3.67 (space $\mathcal{W}_F(\mathcal{M})$),
- $L^2(\Omega)$ by piecewise constant functions (space $\mathcal{Q}_0(\mathcal{M})$).

In the sequel, we will retrict ourselves to conforming simplicial meshes in two dimensions. The focus will be on establishing the Babuška-Brezzi conditions for the pair $\mathcal{W}_F(\mathcal{M}) \times \mathcal{Q}_0(\mathcal{M})$ of trial/test spaces with constants that may depend on the shape regularity measure $\rho_{\mathcal{M}}$ and quasi-uniformity $\mu_{\mathcal{M}}$, but are independent of the meshwidth $h_{\mathcal{M}}$.

To begin with, we recall that Thm. 3.78 also means $\operatorname{div}\mathcal{W}_F(\mathcal{M}) \subset \mathcal{Q}_0(\mathcal{M})$. Thus, it is immediate that

$$\mathcal{N}(\mathsf{B}_n) = \{\mathbf{v}_h \in \mathcal{W}_F(\mathcal{M}) : \ \operatorname{div}\mathbf{v}_h = 0\} \subset \mathcal{N}(\mathsf{B}) = H(\operatorname{div}0;\Omega) \ .$$

Hence, owing to the property (UPD) of $\mathbf{A}$ the "ellipticity on the kernel" (LBB2) with $\alpha_n = \underline{\gamma}$, $\underline{\gamma}$ from (UPD), is obvious

$$\mathsf{a}(\mathbf{v}_n,\mathbf{v}_n) \geq \underline{\gamma}\,\|\mathbf{v}_n\|^2_{L^2(\Omega)} = \underline{\gamma}\,\|\mathbf{v}_n\|^2_{H(\mathrm{div};\Omega)} \quad \forall \mathbf{v}_n \in \mathcal{N}(\mathsf{B}_n) \ .$$

It remains to prove (LBB1) in the form

$$\sup_{\mathbf{v}_n\in\mathcal{W}_F(\mathcal{M})\setminus\{0\}} \frac{\int_\Omega \operatorname{div}\mathbf{v}_n \cdot q_n \,\mathrm{d}\boldsymbol{\xi}}{\|\mathbf{v}_n\|_{H(\mathrm{div};\Omega)}} \geq \beta_n \,\|q_n\|_{L^2(\Omega)} \quad \forall q_n \in \mathcal{Q}_0(\mathcal{M}) \ ,$$

with $\beta_n = \beta_n(\Omega,\mathbf{A},\rho_{\mathcal{M}},\mu_{\mathcal{M}}) > 0$.

We will make use of the finite element interpolation operator I_F belonging to $\mathcal{W}_F(\mathcal{M})$. From Thm. 3.78 we learn that it satisfies

$$\operatorname{div}\circ\mathsf{I}_F = \mathsf{Q}_0 \circ \operatorname{div} \ ,$$

where $\mathsf{Q}_0 : L^2(\Omega) \mapsto \mathcal{Q}_0(\mathcal{M})$ is the $L^2(\Omega)$-orhtogonal projection. Alas, I_F does not qualify as a Fortin projector from Def. 5.14, because

the finite element interpolation operator I_F is not bounded on $H(\mathrm{div};\Omega)$.

Thus, we cannot simply invoke Lemma 5.15 in order to prove (5.2.4), but a profound result from functional analysis comes to our rescue [18, Ch. 1]:

Theorem 5.16. *There is a constant $\gamma = \gamma(\Omega) > 0$ such that*

$$\forall q \in L^2(\Omega), \int_\Omega q \,\mathrm{d}\boldsymbol{\xi} = 0 : \quad \exists \mathbf{v} \in (H_0^1(\Omega))^d : \ \operatorname{div} \mathbf{v} = q \quad \textit{and} \quad \|\mathbf{v}\|_{H^1(\Omega)} \le \gamma \, \|q\|_{L^2(\Omega)} \ .$$

This theorem is very useful, because, because I_F, which relies on the evaluation of edge integrals, is certainly continuous on $(H^1(\Omega))^2$, see Thm. 2.49.

Lemma 5.17. $\quad \exists \gamma = \gamma(\rho_{\mathcal{M}}) : \quad \|\mathsf{I}_F \, \mathbf{v}\|_{H(\operatorname{div};\Omega)} \le \gamma \, |\mathbf{v}|_{H^1(\Omega)} \quad \forall \mathbf{v} \in (H^1(\Omega))^2.$

Proof. For the proof we use a **scaling argument**: Pick an arbitrary $K \in \mathcal{M}$. By the trace theorem Thm 2.49 we conclude

$$\exists \gamma > 0 : \quad \|\mathsf{I}_F \, \mathbf{v}\|_{H^1(\partial K)} \le \gamma \, \|\mathbf{v}\|_{H^1(K)} \quad \forall \mathbf{v} \in (H^1(K))^2 \ . \tag{5.25}$$

Now, let us consider a scaled cell $K' = \boldsymbol{\Phi}(K)$, $\boldsymbol{\Phi}\boldsymbol{\xi} = \alpha\boldsymbol{\xi}$, $\alpha > 0$. Simple transformations show, that (5.25) *will also hold on K' with exactly the same constant γ*. In short, the estimate (5.25) is independent of the size of K.

However, the constant γ in (5.25) will depend on the shape, say the angles, of K, but it does so continuously. By Lemma 4.14 the angles can be bounded by the shape regularity parameter independently of h_K, which involves $\gamma = \gamma(\rho_K)$. □

Given any $q_n \in \mathcal{Q}_0(\mathcal{M})$, by Thm. 5.16 we find $\mathbf{v}_0 \in (H_0^1(\Omega))^2$ such that

$$\operatorname{div} \mathbf{v}_0 = q_0 := q_n - |\Omega|^{-1} \int_\Omega q_n \,\mathrm{d}\boldsymbol{\xi} \quad , \quad \|\mathbf{v}_0\|_{H^1(\Omega)} \le \gamma(\Omega) \, \|q_0\|_{L^2(\Omega)} \ .$$

Then set

$$\mathbf{v}(\boldsymbol{\xi}) := \mathbf{v}_0(\boldsymbol{\xi}) + {}^1\!/_2 \boldsymbol{\xi} \cdot |\Omega|^{-1} \int_\Omega q_n \,\mathrm{d}\boldsymbol{\xi} \ , \quad \boldsymbol{\xi} \in \Omega \ ,$$

which means, with $\gamma_1 = \gamma_1(\Omega)$,

$$\operatorname{div} \mathbf{v} = q_n \quad \text{and} \quad \|\mathbf{v}\|_{H^1(\Omega)} \le \gamma_1 \, \|q_n\|_{L^2(\Omega)} \ . \tag{5.26}$$

$$\begin{aligned}
\|q_n\|_{L^2(\Omega)} = \frac{\int_\Omega \operatorname{div} \mathbf{v} \cdot q_n \,\mathrm{d}\boldsymbol{\xi}}{\|q_n\|_{L^2(\Omega)}} &\overset{\text{Thm. 3.78}}{=} \frac{\int_\Omega \operatorname{div}(\mathsf{I}_F \, \mathbf{v}) \cdot q_n \,\mathrm{d}\boldsymbol{\xi}}{\|q_n\|_{L^2(\Omega)}} \overset{(5.26)}{\le} \gamma_1 \, \frac{\int_\Omega \operatorname{div}(\mathsf{I}_F \, \mathbf{v}) \cdot q_n \,\mathrm{d}\boldsymbol{\xi}}{\|\mathbf{v}\|_{H^1(\Omega)}} \\
&\overset{\mathsf{I}_F \text{ cont.}}{\le} \|\mathsf{I}_F\|_{H^1(\Omega) \mapsto \mathcal{W}_F(\mathcal{M})} \, \gamma_1 \, \frac{\int_\Omega \operatorname{div}(\mathsf{I}_F \, \mathbf{v}) \cdot q_n \,\mathrm{d}\boldsymbol{\xi}}{\|\mathsf{I}_F \, \mathbf{v}\|_{H^1(\Omega)}} \\
&\le \|\mathsf{I}_F\|_{H^1(\Omega) \mapsto \mathcal{W}_F(\mathcal{M})} \, \gamma_1 \sup_{\mathbf{v}_n \in \mathcal{W}_F(\mathcal{M}) \setminus \{0\}} \frac{\int_\Omega \operatorname{div} \mathbf{v}_n \cdot q_n \,\mathrm{d}\boldsymbol{\xi}}{\|\mathbf{v}_n\|_{H(\operatorname{div};\Omega)}} \ .
\end{aligned}$$

This shows $\beta_n = (\|\mathsf{I}_F\|_{H^1(\Omega)\mapsto\mathcal{W}_F(\mathcal{M})}\,\gamma_1)^{-1}$. As is the case for $\|\mathsf{I}_F\|_{H^1(\Omega)\mapsto\mathcal{W}_F(\mathcal{M})}$ and γ_1, the constant β_n will only depend on Ω and $\rho_{\mathcal{M}}$.

We have confirmed (LBB1) for the variational problem (5.11) and the pair $\mathcal{W}_F(\mathcal{M}) \times \mathcal{Q}_0(\mathcal{M})$ of finite element spaces. Thus, Thm. 5.13 teaches the quasi-optimality of the Galerkin solutions and we conclude convergence $O(h_{\mathcal{M}})$ for $h_{\mathcal{M}} \to 0$ and uniform shape-regularity.

Exercise 5.7. The second order elliptic boundary value problem (5.9), (5.10) with homogeneous Dirichlet boundary conditions can also be cast into a saddle point problem by applying integration by parts to (5.10). This leads to the primal variational saddle point problem: seek $\mathbf{j} \in (L^2(\Omega))^d$, $u \in H^1_0(\Omega)$ such that

$$\begin{aligned} \int_\Omega \langle \mathbf{A}^{-1}\mathbf{j}, \mathbf{v}\rangle \,\mathrm{d}\boldsymbol{\xi} \quad - \quad \int_\Omega \langle \mathbf{v}, \mathbf{grad}\, u\rangle \,\mathrm{d}\boldsymbol{\xi} &= 0 \qquad \forall \mathbf{v} \in (L^2(\Omega))^d\,, \\ -\int_\Omega \langle \mathbf{j}, \mathbf{grad}\, q\rangle \,\mathrm{d}\boldsymbol{\xi} &= \int_\Omega f q \,\mathrm{d}\boldsymbol{\xi} \quad \forall q \in H^1_0(\Omega)\,. \end{aligned} \tag{5.27}$$

This is called the **mixed hybrid variational formulation** of the second-order elliptic boundary value problem.

(i) Show existence and uniqueness of solutions of (5.27) by applying the theory of abstract saddle point problems.

On a simplicial triangulation $\mathcal{M}$ of Ω the saddle point problem (5.27) is discretized by approximating

- $(L^2(\Omega))^d$ by $\mathcal{M}$-piecewise constant vectorfields $\in (\mathcal{Q}_0(\mathcal{M}))^d$,and
- $H^1_0(\Omega)$ by piecewise linear Lagrangian finite element functions $\in \mathcal{S}_1(\mathcal{M}) \cap H^1_0(\Omega)$.

(ii) Prove existence and uniqueness of solutions of the discretized saddle point problem.

(iii) Determine the aymptotic order of convergence of the h-version of finite elements, if the above finite element discretization is used and sufficient regularity of the exact solution is assumed.

5.3 Finite elements for the Stokes problem

5.3.1 The Stokes problem

Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, be a computational domain. The Stokes problem describes the slow motion of a viscous, incompressible fluid in Ω. We assume the domain Ω to be bounded and connected. The stationary flow in Ω is described by the velocity field $\mathbf{u}$ and the pressure p.

$$\mathbf{u} : \Omega \to \mathbb{R}^d \text{ and } p : \Omega \to \mathbb{R}.$$

They are governed by following elliptic boundary value problems for a system of partial differential equations (Δ applied to components of $\mathbf{u}$):

$$-\nu\Delta\mathbf{u} - \mathbf{grad}\, p = \mathbf{f} \quad \text{in } \Omega, \tag{5.28}$$

$$\operatorname{div}\mathbf{u} = 0 \quad \text{in } \Omega, \tag{5.29}$$

$$\mathbf{u} = 0 \quad \text{on } \partial\Omega. \tag{5.30}$$

Here, $\nu > 0$ is the kinematic viscosity, $\mathbf{f}$ are body forces per unit mass and $\mathbf{u}_0$ is a given fluid velocity at the boundary. The equation (5.29) describes the **incompressibility** of the fluid, whereas (5.28) is related to a balance of forces.

Remark 5.18. The stationary problem (5.28)–(5.30) is a special case of the time-dependent Stokes problem

$$\frac{\partial\mathbf{u}}{\partial t} - \nu\Delta\mathbf{u} - \mathbf{grad}\, p = \mathbf{f} \text{ in } \Omega, \tag{5.31}$$

$$\operatorname{div}\mathbf{u} = 0 \text{ in } \Omega, \tag{5.32}$$

$$\mathbf{u} = \mathbf{u}_0 \text{ on } \partial\Omega. \tag{5.33}$$

(5.28) follows from the assumption that $\frac{\partial\mathbf{u}}{\partial t} = \mathbf{0}$. If we discretise (5.31) with respect to the time t, e. g. with the implicit Euler scheme

$$\frac{\partial\mathbf{u}}{\partial t} \cong \frac{1}{\Delta t}(\mathbf{u}^{(n+1)} - \mathbf{u}^{(n)}),$$

we get a sequence $\left\{\mathbf{u}^{(n)}, p^{(n)}\right\}_{n=0}^{\infty}$ of approximations

$$-\nu\Delta\mathbf{u}^{(n+1)} + \frac{1}{\Delta t}\mathbf{u}^{(n+1)} + \mathbf{grad}\, p^{(n+1)} = \mathbf{f}(x,(n+1)\Delta t) + \frac{1}{\Delta t}\mathbf{u}^{(n)} \text{ in } \Omega, \tag{5.34}$$

$$\operatorname{div}\mathbf{u}^{(n+1)} = 0 \text{ in } \Omega, \tag{5.35}$$

$$\mathbf{u}^{(n+1)} = \mathbf{u}_0(x,(n+1)\Delta t) \text{ on } \partial\Omega. \tag{5.36}$$

5.3.2 Mixed variational formulation

We follow the usual path for the derviation of variational formulations, similar to the manipulations that yielded the dual variational formulation (2.9) in Sect. 2.5. The boundary conditions on $\mathbf{u}$ will be treated as essential boundary conditions. Thus, the first equation (5.28) is multiplied by a compactly supported smooth vectorfield $\mathbf{v}$, integrated over Ω and integration by parts is performed on both $\Delta\mathbf{u}$ and $\mathbf{grad}\, p$. The second equation is multiplied with a test function $v : \Omega \mapsto \mathbb{R}$ and integrated over Ω. We end up with the variational problem: seek $\mathbf{u} \in (H_0^1(\Omega)^d$, $p \in L_0^2(\Omega)$ such that

$$\begin{aligned} \int_\Omega \langle \nu\boldsymbol{\nabla}\mathbf{u}, \boldsymbol{\nabla}\mathbf{v}\rangle \,\mathrm{d}\boldsymbol{\xi} \;+\; \int_\Omega \operatorname{div}\mathbf{v}\cdot p \,\mathrm{d}\boldsymbol{\xi} &= \int_\Omega \langle \mathbf{f}, \mathbf{v}\rangle \,\mathrm{d}\boldsymbol{\xi} \qquad \forall \mathbf{v} \in (H_0^1(\Omega))^d\,, \\ \int_\Omega \operatorname{div}\mathbf{u}\cdot q \,\mathrm{d}\boldsymbol{\xi} &= 0 \qquad \forall q \in L_0^2(\Omega)\,. \end{aligned} \tag{5.37}$$

This is called the **mixed variational formulation** of the Dirichlet problem for the Stokes equations. Here $\boldsymbol{\nabla}$ refers to the Jacobi matrix, that is, the gradient operator applied to the components of the vectorfield.

As in the case of (5.11), the choice of Sobolev spaces is suggested by the bilinear forms occurring in (5.37). Moreover, in order to eliminate the freedom of adding any constant to p we resorted to the space

$$L_0^2(\Omega) := \{v \in L^2(\Omega) : \int_\Omega v \,\mathrm{d}\boldsymbol{\xi} = 0\} \ .$$

Obviously, (5.37) is a variational saddle point problem and fits the structure (SPP) with

$$\begin{aligned} V = (H_0^1(\Omega))^d \quad , \quad Q = L_0^2(\Omega) \ , \\ \mathsf{a}(\mathbf{u},\mathbf{v}) = \int_\Omega \langle \nu \boldsymbol{\nabla}\mathbf{u}, \boldsymbol{\nabla}\mathbf{v}\rangle \ \mathrm{d}\boldsymbol{\xi} \quad , \quad \mathsf{b}(\mathbf{v},p) = \int_\Omega \operatorname{div}\mathbf{v}\cdot p \,\mathrm{d}\boldsymbol{\xi} \ , \\ \langle g, \mathbf{v}\rangle_{V^*\times V} = \int_\Omega \langle \mathbf{f}, \mathbf{v}\rangle \ \mathrm{d}\boldsymbol{\xi} \qquad , \quad f = 0 \ . \end{aligned}$$

The bilinear forms are continuous (due to the choice of Sobolev spaces), and a is V-elliptic by the Poincaré-Friedrichs inequality Lemma 2.61. It remains to show (5.13) in order to establish existence and uniqueness of solutions of (5.37).

Fix $q \in L_0^2(\Omega)$ and appeal to Thm. 5.16 to conclude the existence of $\mathbf{v} \in (H_0^1(\Omega))^d$ with $\operatorname{div}\mathbf{v} = q$ and $\|\mathbf{v}\|_{H^1(\Omega)} \le \gamma \|q\|_{L^2(\Omega)}$, $\gamma = \gamma(\Omega) > 0$. Hence,

$$\|q\|_{L^2(\Omega)} = \frac{\int \operatorname{div}\mathbf{v}\cdot q \,\mathrm{d}\boldsymbol{\xi}}{\|q\|_{L^2(\Omega)}} \le \gamma \frac{\mathsf{b}(\mathbf{v},q)}{\|\mathbf{v}\|_{H^1(\Omega)}} \le \gamma \sup_{\mathbf{v}\in V} \frac{\mathsf{b}(\mathbf{v},q)}{\|\mathbf{v}\|_V} \ . \tag{5.38}$$

Now, the following result is immediate from Thm. (5.11).

