1 Introduction

1.1 Terminology and classification of PDEs

A differential equation is an equation satisfied by a function $u$, which involves besides $u$ its derivates as well.

If $u$ depends only on one variable, e.g., the time $t$, we call the equation ordinary differential equation (ODE).

If it depends on several variables, e.g., on $z = (z_1, z_2, ..., z_n) \in \mathbb{R}^n$, then there occur partial derivatives

$$\partial_i u(z) := \frac{\partial u(z)}{\partial x_i},$$

and we call the equation partial differential equation (PDE).

A general second-order, linear PDE for a function $u(z)$, $z \in \mathbb{R}^n$ is

$$- \sum_{i,j=1}^{n} \partial_i a_{ij}(z) \partial_j u + \sum_{i=1}^{n} b_i(z) \partial_i u + c(z) u = f(z), \quad (1.1)$$

In case $a_{ij}(z)$, $b_i(z)$ and $c(z)$ are independent of $z$, we have a PDE with constant coefficients.

Note, that for time-dependent problems in $\mathbb{R}^d$ we have $z = (x, t)$ and $n = d + 1$ where $x \in \mathbb{R}^d$. For time-independent problems $z = x$.

If we are searching for $u$ defined in the open set $\mathcal{O} \subset \mathbb{R}^n$ then we need the regularity assumption $u \in C^2(\mathcal{O})$, $a_{ij} \in C^1(\mathcal{O})$, $b_i, c, f \in C^0(\mathcal{O})$ such that all derivatives exist in classical sense. These regularity assumptions will be reduced later when we will formulate the PDE in weak sense.

For $u \in C^2(\mathcal{O})$ the partial derivates can be switched, i.e., $\partial_i \partial_j = \partial_j \partial_i$. If $a_{ij}$ are not symmetric we can symmetrize the coefficients by

$$a_{ij}^{new} := (a_{ij}^{orig} + a_{ji}^{orig})/2,$$

and by adjusting the remaining coefficients $b_i$ such that the form of the PDE remains (1.1).

So, we can assume that the coefficient matrix $A(z) = \{a_{ij}(z)\}_{i,j=1}^{n}$ is symmetric.

PDEs are classified into

- In elliptic equations an incident at a point $z$ influences all point in their neighbourhood.

---

5
In parabolic equations there is in one direction from a point \( z \) for which the influence is only for larger values. This is often the time direction, where only later times are influenced.

In hyperbolic equations there are areas of directions around a point \( z \) which are influenced.

The principal part (Hauptteil) of the PDE

\[
- \sum_{i,j=1}^{n} \partial_i a_{ij}(z) \partial_j u
\]

is mainly responsible for their classification.

**Definition 1.1** (Classification of second-order PDEs). Consider a second-order, linear PDE of the form (1.1) with the symmetric coefficient matrix \( A(z) \).

1. The equation is said to be elliptic at \( z \in \mathcal{O} \) if the eigenvalues of \( A(z) \) are all non-zero and of same sign, so e.g., for positive-definite \( A(z) \).

2. The equation is said to be parabolic at \( z \in \mathcal{O} \) if one eigenvalue of \( A(z) \) vanishes whereas the others are all non-zero and of same sign, e.g., \( A(z) \) is positive semi-definite, but not positive definite, and the rank of \((A(z), b(z))\) is equal to \( n \).

3. The equation is said to be hyperbolic at \( z \in \mathcal{O} \) if \( A(z) \) has only non-zero eigenvalues, whereas \( n - 1 \) of one sign and one of the other sign.

A partial differential equation (1.1) is said to be elliptic, parabolic or hyperbolic in a set \( \mathcal{O} \) of points \( z \), if it has the property for all \( z \in \mathcal{O} \).

The classification covers most (linear) physical models, but is not complete.

**Remark 1.2.** Do not mix second-order elliptic PDEs with elliptic bilinear forms, which we will discuss later.

### 1.2 Partial differential equations as mathematical models

A mathematical model is the description of a system using mathematical language. Mathematical models are obtained by a combination of first principles and constitutive equations.

*First principles* are law of nature like conservative equations, actio = reactio, etc. They are based on basic assumptions.

*Constitutive equations* represent material properties and are based often on measurements. They include empirical parameters.
1.2.1 Heat transfer equation

We have as first principle the conservation of energy which correspond in absence of work to the conservation of temperature

\[
\frac{\partial u}{\partial t}(x, t) + \text{div} \, j(x, t) = f(x, t)
\]  \hspace{1cm} (1.2)

where \( u \) → temperature \quad \[ u \] = 1K
\( j \) → heat flux \quad \[ j \] = 1 \frac{W}{m^2}
\( f \) → heat source/sink \quad \[ f \] = 1 \frac{W}{m^3}

and as constitutive equation Fourier’s law

\[
j(x, t) = -A(x) \text{grad} \, u(x, t),
\]  \hspace{1cm} (1.3)

saying, that the heat flux is proportional to the temperature gradient. The matrix \( A \) represents the heat conductivity tensor (Leitfähigkeitsmatrix). We call the material

- homogen: \( A(x) = A \),
- isotrop: \( A(x) = \alpha(x) \cdot 1 \),

and inhomogen or anisotrop otherwise. \( \alpha(x) \) is the conductivity.

The heat equation is parabolic if \( A \) is positive definite. For a unique solution we have to complete the system by initial conditions

\[ u(x, 0) = u_0(x), \]

and boundary conditions. We may describe the temperature on the boundary of the domain of interest \( \Omega \)

\[ u(x, t) = g(x, t) \text{ for } x \in \partial \Omega \]

or the heat flux

\[ j(x, t) \cdot n(x) = h(x, t) \text{ for } x \in \partial \Omega. \]

Here, \( n(x) \) is the normalised outer normal vector.

1.2.2 Diffusion and Poisson equation

The static limit of the heat equation, \( i.e., \) with \( \frac{\partial u}{\partial t} = 0 \) in (1.3), we get the elliptic system

\[
j = -A(x) \text{grad} \, u,
\]

\[
\text{div} \, j = f,
\]

which has to be equipped with the above boundary conditions.

The same system arises in
• Electrostatics:  
  \( u \rightarrow \) electric potential  
  \( [u] = 1\text{V} \)

  \( j \rightarrow \) displacement current \((\mathbf{D})\)  
  \( [j] = 1\text{As} \)

  \( f \rightarrow \) charge density \((\rho)\)  
  \( [f] = 1\text{As}^{2}\text{m}^{-3} \)

Here, \( \mathbf{A} \) stands for the dielectric tensor, which is usually designated by \( \varepsilon \). The relationship \((\text{EL})\) is Gauss’ law, and \((\text{FL})\) arises from Faraday’s law \( \text{curl} \mathbf{E} = 0 \) and the linear constitutive law \( \mathbf{D} = \varepsilon \mathbf{E} \).

• Stationary electric currents:  
  \( u \rightarrow \) electric potential  
  \( [u] = 1\text{V} \)

  \( j \rightarrow \) electric current  
  \( [j] = 1\text{A} \text{m}^{-2} \)

Inside conductors, where tensor \( \mathbf{A} \) represents the conductivity. The source term \( f \) usually vanishes and excitation is solely provided by non-homogeneous boundary conditions. In this context \((\text{FL})\) arises from Ohm’s and Ampère’s circuit law and \((\text{EL})\) is a consequence of the conservation of charge.

A similar elliptic system is

\[
\mathbf{j} = -\mathbf{A}(\mathbf{x}) \text{grad} u, \quad (\text{FL})
\]

\[
\text{div} \mathbf{j} = f - c(\mathbf{x})u, \quad (\text{EL})
\]

which appear in

• Molecular diffusion:  
  \( u \rightarrow \) concentration  
  \( [u] = 1\text{mol} \text{m}^{-3} \)

  \( j \rightarrow \) flux  
  \( [j] = 1\text{mol} \text{m}^{-2} \text{s}^{-1} \)

  \( f \rightarrow \) production/consumption rate  
  \( [f] = 1\text{mol} \text{m}^{-3} \text{s}^{-1} \)

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

### 1.2.3 Wave equation

For the vertical displacement of a thin membrane can be modelled by the first principle, mass times acceleration is the force,

\[
m(\mathbf{x}) \frac{\partial^2 u}{\partial t^2}(\mathbf{x}, t) = \text{div} \mathbf{\sigma}(\mathbf{x}, t) + f(\mathbf{x}, t),
\]

and Hook’s law as constitutive equation

\[
\mathbf{\sigma}(\mathbf{x}, t) = \mathbf{A}(\mathbf{x}) \text{grad} u(\mathbf{x}, t).
\]

Here, we have  
\( u \rightarrow \) vertical displacement  
\( [u] = 1\text{m} \)

\( \mathbf{\sigma} \rightarrow \) stress vector  
\( [\mathbf{\sigma}] = 1\text{J} \)

\( f \rightarrow \) external force  
\( [f] = 1\text{N} \)

\( \text{grad} u \rightarrow \) (local) deformation of the membrane

\( \frac{\partial u}{\partial t} \rightarrow \) (local) speed of the membrane

\( \frac{\partial^2 u}{\partial t^2} \rightarrow \) (local) acceleration

and the density of mass \( m(\mathbf{x}) \) (\([m] = 1\text{kg} \text{m}^{-3}\)).
The second-order PDE is
\[ m(x) \frac{\partial^2 u}{\partial t^2}(x, t) = \text{div}(A(x) \nabla u) + f, \] (1.4)
and for homogeneous isotropic material with \( A = 1 \) we have
\[ m(x) \frac{\partial^2 u}{\partial t^2}(x, t) = \Delta u + f, \]
which is hyperbolic.

We have to complete the PDE with initial conditions
\[ u(x, 0) = u_0(x), \]
and boundary conditions. We may describe the displacement on the boundary of the membran
\[ u(x, t) = g(x, t) \text{ for } x \in \partial \Omega \]
or the boundary stress
\[ \sigma(x, t) \cdot n(x) = h(x, t) \text{ for } x \in \partial \Omega. \]

### 1.2.4 Time-harmonic wave-equation

Let \( f = \text{Re}(\hat{f}e^{-i\omega t}) \) as well as \( g = \text{Re}(\hat{g}e^{-i\omega t}), h = \text{Re}(\hat{h}e^{-i\omega t}) \) for some (angular) frequency \( \omega \in \mathbb{R}^+ \). With the independance of the coefficient function of \( t \) we can write \( u = \text{Re}(\hat{u}e^{-i\omega t}) \) and insertion into (1.4) leads to the time-harmonic wave-equation
\[ -\text{div}(A(x) \nabla \hat{u}) - \omega^2 m(x)\hat{u} = \hat{f}, \] (1.5)
which is a linear elliptic PDE of second order. The boundary condition transfer to \( \hat{u} \) where \( g \) and \( h \) are replaced by \( \hat{g} \) and \( \hat{h} \).

**Remark 1.3.** Linear parabolic or hyperbolic PDEs of second-order get elliptic in their static limit or in their time-harmonic form.
1.2.5 Further models

Further systems modelled by PDEs are given in Table 1.1.

1.3 Well-posedness

Definition 1.4 (Hadamard’s well-posedness). A problem is said to be well-posed (korrekt gestellt) if

1. it has a unique solution \( u \),
2. the solution depends continuously on the given data \( f \), i.e.,

\[
\|u\| \leq C\|f\|
\]

with a constant independant of \( u \) and \( f \).

Otherwise the problem is ill-posed.

Corollary 1.5. Small pertubations in the data of linear well-posed problems leads to small pertubations of the solution.

Proof. Let \( \tilde{f} = f + \delta f \) with \( \|\delta f\| \ll 1 \) (very small). Then \( \tilde{u} \) is the solution of the system with data \( \tilde{f} \) and \( \delta u = \tilde{u} - u \) the solution of the system with data \( \delta f \). So

\[
\|\delta u\| \leq C\|\delta f\| \ll 1.
\]

Example 1.6 (Ill-posed problem). Let \( \Omega = [0, 1] \times [0, \infty) \) and \( u \) the solution of the PDE

\[
\begin{align*}
\Delta u &= 0 \\
u(x_1, 0) &= \frac{1}{n} \sin(nx_1) \\
\partial_2 u(x_1, 0) &= 0,
\end{align*}
\]

i.e., instead of boundary conditions on all boundaries we prescibe “initial condition” on the boundary \( x_2 = 0 \). For \( n \to \infty \) the boundary data gets smaller and smaller, whereas the solution

\[
u(x_1, x_2) = \frac{1}{n} \sin(nx_1) \cosh(nx_2)
\]

explodes for a fixed \( x_2 > 0 \).
1.4 Some norms and spaces

Repeating the definition of a vector space and a norm.

**Definition 1.7 (Vector space).** A vector space over \( \mathbb{R} \) or \( \mathbb{C} \) is a set \( X \), whose elements are called vectors, for which the operations addition and scalar multiplication are defined, and for any vectors \( x, y \in X \) and \( \lambda \in \mathbb{R} \) or \( \lambda \in \mathbb{C} \) it holds

(i) \( x + y \in X \),    (ii) \( \lambda x \in X \),    (iii) \( 0 \in X \).