Theorem 5.19. *The mixed variational formulation* (5.37) *of the Dirichlet problem for the Stokes equations has a unique solution* $(\mathbf{u},p) \in (H_0^1(\Omega))^d \times L_0^2(\Omega)$.

Remark 5.20. The incompressibility condition (5.29) can also be built into the function space. Therefore we define

$$J_0 := \left\{\mathbf{u} \in H_0^1(\Omega)^d : \operatorname{div}\mathbf{u} = 0 \text{ in } L^2(\Omega)\right\} \ .$$

J_0 is a closed linear subspace of $(H_0^1(\Omega))^d$ and we get the variational problem: seek $\mathbf{u} \in J_0$ such that

$$\int_\Omega \langle \nu\boldsymbol{\nabla}\mathbf{u}, \boldsymbol{\nabla}\mathbf{v}\rangle \ \mathrm{d}\boldsymbol{\xi} = \int_\Omega \langle \mathbf{f}, \mathbf{v}\rangle \ \mathrm{d}\boldsymbol{\xi} \quad \forall \mathbf{v} \in J_0 \ . \tag{5.39}$$

Existence and uniqueness of solutions of (5.39) are clear from Thm. 1.17.

The variational problems (5.37) and (5.39) yield the same solution for the velocity $\mathbf{u}$. However, (5.39) poses difficulties in terms of a conforming finite element discretization, because no finite element subspace of J_0 is known.

Remark 5.21. The kinematic pressure p is often a subject of interest in practical computations; therefore, one tries to deduce an equation for p from (5.28) by formally applying the divergence (div) to (5.28):

$$\begin{aligned} -\nu \operatorname{div} \Delta \mathbf{u} + \operatorname{div} \mathbf{grad}\, p &= \operatorname{div} \mathbf{f} \text{ in } \Omega \\ \Rightarrow -\nu \Delta(\operatorname{div} \mathbf{u}) + \Delta p &= \operatorname{div} \mathbf{f} \text{ in } \Omega \\ \overset{(5.29)}{\Rightarrow} \Delta p &= \operatorname{div} \mathbf{f} \text{ in } \Omega. \end{aligned} \tag{5.40}$$

This is the so-called **pressure Poisson equation**. Its application to numerical simulation has two disadvantages:

1. The deduction was formal and is only valid under strong assumptions on the smoothness of $\mathbf{u}$ and p: $\Delta \mathbf{u} \in L^2(\Omega)^d$ and $\mathbf{grad}\, p \in L^2(\Omega)^d$, and $\mathbf{u} \in H^2(\Omega)^d$ and $p \in H^1(\Omega)$, respectively.

2. Neither from the physical nor from the mathematical point of view it is possible to find any sensible boundary conditions for (5.40) from (5.28)–(5.30). Often, one uses $\mathbf{n} \cdot (5.28)|_{\partial\Omega}$ (e. g. the transport of the momentum over the boundary):

$$\langle \mathbf{grad}\, p, \mathbf{n} \rangle \overset{(5.28)}{=} \langle \mathbf{f} + \nu \Delta \mathbf{u}, \mathbf{n} \rangle , \tag{5.41}$$

i. e. we get (again formally) a *Neumann problem* for p. As explained in Sect. 2.8, the variational formulation looks as follows: find $p \in \tilde{H}^1(\Omega)$ such that

$$\langle \mathbf{grad}\, p, \mathbf{grad}\, q \rangle = l(q) \quad \forall q \in \tilde{H}^1(\Omega), \tag{5.42}$$

where

$$l(q) := -\int_\Omega q \operatorname{div} \mathbf{f} \, \mathrm{d}\boldsymbol{\xi} + \int_{\partial\Omega} q \, \langle \mathbf{f} + \nu \Delta \mathbf{u}, \mathbf{n} \rangle \, \mathrm{d}S.$$

Because of

$$\begin{aligned} l(1) &= -\int_\Omega \operatorname{div} \mathbf{f} \, \mathrm{d}\boldsymbol{\xi} + \int_{\partial\Omega} \langle \mathbf{f} + \nu \Delta \mathbf{u}, \mathbf{n} \rangle \, \mathrm{d}S \\ &= -\int_{\partial\Omega} \langle \mathbf{f}, \mathbf{n} \rangle \, \mathrm{d}S + \int_{\partial\Omega} \langle \mathbf{f} + \nu \Delta \mathbf{u}, \mathbf{n} \rangle \, \mathrm{d}S \\ &= \nu \int_{\partial\Omega} \langle \Delta \mathbf{u}, \mathbf{n} \rangle \, \mathrm{d}S = \nu \int_\Omega \operatorname{div}(\Delta \mathbf{u}) \, \mathrm{d}\boldsymbol{\xi} \\ &= \nu \int_\Omega \Delta(\operatorname{div} \mathbf{u}) \, \mathrm{d}\boldsymbol{\xi} \overset{(5.29)}{=} 0. \end{aligned}$$

the compatibility condition (NCC) holds true. However, the smoothness of $\mathbf{u}$ must again be assumed—only in this case, we have $\langle \Delta \mathbf{u}, \mathbf{n} \rangle \in L^2(\partial\Omega)$ and $|l(q)| \leq C \, \|q\|_1$ (referring to the trace Theorem 2.49, $\mathbf{u} \in (H^3(\Omega))^d$ must be assumed in order to get $\langle \Delta \mathbf{u}, \mathbf{n} \rangle \in L^2(\partial\Omega)$).

(5.42) is only true if $p \in H^1(\Omega)$; but generally, this is not the case. (5.42) holds also true in convex domains; however (5.42) is not anymore valid in domains with reentrant corners.

5.3.3 Unstable finite element pairs

Definition 5.22. *A pair of spaces $V_n \times Q_n$ to be used in the Galerkin discretization of* (SPP) *is said to be* ***unstable****, if the first LBB-condition* (LBB1) *does not hold.*

A family of spaces $\{V_n \times Q_n\}_{n=1}^{\infty}$ with $\dim V_n, \dim Q_n \to \infty$ for $n \to \infty$ is called ***asymptotically unstable****, if* (LBB1) *holds for each n, but $\beta_n \to 0$ for $n \to \infty$ is inevitable.*

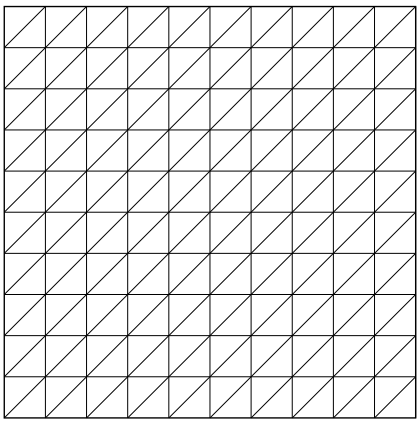
Remark 5.23. Asymptotic instability means that the constants in the estimates of Thm. 5.13 blow up for $n \to \infty$. However, this can be offset by a decrease of the best approximation error, so that we may get overall convergence in the $V \times Q$-norm, albeit worse than the best approximation error of the trial spaces.

Remark 5.24. If one is not interested in the solution for u of (SPP), $f = 0$, and $\mathcal{N}(\mathsf{B}_n) \subset \mathcal{N}(\mathsf{B})$ the estimate (5.19) of Thm. 5.13 shows that $\mathbf{j}_n$ can be a good approximation for $\mathbf{j}$, nevertheless.

Some apparently natural choices of $V_n \subset (H^1(\Omega))^d$ and $Q_n \subset L^2_0(\Omega)$ for the mixed variational formulation (5.37) of the Dirichlet problem for the Stokes equations will turn out to be unstable.

Example 5.25. Consider a triangular mesh of $]0,1[^2$ of the structure sketched beside. We opt for

$$\begin{aligned} V_n &:= (\mathcal{S}_1(\mathcal{M}) \cap (H^1_0(\Omega))^2 \,, \\ Q_n &:= \mathcal{Q}_0(\mathcal{M}) \,. \end{aligned} \tag{5.43}$$



If we have $M \in \mathbb{N}$ mesh cells in one coordinate direction, we find

$$\dim V_n = 2(M-1)^2 \quad , \quad \dim Q_n = 2M^2 \,.$$

Clearly, $\dim Q_n > \dim V_n$, which, according to Remark 5.12, rules out stability of the pair $V_n \times Q_n$.

Example 5.26. Let $\mathcal{M}$ be a tensor product mesh of $]0,1[^2$ with M mesh cells in each coordinate direction. On it the choice (5.43) of finite element spaces for the Stokes problem (5.37) satisfies the dimension condition

$$2(M-1)^2 = \dim V_n > \dim Q_n = M^2 \,.$$

Yet, Nevertheless, also in this case

$$N(B^*) := \{q_n \in Q_n : \ \mathsf{b}(\mathbf{v}_n, q_n) = 0 \ \forall \mathbf{v}_n \in V_n\} \neq \{0\} \,.$$

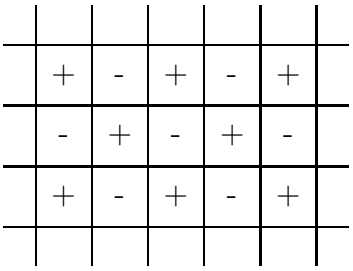


Figure 5.2: Checkerboard instability on quadrilateral meshes.

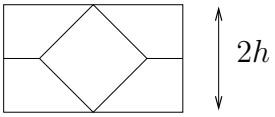


Figure 5.3: Part of a quadrilateral mesh on which $(\mathcal{S}_1(\mathcal{M}))^2 \times \mathcal{Q}_0(\mathcal{M})$ is uniformly stable for $h \to 0$.

More precisely,

$$N(B^*) = \operatorname{span}\{\widetilde{q}_n\} ,$$

where the piecewise constant function $\widetilde{q}_n \in Q_n$ assumes the values ± 1 in the checkerboard fashion depicted in Fig. 5.2. Therefore, this function in the kernel of B^* is called a **checkerboard mode**. In three dimensions the situation is similar, but there are several checkerboard modes.

There are meshes on which the pair of finite element spaces from (5.43) is uniformly stable, see Fig 5.3, but they are mere oddities.

Remark 5.27. If one was only interested in the solution for the velocity $\mathbf{u}$ one might recall Remark 5.24 and wonder, whether one has to worry about the instability at all (apart from the resulting *consistent* algebraic system being singular). However, in the case of (5.37) we never have $\mathcal{N}(\mathsf{B}_n) \subset \mathcal{N}(\mathsf{B})$ so that instability in the pressur will invariably affect the solution for the velocity. To see this examine the constants in the estimates of Thm. 5.13.

5.3.4 A stable non-conforming finite element pair

In order to be able to use finite element spaces with piecewise linear velocity and piecewise constant pressure, we have to switch to a non-conforming approximation of $(H_0^1(\Omega))^d$. We discuss this for simplicial triangulations $\mathcal{M}$ in two dimensions.

In particular, we chose

- the $(H_0^1(\Omega))^2$-non-conforming finite element space

$$V_n = \{\mathbf{v}_n \in (\mathcal{CR}(\mathcal{M}))^2 : \ \mathbf{v}_n = 0 \text{ in midpoints of edges } \subset \Gamma\} ,$$

  see Sect. 5.1.2, and

- $Q_n = \mathcal{Q}_0(\mathcal{M})$.

As explained in Sect. 5.1.1, the use of non-conforming finite element spaces entails a modification of the bilinear forms to render them well definite for arguments in V_n. Similar to the approach in Sect. 5.1.2 we employ the split bilinear forms

$$\widetilde{\mathsf{a}}(\mathbf{u},\mathbf{v}) := \sum_{K\in\mathcal{M}} \int_K \langle \mathbf{grad}\,\mathbf{u}, \mathbf{grad}\,\mathbf{v}\rangle \,\mathrm{d}\boldsymbol{\xi} \quad , \quad \widetilde{\mathsf{b}}(\mathbf{v},p) := \sum_{K\in\mathcal{M}} \int_K \operatorname{div}\mathbf{v}\cdot p\,\mathrm{d}\boldsymbol{\xi}\ .$$

In addition, we have to rely on the split H^1-seminorm:

$$|u|_{1,n} := \left(\sum\nolimits_{K\in\mathcal{M}} |u|^2_{H^1(K)}\right)^{1/2} .$$

It will be this split norm the modified first Babuška-Brezzi condition (LBB1) has to be formulated with: we have to show

$$\exists \beta_n > 0: \qquad \sup_{\mathbf{v}_n\in V_n\setminus\{0\}} \frac{|\widetilde{\mathsf{b}}(\mathbf{v},q)|}{|\mathbf{v}_n|_{1,n}} \geq \beta_n \|q\|_{L^2(\Omega)} \quad \forall q_n \in Q_n\ . \tag{5.44}$$

To prove this we want to rely on a Fortin projector, see Def. 5.14: define $\mathsf{I}_V : (H^1_0(\Omega)) \mapsto V_n$ through the application of the finite element interpolation operator $\mathsf{I}_{\mathcal{CR}}$ associated with $\mathcal{CR}(\mathcal{M})$ to the Cartesian components of the argument.

Lemma 5.28. *The interpolation operator I_V satisfies*

$$\widetilde{\mathsf{b}}(\mathsf{I}_V\,\mathbf{v}, q_n) = \widetilde{\mathsf{b}}(\mathbf{v}, q_n) \quad \forall \mathbf{v}\in (H^1_0(\Omega))^2,\ q_n \in \mathcal{Q}_0(\mathcal{M})\ .$$

Proof. Pick an arbitrary triangle K and denote by I_V^K the local finite element interpolation operator for the space V_n. Recall that for any edge $F\in\mathcal{E}(\mathcal{M})$ the finite element interpolation operator $\mathsf{I}_{\mathcal{CR}}$ satisfies

$$\int_F \mathsf{I}_{\mathcal{CR}}\, v\,\mathrm{d}S = \int_F v\,\mathrm{d}S \quad \forall v\in H^1(\Omega)\ .$$

Thus,

$$\int_K \operatorname{div}\mathsf{I}_V\,\mathbf{v}\,\mathrm{d}\boldsymbol{\xi} = \int_{\partial K} \langle \mathsf{I}_V\,\mathbf{v}, \mathbf{n}\rangle\,\mathrm{d}S = \int_{\partial K} \langle \mathbf{v},\mathbf{n}\rangle\,\mathrm{d}S = \int_K \operatorname{div}\mathbf{v}\,\mathrm{d}\boldsymbol{\xi}\ ,$$

which means gives the assertion of the lemma. □

Moreover, as stated in Lemma 5.9 $\mathsf{I}_V : (H^1_0(\Omega)) \mapsto V_n$ is continuous:

$$|\mathsf{I}_V\,\mathbf{v}|_{1,n} \leq \gamma\,|\mathbf{v}|_{H^1(\Omega)} \quad \forall \mathbf{v}\in(H^1_0(\Omega))^2\ , \tag{5.45}$$

with $\gamma = \gamma(\rho_{\mathcal{M}})$. Summing up, I_V is a Fortin projector for the saddle point problem (5.37) discretized on $V_n \times Q_n$. >From Lemma 5.15 and (5.38) we conclude that $\beta_n =$

$\beta_n(\Omega, \rho_{\mathcal{M}}) > 0$. In other words, the β_n is bounded away from zero independently of the meshwidth.

Hence, Thm. 5.13 gives us quasi-optimal accuracy of the Galerkin solutions and, asymptotically, we have

$$|\mathbf{u} - \mathbf{u}_n|_{1,n} + \|p - p_n\|_{L^2(\Omega)} = O(h_{\mathcal{M}}) \quad \text{for } h_{\mathcal{M}} \to 0 \ ,$$

once uniform shape-regularity is guaranteed.

Remark 5.29. Of course, the Babuška-Brezzi condition (5.38) can readily be satisfied by using piecewise constant pressures in conjunction with a very "large" finite element space for $\mathbf{u}$, for instance, $(\mathcal{S}_3(\mathcal{M}))^2$. However, Thm 5.13 sends the message that in the case $\mathcal{N}(\mathsf{B}_n) \not\subset \mathcal{N}(\mathsf{B})$ a very accurate approximation of the velocity does not gain anything, because $\|\mathbf{u} - \mathbf{u}_n\|_{H^1(\Omega)}$ will also depend on the best approximation error for the pressure space.

Hence, for the sake of efficiency, it is highly desirable to have balanced velocity and pressure finite element spaces that have similar power to approximate $\mathbf{u}$ and p. In the case of the h-version of finite elements we should strive for the same asymptotic decay of the best approximation error in terms of $h_{\mathcal{M}}$.

In the case of the pair $V_n \times Q_n$ discussed in this section, this criterion is met, because, assuming smooth $\mathbf{u}$ and p,

$$\inf_{\mathbf{v}_n \in V_n} |\mathbf{u} - \mathbf{v}_n|_{1,n} = O(h_{\mathcal{M}}) \quad , \quad \int\limits_{q_n \in Q_n} \|p - q_n\|_{L^2(\Omega)} = O(h_{\mathcal{M}}) \ .$$

5.3.5 A stabilized conforming finite element pair

In Sect 5.3.3 we saw that piecewise linear/bilinear velocities are "too small " when combined with piecewise constant pressure. A remedy for this instability is the *augmentation of the finite element space for the velocity.* The augmentation can be so strong that even a piecewise linear continuous approximation of the pressure unknown is feasible. The discussion will be restricted to simplicial triangulations in two dimensions.

Definition 5.30. *For a triangle $K \subset \mathbb{R}^2$ we define the* ***bubble augmented linear Lagrangian finite element*** *("MINI-element") (K, Π_K, Σ_K) by*

- $\Pi_K = \mathcal{P}_1(K) + \mathrm{span}\{\lambda_1\lambda_2\lambda_3\}$, *where λ_i, $i = 1, 2, 3$, is the i-th barycentric coordinate function for K, and*
- $\Sigma_K := \{v \mapsto v(\boldsymbol{\nu}^i),\ v \mapsto v(\boldsymbol{\gamma}^K)\}$, *where $\boldsymbol{\nu}^1, \boldsymbol{\nu}^2, \boldsymbol{\nu}^3$ are the vertices of K, and $\boldsymbol{\gamma}^K$ is its center of gravity.*

Remark 5.31. The function $\lambda_1\lambda_2\lambda_3$ is called a **bubble function** on K, because it is positive everywhere in K and vanishes on ∂K.

Notation: The global finite element space arising from the finite element of Def. 5.30 on a simplicial triangulation $\mathcal{M}$ will be denoted by $\mathcal{BS}_1(\mathcal{M})$.

As $\mathcal{S}_1(\mathcal{M}) \subset \mathcal{BS}_1(\mathcal{M})$, the approximation properties of $\mathcal{BS}_1(\mathcal{M})$ are at least as good as that of the linear Lagrangian finite elements: with $\gamma = \gamma(\rho_{\mathcal{M}}) > 0$

$$\inf_{\mathbf{v}_n \in \mathcal{BS}_1(\mathcal{M})} \|\mathbf{u} - \mathbf{v}_n\|_{H^1(\Omega)} \leq \gamma h_{\mathcal{M}} \, |\mathbf{u}|_{H^2(\Omega)} \quad \forall \mathbf{u} \in (H^2(\Omega))^2 \, . \tag{5.46}$$

The Galerkin discretization of (5.37) will employ the conforming finite element spaces

- $V_n := (\mathcal{BS}_1(\mathcal{M}) \cap H^1_0(\Omega))^2 \subset (H^1_0(\Omega))^2$, and
- $Q_n := \mathcal{S}_1(\mathcal{M})$.

Again, the core task is to verify that the Babuška-Brezzi condition (LBB1) holds with $\beta_n = \beta_n(\Omega, \rho_{\mathcal{M}}) > 0$. This is done via the construction of a Fortin projector, see Def. 5.14, and Lemma 5.15.

A suitable Fortin projector is built in several stages: the first involves a special projector, a so-called **quasi-interpolation operator** $\mathsf{Q}_n : H^1(\Omega) \mapsto \mathcal{S}_1(\mathcal{M})$. Let $\mathcal{M}$ be a simplicial mesh $\mathcal{M}$ of $\Omega \subset \mathbb{R}^2$. To each node $\mathbf{p} \in \mathcal{N}(\mathcal{M})$ we associated an adjacent edge $F_{\mathbf{p}} \in \mathcal{E}(\mathcal{M})$. If $\mathbf{p} \in \Gamma$, we demand that $F_{\mathbf{p}} \subset \Gamma$.

On $F_{\mathbf{p}}$ (with endpoints $\mathbf{p}, \mathbf{q}$) we introduce a linear function

$$\kappa_{\mathbf{p}} \in \mathcal{P}_1(F_{\mathbf{p}}) : \quad \kappa_{\mathbf{p}}(\mathbf{p}) = 4|F_{\mathbf{p}}|^{-1} \quad , \quad \kappa_{\mathbf{p}}(\mathbf{q}) = -2|F_{\mathbf{p}}|^{-1} \, .$$

By elementary computations

$$\int_{F_{\mathbf{p}}} \lambda(\boldsymbol{\xi}) \, \kappa_{\mathbf{p}}(\boldsymbol{\xi}) \, \mathrm{d}S(\boldsymbol{\xi}) = \lambda(\mathbf{p}) \quad \forall \lambda \in \mathcal{P}_1(F_{\mathbf{p}}) \, . \tag{5.47}$$

Then we define $\mathsf{Q}_n : H^1(\Omega) \mapsto \mathcal{S}_1(\mathcal{M})$ by

$$\mathsf{Q}_n(v) := \sum_{\mathbf{p} \in \mathcal{N}(\mathcal{M})} \int_{F_{\mathbf{p}}} v(\boldsymbol{\xi}) \, \kappa_{\mathbf{p}}(\boldsymbol{\xi}) \, \mathrm{d}S(\boldsymbol{\xi}) \cdot b_{\mathbf{p}} \, , \tag{5.48}$$

where $b_{\mathbf{p}}$ is the global shape function of $\mathcal{S}_1(\mathcal{M})$ associated with node $\mathbf{p}$.

Lemma 5.32. *The operator Q_n defined by (5.48) is a continuous projector $H^1_0(\Omega) \mapsto \mathcal{S}_1(\mathcal{M}) \cap H^1_0(\Omega)$, and satisfies*

$$|\mathsf{Q}_n \, v|_{H^1(\Omega)} \leq \gamma \, |v|_{H^1(\Omega)} \quad \forall v \in H^1(\Omega) \, , \tag{5.49}$$

$$\|v - \mathsf{Q}_n \, v\|_{H^1(\Omega)} \leq h_{\mathcal{M}} \, |v|_{H^1(\Omega)} \quad \forall v \in H^1(\Omega) \, . \tag{5.50}$$

Proof. The property $\mathsf{Q}_n v_n = v_n$ for $v_n \in \mathcal{S}_1(\mathcal{M})$ is clear from (5.47). If $v \in H^1_0(\Omega)$ it vanishes on all edges in Γ, so that no global shape function associated with a node in Γ will contribute to $\mathsf{Q}_n v$.

In order to prove (5.49) and (5.50) we use a scaling technique. Pick a triangle $K \in \mathcal{M}$ and denote

$$U := \bigcup \{\overline{K}' : \ K' \in \mathcal{M}, \overline{K} \cap \overline{K}' \neq \emptyset\} \ .$$

Scale U (*ie.* use an affine mapping $\boldsymbol{\Phi} : \boldsymbol{\xi} \mapsto \alpha\boldsymbol{\xi}$, $\alpha > 0$) such that, after scaling $h_{\widehat{K}} = 1$ (K is mapped to $\widehat{K}$ by scaling). If $\widehat{F}$ is an edge in $\widehat{U}$, then Thm. 2.49 teaches us that

$$\exists \gamma = \gamma(\widehat{F}) : \quad \int_{\widehat{F}} \widehat{v}(\boldsymbol{\xi}) \widehat{\kappa}_{\mathbf{p}}(\boldsymbol{\xi}) \, \mathrm{d}S(\boldsymbol{\xi}) \leq \gamma \, \|\widehat{v}\|_{H^1(\widehat{U})} \quad \forall \widehat{v} \in H^1(\widehat{U}) \ .$$

By the definition of Q_n in (5.48) we conclude that with $\gamma_1 = \gamma_1(\widehat{U}) > 0$

$$\left| \widehat{Q}_n \widehat{v} \right|_{H^1(\widehat{K})} \leq \gamma_1 \, |\widehat{v}|_{H^1(\widehat{U})} \quad \forall \widehat{v} \in H^1(\widehat{U}) \ . \tag{5.51}$$

We can use the semi-norms, because $\widehat{Q}_n$ preserves constants on $\widehat{U}$.

This constant preserving property together with (5.51) enables us to apply the Bramle-Hilbert lemma Lemma 4.7, which gives

$$\exists \gamma_2 = \gamma_2(\widehat{U}) > 0 : \quad \left\| \widehat{Q}_n \widehat{v} \right\|_{L^2(\widehat{K})} \leq \gamma_2 \, |\widehat{v}|_{H^1(\widehat{U})} \quad \forall \widehat{v} \in H^1(\widehat{U}) \ . \tag{5.52}$$

The constants in (5.52) and (5.51) depend on the shape of all the triangles in $\widehat{U}$.

Next comes the crucial scaling argument: switching back to K/U and using simple transformation formulas for the norms, we obtain

$$\begin{aligned} |Q_n v|_{H^1(K)} &\leq \gamma_1 \, |v|_{H^1(U)} \qquad \forall v \in H^1(U) \ , \\ \|Q_n v\|_{L^2(K)} &\leq h_K \gamma_2 \, |v|_{H^1(U)} \quad \forall v \in H^1(U) \ . \end{aligned}$$

The constants γ_1 and γ_2 are continuous functions of all the angles of the triangles in U. These angles lie in a compact hypercube determined by $\rho_{\mathcal{M}}$, see Lemma 4.14. On it γ_1 and γ_2 are bounded so that they can be chosen independently of K.

The proof is finished by the observation that the maximum number of other cell neighborhoods whose interiors intersect U can be bounded in terms of $\rho_{\mathcal{M}}$. □

>From Q_n we construct $\underline{\mathsf{Q}}_n : (H^1_0(\Omega))^2 \mapsto V_n$ by componentwise application.

The second ingredient to the desired Fortin projector will be a projection onto the space spanned by local bubble functions. For $K \in \mathcal{M}$ set

$$\beta_K \in \operatorname{span}\{\lambda_1^K \lambda_2^K \lambda_3^K\} \quad , \quad \int_K \beta_K \, \mathrm{d}\boldsymbol{\xi} = 1 \ .$$

Define

$$\mathcal{B}_n := \{v \in H^1(\Omega) : \ v_{|K} \in \operatorname{span}\{\lambda_1^K \lambda_2^K \lambda_3^K\} \ \forall K \in \mathcal{M}\} \ ,$$

and $\mathsf{B}_n : L^2(\Omega) \mapsto \mathcal{B}_n$ by

$$\mathsf{B}_n \, v_{|K} := \int_K v \, \mathrm{d}\boldsymbol{\xi} \cdot \beta_K \ , \quad K \in \mathcal{M} \ .$$

It is immediate that

$$\int_\Omega (v - \mathsf{B}_n \, v) \, \mathrm{d}\boldsymbol{\xi} = 0 \quad \forall v \in L^2(\Omega) \ . \tag{5.53}$$

Lemma 5.33. *With a constant $\gamma = \gamma(\rho_{\mathcal{M}}, \mu_{\mathcal{M}})$*

$$|\mathsf{B}_n \, v|_{H^1(\Omega)} \leq \gamma h_{\mathcal{M}}^{-1} \, \|v\|_{L^2(\Omega)} \quad \forall v \in H^1(\Omega) \ .$$

Proof. By an elementary scaling argument we find

$$\exists \gamma = \gamma(\rho_{\mathcal{M}}) : \quad |\beta_K|_{H^1(K)} \leq \gamma h_K^{-2} \ . \tag{5.54}$$

Using this, the estimate of the lemma can be accomplished by the Cauchy-Schwarz inequality:

$$\begin{aligned} |\mathsf{B}_n \, v|^2_{H^1(\Omega)} &= \sum_{K \in \mathcal{M}} |\mathsf{B}_n \, v|^2_{H^1(K)} = \sum_{K \in \mathcal{M}} \left(\int_K v \, \mathrm{d}\boldsymbol{\xi} \right)^2 |\beta_K|^2_{H^1(K)} \\ &\leq \gamma h_{\mathcal{M}}^{-4} \sum_{K \in \mathcal{M}} |K| \int_K v^2 \, \mathrm{d}\boldsymbol{\xi} \leq \gamma h_{\mathcal{M}}^{-2} \, \|v\|^2_{L^2(\Omega)} \ . \end{aligned}$$

□

Remark 5.34. Hidden in the estimate of Lemma 5.33 is an inverse estimate, *cf.* Lemma 4.21. Therefore, the quasi-uniformity measure of $\mathcal{M}$ will come into play.

Componentwise application of B_n yields $\underline{\mathsf{B}}_n : (H_0^1(\Omega))^2 \mapsto (\mathcal{B}_n)^2 \subset V_n$, which is obviously continuous.

Lemma 5.35. *The operator $\underline{\mathsf{B}}_n$ satisfies*

$$\mathsf{b}(\mathbf{v} - \underline{\mathsf{B}}_n \mathbf{v}, q_n) = 0 \quad \forall q_n \in Q_n \ .$$

Proof. First, we use integration by parts according to (FGF), and observe that $\mathbf{v}, \underline{\mathsf{B}}_n \mathbf{v}$ vanish on Γ, and that $\mathbf{grad}\, q_n$ is $\mathcal{M}$-piecewise constant

$$\mathsf{b}(\mathbf{v} - \underline{\mathsf{B}}_n v, q_n) = \int_\Omega \operatorname{div}(\mathbf{v} - \underline{\mathsf{B}}_n \mathbf{v}) \cdot q_n \, \mathrm{d}\boldsymbol{\xi} = - \int_\Omega \langle \mathbf{v} - \underline{\mathsf{B}}_n \mathbf{v}, \mathbf{grad}\, q_n \rangle \, \mathrm{d}\boldsymbol{\xi} = 0 \ ,$$

where the last equality is clear from (5.53). □

Remark 5.36. The previous lemma tells us that the operator $\underline{\mathsf{B}}_n : H_0^1(\Omega) \mapsto H_0^1(\Omega)$ is not uniformly continuous in the meshwidth $h_{\mathcal{M}}$. Therefore, it cannot be used as Fortin-projector to show that the constant β_n in (LBB1) is uniformly bounded away from zero, even when $h_{\mathcal{M}} \to 0$.

However, the operator $\underline{\mathsf{B}}_n$ is a key building block for the the desired $h_{\mathcal{M}}$-uniformly continuous Fortin projector

$$\mathsf{F}_n := \mathsf{Q}_n + \mathsf{B}_n(Id - \mathsf{Q}_n) \ . \tag{5.55}$$

This operator is readily seen to be a Fortin projector according to Def. 5.14: by Lemma 5.35 we conclude for any $q_n \in Q_n$

$$\mathsf{b}(\mathsf{F}_n\,\mathbf{v}, q_n) = \mathsf{b}(\mathsf{Q}_n\,\mathbf{v}) + \mathsf{b}(\mathsf{B}_n(Id - \mathsf{Q}_n)\mathbf{v}) \overset{\text{Lemma 5.35}}{=} \mathsf{b}(\mathbf{v} - \mathsf{Q}_n\,\mathbf{v}, q_n) + b(\mathsf{Q}_n\,\mathbf{v}) = \mathsf{b}(\mathbf{v}, q_n) \ ,$$

and continuity follows from Lemma 5.32 and Lemma 5.33:

$$\begin{aligned} |\mathsf{F}_n\,\mathbf{v}|_{H^1(\Omega)} &\leq |\mathsf{Q}_n\,\mathbf{v}|_{H^1(\Omega)} + |\mathsf{B}_n(Id - \mathsf{Q}_n)\mathbf{v}|_{H^1(\Omega)} \\ &\overset{(5.49)}{\leq} |\mathbf{v}|_{H^1(\Omega)} + \|\mathsf{B}_n\|_{L^2(\Omega)\mapsto H^1(\Omega)}\, \|(Id - \mathsf{Q}_n)\mathbf{v}\|_{L^2(\Omega)} \\ &\overset{(5.50)}{\leq} |\mathbf{v}|_{H^1(\Omega)} + \gamma h_{\mathcal{M}}^{-1}\, h_{\mathcal{M}}\, |\mathbf{v}|_{H^1(\Omega)} \leq |\mathbf{v}|_{H^1(\Omega)} \ , \end{aligned}$$

with $\gamma = \gamma(\rho_{\mathcal{M}}, \mu_{\mathcal{M}})$. Appealing to Lemma 5.15 the first Babuška-Brezzi condition is straightforward. The uniform stability of the pair $V_n \times Q_n$ of finite element spaces for the mixed variational formulation of the Stokes problem is established.

6 Adaptive Finite Elements

In this chapter we only consider the primal variational formulation of a second order elliptic boundary value problem

$$-\Delta u = f \quad \text{in } \Omega \quad , \quad u = 0 \quad \text{or} \quad \langle \mathbf{grad}\, u, \mathbf{n} \rangle = 0 \quad \text{on } \Gamma \, . \tag{6.1}$$

in a bounded polygon $\Omega \subset \mathbb{R}^2$ with Lipschitz boundary Γ.

Definition 6.1. *A Galerkin discretization of a variational problem is called* ***adaptive****, if it employs a trial space V_n that is based on non-uniform meshes or non-uniform polynomial degree of the finite elements. We distinguish*

- ***a priori*** *adapted finite element spaces, which aim to take into account known features of the exact solution.*
- ***a posteriori*** *adapted finite element spaces, whose construction relies on the data of the problem.*

The next example shows that a posteriori adaptivity can dramatically enhance accuracy:

Example 6.2. If we knew the continuous solution $u \in V$ of the linear varitional problem (LVP), we could just choose $V_n := \operatorname{span}\{u\}$ and would end up with a perfect Galerkin discretization.

Three basic policies can be employed to achieve a good fit of the finite element space and the continuous solution:

- adjusting of the mesh $\mathcal{M}$ while keeping the type of finite elements (**h-adaptivity**).
- adjusting the local trial spaces (usually by raising/lowering the local polynomial degree) while retaining a single mesh (**p-adaptivity**).
- combining both of the above approaches (**hp-adaptivity**).

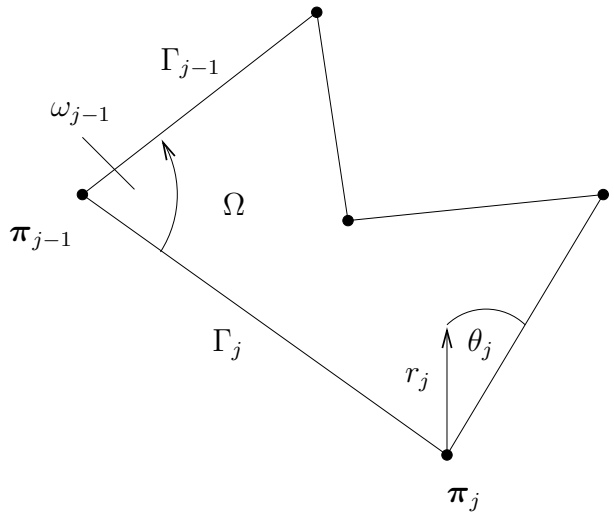


Figure 6.1: Polygon Ω and notation for the corners.

6.1 Regularity of solutions of second-order elliptic boundary value problems

If the geometry does not interfere, the solution of (6.1) is as smooth as the data f permit:

Theorem 6.3. *If $\partial\Omega$ is smooth (i. e. $\partial\Omega$ has a parameter representation with C^∞ functions), then for the solution u of (6.1) it holds*

$$f \in H^k(\Omega) \Longrightarrow u \in H^{k+2}(\Omega) \quad \text{for } k \in \mathbb{N}_0 \,,$$

and

$$\forall k \in \mathbb{N}_0,\ \exists \gamma = \gamma(\Omega, k): \quad \|u\|_{H^{k+2}(\Omega)} \leq \gamma(\Omega, k)\ \|f\|_{H^k(\Omega)} \quad \forall f \in H^k(\Omega)\,.$$

Similar results hold for Neumann boundary conditions on the whole of $\partial\Omega$.