**Definition 1.8 (Norm).** Let \( X \) be a real (or complex) vector space. We call \( \| \cdot \| : X \to \mathbb{R} \) a norm on \( X \) if

\[
\begin{align*}
\|x\| = 0 & \iff x = 0, \quad \text{(definiteness),} & \quad \text{(N1)} \\
\|x\| \geq 0 & \quad \forall x \in X, \quad \quad \text{ (N2)} \\
\|\lambda x\| = |\lambda|\|x\| & \quad \forall \lambda \in \mathbb{R} \text{ (or } \lambda \in \mathbb{C} \text{)}, \quad \quad \text{(homogeneity),} & \quad \text{(N3)} \\
\|x + y\| \leq \|x\| + \|y\| & \quad \forall x, y \in X, \quad \quad \text{(triangle inequality).} & \quad \text{(N4)}
\end{align*}
\]
2 Finite differences scheme for second-order elliptic PDEs

2.1 The strong formulation

We consider the problem the connected bounded open set $\Omega \subset \mathbb{R}^d$ with boundary $\partial \Omega$ and the PDE with $f \in C(\Omega)$, $g \in C(\partial \Omega)$, $0 \leq c \in C(\Omega)$

$$Lu = -\Delta u + cu = f \quad \text{in} \ \Omega,$$
$$u = g \quad \text{on} \ \partial \Omega.$$  \hfill (2.1a)

Lemma 2.1 (Basic Maximum (minimum) principle for $c = 0$). Let $u \in C^2(\Omega) \cap C(\overline{\Omega})$ be solution of (2.1) with $c = 0$. If $f \leq 0$ ($f \geq 0$) in $\Omega$, then the maximum (minimum) of $u$ in $\overline{\Omega}$ is attained on the boundary $\partial \Omega$. Furthermore, if the maximum (minimum) is attained at an interior point of $\Omega$, then the function $u$ is constant.

Proof. First, we carry out the proof for the stronger assumption $f < 0$ in $\Omega$. Suppose that there exists some $\tilde{x} \in \Omega$ (a maximum point in the interior) such that

$$u(\tilde{x}) = \sup_{x \in \Omega} u(x) > \sup_{x \in \partial \Omega} u(x).$$ \hfill (2.2)

With

$$0 > f(\tilde{x}) \quad (2.1a) \Rightarrow (Lu)(\tilde{x}) = -\Delta u(\tilde{x}) = -\sum_{i=1}^{d} \frac{\partial^2 u}{\partial x_i^2}(\tilde{x})$$
which is a contradiction to (2.2), as for a maximum point for all \(1 \leq i \leq d\)
\[
\frac{\partial^2 u}{\partial x_i^2} (\bar{x}) \leq 0.
\]

Now, let \(f \leq 0\), and \(\bar{x}\) is again the maximum of \(u\) in the interior and (2.2) holds. As the \(u\) on the boundary is smaller than the maximal value we can find a sufficiently small \(\beta > 0\) such that the function
\[
w = u + \beta \sum_{i=1}^{d} (x_i - \tilde{x}_i)^2
\]
attains its maximum at an interior point \(x_0\). Since
\[
Lw = f_w = f - d\beta < f
\]
the maximum of \(w\) cannot be attained at an interior point, and we have a contradiction as well, \textit{i.e.}, (2.2) does not hold and \(u\) attains its maximum on the boundary.

Let the maximum is denoted by \(M = \max_{x \in \partial \Omega}\). There might be still points \(\tilde{x}\) in the interior with \(u(\tilde{x}) = M\). Assume that \(u\) is not constant in \(\Omega\). So there exists a ball \(B \subset \Omega\) of positive radius \(R\) and mid-point \(x_0\) with \(\sup_{x \in B} u(x) < M\) for any proper subset \(B_2 \subset B\), but with a boundary point \(\tilde{x} \in \partial B\) for which \(u(\tilde{x}) = M\). Let \(G := B \setminus B_{|x_0|, R/2}\) the open bounded ring domain with outer boundary \(\tilde{\Gamma}\) and inner boundary \(\Gamma\). Let furthermore
\[
v(x) := e^{-\lambda|x-x_0|^2} > 0
\]
with \(\lambda > 0\) large enough such that
\[
Lv = \lambda(d - \lambda|\bar{x} - \bar{x}_0|)e^{-\lambda|x-x_0|^2} < 0
\]
for all \(x \in G\). Note, that
\[
v(x) = 0 \quad \text{on } \tilde{\Gamma}, \tag{2.3}
\]
and so
\[
w_{\varepsilon} := u + \varepsilon v
\]
coincides with \(u\) on \(\tilde{\Gamma}\). Since \(\sup_{x \in \Gamma} u(x) < M\) by assumption we can choose a \(\varepsilon > 0\) so small such that
\[
w_{\varepsilon}(x) - w_{\varepsilon}(\bar{x}) = w_{\varepsilon}(x) - M < 0 \quad \text{in } G. \tag{2.4}
\]
Since \(Lw_{\varepsilon} = f + \varepsilon Lv < 0\) the maximum of \(w_{\varepsilon}\) in \(\bar{G}\) is on \(\partial G\), which is by (2.3) and (2.4) only attained in \(\tilde{x}\). Hence,
\[
\nabla w_{\varepsilon}(\bar{x}) = 0
\]

13
and as \( \text{grad} \, v(\tilde{x}) \neq 0 \) it holds

\[
\text{grad} \, u(\tilde{x}) \neq 0,
\]
i. e., \( \tilde{x} \) is not a maximum point which contradicts the assumptions. So there exists no interior maximum point if \( u \) is not constant.

For \( f \geq 0 \) we consider \( \tilde{u} := -u \) and apply the previous steps.

**Example 2.2.** Consider an open bounded set \( \Omega = (-1, 1)^2 \) and the Poisson equation

\[
-\Delta u = -4 < 0 \quad \text{in} \ \Omega,
\]

which is solved by the function \( u = u(x_1, x_2) = x_1^2 + x_2^2 \) with appropriate boundary condition. Since \( f \leq 0 \) in \( \Omega \) the solution \( u \) attains its maximum on the boundary \( \partial \Omega \). This is indeed true.

![Graph of the solution](image)

**Remark 2.3** (Non-local influence of sources). Let \( f \leq 0 \) and \( g = 0 \). As the maximum of the solution \( u \) of (2.1) is attained on the boundary, there exists no subdomain \( G \subset \Omega \) with \( |G| > 0 \) where \( u \equiv 0 \). Any point in the domain \( \Omega \) “feels” the source.

**Lemma 2.4** (Basic Maximum (minimum) principle for \( c \geq 0 \)). Let \( u \in C^2(\Omega) \cap C(\overline{\Omega}) \) be solution of (2.1) with \( c \geq 0 \). Let furthermore \( f \leq 0 \) (\( f \geq 0 \)) in \( \Omega \). Then if \( u \) has a non-negative maximum (non-positive minimum) in \( \overline{\Omega} \) is attained on the boundary \( \partial \Omega \). Furthermore, if the positive maximum (negative minimum) is attained at an interior point of \( \Omega \), then the function \( u \) is constant.

**Corollary 2.5** (Comparison principle). Let \( u, v \in C^2(\Omega) \cap C(\overline{\Omega}) \) solve the equations \( Lu = f_u \) and \( Lv = f_v \), respectively, and

\[
\begin{align*}
    f_u &\leq f_v \quad \text{in} \ \Omega, \\
    u &\leq v \quad \text{on} \ \partial \Omega.
\end{align*}
\]

Then, \( u \leq v \) in \( \Omega \).
Proof. Let \( w := v - u \) and so \( Lw = f_v - f_u \leq 0 \). If the minimum of \( w \) in \( \overline{\Omega} \) is positive, then \( w \) is positive in \( \overline{\Omega} \) and the corollary is true. If the minimum of \( w \) in \( \overline{\Omega} \) is not positive, then by the minimum principle (Lemma 2.4) it is attained on boundary. But \( w \geq 0 \) on the boundary, i.e., the minimal value is not below 0, and so \( w \geq 0 \) in \( \overline{\Omega} \). \( \square \)

**Corollary 2.6** (Uniqueness). Let \( u \) a solution of the boundary value problem (2.1). Then \( u \) is unique.

Proof. Assume that \( v \neq u \) solves (2.1). Then, \( w := v - u \) solves (2.1) with \( f \equiv 0 \) and \( g \equiv 0 \). Applying the comparison principle we conclude that \( 0 \geq w \geq 0 \) which contradicts the assumption. \( \square \)

**Definition 2.7** (Classical solution). If \( u \in C^2(\Omega) \cap C(\overline{\Omega}) \) satisfy (2.1) (or even (1.1)) in a pointwise sense, and the prescribed boundary conditions, then these functions are called a **classical solution** of the boundary value problem.

In general there does not exist a classical solution. A special case with a classical solution is that of the second-order elliptic PDE with smooth boundary and constant coefficient functions, e.g., (2.1). Here, we can make even a statement about the regularity of the solution.

**Lemma 2.8** (Elliptic shift theorem). Let \( f \in C^k(\Omega) \), \( k \in \mathbb{N}_0 \), \( g \in C^\infty(\partial\Omega) \) and \( \partial\Omega \in C^\infty \). Let \( u \) the unique solution of (2.1). Then \( u \in C^{k+2}(\Omega) \).

**Lemma 2.9** (Continuous extension theorem). Let \( u \) be the solution of (2.1) and \( u \equiv 0 \) in the open bounded subset \( G \subset \Omega \). Then, \( u \equiv 0 \) in \( \Omega \).

Proof. See [2].

### 2.2 Approximation by the Finite Difference Method (FDM)

If the solution is smooth enough we can replace the derivatives by difference quotients.

#### 2.2.1 Difference quotients

Let \( h > 0 \) and \( x \in \mathbb{R} \). Let \( I = [x - h, x + h] \), \( u \in C^{n+1} \) for some \( n \geq 0 \). Taylor’s theorem gives

\[
u(x \pm h) = u(x) \pm hu'(x) + \frac{h^2}{2} u''(x) \pm \frac{h^3}{3!} u'''(x) + \ldots + \frac{(-1)^n h^n}{n!} u^{(n)}(x) + R_n(u; x, h)
\]

with the remainder \( R_n(u; x, h) = \frac{1}{n!} \int_x^{x \pm h} (x - t)^n u^{(n+1)}(t) \, dt = \frac{(\pm h)^{n+1}}{(n+1)!} u^{(n+1)}(\xi) \) for some \( \xi \in I \).

The forward difference for \( u \in C^3(I) \) is

\[
(D^+ u)(x) := \frac{u(x + h) - u(x)}{h} = u'(x) + \frac{h}{2} u''(x) + O(h^2),
\]
whereas the **backward difference** is given by

\[
(D^- u)(x) := \frac{u(x) - u(x-h)}{h} = u'(x) - \frac{h}{2} u''(x) + O(h^2).
\]

The **central difference** for \( u \in C^5(I) \) is

\[
(D^0)(x) := \frac{u(x+h) - u(x-h)}{2h} = u'(x) + \frac{h^2}{3!} u'''(x) + O(h^4).
\]

For the second derivative \( u''(x) \) we get for \( u \in C^4(I) \):

\[
(D^+ D^- u)(x) := \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} = u''(x) + O(h^2).
\]

In two dimensions we have for \( u \in C^4([-h, h] \times [-h, h] + \mathbf{x}) \) the five-point-stencil for the Laplacian

\[
(\Delta u)(\mathbf{x}) = \frac{u(x_1 + h, x_2) + u(x_1 - h, x_2) + u(x_1, x_2 + h) + u(x_1, x_2 - h) - 4u(x_1, x_2)}{h^2} + O(h^2).
\]

Let us also specify the non-equidistant finite difference for \( u \in C^4(I) \)

\[
u''(x) = \frac{2}{h_L + h_R} \left( \frac{u(x + h_R) - u(x)}{h_R} - \frac{u(x) - u(x - h_L)}{h_L} \right) - \frac{h_R - h_L}{3} u'''(x) + O(h^2).
\]

The Laplacian can be approximated for \( u \in C^3([-h, h] \times [-h, h] + \mathbf{x}) \)

\[
(\Delta u)(\mathbf{x}) = \frac{2}{h_L + h_R} \left( \frac{u(x_1 + h_R, x_2) - u(x_1, x_2)}{h_R} - \frac{u(x_1, x_2) - u(x_1 - h_L, x_2)}{h_L} \right)
+ \frac{2}{h_B + h_T} \left( \frac{u(x_1, x_2 + h_T) - u(x_1, x_2)}{h_T} - \frac{u(x_1, x_2) - u(x_1, x_2 - h_B)}{h_B} \right) + O(h).
\]

We call the difference quotient \( \Lambda^* u \), which generalises \( \Lambda u \) (special case \( h = h_L = h_R = h_T = h_B \)).