If $\partial\Omega$ has corners (as in the case of a polygonal domain), the results from the previous section do not hold any longer.

Theorem 6.4. *Let $\Omega \subset \mathbb{R}^2$ be a polygon with J corners $\boldsymbol{\pi}_j$. Denote the polar coordinates in the corner $\boldsymbol{\pi}_j$ by (r_j, θ_j) and the inner angle at the corner $\boldsymbol{\pi}_j$ by w_j as in Figure 6.1. Additionally, let $f \in H^{-1+s}(\Omega)$ with $s \geq 1$ integer[1] and $s \neq \lambda_{jk}$, where the*

[1]The result holds for $s > 0$ non-integer as well. Since we only defined the spaces $H^k(\Omega)$ for k integer, we do not go into the details here.

λ_{jk} are given by the ***singular exponents***

$$\lambda_{jk} = \frac{k\pi}{\omega_j} \quad \text{for } k \in \mathbb{N}\,. \tag{6.2}$$

Then, we have the following decomposition of the solution $u \in H^1_0(\Omega)$ of the ***Dirichlet problem*** (6.1) *into a regular part (i. e. with the regularity one would expect from a smooth boundary according to Thm. 6.3) and finitely many so-called* ***singular functions*** *$s_{jk}(r,\theta)$:*

$$u = u^0 + \sum_{j=1}^{J} \psi(r_j) \sum_{\lambda_{jk}<s} \alpha_{jk} s_{jk}(r_j,\theta_j). \tag{6.3}$$

Here, $u^0 \in H^{1+s}(\Omega)$ and ψ is a C^∞ cut off function ($\psi \equiv 1$ in a neighborhood of 0). The singular functions s_{jk} are explicitly given by

$$\begin{aligned} &\lambda_{jk} \text{ non-integer:} && s_{jk}(r,\theta) = r^{\lambda_{jk}} \sin(\lambda_{jk}\theta), \\ &\lambda_{jk} \in \mathbb{N}: && s_{jk}(r,\theta) = r^{\lambda_{jk}} (\ln r)(\sin(\lambda_{jk}\theta) + \theta\cos(\lambda_{jk}\theta)). \end{aligned}$$

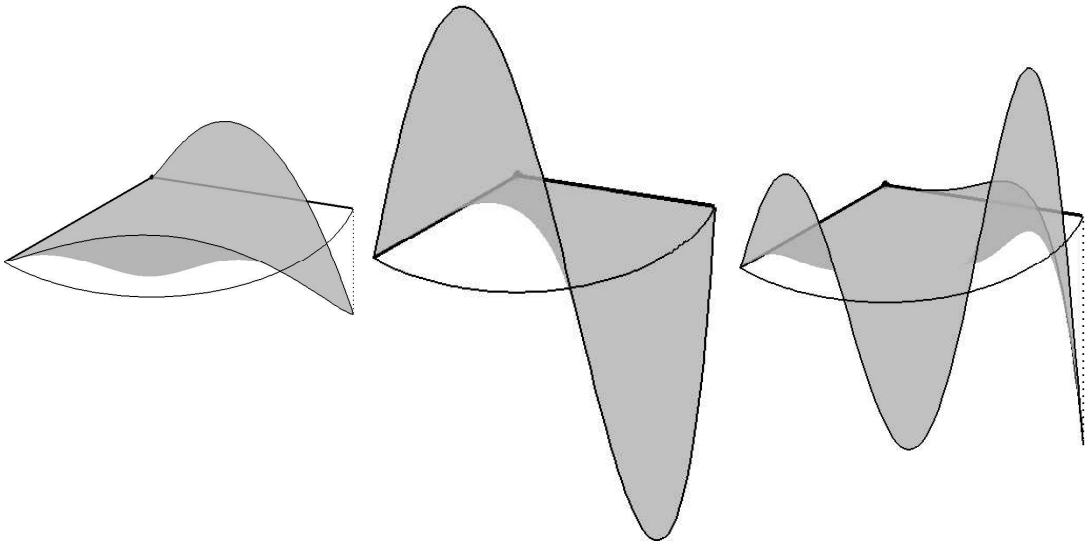


Figure 6.2: Singular functions s_1, s_2, and s_3 at a corner with $\omega = {}^{3\pi}/_{4}$

For the homogeneous Neumann problem in (6.1), sin has to be replaced by cos and vice-versa.

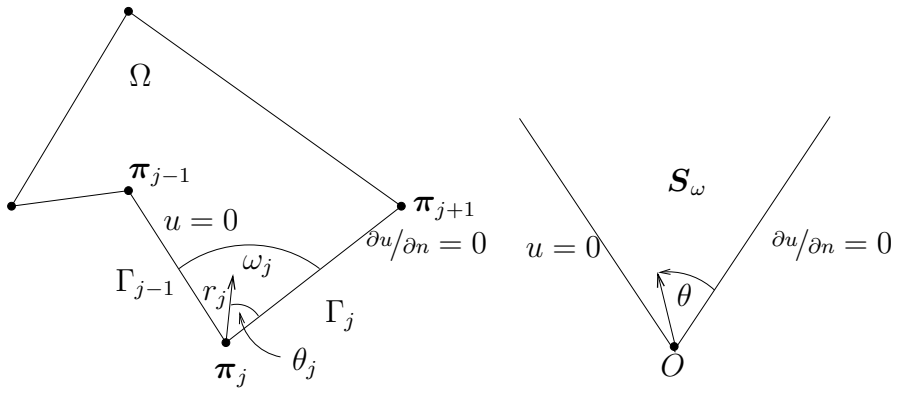


Figure 6.3: Corner $\boldsymbol{\pi}_j$ with changing boundary conditions and the infinite sector $\boldsymbol{S}_\omega$.

Remark 6.5. The coefficients α_{jk} in (6.3) depend only on f and are called (generalised) **stress intensity factors**.

Remark 6.6. At first glance, the decomposition (6.3) appears to be very special and restricted to the problem (6.1). Yet, similar decompositions with suitable $s_{jk}(r,\theta)$ hold for all elliptic boundary value problems of the form

$$-\operatorname{div}(\mathbf{A}\,\mathbf{grad}\,u) = f \quad \text{in } \Omega \quad , \quad u = 0 \quad \text{on } \Gamma_D \ , \ \langle \mathbf{A}\,\mathbf{grad}\,u, \mathbf{n}\rangle = 0 \quad \text{on } \Gamma_N \ .$$

Generally, $s_{jk}(r,\theta) = r^{\lambda_{jk}}\,\Theta_{jk}(\theta)$ is a non-trivial solution of the homogeneous differential equation in an infinite sector $\boldsymbol{S}$ with a tip at the singular point: Two examples should clarify this:

Example 6.7. Consider $-\Delta u = f$ in Ω with mixed boundary conditions at $\boldsymbol{\pi}_j$. Let $\boldsymbol{\pi}_j \in \partial\Omega$ be a boundary point where the type of the boundary conditions changes from Dirichlet to Neumann (cf. Figure 6.3).

In the infinite sector

$$\boldsymbol{S}_\omega = \{(r,\theta) : 0 < r < \infty,\ 0 < \theta < \omega\} \ ,$$

we are looking for non-trivial solutions of the homogeneous problem

$$\Delta s = 0 \text{ in } \boldsymbol{S}_\omega, \frac{\partial s}{\partial n}\Big|_{\theta=0} = 0, \quad s\Big|_{\theta=\omega} = 0$$

of the form $s(r,\theta) = r^\lambda\,\Theta(\theta)$. Using $s = r^\lambda\,\Theta(\theta)$, it follows in $\boldsymbol{S}_\omega$:

$$0 = \Delta s = r^{\lambda-2}(\Theta'' + \lambda^2\Theta) \quad \text{for } r > 0,$$

i. e. the pairs $(\lambda, \Theta(\theta))$ are **eigenpairs of a Sturm-Liouville problem**

$$\mathcal{L}\Theta = \Theta'' + \lambda^2\Theta = 0 \text{ in } (0,\omega) \ , \quad \Theta'(0) = 0 \ , \quad \Theta(\omega) = 0.$$

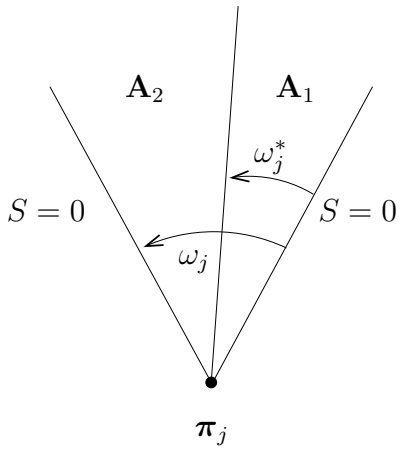


Figure 6.4: Corner P_j where two materials connect.

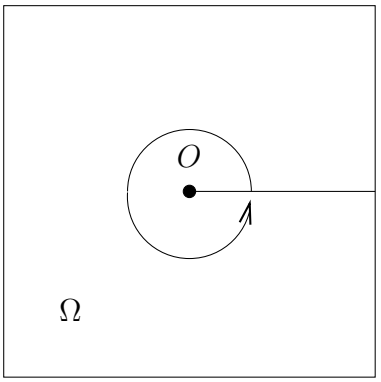


Figure 6.5: Cracked panel.

One recalculates that the eigenpairs are explicitly given by

$$\lambda_k = (k - 1/2)\,\frac{\pi}{\omega} \quad , \quad \Theta_k(\theta) = \cos(\lambda_k \theta), \quad k = 1, 2, 3, \dots \ .$$

Note: even if $\omega = \pi$, *i. e.* for changing boundary conditions on a straight edge, there exists a singularity $r^{1/2}\cos(\theta/2)$ for changing boundary conditions.

Example 6.8. Now, let us assume that $\mathbf{A}$ is piecewise constant with respect to a polygonal sub-division of Ω: let $\Omega \subset \mathbb{R}^2$ be a polygon and $u|_{\partial\Omega} = 0$. Let $\boldsymbol{\pi}_j$ be a corner, where several "materials" connect, *i. e.* $\Omega = \Omega_1 \cup \Omega_2$ with $\mathbf{A}|_{\Omega_i} = \mathbf{A}_i = \text{const}$, $i = 1, 2$, see Fig. 6.4. Then, $s(r, \theta)$ solves

$$Ls = \operatorname{div}(\mathbf{A}\ \mathbf{grad}\, s) = 0 \text{ in } \boldsymbol{S} \text{ and } s = 0 \text{ on } \partial\boldsymbol{S}$$

in the sector $\boldsymbol{S}$, if (λ, Θ) is a solution of the eigenvalue problem

$$\mathcal{L}\Theta := r^{2-\lambda}\, L(r^\lambda\, \Theta) = 0 \text{ in } \boldsymbol{S}, \quad \Theta(0) = \Theta(\omega_j) = 0.$$

By homogeneity, $\mathcal{L}(\partial_\theta)$ does not depend on r but has piecewise constant coefficients in $(0, \omega_j)$. As before, there are eigenpairs (λ_k, Θ_k), $k = 1, \dots, \infty$. However, the eigenfunctions $\Theta_k(\theta)$ are only piecewise smooth. They have a kink at $\theta = \omega_j^* < \omega_j$. As before, on the other hand, the eigenvalues are real and $0 < \lambda_1 \le \lambda_2 \dots$.

Example 6.9. Consider the pure Neumann problem for $-\Delta u = f$ on a domain with a crack (tip of the crack at the origin as in Figure 6.5). Here $\omega = 2\pi$ and therefore $\lambda_k = \frac{k\pi}{2\pi} = \frac{k}{2}$ and

$$u \equiv u^0 + \sum_{k=1}^{\infty} \alpha_k r^{k/2} \cos\left(\frac{k\theta}{2}\right).$$

Remark 6.10. Note that the singular functions $s_{jk}(r,\theta)$ in (6.3) have a singularity at $r = 0$ whereas they are smooth for $r > 0$. Therefore, the solution u of the Poisson problem (6.1) with a smooth right hand side f is smooth in the interior of Ω. The singular behaviour of u is restricted to the corners $\boldsymbol{\pi}_j$.

Remark 6.11. The decomposition of the solution in Theorem 6.4 shows that for $\omega_j > \pi$ the following holds: $\lambda_{j1} = \pi/\omega_j < 1$. Additionally, it follows from $(\partial^{\boldsymbol{\alpha}} s_{j1})(r_j,\theta_j) \sim r_j^{\lambda_{j1}-|\boldsymbol{\alpha}|}$ for $r_j \to 0$ that the derivative $\partial^{\boldsymbol{\alpha}}$ of the singular functions s_{jk} for $|\boldsymbol{\alpha}| = 2$ is not square integrable since $\lambda_{j1} - |\boldsymbol{\alpha}| < -1$, *i. e.* for $|\boldsymbol{\alpha}| = 2$ we have

$$|(\partial^{\boldsymbol{\alpha}} s_{j1})(r_j,\theta_j)|^2 \sim r_j^{-2-\varepsilon} \notin L^1(\Omega).$$

The shift theorem Thm. 6.3 does no longer hold.

Bibliographical notes. Corner and edge singularities for solutions of elliptic problems are discussed in [20, 30, 36].

6.2 Convergence of finite element solutions

Let $u_n \in \mathcal{S}_m(\mathcal{M}_n)$ stand for the Galerkin solution of (6.1) obtained by means of Lagrangian finite elements of uniform polynomial degree $m \in \mathbb{N}$ on the mesh $\mathcal{M}_n$, see Sect. 3.8.1. Temporarily, we will allow $d \in \{1,2,3\}$.

Let $\{\mathcal{M}_n\}_{n=1}^{\infty}$ denote a uniformly shape-regular and quasi-uniform family of triangulations of the polygon Ω such that $h_n := h_{\mathcal{M}_n} \to 0$ as $n \to \infty$. From Sect. 4.5 we know that, if the continuous solution u satisfies $u \in H^t(\Omega)$, $t \geq 2$, we have, as $n \to \infty$, the asymptotic error estimate

$$\|u - u_n\|_{H^1(\Omega)} \leq \gamma h_n^{\min(m+1,t)-1} \, |u|_{H^t(\Omega)} \,, \tag{6.4}$$

with $\gamma > 0$ independent of n and u.

For a unified analysis of the h-version and p-version of finite elements and, in particular, on non-uniform meshes it is no longer meaningful state a-priori error estimates in terms of the meshwidth.

Hence, let us measure the "costs" involved in a finite element scheme by the dimension of the finite element space, whereas the "gain" is gauged by the accuracy of the finite element solution in the H^1-norm. For the h-version we first assume a uniformly shape-regular and quasi-uniform family $\{\mathcal{M}_n\}_{n=1}^{\infty}$ of simplicial meshes. In the case of Lagrangian finite elements of polynomial degree m we have the crude estimates

$$N_n := \dim(\mathcal{S}_m(\mathcal{M}_n)) \leq \binom{d+m}{d} \cdot \sharp\mathcal{M}_n \quad \Rightarrow \quad \sharp\mathcal{M}_n \approx h_{\mathcal{M}_n}^{-d} \,,$$

with constants depending on shape-regtularity and m. Thus, if $t \geq m+1$ we get asymptotically

$$\|u - u_n\|_{H^1(\Omega)} \leq \gamma N_n^{-m/d} \,. \tag{6.5}$$

The constant γ depends on $\Omega, \mathbf{A}$ and the bounds for $\rho_{\mathcal{M}_n}$, $\mu_{\mathcal{M}_n}$. This reveals an **algebraic asymptotic convergence rate** of the h-version of Lagrangian finite elements for second order elliptic problems.

However, even for small m the regularity $u \in H^{m+1}(\Omega)$ cannot be taken for granted. Consider $d = 2$ and remember that from Sect. 6.1 it is merely known that for $f \in H^{k-2}(\Omega)$:

$$u = u^0 + u_{\text{sing}} \tag{6.6}$$

with a smooth part $u^0 \in H^k(\Omega)$, $k \geq 2$, and with a singular part u_{sing}, which is a (finite!) sum of singular functions $s(r_i, \theta_i)$, which have the explicit form

$$s(r, \theta) = r^\lambda \, \Theta(\theta) \,, \tag{6.7}$$

with piecewise smooth Θ, where $0 < \lambda < k-1$ (we assume here that $\log r$ terms are absent). The singular functions (6.7) are only poorly approximated by Lagrangian finite element functions on sequences of quasi-uniform meshes. For the singular functions $s(r, \theta)$ as in (6.7) and with (r, θ) denoting polar coordinates at a vertex of Ω the (optimal) error estimate

$$\min_{v_n \in \mathcal{S}_m(\mathcal{M}_n)} \|s - v_n\|_{H^1(\Omega)} \leq \gamma h_n^{\min(m,\lambda)} \leq \gamma N_n^{-\min(m,\lambda)/2}$$

holds, where again $N_n := \dim \mathcal{S}_m(\mathcal{M}_n) = O(h_n^{-2})$ denotes the number of degrees of freedom.

For a sequence $\{\mathcal{M}_n\}_{n=1}^\infty$ of quasi uniform meshes one therefore observes only the suboptimal convergence rate

$$\|u - u_n\|_{H^1(\Omega)} \leq \gamma h_n^{\min(m,\lambda^*)} \leq \gamma N_n^{-\min(m,\lambda^*)/2} \,, \tag{6.8}$$

where $\lambda^* = \min\{\lambda_{jk} : j = 1, \dots, J,\ k = 1, 2, \dots\}$, as $h_n \to 0$ (or for $N_n \to \infty$), instead of the optimal asymptotic convergence rate (6.5) supported by the polynomial degree of the finite element space.

Since often $\lambda^* < 1$, one observes even for the simple piecewise linear ($m = 1$) elements a reduced convergence rate, and for $m > 1$ we hardly ever get the optimal asymptotic rate $O(N_n^{-m/d})$.