### 2.3 The Finite Difference method

The PDE (2.1) is approximated on a uniform grid

\[
\mathcal{T}_h := \{ \mathbf{x} = (jh, ih) : \mathbf{x} \in \Omega \},
\]

whereas the boundary has the grid

\[
\mathcal{G}_h := \{ \mathbf{x} = (jh, x_2) \text{ or } \mathbf{x} = (x_1, ih) \text{ for some } x_1, x_2 \in \mathbb{R} : \mathbf{x} \in \partial \Omega \}.
\]
The union is $\mathcal{T}_h := \mathcal{T}_h \cup \mathcal{G}_h$ (see Fig. 2.1), and $|\mathcal{T}_h|$ its cardinality.

The cardinalities of $\mathcal{T}_h$ and $\mathcal{G}_h$ behave asymptotically like $|\mathcal{T}_h| \sim h^{-2}$ and $|\mathcal{G}_h| \sim h^{-1}$.

We call a grid function a function $v_h \in \mathcal{T}_h$.

We ask the finite difference approximant $u_h$ to $u$ to fulfill

\begin{align}
(L_h u_h)(x) &= f(x) \quad \text{for all } x \in \mathcal{T}_h, \quad \text{(2.6a)} \\
u_h(x) &= g(x) \quad \text{for all } x \in \mathcal{G}_h, \quad \text{(2.6b)}
\end{align}

with

\[(L_h u_h)(x) := -\Lambda^* u_h(x) + c u_h(x),\]

and $\Lambda^*$ meaning the general difference quotient for $\Delta$ with $h \geq h_L, h_R, h_T, h_B > 0$ the minimal values such that $x - (h_L,0)^\top, x + (h_R,0)^\top, x + (0,h_T)^\top, x - (0,h_B)^\top \in \mathcal{T}_h$.

**2.3.1 Existence and uniqueness**

**Lemma 2.10** (Discrete maximum principle). Let $f \leq 0$ ($f \geq 0$) on $\mathcal{T}_h$. If the maximum (minimum) of the solution $u_h$ of (2.6) is non-negative (non-positive) it is attained on $\mathcal{G}_h$.

**Proof.** Let $L_h u_h \leq 0$ for all $x \in \mathcal{T}_h$, and $\bar{x} \in \mathcal{T}_h$ such that

\[u_h(\bar{x}) = \max_{x \in \mathcal{T}_h} u_h(x)\]

and $u_h(\bar{x}) \geq 0$. 17
As a consequence of the fact that \( \tilde{x} \) is a maximum and an interior point it is

\[
\begin{align*}
\frac{2}{h_L + h_R} \left( u_h(\tilde{x}_1 + h_R, \tilde{x}_2) - u_h(\tilde{x}_1, \tilde{x}_2) - u_h(\tilde{x}_1, \tilde{x}_2 - h_L) \right) & \\
\frac{2}{h_B + h_T} \left( u_h(\tilde{x}_1, \tilde{x}_2 + h_T) - u_h(\tilde{x}_1, \tilde{x}_2) - u_h(\tilde{x}_1, \tilde{x}_2 - h_B) \right) & \\
\end{align*}
\]

\leq 0,$

and so

\[
0 \geq (L_h u_h)(\tilde{x}) = -(\Lambda^* u_h)(\tilde{x}) + c(\tilde{x}) u_h(\tilde{x}) \geq c(\tilde{x}) u_h(\tilde{x}) \geq 0
\]

So, it remains only \((L_h u_h)(\tilde{x}) = 0\), which is possible for

\[
c(\tilde{x}) = 0 \text{ and } u_h(\tilde{x}) = u_h(\tilde{x}) \quad \text{or} \quad c(\tilde{x}) > 0 \text{ and } u_h(\tilde{x}) = u_h(\tilde{x}) = 0
\]

for all points \( x \) in the stencil of \( \tilde{x} \).

If \( \tilde{x} + (h_R, 0)^T \in G_T \) the statement of the lemma is proved, otherwise we repeat the previous steps for \( \tilde{x} + (h_R, 0)^T \) instead of \( \tilde{x} \) until the (right) boundary of the domain is reached, and the lemma holds.

For \( f \geq 0 \) we consider \( \tilde{u}_h := -u_h \) and apply the previous steps.

**Corollary 2.11.** The only solution of (2.6) with \( f \equiv 0 \) on \( T_h \) and \( g \equiv 0 \) on \( G_h \) is \( u_h \equiv 0 \).

**Proof.** As \( f \leq 0 \) there can be only a non-negative maximum of \( u_h \) on the boundary. In view of the boundary conditions \( u_h \leq 0 \). Vice-versa with \( f \geq 0 \) we have \( u_h \geq 0 \) and the proof is complete.

**Corollary 2.12** (Existence and uniqueness). The system of equations (2.6) provides a unique solution \( u_h \).

**Proof.** We have shown in Corollary 2.11 that the null space of the FDM matrix is empty, and as the matrix is finite-dimensional the range is the full and a solutions exists for any \( f \) and \( g \).

**Lemma 2.13** (Discrete comparison principle). Let \( u_{h,1} \) and \( u_{h,2} \) the solutions of (2.6) with \( f = f_1, g = g_1 \) or \( f = f_2, g = g_2 \), respectively. If for the data it hold the inequalities

\[
|f_1(x)| \leq f_2(x) \quad \forall x \in T_h, \quad \quad |g_1(x)| \leq g_2(x) \quad \forall x \in G_h.
\]

Then,

\[
|u_{h,1}(x)| \leq u_{h,2}(x) \quad \forall x \in \overline{T}_h.
\]
Proof. Let $v_h := u_{h,1} - u_{h,2}$, then

$$L_h v_h = f_1 - f_2 \leq 0 \text{ on } T_h,$$

$$v_h = g_1 - g_2 \leq 0 \text{ on } G_h.$$

If there is a non-negative maximum, it is attained on $G_h$, and so $v_h(x) \leq 0$ for all $x \in T_h$. Otherwise, the maximum is negative and so even $v_h(x) < 0$ for all $x \in T_h$.

Repeating the same steps with $w_h := -u_{h,1} - u_{h,2}$ finishes the proof. \qed

### 2.3.2 Convergence

The finite difference method provides approximation on grid points only, therefore we can also only use norms on the grid to measure the error.

Grid functions are measured by the $\ell_\infty$ norm

$$\|v_h\|_\infty := \max_{x \in T_h} |v_h(x)|,$$

or the weighted $\ell_2$ norm defined by

$$\|v_h\|_{2,h} := \frac{1}{\sqrt{|T_h|}} \left( \sum_{x \in T_h} |v_h(x)|^2 \right)^{1/2} \sim h \|v_h\|_2.$$

The weight is to make the norm independent of $h$, so that we can compare different mesh widths. For example, the constant function 1 has same norm on meshes $T_{h_1}$ and $T_{h_2}$ even if $h_1 \neq h_2$. Note, that this holds also true for meshes of different domains in difference to the $L^2$-norm. We use the same norms also on $T_h$.

By the Cauchy-Schwarz inequality we have for any grid function $v_h$

$$\|v_h\|_{2,h} \leq \|v_h\|_\infty. \quad (2.7)$$

Let $v \in C^3(\Omega)$. Then, we call the grid function $(R_h v) \in \overline{T}_h$ its restriction to $T_h$.

With the consistency error we measure the accuracy if the exact solution would be inserted in the numerical scheme.

**Definition 2.14** (Consistency error). Let $u$ be the solution of (2.1). The **consistency error** is defined as

$$\|L u_h - L_h u\| = \|f - L_h u\|$$

with $\|\cdot\|$ some norm on the mesh $T_h$. The consistency error is of order $k$ if for $h \to 0$ it can be bounded by $C h^k$ for some constant $C > 0$.

**Lemma 2.15** (Estimate of the consistency error). Let $u \in C^{3+\ell}(\Omega)$ with $\ell = 0, 1$ the solution of (2.1). There exist two constants $C_\infty, C_2 > 0$ independent of $h$ such that on grids $T_h$ the consistency error can be bounded as

$$\|f - L_h u\|_\infty \leq C_\infty h,$$

$$\|f - L_h u\|_{2,h} \leq C_2 h^{1+\ell/2}.$$
Proof. Since for any \( x \in T_h \)
\[
f - L_h u = -\Delta u + \Lambda^* u,
\]
we have to bound the error of the finite difference approximation of the Laplacian.

The non-equidistant finite difference quotient \( \Lambda^* \), which is needed close to the boundary, approximates the Laplacian point-wise with \( O(h) \), and dominates so the consistency error in the \( L_\infty \) norm.

For \( u \in C^4(\Omega) \) we have on \( O(h^{-2}) \) points with the difference quotient \( \Lambda \) an point-wise error of \( O(h^2) \) and on \( O(h^{-1}) \) points with the difference quotient \( \Lambda^* \) a point-wise error of \( O(h) \) and so
\[
\|\Delta u - \Lambda^* u\|_{2,h}^2 \leq O(h^2) \cdot (O(h^{-2})O(h^4) + O(h^{-1})O(h^2)) = O(h^3).
\]
The error on the boundary dominates again. If \( u \in C^3(\Omega) \) the pointwise error is anyway \( O(h) \) and we loose half an order. \(\square\)

As the numerical scheme works with finite precision we have to consider the stability of the scheme, i.e., the influence of small pertubation of the data to the solution, as well.

**Lemma 2.16** (Stability of the FDM in the \( L_\infty \)-norm). There exists a constant \( C > 0 \) only depending on the size of \( \Omega \) such that for the solution \( u_h \) of (2.6) it holds
\[
\|u_h\|_\infty \leq \frac{\text{diam}(\Omega)^2}{4} \|f\|_\infty + \|g\|_\infty.
\]

**Proof.** Without loss of generality let \( 0 \in \Omega \) and let \( R = \text{diam}(\Omega) \). Then, for any constants \( \alpha, \beta > 0 \) the grid function
\[
v_h = \alpha(R^2 - x_1^2 - x_2^2) + \beta \quad \text{on } T_h,
\]
is positive for all \( x \in \overline{T}_h \). Inserting (2.8) into \( \Lambda^* \) we have
\[
(\Lambda^* v_h)(x) = -\Lambda^*(x_1^2 + x_2^2)
= -\alpha \left( \frac{2}{h_L + h_R} \left( \frac{(x_1 + h_R)^2 - x_1^2}{h_R^2} - \frac{x_2^2 - (x_1 - h_L)^2}{h_L^2} \right) \right.
+ \left. \frac{2}{h_T + h_B} \left( \frac{(x_2 + h_T)^2 - x_2^2}{h_T^2} - \frac{x_1^2 - (x_2 - h_B)^2}{h_B^2} \right) \right)
= -4 \alpha < 0.
\]
Thus, \( v_h \) is by Corollary 2.11 the unique solution of
\[
(L_h v_h)(x) = \alpha \left( 4 + c(x)(R^2 - x_1^2 - x_2^2) \right) + \beta =: \tilde{f} > 0 \quad \text{on } T_h,
\]
\[
v_h(x) = \alpha(R^2 - x_1^2 - x_2^2) + \beta =: \tilde{g} > 0 \quad \text{on } G_h.
\]

20
For $\alpha := \frac{1}{4}\|f\|_{\infty}$ and $\beta := \|g\|_{\infty}$ we have

$$|f(x)| \leq \tilde{f}(x) \quad \forall x \in T_h, \quad |g(x)| \leq \tilde{g}(x) \quad \forall x \in G_h.$$ 

By Lemma 2.13 we have the inequality

$$|u_h(x)| \leq v_h(x) \quad \forall x \in T_h,$$

and so

$$\|u_h\|_{\infty} \leq \|v_h\|_{\infty} \leq \alpha R^2 + \beta = \frac{R^2}{4} \|f\|_{\infty} + \|g\|_{\infty}. \quad \blacksquare$$

**Lemma 2.17** (Stability of the FDM in the weighted $\ell_2$-norm). Let $g \equiv 0$ and $T_h$ a uniform grid. There exists a constant $C > 0$ only depending on the size of $\Omega$ such that for the solution $u_h$ of (2.6) it holds

$$\|u_h\|_{2,h} \leq \frac{\text{diam}(\Omega)^2}{2\pi^2} \|f\|_{2,h}.$$ 

**Proof.** As both grid functions $u_h$ and $f$ have same length it suffices to show the lemma directly for the $\ell_2$ norm.

Since $u_h \equiv 0$ on $G_h$ by (2.6b), it remains the linear system of equations (2.6a) with $N = |T_h|$ unknowns (without values on the boundary) and $N$ equations:

$$Au_h = f,$$

where we give to the points in $T_h$ an order. The matrix $A$ is symmetric (for uniform grid $T_h$), and so its eigenvalues are real.

The eigenvalues of $A$ are in fact even positive. Assume there exists a $\lambda \leq 0$ such that

$$(A - \lambda)v_h = 0,$$

where $v_h$ is the corresponding eigenvector and with the same notation grid function. Then extending it with zero onto $G_h$ the grid function $v_h$ is unique solution of

$$(L_{h,\lambda}v_h(x)) := ((L_h - \lambda)v_h)(x) = 0 \quad \text{for all } x \in T_h,$$

$$v_h(x) = 0 \quad \text{for all } x \in G_h,$$

the with $L_{h,\lambda} = -\Lambda^* + (c-\lambda)$. By uniqueness $v_h$ vanishes, is consequently no eigenvector and we have only positive eigenvalues.

If the inverse of $A$ exists, it holds (without absolute value)

$$\|A^{-1}\|_2 = \lambda_{\max}(A^{-1}) = \max_{1 \leq k \leq N} \lambda_k(A^{-1}).$$
For the corresponding eigenvectors $v_{h,k}$ it is

$$A^{-1}v_{h,k} = \lambda_k(A^{-1})v_{h,k} \iff \lambda^{-1}_k(A^{-1})v_{h,k} = Av_{h,k},$$

i. e., $\lambda_k(A) = \lambda^{-1}_k(A^{-1})$ and so

$$\|A^{-1}\|_2 = \max_{1 \leq k \leq N} \lambda^{-1}_k(A) = \frac{1}{\lambda_{h,1}},$$

where we assume an order of the eigenvalues of $A$.

The discrete eigenvalues $\lambda_{h,k}$ are upper bounds to the eigenvalues $\lambda_k$ of the continuous problem in $\Omega$ which can be proved by the minimax principle for variational problems using piecewise-linear eigenfunctions as shown for example by Kuttler [3].

With the minimax principle it can been shown as well that the continuous eigenvalues increase if the problem is stated in a subdomain. So, the lowest discrete eigenvalue $\lambda_{h,1}$ can be bounded from below by the lowest continuous eigenvalue in the domain $\Omega$, which can be bounded from below by the lowest continuous eigenvalue in the square domain with side lengths $\text{diam}(\Omega)$, which is larger than $2\pi^2/(\text{diam}(\Omega))^2$ (the corresponding eigenfunction for $c = 0$ is $\sin(\pi x_1/R)\sin(\pi x_2/R)$).

 Remark 2.18. In case of non-uniform grids we can rely at least on the bound

$$\|u_h\|_{2,h} \leq \frac{\text{diam}(\Omega)^2}{4} \|f\|_{\infty},$$

which is consequence of Lemma 2.16 and (2.7).

 Remark 2.19 (Minimax principle). For the variational eigenvalue problem $a(u,v) = \lambda \langle u,v \rangle$ with the symmetric positive-semidefinite bilinear form $a$ the $k$-eigenvalue is

$$\lambda_k = \min_{H_k \subset H} \max_{\dim(H_k) = k} \frac{a(u,u)}{\|u\|^2}.$$ 

The finite difference method is stable and so if consistent it convergences for $h \to 0$.

 Lemma 2.20 (Convergence of the finite difference method). Let $u \in C^{3+\ell}(\Omega)$, $\ell = 0, 1$ the solution of (2.1), and $u_h$ the solution of (2.6). Then, there exists $C > 0$ such that

$$\|u_h - u\|_{\infty} \leq C h, \quad \|u_h - u\|_{2,h} \leq C h^{1+\ell/2}.$$ 

Proof. We have the systems

$$(L_h u_h)(x) = f(x) \text{ for all } x \in \mathcal{T}_h, \quad (L_h u)(x) = (L_h u)(x) \text{ for all } x \in \mathcal{T}_h,$$

$$(L_h u)(x) = g(x) \text{ for all } x \in \mathcal{G}_h, \quad u(x) = g(x) \text{ for all } x \in \mathcal{G}_h,$$

and so the grid function $u_h - u$ is unique solution of

$$(L_h(u_h - u))(x) = (f-L_h u)(x) \text{ for all } x \in \mathcal{T}_h,$$

$$(u_h - u)(x) = 0 \text{ for all } x \in \mathcal{G}_h.$$ 

22
With the stability estimates in Lemma 2.16 and 2.17, and the estimates of the consistency error we conclude in

\[\|u_h - u\|_\infty \leq C(\Omega)\|f - L_h u\|_\infty \leq C h,\]
\[\|u_h - u\|_{2,h} \leq C(\Omega)\|f - L_h u\|_{2,h} \leq C h^{1+\ell/2}.\]

\[\square\]

### 2.3.3 Implementational issues

The representation of (2.6) with separation of finite differences and boundary approximation is advantageous for the theory. However, the size of the system can be easily reduced by writing only a system for the internal unknown \(u_h\) on \(T_h\), this is by omitting (2.6b) and by moving the known values of \(u_h\) on \(G_h\) to the right hand side of (2.6a).

For example in 1D with the step sizes \(h_L, h, \ldots, h, h_R\) and \(c \equiv 0\) we have instead of the linear system

\[
\begin{pmatrix}
-2h_L^{-1} + \frac{2(h_L^{-1} + h^{-1})}{h_L + h} & \cdots & \cdots & \cdots \\
\cdots & -h^{-2} & 2h^{-2} & -h^{-2} & \cdots \\
\cdots & \cdots & -h^{-2} & 2h^{-2} & -h^{-2} & \cdots \\
1 & \cdots & \cdots & \cdots & -2h_R^{-1} + \frac{2(h_R^{-1} + h^{-1})}{h_R + h} & 1
\end{pmatrix}
\begin{pmatrix}
u_L \\
u_1 \\
\vdots \\
u_N \\
u_R
\end{pmatrix}
= \begin{pmatrix}
f_1 \\
f_2 \\
\vdots \\
f_{N-1} \\
g_L \\
g_R
\end{pmatrix}.
\]

the reduced one

\[
\begin{pmatrix}
\frac{2(h_L^{-1} + h^{-1})}{h_L + h} & -2h_L^{-1} + \frac{2(h_L^{-1} + h^{-1})}{h_L + h} & \cdots & \cdots & \cdots \\
\cdots & -h^{-2} & 2h^{-2} & -h^{-2} & \cdots \\
\cdots & \cdots & -h^{-2} & 2h^{-2} & -h^{-2} & \cdots \\
1 & \cdots & \cdots & \cdots & -2h_R^{-1} + \frac{2(h_R^{-1} + h^{-1})}{h_R + h} \\
\cdots & \cdots & \cdots & \cdots & 1
\end{pmatrix}
\begin{pmatrix}
u_1 \\
\vdots \\
u_N
\end{pmatrix}
= \begin{pmatrix}
f_1 + g_L \frac{2h_L^{-1}}{h_L + h} \\
f_2 \\
\vdots \\
f_{N-1} \\
2h_R^{-1} + \frac{2(h_R^{-1} + h^{-1})}{h_R + h}
\end{pmatrix}.
\]

The system becomes in case of uniform grids symmetric, where linear solvers for symmetric matrices, like conjugate gradients (CG), could be used.

For writing down a linear system of equations we have to order the unknowns and equations. This is in one space dimension straightforward (see last example). In two space dimensions often rectangular domains are used in practice, i.e., \(0 \leq j < N_1\), \(0 \leq i < N_2\), and global index is

\[I(i,j) = iN_1 + j.\]
More general grids (for example that in Fig. (2.1)) may we cut into layers with \( j_{\min}(i) \leq j \leq j_{\max}(i) \), \( 0 \leq i < N_2 \), and the global index is defined recursively

\[
I(0, j_{\min}(0)) = 0, \\
I(i, j_{\min}(i)) = I(i - 1, j_{\max}(i - 1)) + 1, \\
I(i, j) = I(i, j_{\min}(i)) + j - j_{\min}(i).
\]

### 2.3.4 Discussion

**Pros**

- The FDM is easy to implement.
- The system matrices are sparse with \( O(h^{-d}) \) non-zero entries out of \( O(h^{-2d}) \) entries which are even locally repeating in almost the whole domain.
- Convergent solutions in the \( \ell_\infty \) and weighted \( \ell_2 \) norm.

**Contrats**

- Limitation of the application to \( C^2 \) continuous solutions (of the strong problem).
- Limitations of the grid to simple geometries and continuous material function \( \sigma(x) \) in case of a \( \text{div} \sigma(x) \text{grad} \) operator.
- The system matrices are for non-uniform grids not symmetric.
- Only point-wise approximation.

**Extension**

- The FDM can be extended to higher orders by finite difference approximation of higher order which leads to larger stencils and so denser matrices. The FDM of higher orders convergence with a higher rate if the solution has higher regularity.

- \( C^2 \) continuous approximations can be obtained by interpolating the solution, e.g., with cubic B-splines [4] (small support) or cubic Z-splines [5] (derivatives on grid points coincides with finite differences).

A cubic B-spline basis function for equidistant points (see Fig. 2.2(a)) is given by

\[
S(x) = \begin{cases}
1 - 6x^2 - 6|x|^3 & \text{for } |x| \leq \frac{1}{2}, \\
2(1 - |x|)^3 & \text{for } \frac{1}{2} \leq |x| \leq 1, \\
0 & \text{for } 1 \leq |x|.
\end{cases}
\]

The cubic Z-spline basis function for equidistant points (see Fig. 2.2(a)) is given by

\[
Z_2(x) = \begin{cases}
1 - \frac{5}{2}x^2 + \frac{3}{2}|x|^3 & \text{for } |x| \leq 1, \\
\frac{1}{2}(2 - |x|)^2(1 - |x|) & \text{for } 1 \leq |x| \leq 2, \\
0 & \text{for } 2 \leq |x|.
\end{cases}
\]
Figure 2.2: Cubic B-spline (a) and Z-spline (b) basis functions for equidistant points.

Figure 2.3: Finite difference grids with two materials, one with a circular shape.

References

3 Elliptic Boundary Value Problems

We consider the problem the connected bounded open set $\Omega \subset \mathbb{R}^d$ with boundary $\partial \Omega$ and the PDE

$$Lu = -\text{div} \sigma \text{grad} u + cu = f \quad \text{in } \Omega, \quad (3.1a)$$
$$u = g \quad \text{on } \Gamma_D, \quad (3.1b)$$
$$\sigma \text{grad} u \cdot n = h \quad \text{on } \Gamma_N, \quad (3.1c)$$
$$\sigma \text{grad} u \cdot n + \beta u = h \quad \text{on } \Gamma_R, \quad (3.1d)$$

where $\partial \Omega = \Gamma_D \cup \Gamma_N \cup \Gamma_R$, $c \geq 0$ and $\infty > \sigma_1 \geq \sigma \geq \sigma_0 > 0$. The last boundary condition (3.1c) is the Robin boundary condition which models an resistant (absorbing) outer medium in $\mathbb{R}^d \setminus \Omega$.

3.1 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 $\Omega$. The topological closure of $\Omega$ is $\overline{\Omega}$ and its boundary $\partial \Omega := \overline{\Omega} \setminus \Omega$. A domain has an unbounded open complement $\Omega' := \mathbb{R}^d \setminus \overline{\Omega}$.

Example 3.1. For $d = 1$ admissible domains will be open intervals $]a, b[$, $a > b$, and $\partial \Omega = \{a, b\}$.

The minimum requirement for a boundary of a domain is that it is $C^0$ continuous and closed, i.e., the boundary of the boundary is empty.

However, meaningful boundary values for solutions of partial differential equations can only be imposed if we make additional assumptions on $\partial \Omega$. 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

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

Definition 3.2. A domain $\Omega \subset \mathbb{R}^d$ is called a Lipschitz domain, if there exists a finite covering of $\partial \Omega$ of open $d$-dimensional rectangles $\mathcal{U}$ such that for every $x \in \partial \Omega$ we can find an open neighborhood $U \in \mathcal{U}$ such that there is a bijective mapping

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

which satisfies
1. Both \( \Phi \) and \( \Phi^{-1} \) are Lipschitz continuous.

2. \( U \cap \partial \Omega = \Phi(\{\xi \in R : x_d = 0\}) \).

3. \( U \cap \Omega = \Phi(\{\xi \in R : \xi_d < 0\}) \).

4. \( U \cap \Omega' = \Phi(\{\xi \in R : \xi_d > 0\}) \).

We call the boundary of a Lipschitz domain a Lipschitz boundary. If \( \Phi \) can be chosen to be \( k \)-times continuously differentiable, \( k \in \mathbb{N} \), then \( \Omega \) is said to be of class \( C^k \).

Lipschitz domains have a boundary which is “slightly” smoother than only continuous. This excludes especially domains with fractal boundaries for which a finite covering of \( \partial \Omega \) is not possible. For those domains we cannot assume the statements in following of the lecture.

![Figure 3.1: Domain whose boundary is the limit of a Koch curve is not Lipschitz (bounded domain, but boundary of infinite length).](image)

There are a few examples of simple domains that do not qualify as Lipschitz domains.

![Figure 3.2: Domains that are not Lipschitz: slit domain (left), cusp domain (middle), crossing edges (right).](image)

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 \( L^\infty \).
Therefore we can be define the exterior unit vectorfield $\mathbf{n} : \partial \Omega \mapsto \mathbb{R}^d$ by

$$n(x) := \frac{\frac{\partial \Phi}{\partial \xi}(\xi)}{|\frac{\partial \Phi}{\partial \xi}(\xi)|} \text{ for almost all } x \in \partial \Omega \text{ and } \xi = \Phi^{-1}x . \quad (3.2)$$

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 (computer-aided design) software packages.