Remark 6.12. If the exact solution u is very smooth, that is, $t \gg 1$, raising the polynomial degree m is preferable (p-version), because asymptotically for $m \geq t$ we have $N_n \approx m^d h^{-d}$ and, thus the estimate (see Remark 4.28)

$$\inf_{v_n \in \mathcal{S}_m(\mathcal{M})} \|u - v_n\|_{H^1(\Omega)} \leq \gamma(\rho_{\mathcal{M}}, \mu_{\mathcal{M}}) \left(\frac{h_{\mathcal{M}}}{m}\right)^{\min\{m+1,t\}-1} \|u\|_{H^t(\Omega)} \,. \tag{4.21}$$

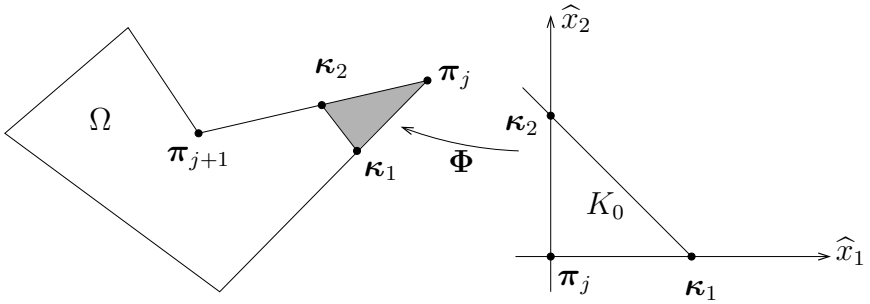


Figure 6.6: Polygon Ω with corner $\boldsymbol{\pi}_j$, subdomain $\mathrm{conv}(\boldsymbol{\pi}_j, \boldsymbol{\kappa}_1, \boldsymbol{\kappa}_2)$ adjacent to it and its affine map F to the standard vertex K_0.

gives

$$\|u - u_n\|_{H^1(\Omega)} \leq \gamma N_n^{(t-1)/d} \,. \tag{6.9}$$

For large t this is clearly superior to (6.5).

The bottom line is that low Sobolev regularity of the exact solution suggests the use of the h-version of finite elements, whereas in the case of very smooth solutions the p-version is more efficient (w.r.t. the dimension of the finite element space).

Remark 6.13. If the exact solution is **analytic** in $\overline{\Omega}$, that is, it is C^∞ and can be expanded into a locally convergent power series in each point of Ω, then the p-version yields an **exponential asymptotic convergence rate**

$$\|u - u_n\|_{H^1(\Omega)} \leq \gamma \exp(-\gamma' N_n^\beta) \,, \tag{6.10}$$

with $\gamma, \gamma', \beta > 0$ only depending on problem parameters and the fixed triangulation, but independent of the polynomial degree m of the Lagrangian finite elements.

6.3 A priori adaptivity

The developments of Sect. 6.1 give plenty of information about the structure of the solutions of (6.3) for smooth data f. It is the gist of **a priori adaptive** schemes to take into account this information when picking the finite element space.

This can overcome the poor performance of Lagrangian finite elements on quasi-uniform meshes pointed out in Sect. 6.2.

6.3.1 A priori graded meshes

One option is **judicious mesh refinement towards the vertices of the polygon**. Consider the polygon Ω shown in Fig. 6.6. In Ω, consider any vertex $\boldsymbol{\pi}_j$ (In Fig. 6.6

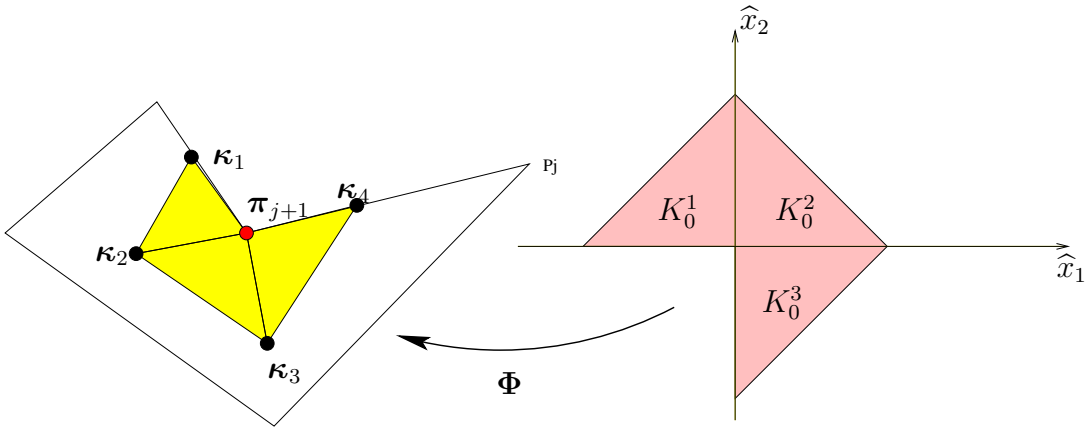


Figure 6.7: Mapping for a re-entrant corner at $\boldsymbol{\pi}_{j+1}$

we chose a convex corner, the approach to a re-entrant corner at $\boldsymbol{\pi}_{j+1}$ is indicated in Fig. 6.7). We denote again by (r, θ) polar coordinates at vertex $\boldsymbol{\pi}_j$, and by $s(r, \theta)$ a singular function as in (6.7)

$$s(r, \theta) = r^\lambda \, \Theta(\theta)$$

with a smooth $\Theta(\theta)$. The triangle $K = \mathrm{conv}(\boldsymbol{\pi}_j, \boldsymbol{\kappa}_1, \boldsymbol{\kappa}_2)$ denotes a neighbourhood of vertex $\boldsymbol{\pi}_j$ in Ω (shown shaded in Fig. 6.6). By means of an affine map $\boldsymbol{\Phi}$ the triangle K is mapped onto the reference triangle K_0 with polar coordinates $(\widehat{r}, \widehat{\theta})$. The singular function $s(r, \theta)$ in Ω is transformed by $\boldsymbol{\Phi}$ into

$$\widehat{s}(\widehat{r}, \widehat{\theta}) = \widehat{r}^\lambda \, \widehat{\Theta}(\widehat{\theta}) \quad \text{in } K_0 \, ,$$

with the same exponent λ but with another C^∞-function $\widehat{\Theta}(\widehat{\theta})$:

Example 6.14. Let $\boldsymbol{\pi}_j = (0, 0)$, (x_1, x_2) stand for the coordinates in Ω,

$$x_1 = r \cos\theta \quad , \quad x_2 = r \sin\theta$$
$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \mathbf{F} \begin{pmatrix} \widehat{x}_1 \\ \widehat{x}_2 \end{pmatrix} = \widehat{r} \begin{pmatrix} \cos\widehat{\theta} \\ \sin\widehat{\theta} \end{pmatrix}$$

and

$$s(r, \theta) = r^\lambda \, \Theta(\cos\theta, \sin\theta)$$

denote the singular function in (6.7). To prove that $s(r, \theta)$ is, in the coordinates $\widehat{x}_1, \widehat{x}_2$, once again of the form (6.7) let $\mathbf{F} = (f_{ij})_{1 \le i,j \le 2}$. Then

$$\begin{aligned} r^2 = x_1^2 + x_2^2 = \; & (f_{11}\,\widehat{x}_1 + f_{12}\,\widehat{x}_2)^2 + (f_{21}\,\widehat{x}_1 + f_{22}\,\widehat{x}_2)^2 \\ = \; & \widehat{r}^2 \, \{(f_{11}\cos\widehat{\theta} + f_{12}\sin\widehat{\theta})^2 + (f_{21}\cos\widehat{\theta} + f_{22}\sin\widehat{\theta})^2\} \, , \end{aligned}$$

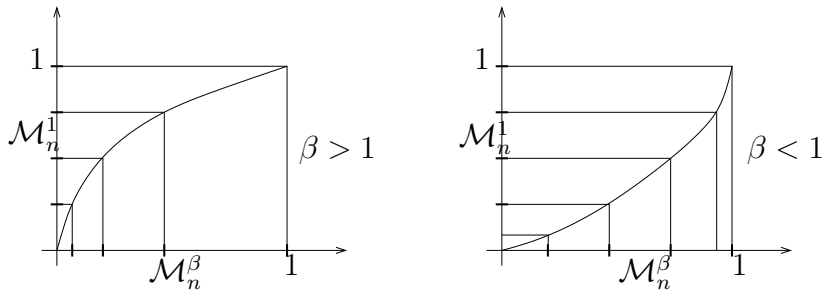


Figure 6.8: Graded meshes $\mathcal{M}_n^\beta$ in $\Omega =]0,1[$, cases $\beta > 1$ and $\beta < 1$.

and

$$r^\lambda = \widehat{r}^\lambda \{(f_{11}\cos\widehat{\theta} + f_{12}\sin\widehat{\theta})^2 + (f_{21}\cos\widehat{\theta} + f_{22}\sin\widehat{\theta})^2\}^{\frac{\lambda}{2}} = \widehat{r}^\lambda\, \Theta_1(\widehat{\theta}) \ ,$$

with a smooth (analytic) function $\Theta_1(\widehat{\theta})$. Analogously, we have that $\Theta(\theta) = \widehat{\Phi}_2(\widehat{\theta})$ with a smooth function $\widehat{\Theta}_2(\widehat{\theta})$.

Due to the transformation theorem it is therefore sufficient to investigate the finite element approximation of $s(r,\theta)$ in (6.7) in the reference domain K_0 as shown in Figure 6.9. In the case of a re-entrant corner, the reference domain consists of three triangles, see Fig. 6.7, and the ensuing considerations can be applied to each of them.

In what follows we show that by using so-called **algebraically graded meshes** $\mathcal{M}_n^\beta$ at the vertices of Ω the optimal asymptotic behavior $O(N^{-m/2})$ of the best approximation error of Lagrangian finite elements of uniform global degree m can be retained for singular functions as well.

Definition 6.15. *A family $\{\mathcal{M}_n^\beta\}_{n=1}^\infty$ of meshes of a computational domain $\Omega \subset \mathbb{R}^2$ is called **algebraically graded** with respect to $\boldsymbol{\pi} \in \overline{\Omega}$ and grading factor $\beta \geq 1$ if*

(i) the meshes are uniformly shape-regular, and

(ii) with constants independent of n and $h_n := h_{\mathcal{M}_n^\beta}$,

$$\forall K \in \mathcal{M}_n^\beta,\ \boldsymbol{\pi} \notin \overline{K}: \quad h_K \approx n^{-1}\,\mathrm{dist}(\boldsymbol{\pi}, K)^{1-1/\beta} \ .$$

We will describe the concrete construction of algebraically graded meshes $\mathcal{M}_n^\beta$,$n \in \mathbb{N}$, with grading factor $\beta \geq 1$ $n \in \mathbb{N}$ on the reference domain K_0 with respect to the vertex $\binom{0}{0}$, see Fig. 6.9:

Construction 6.16 (Graded mesh on reference triangle). *On $K_0 = \mathrm{convex}\{\binom{0}{0}, \binom{1}{0}, \binom{0}{1}\}$ we proceed as follows:*

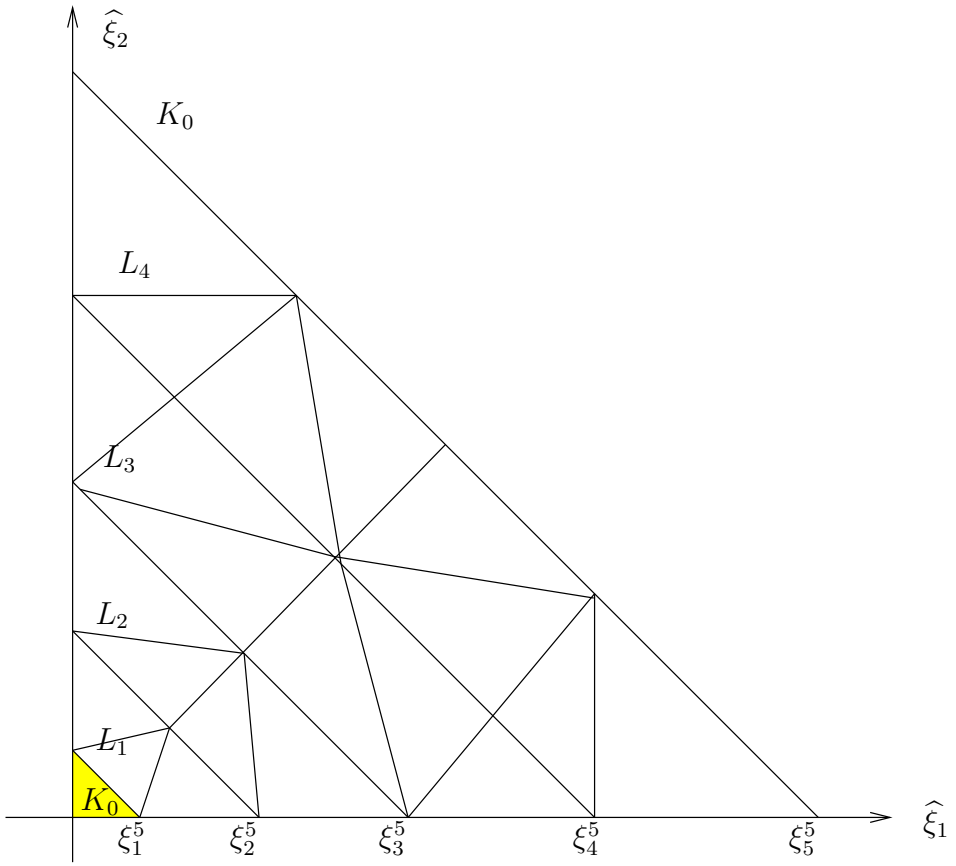


Figure 6.9: Graded mesh for grading factor $\beta = 3/2$ on reference domain K_0 that is $\tau_j^5 = (j/5)^\beta$.

1. *Construct a partition* $0 = \tau_0^n < \tau_1^n < \cdots < \tau_n^n = 1$ *of* $]0;1[$ *by setting* $\tau_j^n := (j/n)^\beta$, $j = 1, \ldots, n$.

2. *Use this partition to define the layers*

$$L_j = \{\widehat{\boldsymbol{\xi}} \in K_0 : \ \tau_{j-1}^n < \xi_1 + \xi_2 < \tau_j^n\}\,, \quad j = 1, \ldots, n\,.$$

3. *Equip each layer* L_j, $j = 1, \ldots, n$ *with a simplicial triangulation* $\mathcal{M}^\beta_{n|L_j}$ *such that*

   a) *their union yields a simplicial triangulation of* K_0,

   b) *the shape regularity measure of* $\mathcal{M}^\beta_{n|L_j}$ *(see Def. 4.15) is uniformly bounded idependently of* j *and* n,

   c) *for each* $K \in \mathcal{M}^\beta_{n|L_1}$ *we have* $h_K \approx \tau_j - \tau_{j-1}$ *with constants idependent of* j *and* n, *and*

   d) $\mathcal{M}^\beta_{n|L_1}$ *consists of a single triangle* K^* *adjacent to* $\binom{0}{0}$

Exercise 6.1. Write a small program that creates an algebraically graded mesh $\mathcal{M}_n^\beta$, $\beta \geq 1$, $n \in \mathbb{N}$, of the reference triangle from (4.9) and stores it in a file in the format described in Remark 3.14.

Remark 6.17. For $\beta = 1$ the meshes $\mathcal{M}_n^\beta$ are quasi-uniform with meshwidth $1/n$.

Lemma 6.18. *Fix* $\beta > 1$. *Then, with constants only depending on the bounds on the shape-regularity measure* $\rho(\mathcal{M}^\beta_{n|L_j})$ *and quasi-uniformity measure* $\mu(\mathcal{M}^\beta_{n|L_j})$ *we find*

$$h_K \approx \frac{\beta}{n}\left(\frac{j}{n}\right)^{\beta-1} \quad \forall K \in \mathcal{M}^\beta_{n|L_1}\,, \tag{6.11}$$

and

$$\sharp\mathcal{M}^\beta_{n|L_1} \approx j\,. \tag{6.12}$$

Proof. Pick $n \in \mathbb{N}$ and $j \in \{2, \ldots, n\}$. By the mean value theorem we find

$$\frac{\beta}{n}\left(\frac{j-1}{n}\right)^{\beta-1} \leq \tau_j - \tau_{j-1} \leq \frac{\beta}{n}\left(\frac{j}{n}\right)^{\beta-1}\,.$$

Together with $h_{K^*} = n^{-\beta}$ we conclude the first assertion of the lemma.

To confirm the second, we start with the volume formula

$$2|L_j| = \left(\frac{j}{n}\right)^{2\beta} - \left(\frac{j-1}{n}\right)^{2\beta} \approx \frac{2\beta}{n}\left(\frac{j}{n}\right)^{2\beta-1}\,. \tag{6.13}$$

As a consequence of (6.11), the area of a triangle $\subset L_j$ is

$$2|K| \approx h_K^2 \approx \frac{\beta^2}{n^2}\left(\frac{j}{n}\right)^{2\beta-2} \quad \forall K \in \mathcal{M}^\beta_{n|L_j} \tag{6.14}$$

None of the constants depends on n and j. Dividing (6.13) by (6.14) yields (6.12). □

Corollary 6.19. *The family* $\{\mathcal{M}_n^\beta\}$, $n \in \mathbb{N}$, $\beta \geq 1$, *of meshes emerging from construction 6.16 is algebraically graded with respect to* $\binom{0}{0}$ *and grading factor* β.

Corollary 6.20. *The algebraically graded meshes* $\mathcal{M}_n^\beta$, $n \in \mathbb{N}$, *of* K_0 *constructed as above for* $\beta \geq 1$ *satisfy*

$$h_n := h_{\mathcal{M}_n^\beta} \approx {}^{\beta}\!/_{n} \quad , \quad \sharp\mathcal{M}_n^\beta \approx n^{-2} \, ,$$

with constants independent of n.

As a consequence, $h(\mathcal{M}_n^\beta) \to 0$ for $n \to \infty$, if $\beta \geq 1$. Moreover, for fixed $m \in \mathbb{N}$, we get from Cor. 6.20 that

$$N_n = \dim \mathcal{S}_m(\mathcal{M}_n^\beta) \leq \gamma \sharp\mathcal{M}_n^\beta \leq \gamma n^2$$

holds with a constant independent of n. As, again by Cor. 6.20, $n^{-1} \leq \gamma N_n^{-\frac{1}{2}}$, we deduce from Thm. 4.24 that for the regular part $u^0 \in H^{m+1}(K_0)$ of the decomposition (6.6) of the solution $u \in H_0^1(\Omega)$ of $-\Delta u = f$ holds

$$\min_{v_n \in \mathcal{S}_m(\mathcal{M}_n^\beta)} \left\| u^0 - v_n \right\|_{H^1(K_0)} \leq \left\| u^0 - \mathsf{I}\, u^0 \right\|_{H^1(K_0)} \leq \gamma h_n^m \leq \gamma N_n^{-m/2} \, , \tag{6.15}$$

with $\gamma = \gamma(m, \rho_{\mathcal{M}_n^\beta})$. Here I is the finite element interpolation operator for $\mathcal{S}_m(\mathcal{M}_n^\beta)$, *cf.* Sect. 3.6.