**Definition 3.3.** In the case $d = 2$ a connected domain $\Omega$ is called a **curvilinear Lipschitz polygon**, if $\Omega$ is a Lipschitz domain, and there are a finite number of open subsets $\Gamma_k \subset \partial \Omega$, $k = 1, \ldots, P$, $P \in \mathbb{N}$, such that

$$\partial \Omega := \Gamma_1 \cup \cdots \cup \Gamma_P , \quad \Gamma_k \cap \Gamma_l = \emptyset \text{ if } k \neq l ,$$

and for each $k \in \{1, \ldots, P\}$ there is a $C^1$-diffeomorphism $\Phi_k : [0, 1] \mapsto \Gamma_k$.

The boundary segments are called edges, their endpoints are the vertices of $\Omega$. A tangential direction can be defined for all points of an edge including the endpoints, which gives rise to the concept of an angle at a vertex, see Fig. 3.3. The mappings $\Phi_k$ can be regarded as smooth parametrizations of the edges. An analogous 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 3.4.** A connected domain $\Omega \subset \mathbb{R}^3$ is called a **curved Lipschitz polyhedron**, if

Figure 3.3: Curvilinear polygon with added angles at vertices
1. Ω is a Lipschitz domain,

2. there is a continuous bijective mapping \( \Phi : \partial \Pi \mapsto \partial \Omega \), where \( \Pi \) is a polyhedron with flat faces \( F_1, \ldots, F_P \), \( P \in \mathbb{N} \).

3. \( \Phi : F_k \mapsto \Phi(F_k) \) is a \( C^1 \)-diffeomorphism for every \( k = 1, \ldots, P \).

We call \( \Phi(F_k) \) the (open) face \( \Gamma_k \) of \( \Omega \). Further, \( \Phi \) takes edges and vertices of \( \Pi \) to edges and vertices of \( \Omega \).

Based on the parametrization we can introduce the surface measure \( dS \) and define integrals of measurable functions on \( \Gamma \)

\[
\int \Gamma f(x) dS(x) := \sum_{k=1}^{P} \int_{F_k = \Phi^{-1} \left( \Gamma_k \right)} f(\Phi(\xi)) \sqrt{|\det(D\Phi(\xi)D\Phi(\xi)^\top)|} d\xi , \quad (3.3)
\]

where \( d\xi \) is the \((d-1)\)-dimensional Lebesgue measure on the flat face \( F_k \).

**Definition 3.5.** 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 \( d = 3 \).

### 3.2 Linear differential operators

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

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

Set \( |\alpha| := \alpha_1 + \cdots + \alpha_n \) and denote by

\[
\partial^\alpha := \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_n}}{\partial x_n^{\alpha_n}}
\]

the partial derivative of order \( |\alpha| \). 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

\[
\text{grad } f(x) := \left( \frac{\partial f}{\partial x_1}(x), \ldots, \frac{\partial f}{\partial x_1}(x) \right)^\top , \quad x \in \Omega .
\]

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

\[
\text{div } f(x) := \sum_{k=1}^{d} \frac{\partial f_k}{\partial x_k}(x) , \quad x \in \Omega .
\]
The differential operator $\Delta := \text{div} \circ \text{grad}$ is known as \textbf{Laplacian}. In the case $d = 3$ the \textbf{rotation} of a vectorfield $f : \Omega \subset \mathbb{R}^3 \mapsto \mathbb{R}^3$ is given by
\[
\text{curl } f(x) := \begin{pmatrix}
\frac{\partial f_3}{\partial x_2}(x) - \frac{\partial f_2}{\partial x_3}(x) \\
\frac{\partial f_1}{\partial x_3}(x) - \frac{\partial f_3}{\partial x_1}(x) \\
\frac{\partial f_2}{\partial x_1}(x) - \frac{\partial f_1}{\partial x_2}(x)
\end{pmatrix}, \quad x \in \Omega.
\]

\textbf{Notation 3.6.} \textit{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 $f$ we write $f_k$ or, in order to avoid ambiguity, $(f)_k$.

### 3.3 Integration by parts

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 $n = (n_1, \ldots, n_d)^T$ for the exterior unit normal vectorfield that is defined almost everywhere on $\Gamma := \partial \Omega$, see Sect. 3.1.

\textbf{Theorem 3.7} (Gauß' theorem). If $f \in (C^1(\Omega))^d \cap (C(\overline{\Omega}))^d$, then
\[
\int_{\Omega} \text{div } f \, dx = \int_{\Gamma} \langle f, n \rangle \, dS.
\]

\textit{Proof.} Please consult [5, § 15] and [6].

By the product rule
\[
\text{div}(uf) = u \text{div } f + \langle \text{grad } u, f \rangle
\]
for $u \in C^1(\Omega)$, $f \in (C^1(\Omega))^d$, we deduce the \textbf{first Green formula}
\[
\int_{\Omega} \langle f, \text{grad } u \rangle + \text{div } f \, u \, dx = \int_{\Gamma} \langle f, n \rangle \, u \, dS \quad \text{(FGF)}
\]
for all $f \in (C^1(\Omega))^d \cap (C^0(\overline{\Omega}))^d$ and all $u \in C^1(\Omega) \cap C^0(\overline{\Omega})$. Plugging in the special $f = fe_k$, $k = 1, \ldots, d$, $e_k$ the $k$-th unit vector, we get
\[
\int_{\Omega} f \frac{\partial u}{\partial \xi_k} + \frac{\partial f}{\partial \xi_k} \, u \, dx = \int_{\Gamma} fu_n \, dS \quad \text{(IPF)}
\]
for $f, u \in C^1(\Omega) \cap C^0(\overline{\Omega})$. We may also plug $f = \text{grad } v$ into (FGF), which yields
\[
\int_{\Omega} \langle \text{grad } v, \text{grad } u \rangle + \Delta v \, u \, dx = \int_{\Gamma} \langle \text{grad } v, n \rangle \, u \, dS \quad \text{(3.4)}
\]
for all $v \in C^2(\Omega) \cap C^1(\Omega)$, $u \in C^1(\Omega) \cap C^0(\Omega)$.

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

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

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

$$\int_{\Omega} \langle \text{curl} \, u, f \rangle - \langle u, \text{curl} \, f \rangle \, dx = \int_{\Gamma} \langle u \times f, n \rangle \, dS \quad \text{(CGF)}$$

**Remark 3.8.** For $d = 1$ Gauss’ theorem boils down to the fundamental theorem of calculus, and (FGF) becomes the ordinary integration by parts formula.

### 3.4 Distributional derivatives

**Example 3.9.** Consider heat conduction in a plane wall composed of two layers of equal thickness and with heat conductivity coefficients $\kappa_1$ and $\kappa_2$. 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

$$j = -\kappa(x) \frac{d}{dx} u \quad , \quad \frac{d}{dx} j = 0 \quad , \quad u(0) = 0 \quad , \quad u(1) = u_1 \quad , \quad (3.5)$$

where

$$\kappa(x) = \begin{cases} 
\kappa_1 & \text{for } 0 < x < \frac{1}{2} , \\
\kappa_2 & \text{for } \frac{1}{2} < x < 1 .
\end{cases}$$

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

$$u(x) = \begin{cases} 
\frac{2u_1\kappa_2 x}{\kappa_1 + \kappa_2} & \text{for } 0 < x < \frac{1}{2} , \\
\frac{2u_1\kappa_1(x - 1)}{\kappa_1 + \kappa_2} + u_1 & \text{for } \frac{1}{2} < x < 1 .
\end{cases} \quad (3.6)$$

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

**Definition 3.10** (Test function space). For a non-empty open set $\Omega \subset \mathbb{R}^d$ we denote

$$C^\infty_0(\Omega) := \{ v \in C^\infty(\Omega) : \text{supp} \, v := \{ x \in \Omega : v(x) \neq 0 \} \subset \Omega \}$$

the space of test functions, which is the space of functions $C^\infty(\overline{\Omega})$ with compact support.
Lemma 3.11. Two integrable functions $f$ and $g$ defined in the bounded set $\Omega$ are almost everywhere equal if and only if
\[
\int_{\Omega} f v \, dx = \int_{\Omega} g v \, dx
\]
for all test functions $v \in C_0^\infty(\Omega)$.

We consider in the following functions to be equivalent if they are equal almost everywhere.

Definition 3.12. Let $u \in L^2(\Omega)$ and $\alpha \in \mathbb{N}_0^n$. A function $w \in L^2(\Omega)$ is called the weak derivative or distributional derivative $\partial^\alpha u$ (of order $|\alpha|$) of $u$, if
\[
\int_{\Omega} w v \, dx = (-1)^{|\alpha|} \int_{\Omega} u \partial^\alpha v \, dx \quad \forall v \in C_0^\infty(\Omega) .
\]

Based on this definition, all linear differential operators introduced in Sect. 3.2 can be given a weak/distributional interpretation. For example, the “weak” gradient $\text{grad} u$ of a function $u \in L^2(\Omega)$ will be vectorfield $w \in (L^2(\Omega))^d$ with
\[
\int_{\Omega} \langle w, v \rangle \, dx = - \int_{\Omega} u \, \text{div} v \, dx \quad \forall v \in (C_0^\infty(\Omega))^d .
\] (3.7)

This can be directly concluded from (FGF). The same is true of the “weak divergence” $w \in L^2(\Omega)$ of a vectorfield $u \in (L^2(\Omega))^d$
\[
\int_{\Omega} w v \, dx = - \int_{\Omega} \langle u, \text{grad} v \rangle \, dx \quad \forall v \in C_0^\infty(\Omega) .
\] (3.8)

The term “weak derivatives” is justified, because this concept is a genuine generalization of the classical derivative.

Theorem 3.13. If $u \in C^m(\Omega)$, then all weak derivatives of order $\leq m$ agree in $L^2(\Omega)$ with the corresponding classical derivatives.

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

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

**Lemma 3.15.** Let $\Omega \subset \mathbb{R}^d$ be bounded with Lipschitz boundary and assume a partition $\overline{\Omega} = \Omega_1 \cup \Omega_2$, $\Omega_1 \cap \Omega_2 = \emptyset$, 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 $\Omega_l$, $l = 1, 2$, belongs to $C^1(\Omega_l)$ and that $u|_{\Omega_l}$ can be extended to a function in $C^0(\overline{\Omega})$.

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

$$\frac{\partial u}{\partial x_k}(x) = \begin{cases} \frac{\partial}{\partial x_k} u|_{\Omega_1} & \text{if } x \in \Omega_1, \\ \frac{\partial}{\partial x_k} u|_{\Omega_2} & \text{if } x \in \Omega_2 \end{cases} \in L^2(\Omega). \tag{3.9}$$

**Proof.** Using locally supported test functions in the definition of the weak derivative, it is clear that (3.9) supplies the only meaningful candidate for the weak derivative of $u$. Then we appeal to (FGF) and the fact that any crossing direction $n_\Sigma$ 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

$$\int_\Omega \langle \text{grad}_{cl} u, v \rangle \, dx = \int_{\Omega_1} \langle \text{grad}_{cl} u, v \rangle \, dx + \int_{\Omega_2} \langle \text{grad}_{cl} u, v \rangle \, dx = - \int_\Omega u \text{div } v \, dx + \int_\Sigma [u]_\Sigma \langle \{v\}_\Sigma, n_\Sigma \rangle \, dS + \int_\Sigma \{u\}_\Sigma \langle [v]_\Sigma, n_\Sigma \rangle \, dS,$$

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

$$\int_\Sigma [u]_\Sigma \langle v, n_\Sigma \rangle \, dS = 0 \quad \forall v \in C^\infty_0(\Omega) \quad \Leftrightarrow \quad [u]_\Sigma = 0,$$

the assertion of the lemma follows from the definition of the weak gradient. \hfill \Box

**Example 3.16.** The weak derivative of the temperature distribution from Example 3.9 is given by (cf. Figure 3.4)

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

33
Remark 3.17. The fact that a function has classical derivatives in subdomains and even the fact that the classical derivatives can be extended to a function in $\Omega$ which is in $L^2(\Omega)$, does not guarantee that it has a weak derivative in the whole domain $\Omega$. For example a function which is piecewise $C^\infty(\Omega_l)$, $l = 1, 2$ (notation like in Lemma 3.9), but discontinuous over $\Sigma$, has not a weak gradient.

Corollary 3.18. Under the geometric assumptions of the previous lemma let $u|_{\Omega_l}$ belong to $C^m(\Omega_l)$ with possible extension to $C^{m-1}(\Omega_l)$. Then

$$\partial^\alpha u \text{ exists and } \partial^\alpha u \in L^2(\Omega) \quad \forall \alpha \in \mathbb{N}^d_0, |\alpha| \leq m \quad \Leftrightarrow \quad u \in C^{m-1}(\Omega).$$

Lemma 3.19. We retain the assumptions of Lemma 3.15 with the exception that $u$ is replaced by a vectorfield $u \in (L^2(\Omega))^d$ with restrictions $u|_{\Omega_l} \in (C^1(\Omega_l))^d$ that can be extended to continuously differentiable functions on $\overline{\Omega_l}$, $l = 1, 2$.