This implies that the regular part u^0 of the solution u can also be approximated on algebraically β-graded meshes at the optimal rate (6.4), independently of the size of $\beta \geq 1$ (the size of the constant γ in the error estimates (6.15) depends of course on β and possibly grows strongly with $\beta > 1$; for fixed β and $n \to \infty$ the convergence rate (6.15) is optimal, however).

Let us now consider the singular part of u in the decomposition (6.6). According to (6.3) the solution u_{sing} is a finite sum of terms of the form (6.7) (the treatment of terms of the form $r^\lambda |\log r|\, \Theta(\theta)$ is left to the reader as an exercise), where $\Theta(\theta) \in C^\infty([0, \omega])$ is assumed without loss of generality, and where $\lambda > 0$.

Theorem 6.21. *Let* $s(r, \theta) = r^\lambda\, \Theta(\theta)$ *with* $\lambda > 0$ *and* $\Theta \in C^\infty([0, {}^{\pi}\!/_{2}])$ *for* $(r, \theta) \in K_0$ *as in Fig. 6.9. Let further*

$$\beta > \max\{m/\lambda, 1\}.$$

Then there holds, as $N_n = \dim \mathcal{S}_m(\mathcal{M}_n^\beta) \to \infty$

$$\min_{v_n \in \mathcal{S}_m(\mathcal{M}_n^\beta)} |s - v_n|_{H^1(K_0)} \leq \gamma(m, \lambda, \beta)\, N_n^{-\frac{m}{2}},$$

i. e. for $\beta > m/\lambda$ *the optimal asymptotic convergence rate (6.4) for smooth solutions* u *is recovered.*

Proof. The proof relies on local estimates for the interpolation error $s - \mathsf{I}\, s$ For the sake of simplicity we restrict the discussion to the case of $m = 1$ and leave the generalization to arbitrary polynomial degree to the reader. Hence, I designates linear interpolation on $\mathcal{M}_n^\beta$.

Let $K^* \in \mathcal{M}_n^\beta$ denote the triangle which contains the origin $\binom{0}{0}$ in its closure. We are going to demonstrate that the contribution of this triangle to the interpolation error is negligible.

First, using polar coordinates, observe that

$$\begin{aligned}|s|^2_{H^1(K^*)} &\le \int_0^{\tau_1}\int_0^{\pi/2} \left(\left(\frac{\partial s}{\partial r}\right)^2 + \left(\frac{1}{r}\frac{\partial s}{\partial \theta}\right)^2\right) r\,\mathrm{d}\theta\,\mathrm{d}r = \int_0^{\tau_1}\int_0^{\pi/2} \left(\left(\lambda r^{\lambda-1}\Theta(\theta)\right)^2 + \left(\frac{1}{r}\, r^\lambda \Theta'(\theta)\right)^2\right) r\,\mathrm{d}\theta\,\mathrm{d}r \\ &\le \gamma(\Theta) \int_0^{\tau_1} \lambda^2 r^{2\lambda-1} + r^{2\lambda-1}\,\mathrm{d}r = \gamma(\Theta)\,(1+\lambda^2)\,\left(\frac{1}{n}\right)^{2\beta\lambda}\,.\end{aligned}$$

Since $\mathsf{I}(s)$ is linear on K^*, we find

$$\begin{aligned}|\mathsf{I}\, s|^2_{H^1(K^*)} &= |K^*|\tau_1^{-2}\left(s(\tau_1,0)^2 + s(0,\tau_1)^2\right) = {}^1\!/_2\tau_1^{2\lambda}(\Theta(0)^2 + \Theta({}^\pi\!/_2)^2) \\ &= \gamma(\Theta)\,\tau_1^{2\lambda} = \gamma(\Theta)\,\left(\frac{1}{n}\right)^{2\beta\lambda}\,.\end{aligned}$$

Thus, a simple application of the triangle inequality yields

$$|s - \mathsf{I}\, s|^2_{H^1(K^*)} \le \gamma(\Theta)\,\left(\frac{1}{n}\right)^{2\beta\lambda} \le \gamma N_n^{-\beta\lambda}\,.$$

Next, consider $K_0 \setminus K^*$. For $x \in K_0$ define the piecewise constant function $h(x)$ by

$$h(x)|_K = h_K \quad \forall K \in \mathcal{M}_n^\beta\,.$$

Then, the local interpolation error estimate of Thm. 4.24 (with $r = 1$, $m = 1$, $t = 2$) yields

$$\begin{aligned}|s - \mathsf{I}\, s|^2_{H^1(K_0\setminus K^*)} &= \sum_{\substack{K\in\mathcal{M}_n^\beta \\ K\neq K^*}} |s - \mathsf{I}\, s|^2_{H^1(K)} \\ &\le \gamma \sum_{\substack{K\in\mathcal{M}_n^\beta \\ K\neq K^*}} h_K^2\,|s|^2_{H^2(T)} = \gamma \int_{K_0\setminus K^*} h^2\,|D^2 s|^2\,\mathrm{d}\boldsymbol{\xi}\,.\end{aligned}$$

The construction of the graded mesh implies that at the point $\boldsymbol{\xi} \in K_0 \setminus K^*$ holds

$$\begin{gathered} r = |\boldsymbol{\xi}| > \gamma n^{-\beta} \text{ with } \gamma > 0 \text{ independent of } n, \\ |h(\boldsymbol{\xi})| \le \gamma n^{-1}\, r^{1-1/\beta},\ |D^2 s| \le \gamma r^{\lambda-2}\,.\end{gathered}$$

Hence, for $n \gg 1$,

$$\int_{K_0 \setminus T_0} (h(\boldsymbol{\xi}))^2 \, |D^2 s|^2 \, \mathrm{d}\boldsymbol{\xi} \leq \gamma \int_{1/2\sqrt{2}n^{-\beta}}^{1} n^{-2} \, r^{2(1-1/\beta)+2\lambda-4} \, r \, \mathrm{d}r$$
$$\leq \left[\gamma n^{-2} \, r^{2\lambda-2/\beta}\right]_{1/2\sqrt{2}n^{-\beta}}^{1} \leq \gamma n^{-2} \leq \gamma N_n^{-1} \, ,$$

which implies the assertion. Here, $\lambda > \beta^{-1}$ was used. *i. e.* , either $\lambda > p$ and $\beta = 1$ or $\lambda < p$ and $\beta > p/\lambda$. □

Exercise 6.2. Examine what happens for $m = 1$ and $\lambda = m/\beta$ in the setting of the previous theorem.

Remark 6.22. From Thm. 6.21 we learn that either $\lambda > m$ and $\beta = 1$ or $\lambda < m$ and $\beta > m/\lambda$ will ensure the optimal rate of convergence of the finite element solution in terms of the dimension of the finite element space.

Corollary 6.23. *For K_0 as shown in Fig. 6.9 and $k_{\max}$ such that $\lambda_{k_{\max}} = \{\max \lambda_k : \lambda_k < p\}$, we have $u = u^0 + u_{\text{sing}}$ with $u^0 \in H^{p+1}(K_0)$ and*

$$u_{\text{sing}} = \sum_{k=1}^{k_{\max}} \alpha_k \, r^{\lambda_k} \, \Phi_k(\theta) \, ,$$

where we assume that $\alpha_1 \neq 0$, and $0 < \lambda_1 \leq \lambda_2 \leq \ldots$, and $\Phi_k \in C^\infty([0, \omega])$. Then for $m \geq 1$, $\beta > \max\{1, m/\lambda_1\}$ it holds

$$\min_{v \in \mathcal{S}_m(\mathcal{M}_n^\beta)} |u_{\text{sing}} - v|_{H^1(K_0)} \leq \gamma N_n^{-m/2} \quad \textit{with } \gamma = \gamma(p, \alpha_k, \beta) \, . \tag{6.16}$$

Remark 6.24.

1. If $m/\lambda_1 > 1$ we achieve with the grading factor

$$\beta > p/\lambda_1 \tag{6.17}$$

the same convergence rate as for smooth solutions u with a quasi uniform triangulation.

2. For fixed N we have to increase β (*i. e.* the grading must be more pronounced) if a) m is raised and b) if the singular exponent λ_1 is reduced.

3. Usually the singular exponent $\lambda_1 > 0$ is unknown. Thm. 6.21 shows, however, that $\beta > 1$ must only be chosen sufficiently large in order to compensate for the effect of the corner singularity on the convergence rate of the FEM. The precise value of λ_1 is not necessary.

4. No refinement is required, if $\lambda_1 \geq m$. This is the case e. g. for $m = 1$ and the Laplace equation in convex polygons Ω, where $\lambda_j := \pi/\omega_j > 1$. For $m > 1$ mesh refinement near vertices is also required in convex domains, in general.

Exercise 6.3. Let $m > 1$. Discuss the estimate $|s - I_m s|_{H^1(K_0)}$ in detail, in particular the treatment on K_0 of the error $(\mathsf{I}_m - \mathsf{I}_1)s$, where I_m is the finite element interpolation operator for $\mathcal{S}_m(\mathcal{M}_n^\beta)$.

The preceding analysis at a single vertex can be transferred to the general polygon: let $\Omega \subset \mathbb{R}^2$ denote a polygon with straight sides and

$$f \in H^{k-2}(\Omega), \quad k \geq 2 \, .$$

Let further $u \in H_0^1(\Omega)$ denote the solution of the homogeneous Dirichlet problem for $-\Delta u = f$. Then, for every $m \geq 1$ there exists a $\beta > 1$ such that for $k \geq m + 1$ it holds

$$\inf_{v_n \in \mathcal{S}_m(\mathcal{M}_n^\beta)} |u - v_n|_{H^1(\Omega)} \leq \gamma N^{-m/2} \, ,$$

for $N = \dim \mathcal{S}_m(\mathcal{M}_n^\beta) \to \infty$.

Bibliographical notes. More details on approximation on graded meshes are given in [36, Sect. 4.3].

6.4 A posteriori error estimation

The error estimates (6.5) and also those of (6.23) are **a-priori**, *ie.* the error bound already holds **before** calculating u_n and without detailed knowledge of u_n. Two main drawbacks of these estimates are that

1. unless restrictive assumptions on the data are made, the regularity of the exact solution is unknown,

2. the constants in the estimates are either unknown or only given by very pessimistic upper bounds.

Therefore, a-priori error estimates allow only to estimate the ratio accuracy vs. number of degrees of freedom asymptotically under some regularity assumptions on the (unknown) exact solution, *but they are not suitable for a stopping criterion indicating that a prescribed accuracy has been reached.*

In practice, once the finite element solution u_n is given, the following questions arise:

1. Is it possible to use u_n for **computing** a bound for a suitable norm of the error $u - u_n$?

2. Can the program itself refine the mesh $\mathcal{M}$ (h-refinement) and increase the polynomial degree m (p-refinement) respectively, in an optimal fashion?

3. When should this process stop? (stopping criterion)

We will investigate these questions for the model problem (6.1) in two dimensions, the case of the h-version of lowest order Lagrangian finite elements on triangular meshes, and for the H^1-norm.

In this setting these questions can be answered by "yes": the so-called **local a-posteriori error estimators** use the finite element solution $u_n \in \mathcal{S}_m(\mathcal{M})$ for computing bounds for $\|u - u_N\|_{H^1(\Omega)}$. The basic idea behind every a-posteriori error estimate is the following: Because the exact solution u is unknown, we can only use the differential equation. Therefore, an error estimate always quantifies how well the finite element solution u_n solves the differential equation.

Definition 6.25. *Given a mesh $\mathcal{M}$ and an associated finite element solution u_n of boundary value problem with exact solution u, a **local a-posteriori error estimator** is a mapping $\eta : \mathcal{M} \mapsto \mathbb{R}^+$. Individual numbers $\eta^K := \eta(K)$, $K \in \mathcal{M}$, are known as **elemental error indicators** and the sum*

$$\mathrm{EST} := \left(\sum\nolimits_{K\in\mathcal{M}} (\eta^K)^2 \right)^{1/2}$$

*is the **global error estimate**.*

*The local a posteriori error estimator η is called **reliable**, if, with a constant $\gamma > 0$ only depending on the parameters of the continuous problem and the shape-regularity measure of $\mathcal{M}$ (and independent of the data of the problem, $h_{\mathcal{M}}$ and $\mu_{\mathcal{M}}$),*

$$\|u - u_n\| \leq \gamma \, \mathrm{EST}$$

The local a-posteriori error estimator η qualifies as efficient, if, with a constant $\gamma > 0$ as above

$$\eta^K \leq \gamma \, \|u - u_n\|_{\omega_K} \quad \forall K \in \mathcal{M} \, .$$

Here ω_K is a local $\mathcal{M}$-neighborhood of K.

Remark 6.26. The norm $\|\cdot\|$ that occurs in the above definition will often be the energy norm arising from an s.p.d. linear variational problem. In the concrete case of problem (6.1) it is the H^1-norm.

Remark 6.27. Efficiency of a local a posteriori error estimator implies

$$\mathrm{EST}^2 = \sum\nolimits_{K\in\mathcal{M}} (\eta^K)^2 \leq \gamma \sum\nolimits_{K\in\mathcal{M}} \|u - u_n\|^2_{\omega_K} \leq \gamma \, \|u - u_n\|^2 \, ,$$

if the norm $\|\cdot\|$ arises from local contributions and the $\mathcal{M}$-neighborhoods ω_K enjoy a finite overlap property.

6.4.1 Residual error estimators

Consider a linear variational problem (LVP) with bilinear form b and its Galerkin discretization based on $V_n \subset V$, see Sect. 1.4. The functional $r \in V^*$ defined by

$$r(v) = \langle r, v\rangle_{V^*\times V} := \langle f, v\rangle_{V^*\times V} - \mathsf{b}(u_n, v) \quad v \in V$$

is called the **weak residual**. It is connected with the discretization error

$$e := u - u_n \in V$$

by the **error equation**

$$\mathsf{b}(e, v) = \langle r, v\rangle_{V^*\times V} \quad \forall v \in V \ . \tag{6.18}$$

Let the assumptions of Thms. 1.17 and 1.30 be satisfied. Then, the V-norm discretization error can be estimated by

$$\gamma_s \left\|e\right\|_V \leq \sup_{v\in V\setminus\{0\}} \frac{|b(e,v)|}{\left\|v\right\|_V} = \left\|r\right\|_{V^*} \leq \left\|\mathsf{b}\right\| \left\|e\right\|_V \ . \tag{6.19}$$

Thus, *the dual norm of the weak residual functional provides an upper and lower bound for the V-norm of the Galerkin discretization error.*

In turns,

1. to get an upper bound for $\left\|r\right\|_{V^*}$ exploit Galerkin orthogonality, (1.16)

   $$\mathsf{b}(u - u_n, v_n) = 0 \quad \forall v_n \in V_n \ , \tag{1.16}$$

   which means that

   $$r(v) = r(v - v_n) \quad \forall v \in V,\ v_n \in V_n \ ,$$

   and plug in a suitable "interpolant" v_n for v.

2. to estimate $\left\|r\right\|_{V^*}$ from below use an appropriate "candidate function" v in the definition

   $$\left\|r\right\|_{V^*} = \sup_{v\in v\setminus\{0\}} \frac{r(v)}{\left\|v\right\|_V} \ .$$

For the model problem we have

$$\mathsf{b}(u, v) = \int_\Omega \langle \mathbf{grad}\, v, \mathbf{grad}\, u\rangle \ \mathrm{d}\boldsymbol{\xi} \quad , \quad V = H^1_0(\Omega) \quad , \quad V_n := \mathcal{S}_1(\mathcal{M}) \cap H^1_0(\Omega) \ .$$

Given the finite element solution $u_n \in V_n$, the weak residual will be

$$r(v) = \sum\nolimits_{K\in\mathcal{M}} \int_K fv - \langle \mathbf{grad}\, u_n, \mathbf{grad}\, v\rangle \,\mathrm{d}\boldsymbol{\xi} \,.$$

Next, we apply integration by parts on each element $K \in \mathcal{M}$:

$$\begin{aligned} r(v) &= \sum\nolimits_{K\in\mathcal{M}} \int_K fv \,\mathrm{d}\boldsymbol{\xi} - \int_{\partial K} \langle \mathbf{grad}\, u_n, \mathbf{n}_{\partial K}\rangle \, v \,\mathrm{d}S \\ &= \sum\nolimits_{K\in\mathcal{M}} \int_K fv \,\mathrm{d}\boldsymbol{\xi} - \sum_{\substack{F\in\mathcal{E}(\mathcal{M})\\ F\not\subset\Gamma}} \left[\langle \mathbf{grad}\, u_n, \mathbf{n}_F\rangle\right]_F \, v \,\mathrm{d}S \,, \end{aligned}$$

where $[\cdot]_F$ stands for the jump of a globally discontinuous functions across the edge F, and $\Delta u_n = 0$ on K was used.