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

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

3.5 Weak formulations

3.5.1 Pure Dirichlet boundary conditions

Interpreting the div in (3.1a) as weak derivative we can equivalently write

$$\int_\Omega \sigma \langle \text{grad } u, \text{grad } v \rangle + c uv \, dx = \int_\Omega f v \, dx \quad (3.10)$$

for test functions $v \in C_0^\infty(\Omega)$. Additionally the solution $u$ has to fulfill the Dirichlet boundary conditions (3.1b), which all trial functions (dt. Ansatzfunktionen) – functions to try if they solve (3.10) – have to fulfill.

Boundary conditions which are present in constraint to the space of trial or test functions are called essential.

3.5.2 Neumann boundary conditions

Multiplying (3.1a) with a test function $v \in C^\infty(\Omega) \cap C(\overline{\Omega}) \subset C_0^\infty(\Omega)$ we get

$$\int_\Omega \sigma \langle \text{grad } u, \text{grad } v \rangle + c uv \, dx - \int_{\partial\Omega} \sigma \langle \text{grad } u, n \rangle \, v \, dS = \int_\Omega f v \, dx, \quad (\text{FWP})$$

and we can incorporate the Neumann boundary condition (3.1c) in the variational formulation. For pure Neumann boundary conditions, i. e., $\Gamma_N = \partial\Omega$, this reads

$$\int_\Omega \sigma \langle \text{grad } u, \text{grad } v \rangle + c uv \, dx = \int_\Omega f v \, dx + \int_{\partial\Omega} h v \, dS. \quad (3.11)$$
Boundary conditions which are present in the variational formulation are called natural.

To see that (3.12) is equivalent to (3.1a) and (3.1c) we choose first test functions \( v \in C_0^\infty(\Omega) \). Then, the term on \( \partial \Omega \) disappears, and with the definition of the weak divergence we get (3.1a). Now, let \( v \in C^\infty(\Omega) \cap C^\infty(\partial \Omega) \). Integrating by parts and using the fact that (3.1a) holds we get

\[
\int_{\partial \Omega} \sigma \langle \text{grad} u, n \rangle \ v dS = \int_{\partial \Omega} hv dS
\]

which is by Lemma 3.11 equivalent to (3.1c).

### 3.5.3 Robin boundary conditions

For the case of pure Robin boundary conditions, where Neumann b.c. are included with \( \beta = 0 \), inserting (3.1d) into (FWP) leads to

\[
\int_{\Omega} \sigma \langle \text{grad} u, \text{grad} v \rangle + cuv dx + \int_{\partial \Omega} \beta uv dS = \int_{\Omega} fv dx + \int_{\partial \Omega} hv dS. \tag{3.12}
\]

Also Robin boundary conditions are natural.

### 3.5.4 Primal formulation for the first order system

The static limit of the heat equation was given also as first order system

\[
\begin{align*}
\mathbf{j} &= -\sigma(x) \ \text{grad} \ u, \quad \text{(FL)} \\
\text{div} \ j &= f, \quad \text{(EL)}
\end{align*}
\]

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 \( \Omega \), 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.

If we cast (EL) in weak form and use the strong form of (FL) we get

\[
\begin{align*}
\int_{\Omega} \langle \mathbf{j}, \mathbf{q} \rangle dx &= -\int_{\Omega} \langle \sigma \ \text{grad} u, \mathbf{q} \rangle dx \quad \forall \mathbf{q}, \\
-\int_{\Omega} \langle \mathbf{j}, \text{grad} v \rangle + \int_{\Gamma} \langle \mathbf{j}, \mathbf{n} \rangle \ v dS &= \int_{\Omega} fv dx \quad \forall v.
\end{align*} \tag{3.13}
\]

Choosing \( \mathbf{q} = \text{grad} v \) we can merge both equations, and the result will coincide with (FWP). This formal weak formulation is called primal.
3.5.5 Dual formulation for the first order system

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

\[-\int_{\Omega} \langle \sigma^{-1} j, q \rangle + u \text{ div } q \, dx = \int_{\partial \Omega} u \langle q, n \rangle \, dS \quad \forall q, \]
\[\int_{\Omega} \text{div } j v \, dx = \int_{\Omega} f v \, dx \quad \forall v. \tag{3.14}\]

Here, elimination of \(u\) is not possible.

Note, that the Dirichlet trace of \(u\) appear in the variational formulation, i.e., here the Dirichlet boundary conditions are natural. Contrary, the trace of \(j\) is absent and Neumann boundary conditions are essential. For \(\beta > 0\) the Robin boundary conditions (3.1d) can be inserted in the variational formulation as \(u = \beta^{-1}(h - \langle j, n \rangle)\), hence, they are natural.

3.6 Linear and bilinear forms

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 3.20.** Let \(V, W\) be real vector spaces. A mapping

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

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

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

**Definition 3.21.** The dual \(V'\) of a normed vector space \(V\) is the normed vector space \(L(V, \mathbb{R})\) of continuous linear forms on \(V\). The dual space is equipped with the (operator) norm

\[\|f\|_{V'} = \sup_{0 \neq v \in V} \frac{|f(v)|}{\|v\|}.\]

**Notation 3.22.** For \(f \in V'\) we will usually write \((f, v)_{V' \times V}\) instead of \(f(v), v \in V\) (“duality pairing”). The notation \(\langle \cdot, \cdot \rangle\) is reserved for the Euclidean inner product in \(\mathbb{R}^n\) and \(|\cdot|\) will designate the Euclidean norm.

Given some linear form \(f\) a linear variational problem seeks \(u \in V\) such that

\[b(u, v) = (f, v)_{V' \times V} \quad \forall v \in V. \tag{LVP}\]
Definition 3.23. A bilinear form \( b \) on a vector space \( V \) is called symmetric, if

\[
b(v, w) = b(w, v) \quad \forall v, w \in V.
\]

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

Definition 3.24. A bilinear form \( b \) on a real vector space is positive definite, if for all \( v \in V \)

\[
b(v, v) > 0 \quad \iff \quad v \neq 0.
\]

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

Definition 3.25. A continuous bilinear form \( b \) on a normed space \( V \) is called \( V \)-elliptic with ellipticity constant \( \gamma_e > 0 \), if

\[
|b(v, v)| \geq \gamma_e \|v\|^2_V \quad \forall v \in V.
\]

3.7 Definition of Sobolev spaces

3.7.1 Banach spaces

Definition 3.26. 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 3.27. The function spaces \( L^p(\Omega), \ 1 \leq p \leq \infty \), and \( C^m(\Omega), \ m \in \mathbb{N}_0 \), are Banach spaces.

3.7.2 Hilbert spaces

For a symmetric positive definite bilinear form \( a \) an inner product \( a \) induces a norm through

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

The fundamental Cauchy-Schwarz-inequality

\[
a(v, w) \leq \|v\|_a \|w\|_a \quad \forall v, w \in V \quad \text{(CSI)}
\]

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

\[
a(v, w) = 0 \quad \iff \quad \|v\|_a^2 + \|w\|_a^2 = \|v + w\|_a^2.
\]

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

Lemma 3.29. The dual of a Hilbert space is a Hilbert space as well.

Exercise 3.1. If an inner product $a$ is $V$-elliptic and continuous in Banach space $V$ then $(V, \|\cdot\|_a)$ is a Hilbert space.

Notation 3.30. 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.

3.7.3 Sobolev spaces

In Sect. 3.5 we learned that the formal variational problem associated with the pure homogeneous Neumann problem for (3.1) is: seek $u: \Omega \mapsto \mathbb{R}$ such that

$$
\int_\Omega \langle \sigma \text{grad} u, \text{grad} v \rangle + cuv \, dx = \int_\Omega fv \, dx \quad \forall v.
$$

(3.15)

We already know that $\text{grad}$ has to be used in distributional sense. The concrete spaces have deliberately been omitted in (3.15), because we want to heed the guideline formulated in the context of Example 3.9 and set out from (3.15) and design the “ideal” space. It goes without saying that the investigation of (3.15) is easiest, when we regard a Hilbert space $H$ equipped whose inner product coincides with the bilinear form, and consequently which is equipped with the energy norm

$$
\|v\|_e^2 := \int_\Omega \langle \sigma \text{grad} v, \text{grad} v \rangle + c|v|^2 \, dx
$$

(3.16)

as its norm. So we arrive at the preliminary “definition”

$$
H := \{v: \Omega \mapsto \mathbb{R} : \text{weak grad} v \text{ exists and energy norm (3.16) of } v < \infty\}.
$$

Definition 3.31 (Sobolev space $H^1(\Omega)$). The Sobolev space $H^1(\Omega)$ is the space of square integrable functions $\Omega \to \mathbb{R}$ with square integrable weak gradients:

$$
H^1(\Omega) := \{v \in L^2(\Omega) : (\text{the weak gradient}) \text{ grad} v \text{ exists and } \text{grad} v \in L^2(\Omega)\}
$$

with norm

$$
\|v\|_{H^1(\Omega)}^2 := \|v\|_{L^2(\Omega)}^2 + D^2 \|\text{grad} v\|_{L^2(\Omega)}^2
$$

where $D = \text{diam}(\Omega)$ is introduced to match units, which is often omitted in the definition.

Note, that the $H^1(\Omega)$-norm and the energy norm are equivalent if the latter is based on an $H^1(\Omega)$-elliptic and $H^1(\Omega)$-continuous bilinear form, i.e., there exists two constants $C_1, C_2 > 0$ such that

$$
C_1\|v\|_e \leq \|v\|_{H^1(\Omega)} \leq C_2\|v\|_e \quad \forall v \in H^1(\Omega).
$$

38
**Definition 3.32.** For $m \in \mathbb{N}_0$ and $\Omega \subset \mathbb{R}^d$ we define the **Sobolev space** of order $m$ as

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

equipped with the norm

$$\|v\|_{H^m(\Omega)} := \left( \sum_{|\alpha| \leq m} \|\partial^\alpha v\|^2_{L^2(\Omega)} \right)^{\frac{1}{2}}.$$

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

**Notation 3.33.** For all $m \in \mathbb{N}_0$ and $\Omega \subset \mathbb{R}^d$

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

denotes a semi-norm on $H^m(\Omega)$.

The Sobolev spaces are a promising framework for variational problems, see [7, Thm. 3.1]:

**Theorem 3.34.** The Sobolev spaces $H^m(\Omega)$, $m \in \mathbb{N}_0$, are Hilbert spaces with the inner product

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

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

**Definition 3.35.** If $D : (C^\infty(\Omega))^l \to (C^\infty(\Omega))^k$, $l, k \in \mathbb{N}$, is a linear differential operator of order $m$, $m \in \mathbb{N}$, we write

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

where the corresponding norm on this space is given by

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

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

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

An analogue of Thm. 3.34 holds true for such spaces $H(D; \Omega)$. 

39
Example 3.36. The most important representatives of spaces covered by Def. 3.35 are

\[ H(\text{div}; \Omega) := \{ u \in (L^2(\Omega))^d : \text{div} u \in L^2(\Omega) \} , \]  

\[ H(\text{curl}; \Omega) := \{ u \in (L^2(\Omega))^3 : \text{curl} u \in (L^2(\Omega))^3 \} , \]  

\[ H(\Delta, \Omega) := \{ v \in H^1(\Omega) : \Delta u \in L^2(\Omega) \} . \]  

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

Remark 3.37. On an intersection of Hilbert spaces we use the product norm:

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

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

From now we confine \( \Omega \subset \mathbb{R}^3 \) to the class of computational domains according to Def. 3.5 from Sect. 3.1. We recall a definition from functional analysis

Definition 3.38. A subspace \( U \) of a normed space \( V \) is called dense, if

\[ \forall \epsilon > 0, \; v \in V : \; \exists u \in U : \| v - u \|_V \leq \epsilon . \]  

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 [7, Thm. 3.6]:

Theorem 3.39. The space \( C^\infty(\Omega) \) is a dense subspace of \( H^m(\Omega) \) for all \( m \in \mathbb{N}_0 \). Moreover, the space \((C^\infty(\Omega))^d\) is a dense subspace of \( H(\text{div}; \Omega) \) and \( H(\text{curl}; \Omega) \) (for \( d = 3 \) in the latter case).

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 [8, 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.

Thanks to Thm. 3.39 we can give an alternative definition of the Sobolev spaces.

Corollary 3.40. The spaces \( H^m(\Omega) \), \( H(\text{div}; \Omega) \), and \( H(\text{curl}; \Omega) \) (for \( d = 3 \)) can be obtained by the completion of spaces of smooth functions with respect to the corresponding norms.

Remark 3.41. Thm. 3.39 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

\[ A(v) \leq B(v) \quad \text{or} \quad A(v) = B(v) \]  

claimed for all functions \( u \) of a Sobolev space and involving continuous expressions \( A, B \), then it suffices to prove (3.21) for the dense subspace of smooth functions.
That means for the variational formulation, \( e.g., (3.15) \), we can replace the space of test function by a Sobolev space.

Furthermore, we define as \( H^1_0(\Omega) \) as the completion of \( C^\infty_0(\Omega) \) with respect to the \( H^1(\Omega) \)-norm.