Now, recall the quasi-interpolation operator $\mathsf{Q}_n : H^1(\Omega) \mapsto \mathcal{S}_1(\mathcal{M})$ introduced in Sect 5.3.5, see (5.48). According to Lemma 5.32, it is continuous on $H^1(\Omega)$ and a projection $H^1_0(\Omega) \mapsto V_n$. Unfortunately, Q_n is not perfectly local, which forces us to rule out meshes, for which the size of cells varies strongly over small distances:

Definition 6.28. *Given a mesh $\mathcal{M}$ its **local shape regularity measure** $\mu^{\mathrm{loc}}_{\mathcal{M}}$ is defined by*

$$\mu^{\mathrm{loc}}_{\mathcal{M}} := \max\{{}^{h_K}\!/_{h_{K'}} : \; K, K' \in \mathcal{M},\, \overline{K} \cap \overline{K'} \neq \emptyset\} \,.$$

To begin with, checking the details of the proof of Lemma 5.32, we find that, with $\gamma = \gamma(\rho_{\mathcal{M}}, \mu^{\mathrm{loc}}_{\mathcal{M}}) > 0$,

$$\|v - \mathsf{Q}_n\, v\|_{L^2(K)} \leq \gamma h_K \, |v|_{H^1(\omega_K)} \quad \forall v \in H^1(\Omega),\, K \in \mathcal{M} \,, \tag{6.20}$$

where

$$\omega_K := \bigcup \{\overline{K'} : \; K \in \mathcal{M},\, \overline{K'} \cap \overline{K} \neq \emptyset\} \,, \quad K \in \mathcal{M} \,.$$

We need a similar result for edges:

Lemma 6.29. *Let F be an edge of the triangular mesh $\mathcal{M}$ of Ω. The quasi-interpolation operator Q_n from (5.48) allows the estimate*

$$\|v - \mathsf{Q}_n\, v\|_{L^2(F)} \leq \gamma |F|^{1/2} \, |v|_{H^1(\omega_F)} \quad \forall v \in H^1(\Omega) \,,$$

with $\gamma = \gamma(\rho_{\mathcal{M}}, \mu^{\mathrm{loc}}_{\mathcal{M}}) > 0$ independent of F, and

$$\omega_F := \bigcup \{\overline{K} : \; K \in \mathcal{M},\, \overline{K} \cap F \neq \emptyset\} \,.$$

Proof. The proof is based on a scaling argument similar to that of Lemma 5.32. □
Use the representation derived above combined with Galerkin orthogonality

$$r(v) = \sum\nolimits_{K\in\mathcal{M}} \int_K f(v - \mathsf{Q}_n\, v)\,\mathrm{d}\boldsymbol{\xi} - \sum_{\substack{F\in\mathcal{E}(\mathcal{M})\\ F\not\subset\Gamma}} [\langle \mathbf{grad}\, u_n, \mathbf{n}_F\rangle]_F\ (v - \mathsf{Q}_n\, v)\,\mathrm{d}S$$

$$\leq \sum\nolimits_{K\in\mathcal{M}} \|f\|_{L^2(K)}\, \|v - \mathsf{Q}_n\, v\|_{L^2(K)} + \sum_{\substack{F\in\mathcal{E}(\mathcal{M})\\ F\not\subset\Gamma}} \|[\langle \mathbf{grad}\, u_n, \mathbf{n}_F\rangle]_F\|_{L^2(F)}\, \|v - \mathsf{Q}_n\, v\|_{L^2(F)}\ ,$$

the estimates (6.20) and of Lemma 6.29

$$\leq \gamma\Big(\sum\nolimits_{K\in\mathcal{M}} h_K\, \|f\|_{L^2(K)}\, |v|_{H^1(\omega_K)} + \sum_{\substack{F\in\mathcal{E}(\mathcal{M})\\ F\not\subset\Gamma}} |F|^{1/2}\, \|[\langle \mathbf{grad}\, u_n, \mathbf{n}_F\rangle]_F\|_{L^2(F)}\, |v|_{H^1(\omega_F)}\Big)\ ,$$

and the Cauchy-Schwarz inequality

$$\leq \gamma\ \Bigg\{ \Big(\sum\nolimits_{K\in\mathcal{M}} h_K^2\, \|f\|^2_{L^2(K)}\Big)^{1/2} \Big(\sum\nolimits_{K\in\mathcal{M}} |v|^2_{H^1(\omega_K)}\Big)^{1/2}$$

$$+ \Big(\sum_{\substack{F\in\mathcal{E}(\mathcal{M})\\ F\not\subset\Gamma}} |F|\, \|[\langle \mathbf{grad}\, u_n, \mathbf{n}_F\rangle]_F\|^2_{L^2(F)}\Big)^{1/2} \Big(\sum_{\substack{F\in\mathcal{E}(\mathcal{M})\\ F\not\subset\Gamma}} |v|^2_{H^1(\omega_K)}\Big)^{1/2}\Bigg\}$$

$$= \gamma\ \Bigg\{ \Big(\sum\nolimits_{K\in\mathcal{M}} h_K^2\, \|f\|^2_{L^2(K)}\Big)^{1/2} \Big(\sum\nolimits_{K\in\mathcal{M}} \sum\nolimits_{K'\in\omega_K} |v|^2_{H^1(K')}\Big)^{1/2}$$

$$+ \Big(\sum_{\substack{F\in\mathcal{E}(\mathcal{M})\\ F\not\subset\Gamma}} |F|\, \|[\langle \mathbf{grad}\, u_n, \mathbf{n}_F\rangle]_F\|^2_{L^2(F)}\Big)^{1/2} \Big(\sum_{\substack{F\in\mathcal{E}(\mathcal{M})\\ F\not\subset\Gamma}} \sum\nolimits_{K'\in\omega_F} |v|^2_{H^1(K')}\Big)^{1/2}\Bigg\}$$

$$\leq \gamma\ \Bigg\{ \Big(\sum\nolimits_{K\in\mathcal{M}} h_K^2\, \|f\|^2_{L^2(K)}\Big)^{1/2} + \Big(\sum_{\substack{F\in\mathcal{E}(\mathcal{M})\\ F\not\subset\Gamma}} |F|\, \|[\langle \mathbf{grad}\, u_n, \mathbf{n}_F\rangle]_F\|^2_{L^2(F)}\Big)^{1/2}\Bigg\}\ |v|_{H^1(\Omega)}\ .$$

For the final step we resorted to the finite overlap property of the neighborhoods ω_K and ω_F that results from Lemma 4.18. Summing up, we have

$$\|r\|_{H^{-1}(\Omega)} = \|r\|_{V^*} \leq \gamma\ \Bigg\{\ \Big(\sum\nolimits_{K\in\mathcal{M}} h_K^2\, \|f\|^2_{L^2(K)}\Big)^{1/2} + \Big(\sum_{\substack{F\in\mathcal{E}(\mathcal{M})\\ F\not\subset\Gamma}} |F|\, \|[\langle \mathbf{grad}\, u_n, \mathbf{n}_F\rangle]_F\|^2_{L^2(F)}\Big)^{1/2}\Bigg\}\ . \tag{6.21}$$

This motivates the following choice of the elemental error indicators

$$\big(\eta^K\big)^2 := \sum\nolimits_{K\in\mathcal{M}} \eta_K^2 + \frac{1}{2} \sum_{\substack{F\in\mathcal{E}(K)\\ F\not\subset\Gamma}} \eta_F^2\ , \tag{6.22}$$

where

$$\eta_K^2 := h_K^2 \, \|f\|_{L^2(K)}^2 \quad , \quad \eta_F^2 := h_F \, \|[\langle \mathbf{grad}\, u_n, \mathbf{n}_F \rangle]_F\|_{L^2(F)}^2 \tag{6.23}$$

The resulting **residual a posteriori error estimator** will be reliable:

Theorem 6.30. *With a constant $\gamma = \gamma(\rho_{\mathcal{M}}, \mu_{\mathcal{M}}^{\mathrm{loc}}) > 0$ the local a posteriori error estimator defined in* (6.22) *and* (6.23) *satisfies*

$$|e|_{H^1(\Omega)}^2 = |u - u_n|_{H^1(\Omega)}^2 = \|r\|_{H^{-1}(\Omega)}^2 \leq \gamma \sum\nolimits_{K \in \mathcal{M}} (\eta^K)^2 \, .$$

Proof. Straightforward from (6.21). □

Remark 6.31. For the residual error estimator the exact value of the "reliability constant" γ is unknown. Only a crude (often very pessimistic) estimate is avialable. Therefore, EST $< \tau$ cannot guarantee that the discretisation error (in H^1-norm) is really smaller than τ.

Theorem 6.30 only guarantees that a smaller tolerance τ implies a smaller discretisation error. Furthermore, the dependence of γ on the shape regularity of $\mathcal{M}$ entails preserving the shape-regularity measure of the triangles $\in \mathcal{M}$ has to be controlled during any adaptive refinement.

The next theorem asserts that the a posteriori error estimator constructed above is also efficient:

Theorem 6.32. *For all $K \in \mathcal{M}$, $F \in \mathcal{E}(\mathcal{M})$, $F \subset \Omega$, we have for the local error constributions from* (6.23)

$$\eta_K \leq \gamma \, \left(|e|_{H^1(\omega_K)} + h_K \left\| f - \overline{f}_K \right\|_{L^2(K)} \right) \, , \tag{6.24}$$

$$\eta_F \leq \gamma \, \left(|e|_{H^1(\omega_F)} + h_F \left\| f - \overline{f}_K \right\|_{L^2(\omega_F)} \right) \, , \tag{6.25}$$

with constants $\gamma = \gamma(\rho_{\mathcal{M}}, \mu_{\mathcal{M}}^{\mathrm{loc}}) > 0$, and

$$\overline{f}_K := |K|^{-1} \int_K f \, \mathrm{d}\boldsymbol{\xi} \, .$$

Proof. The main idea is to plug particular "locally supported test functions" into the weak residual. These particular test functions will be *bubble functions*, *cf.* Sect. 5.3.5.

- Element bubble functions $\beta_K = \lambda_1 \lambda_2 \lambda_3$, where λ_i, $i = 1, 2, 3$, are the barycentric coordinate functions for triangle K.
- Edge buble functions $\beta_F = b_{\mathbf{p}} b_{\mathbf{q}}$, $F \in \mathcal{E}(\mathcal{M})$, where $\mathbf{p}, \mathbf{q} \in \mathcal{N}(\mathcal{M})$ are the endpoints of the edge $F = [\mathbf{p}, \mathbf{q}]$ and $b_{\mathbf{p}}$ stands for "hat function" associated with node $\mathbf{p}$.

Note that $\beta_K \in \mathcal{P}_3(K)$ and $\beta_{F|K} \in \mathcal{P}_2(K)$ for each triangle K adjacent to the edge F. Moreover,

$$\operatorname{supp}\beta_K = \overline{K} \quad , \quad \operatorname{supp}\beta_F = \bigcup\{\overline{K} : \ K \in \mathcal{M}, \ F \subset \overline{K}\} \ .$$

First, we tackle (6.24) and take a close look at β_K. Simple scaling arguments confirm that with constants only depending on ρ_K

$$\|\beta_K\|_{L^2(K)} \approx h_K \quad , \quad |\beta_K|_{H^1(K)} \approx 1 \quad \forall K \in \mathcal{M} \ .$$

Alternatively, these estimates can be obtained by applying the formulas of Sect. 3.9.2. By the Bramble-Hilbert Lemma lemma 4.7 and the transformation techniques of Sect. 4.2, we also get

$$\exists \gamma = \gamma(\rho_K) > 0 : \quad \left\|\beta - \overline{\beta}_K\right\|_{L^2(K)} \leq \gamma h_K \quad \forall K \in \mathcal{M} \ , \tag{6.26}$$

with the mean value

$$\overline{\beta}_K := |K|^{-1} \int_K \beta_K \, \mathrm{d}\boldsymbol{\xi} \ .$$

We pick some $K \in \mathcal{M}$ and continue by observing that

$$\Delta u_n = 0 \quad , \quad \int_K (f - \overline{f}_K) \, \mathrm{d}\boldsymbol{\xi} = 0 \ .$$

This means

$$\eta_K^2 = h_K^2 \, \|f\|_{L^2(K)}^2 = h_K^2 \left(\left\|f - \overline{f}_K\right\|_{L^2(K)}^2 + \left\|\overline{f}_K\right\|_{L^2(K)}^2 \right) \ . \tag{6.27}$$

Then integration by parts introduces the residual

$$\begin{aligned} h_K^2 \left\|\overline{f}_K\right\|_{L^2(K)}^2 &= h_K^2 \overline{f}_K^2 \frac{1}{\overline{\beta}_K} \int_K \beta_K \, \mathrm{d}\boldsymbol{\xi} = h_K^2 \overline{f}_K \frac{1}{\overline{\beta}_K} \int_K \overline{f}_K \beta_K \, \mathrm{d}\boldsymbol{\xi} \\ &= \frac{h_K^2 \overline{f}_K}{\overline{\beta}_K} \left(\int_K (f + \Delta u_n) \beta_K \, \mathrm{d}\boldsymbol{\xi} + \int_K (\overline{f}_K - f) \beta_K \, \mathrm{d}\boldsymbol{\xi} \right) \\ &= \frac{h_K^2 \overline{f}_K}{\overline{\beta}_K} \left(\int_K f \beta_K - \langle \mathbf{grad}\, u_n, \mathbf{grad}\, \beta_K \rangle \, \mathrm{d}\boldsymbol{\xi} + \int_K (\overline{f}_K - f) \beta_K \, \mathrm{d}\boldsymbol{\xi} \right) \end{aligned}$$

and using the error equation $\mathsf{b}(e, v) = r(v)$ yields

$$\begin{aligned} &= \frac{h_K^2 \overline{f}_K}{\overline{\beta}_K} \left(\int_K \langle \mathbf{grad}\, e, \mathbf{grad}\, \beta_K \rangle \, \mathrm{d}\boldsymbol{\xi} + \int_K (\overline{f}_K - f) \beta_K \, \mathrm{d}\boldsymbol{\xi} \right) \\ &\leq \frac{h_K^2 \overline{f}_K}{\overline{\beta}_K} \left(|e|_{H^1(K)} \cdot |\beta_K|_{H^1(K)} + \left\|f - \overline{f}_K\right\|_{L^2(K)} \|\beta_K\|_{L^2(K)} \right) \\ &\overset{(6.26)}{\leq} \gamma \frac{h_K^2 \overline{f}_K}{\overline{\beta}_K} \left(|e|_{H^1(K)} + h_K \left\|f - \overline{f}_K\right\|_{L^2(K)} \right) \end{aligned}$$

For $\overline{f}_K \neq 0$ this implies

$$\overline{f}_K \leq \gamma \frac{1}{|K|\overline{\beta}_K} \left(|e|_{H^1(K)} + h_K \left\| f - \overline{f}_K \right\|_{L^2(K)} \right) ,$$

which can be plugged into (6.27).:

$$\begin{aligned} \eta_K^2 &\leq h_K^2 \left\| f - \overline{f}_K \right\|_{L^2(K)}^2 + \gamma \left(|e|_{H^1(K)} + h_K \left\| f - \overline{f}_K \right\|_{L^2(K)} \right)^2 \\ &\leq \gamma \left(h_K^2 \left\| f - \overline{f}_K \right\|_{L^2(K)}^2 + |e|_{H^1(K)}^2 \right) . \end{aligned}$$

All the constants merely depend on ρ_K. This proves (6.24).

To prove (6.25), we pick an edge $F \in \mathcal{E}(\mathcal{M})$ and use the edge bubble function $\beta_F \in H_0^1(\omega_F)$, which is piecewise quadratic in elements $K \subset \omega_F$ with values 0 at the boundary nodes of ω_F and $\beta_F(\boldsymbol{\mu}_F) = 1$ in the midpoint $\boldsymbol{\mu}_F$ of F. Then,

$$\int_F \beta_F \, \mathrm{d}S \leq \gamma_1 h_F \, |\beta_F|_{H^1(\omega_F)} = \gamma_1 \quad \| \operatorname{\mathbf{grad}} \beta_F \|_{L^2(\omega_F)} \tag{6.28}$$

with constants only depending on the shape regularity of $\mathcal{M}$ and the local shape regularity measure. Since $u_n \in \mathcal{S}_1(\mathcal{M})$, obviously

$$\left[\frac{\partial u_n}{\partial \mathbf{n}_F} \right] = \text{ constant on edge } F ,$$

and, for $\overline{\omega}_F = \overline{K_+ \cup K_-}$, we have

$$\begin{aligned} \left[\frac{\partial u_n}{\partial \mathbf{n}_F} \right]_{|F} &= \frac{1}{\gamma_2 \, h_F} \int_F \left[\frac{\partial u_n}{\partial \mathbf{n}_F} \right] \beta_F \, ds \\ &= \frac{1}{\gamma_2 \, h_F} \int_F \beta_F \left\langle \operatorname{\mathbf{grad}} u_n \big|_{K_+} - \operatorname{\mathbf{grad}} u_n \big|_{K_-}, \mathbf{n}_F \right\rangle \mathrm{d}S, \\ &= \frac{1}{\gamma_2 \, h_F} \int_F \beta_F \left\langle \operatorname{\mathbf{grad}} e \big|_{K_+} - \operatorname{\mathbf{grad}} e \big|_{K_-}, \mathbf{n}_F \right\rangle \mathrm{d}S, \end{aligned}$$

with $\mathbf{n}_F = \mathbf{n}_{K_+} = -\mathbf{n}_{K_-}$. Because $\beta_F \in H_0^1(\omega_F)$, it follows by integration by parts in K_+

$$\begin{aligned} \int_F \beta_F \left\langle \operatorname{\mathbf{grad}} e \big|_{K_+}, \mathbf{n}_F \right\rangle ds &= \int_{K_+} (\beta_F \Delta e + \langle \operatorname{\mathbf{grad}} \beta_F, \operatorname{\mathbf{grad}} e \rangle) \, \mathrm{d}\boldsymbol{\xi} \\ &\leq \|\beta_F\|_{L^2(K_+)} \|\Delta e\|_{L^2(K_+)} + \| \operatorname{\mathbf{grad}} \beta_F \|_{L^2(K_+)} \| \operatorname{\mathbf{grad}} e \|_{L^2(K_+)} \\ &\leq |K_+|^{1/2} \|\beta_F\|_{L^\infty(K_+)} \|f\|_{L^2(K_+)} + \gamma_3 \| \operatorname{\mathbf{grad}} e \|_{L^2(K_+)} \end{aligned}$$

since $\Delta e = \Delta u = -f$ on K_+. An analogous estimate is valid in K_-. It implies

$$\begin{aligned}\left|\left[\frac{\partial u_n}{\partial \mathbf{n}_F}\right]_{|F}\right| &\le \frac{1}{\gamma_2 h_2}\big\{|K_+|^{1/2}\|f\|_{L^2(K_+)} + |K_-|^{1/2}\|f\|_{L^2(K_-)} + 2\gamma_3\|\operatorname{\mathbf{grad}} e\|_{L^2(\omega_F)}\big\}\\ &\le \frac{\gamma_4}{\gamma_2 h_F}\big\{h_{K_+}\|f\|_{L^2(K_+)} + h_{K_-}\|f\|_{L^2(K_-)} + 2\gamma_3\|\operatorname{\mathbf{grad}} e\|_{L^2(\omega_F)}\big\}\\ &\overset{(6.23)}{\le} \frac{\gamma_4}{\gamma_2 h_F}\big\{\eta_{K_+} + \eta_{K_-} + 2\gamma_3\|\operatorname{\mathbf{grad}} e\|_{L^2(\omega_F)}\big\}.\end{aligned}$$

By the definition of the edge indicators η_F, *i. e.*

$$(\eta_F)^2 = h_F \,\|\,[\langle \mathbf{n}_F, \operatorname{\mathbf{grad}} u_n\rangle]_F\,\|^2_{L^2(F)} = h_F^2 \,\left|\,\left[\frac{\partial u_n}{\partial \mathbf{n}_F}\right]_{|F}\right|^2,$$

it follows that

$$(\eta_F)^2 \le \frac{\gamma_4^2}{\gamma_2^2}\left\{\eta_{K_+} + \eta_{K_-} + 2\gamma_3\,\|\operatorname{\mathbf{grad}} e\|_{L^2(\omega_F)}\right\}^2$$

and with (6.24) in K_+ and K_-, (6.25) follows. □

The estimates (6.24) and (6.25) mean that

$$\sum_{K\in\mathcal{M}} \eta_K^2 + \sum_{\substack{F\in\mathcal{E}(\mathcal{M})\\ F\not\subset\Gamma}} \eta_F^2 \le c\left\{\,|e|^2_{H^1(\Omega)} + \sum_{K\in\mathcal{M}} h_K^2\,\|f - \overline{f}_K\|^2_{L^2(K)}\right\}.$$

Together with the assertion of Thm. 6.30 we conclude that EST is equivalent to the H^1-norm of the discretization error up to (higher order) approximation errors of f. However, the equivalence constants depend on the shape-regularity measure of the mesh, and will, in general, be elusive.