### 3.8 The Dirichlet principle

Consider the homogeneous Neumann boundary value problem for (3.1), i.e., \( \Gamma_N = \partial \Omega \) 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 \( u \in H^1(\Omega) \) such that

\[
\int_\Omega \sigma \langle \text{grad } u, \text{grad } v \rangle + cuv \, dx = \int_\Omega f v \, dx \quad \forall v \in H^1(\Omega) .
\]

(3.22)

Evidently, its associated bilinear form

\[
a(u, v) = \int_\Omega \sigma \langle \text{grad } u, \text{grad } v \rangle + cuv \, dx
\]

is symmetric positive definite.

**Lemma 3.42.** The solution \( u \) of (3.22) with \( V = H^1(\Omega) \) can be characterised by

\[
u = \arg \min_{v \in V} J(v) \quad \text{with} \quad J(v) = \frac{1}{2} a(v, v) - \langle f, v \rangle_{V' \times V}.
\]

**Proof.** If \( u \) denotes the solution of (3.22), a simple calculation shows

\[
J(v) - J(u) = \frac{1}{2} a(v, v) - \langle f, v \rangle_{V' \times V} - \frac{1}{2} a(u, u) + \langle f, u \rangle_{V' \times V}
\]

\[
= \frac{1}{2} a(v, v) - a(u, v) - \frac{1}{2} a(u, v) + \frac{1}{2} a(u, u)
\]

\[
= \frac{1}{2} a(v - u, v - u) =: \frac{1}{2} \| v - u \|_a^2 \quad \forall v \in V .
\]

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

It is easy to establish that \( J \) is strictly convex and coercive (\( J(v) \) tends to \( +\infty \) for \( \|v\| \to \infty \)), which implies existence and uniqueness of a global minimizer \( u \). Now, consider the function

\[
h_v(\tau) := J(u + \tau v) \quad \tau \in \mathbb{R}, \ v \in V .
\]

Since \( h_v \) is smooth (it is a quadratic function) and, since \( u \) is a global minimizer

\[
\frac{d}{d\tau} h_v(\tau) \big|_{\tau=0} = 0 ,
\]

which is equivalent to (3.22), since any \( v \) can be chosen. \( \square \)
This means that a solution of (3.22) will be a global minimizer of the energy functional $J(v)$. This hints at the general fact that

\[
\text{Selfadjoint elliptic boundary value problems are closely related to minimization problems for convex functionals on a function space.}
\]

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 the second-order scalar elliptic boundary value problem (3.1). For $g \in C^0(\Gamma)$ define the affine subset of $H^1(\Omega)$

\[
H^1_{\Gamma_D,g}(\Omega) := \{ u \in H^1(\Omega) : u = g \text{ on } \Gamma_D \}.
\]

Consider the strictly convex functional $J : H^1_{\Gamma_D,g}(\Omega) \to \mathbb{R}$

\[
J(v) := \frac{1}{2} \int_\Omega \langle \sigma \nabla v, \nabla v \rangle + c |v|^2 \, dx - \int_\Omega fv \, dx + \frac{1}{2} \int_{\Gamma_R} \beta v^2 \, dS - \int_{\Gamma_N \cup \Gamma_R} hv \, dS.
\]

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

\[
\frac{d}{d\tau} J(u + \tau v) \bigg|_{\tau=0} = 0 \quad \forall v \in H^1_{\Gamma_D,0}(\Omega),
\]

which is equivalent to the linear variational problem: seek $u \in H^1_{\Gamma_D,0}(\Omega)$ such that

\[
\int_\Omega \langle \sigma \nabla u, \nabla v \rangle + cuv \, dx + \int_{\Gamma_R} \beta uv \, dS = \int_\Omega fv \, dx + \int_{\Gamma_N \cup \Gamma_R} hv \, dS \quad \forall v \in H^1_{\Gamma_D,0}(\Omega).
\]

(3.23)

Therefore, numerical methods for variational problems are also suitable for solving a certain class of optimization problems.

In Section 3.5 we have shown also the equivalence of the variational formulation and the second order elliptic boundary value problem (3.1). Thus, the variational problem (3.23) emerges as the link between the PDE and the minimization problem, see Fig. 3.5.

---

**Figure 3.5: Relationship between minimization problems, variational problems, and elliptic boundary value problems**

---

42
3.9 Theory of variational formulations

3.9.1 The Riesz representation theorem

**Theorem 3.43** (Riesz representation theorem). Let $H$ be a Hilbert space and $\varphi \in H'$ an arbitrary linear form on $H$. Then there exists a unique element $u \in H$ such that

$$\varphi(v) = (u,v)_H \quad \text{for all } v \in H.$$ 

Moreover, $\|\varphi\|_{H'} = \|u\|_H$.

The Riesz representation theorem tell that for any element in $H'$ an element in $H$ can be assigned, where the norms coincide, i.e., there is an isometric isomorphism between Hilbert spaces and its dual, “$H = H'$” – the **Riesz isomorphism**.

For variational formulations with symmetric, positive definite bilinear forms the Riesz representation theorem guarantees existence and uniqueness of the solution, and its continuous dependancy by the functional on the right in the energy norm.

If the bilinear form is elliptic and continuous in the test and trial space $H$, and with equivalence to the energy norm, the solution is also bounded in the $H$-norm itself.

$$\sqrt{\gamma} \|u\|_H \leq \frac{\|f\|_{H'}}{\sqrt{(u,u)_a}} = \frac{\sup_{v \in H} \frac{|f(v)|}{\sqrt{(v,v)_a}}}{\sup_{v \in H} \frac{|v|_{H'}}{\sqrt{(v,v)_a}}} \leq \frac{\|f\|_{H'}}{\sqrt{\gamma}}.$$ 

3.9.2 The inf-sup conditions

Recall that the dual $V'$ of $V$ includes all bounded linear forms on $V$. We can also consider the bidual of $V$, the dual $V''$ of $V'$, that is the space of functions where the duality pairing with linear forms in $V'$ is bounded. It is equipped with the operator norm

$$\|v\|_{V''} = \sup_{g \in V' \setminus \{0\}} \frac{|v(g)|}{\|g\|_{V'}} = \sup_{g \in V' \setminus \{0\}} \frac{|\langle v, g \rangle_{V' \times V'}|}{\|g\|_{V'}}.$$ 

(3.24)

Obviously $V \subset V''$. If any element $v'' \in V''$ can be identified with an element $v \in V$ such that for all $g \in V'$

$$v''(g) = \langle v'', g \rangle_{V' \times V'} = \langle g, v \rangle_{V' \times V} = g(v)$$

and $\|v''\|_{V''} = \|v\|_V$, i.e., $V$ and $V''$ are isomorphic, then $V$ is called **reflexive** and we write $V \simeq V''$.

**Example 3.44.** The space $L^p(\Omega)$ is reflexive for any $1 < p < \infty$. However, $L^\infty(\Omega)$ and $L^1(\Omega)$ are not reflexive.

**Corollary 3.45.** All Hilbert spaces are reflexive.
Proof. By Lemma 3.29 the dual $H'$ of a Hilbert space $H$ is a Hilbert again, and so $H''$ as well. By Riesz representation theorem we can identify with a $u'' \in H''$ a linear form $f \in H'$ with the same norm by

$$(f, w)_{H'} = \langle u'', w \rangle_{H'' \times H'} \quad w \in H'$$

and, applying once more, a function $u \in H$, again with the same norm.

Throughout this section $U, V$ will stand for reflexive Banach spaces with norm $\| \cdot \|_U, \| \cdot \|_V,$ and $b$ will designate a continuous bilinear form on $U, V,$ that is $b \in L(U \times V, \mathbb{R}),$ or a continuous sesquilinear form on $U, V,$ that is $b \in L(U \times V, \mathbb{C}).$

Given some $f \in W'$ a linear variational problem seeks $u \in U$ such that

$$b(u, v) = \langle f, v \rangle_{V' \times V} \quad \forall v \in V.$$ 

(LVP)

Theorem 3.46. The following statements are equivalent:

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

$$\| u_f \|_U \leq \frac{1}{\gamma_s} \| f \|_{V'},$$

with $\gamma_s > 0$ independent of $f$.

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

$$\exists \gamma_s > 0 : \inf_{w \in U \setminus \{0\}} \sup_{v \in V \setminus \{0\}} \frac{|b(w, v)|}{\| v \|_V \| w \|_U} \geq \gamma_s \quad \text{(IS1)}$$

$$\forall v \in V \setminus \{0\} : \sup_{w \in U \setminus \{0\}} |b(w, v)| > 0. \quad \text{(IS2)}$$

Proof.

1. (ii) $\Rightarrow$ (i): Injectivity (Uniqueness)

Let $u_1, u_2 \in U$ be two solutions of (LVP) for the same $f \in V'$. Then $b(u_1 - u_2, v) = 0$ for all $v \in V$, and from (IS1) we immediately infer $u_1 = u_2$.

2. (ii) $\Rightarrow$ (i): Closedness of the range

To prove existence of solutions of (LVP) we define the following subspace of $V'$, which correspond to the to $b$ associate operator:

$$V'_b := \{ g \in V' : \exists w \in U : b(w, v) = \langle g, v \rangle_{V' \times V} \quad \forall v \in V \}.$$

Let $\{ g_k \}_{k=1}^\infty$ be a Cauchy-sequence in $V'_b$. Since $V'$ is a Banach space and hence complete, $g_k$ will converge to some $g \in V'$. By definition

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

(3.26)
Thanks to the inf-sup condition (IS1), we have for any \( k, m \in \mathbb{N} \)
\[
\|w_k - w_m\|_U \leq \frac{1}{\gamma_s} \sup_{v \in V \setminus \{0\}} \frac{|\langle g_k - g_m, v \rangle_{V' \times V}|}{\|v\|_V} = \frac{1}{\gamma_s} \|g_k - g_m\|_{V'}.
\]
Hence, \( \{w_k\}_{k=1}^\infty \) is a Cauchy-sequence, too, and will converge to some \( w \in U \). The continuity of \( b \) and of the duality pairing makes it possible to pass to the limit on both sides of (3.26)
\[
b(w, v) = \langle g, v \rangle_{V' \times V} \quad \forall v \in V,
\]
which reveals that \( g \in V'_b \). Since \( g_k \in V'_b \) has its limit in \( V'_b \), it is a closed subspace of \( V' \).

\( \text{(ii) } \Rightarrow \text{(i): Surjectivity (Existence)} \)
Now, assume that \( V'_b \neq V' \). As \( V'_b \subset V' \) is closed, a corollary of the Hahn-Banach theorem, see [9, Satz 4.1], confirms the existence of \( v_0 \in V'' \simeq V \) (\( V \) reflexive!) such that
\[
0 = v_0(g) = g(v_0) = \langle g, v_0 \rangle_{V' \times V} \quad \forall g \in V'_b.
\]
By definition of \( V'_b \) this means \( b(w, v_0) = 0 \) for all \( w \in U \) and contradicts (IS2).

\( \text{(i) } \Rightarrow \text{(ii): Stability estimate} \)
Fix some \( w \in V \) and denote by \( g_w \in V' \) the continuous functional \( v \mapsto b(u, v) \), i.e., \( w \) is the unique solution of
\[
b(w, v) = \langle g_w, v \rangle_{V' \times V} \quad \forall v \in V
\]
from which we conclude (IS1) by (3.25)
\[
\|w\|_V \leq \frac{1}{\gamma_s} \|g_w\|_{V'} = \frac{1}{\gamma_s} \sup_{v \in V \setminus \{0\}} \|b(w, v)\|_V.
\]
Now, fix some \( v \in V \setminus \{0\} \). Then,
\[
0 < \|v\|_{V''} = \sup_{g \in V' \setminus \{0\}} \frac{|\langle v, g \rangle_{V'' \times V'}|}{\|g\|_{V'}} \quad \text{and so} \quad 0 < \sup_{g \in V' \setminus \{0\}} |\langle v, g \rangle_{V'' \times V'}|,
\]
By the reflexivity of \( V \) we have
\[
0 < \sup_{g \in V' \setminus \{0\}} |\langle g, v \rangle_{V' \times V}| = \sup_{g \in V' \setminus \{0\}} |b(w, v)|,
\]
where \( w \) is the unique solution of \( b(w, v) = \langle g, v \rangle_{V' \times V} \), \( \forall v \in V \), and (IS2) follows.
Definition 3.47 (Ellipticity of sesquilinear forms). A sesquilinear form $b(\cdot, \cdot)$ is called $V$-elliptic, if there exists a $\gamma > 0$ and a $\sigma \in \mathbb{C}$ with $|\sigma| = 1$, such that for all $v \in V$
\[ \text{Re}(\sigma b(v, v)) \geq \gamma \|v\|^2_V. \]

Lemma 3.48 (Lax-Milgram). Let $V$ be a reflexive Banach space. Let the bilinear form $b : V \times V \to \mathbb{R}$ or sesquilinear form $b : V \times V \to \mathbb{C}$ be $V$-elliptic (see Def. 3.25 and Def. 3.47). Then, the variational problem (LVP) has for any $f \in V'$ a unique solution $u \in V$ with
\[ \|u\|_V \leq \frac{1}{\gamma} \|f\|_{V'}. \]

Proof. The $V$-ellipticity of the sesquilinear form $b$ implies for any $v \in V$
\[ \gamma \|v\|^2_V \leq \text{Re}(\sigma b(v, v)) \leq |\sigma b(v, v)| = |b(v, v)|, \tag{3.28} \]
which holds directly for the bilinear form $b$. Then, (IS1) holds as
\[ \sup_{w \in V \setminus \{0\}} \frac{|b(w, v)|}{\|v\|_V} \geq \frac{|b(w, w)|}{\|w\|_V} \geq \gamma \|w\|_V. \]
Furthermore, any $v \in V$
\[ \sup_{w \in V \setminus \{0\}} |b(w, v)| \geq |b(v, v)| \geq \gamma \|v\|^2_V > 0. \]
and (IS2) holds. Applying Thm. 3.46 the proof is complete.