Bibliographical notes. There are many books on a posteriori error estimation, e.g. [2, 40].

6.5 Adaptive mesh refinement

The considerations in Sect. 6.3.1 demonstrate that judicious mesh refinement can compensate the reduced convergence rate of the FEM caused by singularities of the exact solution u at the corners of Ω. The grading exponent β in the graded mesh depends only on the smallest singular exponent λ in (6.7), but not on the precise form (6.7) of the corner singularities. Nevertheless, in practice it is desirable to choose the refined meshes not a-priori, but depending on information about u that is gleaned *during the computation*. Thus, for a given specific problem the mesh can precisely be tailored to the solution u by *a posteriori adaptive mesh refinement*.

Input:
- Coefficients and right hand side for boundary value problem
- Initial triangulation $\mathcal{M}_0$ with good shape regularity measure
- Tolerance $\tau > 0$

1. Set $k := 1$.
2. Compute the finite element solution u_k on the triangulation $\mathcal{M}_0$ by assembling and solving a sparse linear system of eqations.
3. Compute the local error indicators η^K for each $K \in \mathcal{M}_k$.
4. `IF` EST $< \tau$ for EST according to (6.29), `THEN STOP`.
5. *Mark* all cells $\in \mathcal{M}_k$ for which
$$\eta^K > \theta \max\{\eta^K,\ K \in \mathcal{M}_k\} \quad \text{for some} \quad \theta \in]0,1[\ .$$
6. Create a new triangulation $\mathcal{M}_{k+1}$ by *refining* all marked cells of $\mathcal{M}_k$.
7. $k := k + 1$ and `GOTO STEP 2`.

Algorithm 6.1: Basic adaptive algorithm for the h-version of finite elements for a linear elliptic boundary value problem

6.5.1 Adaptive strategy

The one main ingredient of of adaptive mesh refinement methods is a-posteriori error estimation as explained Sect. 6.4. There we derived an efficient and reliable local a posteriori residual error estimator $\eta : \mathcal{M} \mapsto \mathbb{R}^+$ (see Def. 6.25) for the model problem (6.1) discretized by means of lowest order Lagrangian finite elements.

The local a posteriori error estimator provides elemental error indicators η^K, $K \in \mathcal{M}$, which will steer the adaptive construction of a sequence of meshes. Moreover, provided that η is reliable,

$$\text{EST} := \Big(\sum\nolimits_{K\in\mathcal{M}} (\eta^K)^2\Big)^{1/2} \tag{6.29}$$

will supply information about the norm of the discretization error. Hence, checking EST against a prescribed tolerance τ can serve as a *stopping criterion* for the adaptation loop, see Algorithm 6.1.

Remark 6.33. The selection factor θ in Algorithm 6.1 is usually chosen to be slightly less than 1, e.g., $\theta \approx 0.9$.

Other *marking strategies* are conceivable, for instance, to mark all cells for which

$$(\eta^K)^2 > \theta \frac{1}{\sharp\mathcal{M}_k} \sum\nolimits_{K\in\mathcal{M}} (\eta^K)^2 \ , \quad \theta \approx 0.9 \ .$$

The heuristics behind Algorithm 6.1 is that one should

strive for the equidistribution of the local errors.

The hope is that this will lead to the most economical way to compute the finite element solution.

The adaptive algorithm is based on the fundamental premises that

locally enhancing the resolution of the finite element space can cure local errors.

This is generally true for elliptic boundary value problems that arise from a minimization principle, see Sect. 2.6. However, whenever transport or wave propagation play a crucial role in a mathematical model, errors in one part of the computational domain can have causes in another part. In this case the policy of Algorithm 6.1 is pointless, even if the η^K are closely related to a localized error norm.

6.5.2 Algorithms

Step 6 of Algorithm 6.1 still needs to be specified in detail. We will discuss it for simplicial triangulations in two dimensions, the type of meshes for which we have presented the error estimator in Sect. 6.4.

When refining a triangulation we have to take pains that

- the new triangulation remains a *conforming* triangulation of the computational domain Ω, see Def. 3.8,
- iterating the procedure the shape regularity measures of the resulting triangulations remain below a bound that is only slightly larger than the shape-regularity measure of the intial triangulation,
- the local quasi-uniformity measures of the created meshes remain uniformly small.

Otherwise the performance of the local a posteriori error estimator can no longer be guaranteed.

In the sequel we will sketch the **red-blue-green refinement scheme** for a two-dimensional triangulation consisting of triangles that meets the above requirements.

Definition 6.34. ***Red refinement*** *splits a triangle into four congruent triangles by connecting the midpoints of the edges, see Fig. 6.10.*

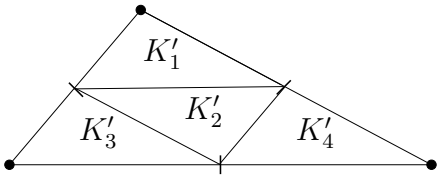


Figure 6.10: Subdivision of triangle K into 4 congruent sub-triangles K_i', $i = 1, 2, 3, 4$.

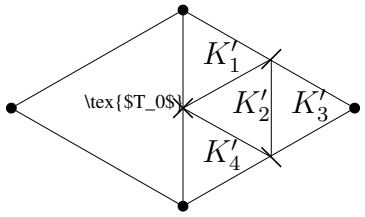


Figure 6.11: Two triangles K_0, K_1: red refinement of T_1 alone results in hanging node "×".

In order to avoid hanging nodes, types of refinement other than red refinement have to be used, *cf.* Fig. 6.11. Therefore, in two dimensions there are two more types of refinement:

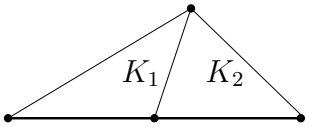


Figure 6.12: Green refinement of K_0 results in K_1, K_2.

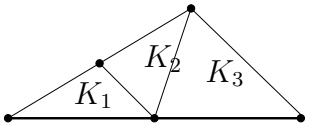


Figure 6.13: Blue refinement of K_0 spawns K_1, K_2, K_3.

Definition 6.35. ***Green refinement*** *(longest edge bisection) of a triangle splits it into two by connecting the midpoint of its longest edge with the opposite vertex.*

Definition 6.36. ***Blue refinement*** *partitions a triangle K into three sub-triangles by first carrying out green refinement, and thereafter halving one of the edges of T_0 that has not been subdivided in the green refinement by connecting its midpoint with the midpoint of the longest side of K (see Figure 6.13).*

These local refinement patterns can be combined into a mesh refinement algorithm, see Algorithm 6.2.

Theorem 6.37. *If the angles of the triangles of the initial mesh $\mathcal{M}_0$ are bounded from below by $\alpha > 0$, then all angles of all triangles of all meshes arising from repeatedly applying Algorithm 6.2 are bounded from below by $\alpha/2$.*

Proof. See [32]. □

Definition 6.38. *Two meshes $\mathcal{M}$ and $\mathcal{M}'$ are called* ***nested*** *and we write $\mathcal{M} \prec \mathcal{M}'$, if*

$$\forall K \in \mathcal{M}: \quad \exists S \subset \mathcal{M}': \quad \overline{K} = \bigcup_{K' \in S} \overline{K'} \,.$$

Input:
- Conforming simplicial triangulation $\mathcal{M}$ of polygonal computational domain $\Omega \subset \mathbb{R}^2$.
- Set $\mathcal{R} \subset \mathcal{M}$ of marked cells of $\mathcal{M}$.

1. Carry out red refinement according to Def. 6.34 of all triangles in $\mathcal{R}$. This will create a new non-conforming triangulation $\mathcal{M}^{(0)}$. Set $j := 0$.
2. Define
$$\mathcal{H} := \{K \in \mathcal{M}^{(j)} : \ K \text{ has a hanging node}\} .$$
3. Set $j := j + 1$. Create $\mathcal{M}^{(j)}$ by red refinement of all triangles of $\mathcal{M}^{(j-1)}$ that have three hanging nodes.
4. Carry out green refinement according to Def. 6.35 of all triangles in $\mathcal{M}^{(j)}$ that have a single hanging node on their longest edge $\Rightarrow \quad \mathcal{M}^{(j+1)}$. Set $j := j + 1$.
5. Carry out blue refinement according to Def. 6.36 of all triangles in $\mathcal{M}^{(j)}$ with hanging nodes. If one of those had been located on the longest edge, then no hanging nodes must remain. This yields $\mathcal{M}^{(j+1)}$. Set $j := j + 1$.
6. `IF` $\mathcal{M}^{(j)}$ is non-conforming, `THEN GOTO STEP 2`

Algorithm 6.2: Red-green-blue refinement of a simplicial triangulation in two dimensions
.

It is obvious that successive application of Algorithm 6.2 creates a nested sequence of conforming simplicial triangulations of Ω.

Theorem 6.39. *Let $\mathcal{M}, \mathcal{M}'$ be two conforming triangulations of a single computational domain Ω. Then, for any $m \in \mathbb{N}$,*
$$\mathcal{M} \prec \mathcal{M}' \quad \Rightarrow \quad \mathcal{S}_m(\mathcal{M}) \subset \mathcal{S}_m(\mathcal{M}') .$$

Proof. This is a straightforward consequence of the alternative definition of the Lagrangian finite element space on simplicial meshes, e.g.,
$$\mathcal{S}_m(\mathcal{M}) = \{v \in H^1(\Omega) : \ v_{|K} \in \mathcal{P}_m(K) \ \forall K \in \mathcal{M}\} ,$$
see Sect. 3.10, because, when the (closed) simplex K is the union of (closed) simplices $K_1, \ldots, K_L$, $L \in \mathbb{N}$, then
$$\mathcal{P}_m(K) \subset \{v \in C^0(K) : \ v_{|K} \in \mathcal{P}_m(K) \ \forall K \in S\} ,$$

where $S \subset \mathcal{M}'$ such that $\overline{K} = \bigcup_{K' \in S} \overline{K'}$. □

As a consequence of Thm. 6.39 and the quasi-optimality of the finite element solutions, adaptive refinement according to Algorithm 6.2 will, eventually, lead to a more accurate solution, see [14].

Bibliographical notes. Algorithmic details about mesh refinements in two and three dimensions can be found in [32, 4, 29].

Bibliography

[1] J. Alberty, C.Carstensen, and A. Funken, *Remarks around 50 lines of Matlab: short finite element implementation*, Numerical Algorithms, 20 (1999), pp. 117–137.

[2] W. Bangerth and R. Rannacher, *Adaptive finite element methods for differential equations*, Lectures in Mathematics, ETH Zürich, Birkhäuser, Basel, 2003.

[3] R. Bank, *PLTMG: A Software Package for Solving Elliptic Partial Differential Equations, User's Guide 6.0*, SIAM, Philadelphia, 1990.

[4] E. Bänsch, *Local mesh refinement in 2 and 3 dimensions*, IMPACT Comput. Sci. Engrg., 3 (1991), pp. 181–191.

[5] R. Barrett, M. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. V. der Vorst, *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*, SIAM, Philadelphia, PA, 2nd ed., 1994.

[6] P. Bastian, K. Birken, K. Johannsen, S. Lang, N. Neuss, H. Rentz-Reichert, and C. Wieners, *UG - A flexible software toolbox for solving partial differential equations*, Computing and Visualization in Science, 1 (1997), pp. 27–40.

[7] C. Bernardi and Y. Maday, *Spectal methods*, in Handbook of Numerical Analysis Vol. V, P. Ciarlet and J.-L. Lions, eds., North-Holland, Amsterdam, 1997, pp. 209–485.

[8] D. Braess, *Finite Elemente*, Springer–Verlag, Berlin, 1992.

[9] S. Brenner and R. Scott, *Mathematical theory of finite element methods*, Texts in Applied Mathematics, Springer–Verlag, New York, 1994.

[10] F. Brezzi and M. Fortin, *Mixed and hybrid finite element methods*, Springer, 1991.

[11] H. Cartan, *Differentialformen*, Bibliographisches Institut, Zürich, 1974.

[12] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, vol. 4 of Studies in Mathematics and its Applications, North-Holland, Amsterdam, 1978.

[13] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, vol. 4, Springer, Berlin, 1990.

[14] W. Dörfler, *A convergent adatpive algorithm for poisson's equation*, SIAN J. Numer. Anal., 33 (1996), pp. 1106–1124.

[15] O. Forster, *Analysis 3. Integralrechung im $\mathbb{R}^n$ mit Anwendungen*, Vieweg–Verlag, Wiesbaden, 3rd ed., 1984.

[16] P. Frauenfelder and C. Lage, *Concepts – An object-oriented software package for partial differential equations*, Research Report 2002-09, Seminar für Angewandte Mathematik, ETH Zürich, Zürich, Switzerland, July 2002.

[17] P. Frey and P.-L. George, *Mesh generation. Application to finite elements*, Hermes Science Publishing, Oxford, UK, 2000.

[18] V. Girault and P. Raviart, *Finite element methods for Navier–Stokes equations*, Springer, Berlin, 1986.

[19] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[20] ——, *Singularities in boundary value problems*, vol. 22 of Research Notes in Applied Mathematics, Springer–Verlag, New York, 1992.

[21] W. Hackbusch, *Theorie und Numerik elliptischer Differentialgleichungen*, B.G. Teubner–Verlag, Stuttgart, 1986.

[22] ——, *Elliptic Differential Equations. Theory and Numerical Treatment*, vol. 18 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, Heidelberg, New York, 1992.

[23] H. Heuser, *Funktionalanalysis*, Teubner–Verlag, Stuttgart, 2 ed., 1986.

[24] R. Hiptmair, *Concepts for an object oriented finite element code*, Tech. Report 335, Math.–Nat. Fakultät, Universität Augsburg, 1995.

[25] ——, *Finite elements in computational electromagnetism*, Acta Numerica, (2002), pp. 237–339.

[26] F. Hirzebruch and W. Scharlau, *Einführung in die Funktionalanalysis*, vol. 296 of BI Hochschultaschenbücher, Bibliographisches Institut, Mannheim, 1971.

[27] T. Kato, *Estimation of iterated matrices with application to von Neumann condition*, Numer. Math., 2 (1960), pp. 22–29.

[28] H. König, *Ein einfacher Beweis des Integralsatzes von Gauss*, Jahresber. Dtsch. Math.-Ver., 66 (1964), pp. 119–138.

[29] P. Leinen, *Data structures and concepts for adaptive finite element methods*, Computing, 55 (1995), pp. 325–354.

[30] S. Nazarov and B. Plamenevskii, *Elliptic Problems in Domains with Piecewise Smooth Boundaries*, vol. 13 of Expositions in Mathematics, Walter de Gruyter, Berlin, 1994.

[31] J. Nédélec, *Mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 35 (1980), pp. 315–341.

[32] M. Rivara, *Algorithms for refining triangular grids suitable for adaptive and multigrid techniques*, Int. J. Numer. Meth. Engr., 20 (1984), pp. 745–756.

[33] W. Rudin, *Functional Analysis*, McGraw–Hill, 1st ed., 1973.

[34] ——, *Real and Complex Analysis*, McGraw–Hill, 3rd ed., 1986.

[35] J. Ruppert, *A Delaunay refinement algorithm for quality 2-dimensional mesh generation*, J. Algorithms, 18 (1995), pp. 548–585.

[36] C. Schwab, *p- and hp-Finite Element Methods. Theory and Applications in Solid and Fluid Mechanics*, Numerical Mathematics and Scientific Computation, Clarendon Press, Oxford, 1998.

[37] J. Shewchuk, *Triangle: Engineering a 2D Quality Mesh Generator and Delaunay Triangulator*, in Applied Computational Geometry: Towards Geometric Engineering, M. C. Lin and D. Manocha, eds., vol. 1148 of Lecture Notes in Computer Science, Springer-Verlag, May 1996, pp. 203–222.

[38] K. Smith, *Inequalities for formally positive integro-differential forms*, Bull. Am. Math. Soc., 67 (1961), pp. 368–370.

[39] A. Stroud, *Approximate Calculation of Multiple Integrals*, Prentice-Hall Series in Automatic Computation, Prentice-Hall, Englewood Cliffs, NJ, 1971.

[40] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh–Refinement Techniques*, Wiley–Teubner, Chichester, Stuttgart, 1996.

[41] D. Werner, *Funktionalanalysis*, Springer, Berlin, 1995.

[42] H. Whitney, *Geometric Integration Theory*, Princeton University Press, Princeton, 1957.

[43] J. Wloka, *Partielle Differentialgleichungen*, Teubner–Verlag, Stuttgart, 1982.

[44] J. Xu and L. Zikatanov, *Some oberservations on Babuška and Brezzi theories*, Report AM222, PennState Department of Mathematics, College Park, PA, September 2000. To appear in Numer. Math.