3.10 Wellposedness for variational problem of second order elliptic PDEs

3.10.1 Boundary value problems with Dirichlet data

In Sec. 3.8 we have derived a variational problem for the second order BVP (3.1).
Strictly speaking, (3.23) 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, (3.23) can easily be converted into the form (LVP) by using an extension $u_g \in H^1(\Omega)$ of the Dirichlet data, that is, $u_g = g$ on $\Gamma_D$, and plugging $u := u_g + \delta u$ into (3.23), where, now, the offset $\delta u \in H^1_{\Gamma_D, 0}(\Omega)$ assumes the role of the unknown function and $u_g$ will show up in an extra contribution to the right hand side functional.
This throws into question:
- In which we ask for the equality $u_g = g$ ?
- For which functions $g$ such an extension exists ?
- How can we prescribe the zero trace on $\Gamma_D$ in $H^1_{\Gamma_D, 0}(\Omega)$ ?
3.10.2 Traces

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 \( \partial \Omega \) 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 \( L^2(\partial \Omega) \), such that

\[
(R_m u)(x) = u(x) \quad \forall x \in \partial \Omega, \forall u \in C^\infty(\overline{\Omega}) .
\]

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

**Example 3.49.** For \( u \in L^2(\Omega) \) a continuous trace operator cannot be defined. In particular, a trace inequality of the form

\[
\exists \gamma_t > 0 : \| u|_{\partial \Omega} \|_{L^2(\partial \Omega)} \leq \gamma_t \| u \|_{L^2(\Omega)} \quad \forall u \in C^\infty(\overline{\Omega}) \quad (3.29)
\]

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

\[
v(x) := \begin{cases} 
0 & \text{if } h \leq x_1 \leq 1, 0 \leq x_2 \leq 1 , \\
1 - \frac{x_1}{h} & \text{if } 0 \leq x_1 \leq h, 0 \leq x_2 \leq 1 .
\end{cases}
\]

Then we can compute

\[
1 = \int_0^1 |v(0, x_2)|^2 \, dx_2 \leq \| v|_{\partial \Omega} \|_{L^2(\partial \Omega)}^2 ,
\]

and

\[
\| v \|_{L^2(\Omega)}^2 = \int_0^1 \left( 1 - \frac{x_1}{h} \right) \, dx_1 \int_0^1 \left( 1 - y \right)^2 \, dy = \frac{h^3}{3} .
\]

If (3.29) 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 3.50.** The trace operator \( R_1 \) is continuous from \( H^1(\Omega) \) into \( L^2(\partial \Omega) \), that is,

\[
\exists \gamma_t > 0 : \| u|_{\partial \Omega} \|_{L^2(\partial \Omega)} \leq \gamma_t \| u \|_{H^1(\Omega)} \quad \forall u \in C^\infty(\overline{\Omega}) .
\]

More precisely, the following multiplicative trace inequality holds true:

\[
\exists \gamma(\Omega) > 0 : \| R_1 u \|_{L^2(\partial \Omega)} \leq \gamma(\Omega) \| u \|_{L^2(\Omega)} \left\{ \| u \|_{L^2(\Omega)} + \| \text{grad } u \|_{L^2(\Omega)} \right\} \quad \forall u \in H^1(\Omega) .
\]

**Notation 3.51.** Writing \( \gamma(\Omega) \) we hint that the "constant" \( \gamma \) may only depend on the domain \( \Omega \).
Notation 3.52. The trace operator $R_1$ is often suppressed in expressions like $\int_\Gamma v \ldots dS$, when it is clear that the restriction of the function $v \in H^1(\Omega)$ to a part of the boundary $\Gamma$ is used.

If $\Gamma$ denotes a part of $\partial \Omega$ with positive measure, we can restrict $R_1 u, u \in H^1(\Omega)$, to $\Gamma$, write $R_\Gamma$ for the resulting operator, and trivially have the continuity
\[ \exists \gamma > 0 : \|R_\Gamma u\|_{L^2(\Gamma_0)} \leq \gamma \|u\|_{H^1(\Omega)} \quad \forall u \in H^1(\Omega). \]

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

**Definition 3.53.** For $m \in \mathbb{N}$ and any part $\Gamma$ of the boundary $\partial \Omega$ of $\Omega$ with $|\Gamma| > 0$ we define
\[ H^1_m(\Omega) := \{ v \in H^1(\Omega) : R_\Gamma(\partial^\alpha v) = 0 \text{ in } L^2(\Gamma), \forall \alpha \in \mathbb{N}_0^d, |\alpha| < m \}. \]

If $\Gamma = \partial \Omega$ we write $H^1_m(\Omega) = H^1_m(\Omega)$.

Obviously, the spaces $H^1_m(\Omega)$ are closed subspaces of $H^1(\Omega)$. Another important density result holds true, see [7, Thm. 3.7].

**Theorem 3.54.** The functions in $C^\infty(\overline{\Omega})$ whose support does not intersect $\Gamma$ form a dense subspace of $H^1_m(\Omega)$, $m \in \mathbb{N}$. In particular, $C^\infty(\Omega)$ is a dense subspace of $H^1_m(\partial \Omega)$.

By Thm. 3.50 the trace operator $R_1$ maps continuously into $L^2(\Gamma)$. This raises the issue, whether it is also onto. The answer is negative.

The question is, how we can characterize the range of the trace operator $R_1$. Let $\Gamma$ temporarily stand for a connected component of the boundary of $\Omega$. We start by introducing a norm
\[ \|v\|_{H^{1/2}(\Gamma)} = \inf \|w\|_{H^1(\Omega)} : w \in C^\infty(\overline{\Omega}), w|_{\Gamma} = v \] (3.30)
on the space of restrictions of smooth functions to $\Gamma$. It is highly desirable that this norm is *intrinsic* to $\Gamma$, that is, switching to another domain $\tilde{\Omega}$, for which $\Gamma$ is also a connected component of the boundary, and using (3.30) produces an equivalent norm.

**Definition 3.55.** 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 [7, § 3].

**Theorem 3.56.** The space $H^{1/2}(\Gamma)$ is a Hilbert space and can be equipped with the (equivalent) Sobolev-Slobodeckij-norm
\[ \|v\|^2_{H^{1/2}(\Gamma)} := \int_{\Gamma} \int_{\Gamma} \frac{|v(x) - v(y)|^2}{|x - y|^d} dS(x)dS(y). \]

**Theorem 3.57.** The trace operator $R_1 : H^1(\Omega) \mapsto H^{1/2}(\Gamma)$ is continuous and surjective and has a bounded right inverse $F_1 : H^{1/2}(\Gamma) \mapsto H^1(\Omega)$, ie., $R_1 \circ F_1 = Id$. 

48
3.10.3 Dual spaces

By definition $H^1(\Omega) \subset L^2(\Omega)$ and so for $f \in L^2(\Omega)$ by the Cauchy-Schwarz-inequality the linear form $H^1(\Omega) \to \mathbb{R}$

$$ v \mapsto \int_{\Omega} f v \, dx $$

is bounded in $(H^1(\Omega))^\prime$. The same holds for subspaces $H_{Γ,D,0}^1(\Omega)$ or $H_0^1(\Omega)$.

It is easy to verify that

$$ \| u \|_{H^{-1}(\Omega)} := \sup \{ v \in H_0^1(\Omega) \setminus \{0\} : \|v\|_{H^1(\Omega)} \} \int_{\Omega} u v \, dx $$

(3.31)

defines a norm on $L^2(\Omega)$.

**Definition 3.58.** The completion of $L^2(\Omega)$ with respect to the norm given by (3.31) is called $H^{-1}(\Omega)$.

**Lemma 3.59.** The space $H^{-1}(\Omega)$ is a Hilbert space, which is isometrically isomorphic to $(H_0^1(\Omega))^\prime$.

**Remark 3.60.** By construction we have the continuous and dense embeddings

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

Sometimes, such an arrangement is called a Gelfand triple.

**Notation 3.61.** Often the integral $\int_{\Omega} \ldots \, dx$ is used to denote the duality pairing of $H^{-1}(\Omega)$ and $H_0^1(\Omega)$.

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 $\int_{\Gamma} \ldots \, dS$ is often used to indicate the corresponding duality pairing.

Dual spaces play a crucial role when it comes to defining traces of vectorfields.

**Lemma 3.62.** The normal components trace $R_n$ for $u \in (C^\infty(\Omega))^d$, defined by $R_n u(x) := \langle u(x), n(x) \rangle$ for all $x \in \partial \Omega$, can be extended to a continuous and surjective operator $R_n : H(\text{div}; \Omega) \to H^{-1/2}(\partial \Omega)$.

**Proof.** Pick some $u \in (C^\infty(\Omega))^d$. By (FGF) and the Cauchy-Schwarz inequality in $L^2(\Omega)$ we find

$$ \| \langle u, n \rangle \|_{H^{-1/2}(\Gamma)} = \sup_{v \in H^{1/2}(\Gamma) \setminus \{0\}} \frac{\int_{\Gamma} \langle u, n \rangle v \, dS}{\| v \|_{H^{1/2}(\Gamma)}} $$

$$ = \sup_{v \in H^{1/2}(\Gamma) \setminus \{0\}} \frac{1}{\| v \|_{H^{1/2}(\Gamma)}} \int_{\Omega} \langle u, \text{grad} F_1 v \rangle + \text{div} u \, F_1 v \, dx $$

$$ \leq \frac{\| F_1 v \|_{H^1(\Omega)}}{\| v \|_{H^{1/2}(\Gamma)}} \| u \|_{H(\text{div}; \Omega)} \leq \frac{\| F_1 \|_{H^{1/2}(\Gamma) \to H^1(\Omega)}}{\| u \|_{H(\text{div}; \Omega)}} \| u \|_{H(\text{div}; \Omega)}, $$

49
where $\tilde{v} := F_1 v$, which means that $R_1 \tilde{v} = v$, see Thm. 3.57. Applying the principle explained in Remark 3.41 shows the continuity of $R_n$ asserted in the Lemma.

To confirm that $R_n$ is onto $H^{-1/2}(\Gamma)$, we rely on the symmetric positive definite linear variational problem: seek $u \in H^1(\Omega)$ such that

$$\int_{\Omega} \langle \nabla u, \nabla v \rangle + uv \, dx = \int_{\Gamma} h_1 \, v \, dS \quad \forall v \in H^1(\Omega).$$

(3.32)

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. ?? the variational problem (3.32) has a unique solution in $H^1(\Omega)$.

Testing with $v \in C^\infty_0(\Omega)$ and recalling the definition of weak derivatives, see Def. 3.12, we find that

$$-\text{div}(\nabla u) + u = 0 \quad \text{in } L^2(\Omega).$$

Note that div has to be understood as differential operator in the sense of distributions according to (3.8). This shows div$(\nabla u) \in L^2(\Omega)$, which allows to apply (FGF) to (3.32):

$$\int_{\Omega} \left( -\text{div}(\nabla u) + u \right) v \, dx + \int_{\Gamma} \langle \nabla u, n \rangle R_1 v \, dS = \int_{\Gamma} h_1 \, v \, dS$$

for all $v \in C^\infty(\Omega)$. By a density argument, this amounts to $h = \langle \nabla u, n \rangle$ in $H^{-1/2}(\Gamma)$.

Moreover, the trace theorems for $H^1(\Omega)$ and $H(\text{div}; \Omega)$ enable us to apply the argument elaborated in Remark 3.41 to (FGF). Hence this integration by parts formula is seen to hold for all $f \in H(\text{div}; \Omega)$ and $u \in H^1(\Omega)$:

$$\int_{\Omega} \langle f, \nabla u \rangle + \text{div} f \, u \, dx = \int_{\Gamma} \langle f, n \rangle \, u \, dS \quad \forall f \in H(\text{div}; \Omega), u \in H^1(\Omega).$$

(FGF)

Lemma 3.63. The tangential components trace $R_t$ for $u \in (C^\infty(\Omega))^3$, defined by $R_t \, u(x) := u(x) \times n(x)$ for all $x \in \partial \Omega$, can be extended to a continuous and surjective operator $R_t : H(\text{curl}; \Omega) \mapsto (H^{-1/2}(\partial \Omega))^3$, where $(H^{-1/2}(\Gamma))^3$ is understood as the dual of $(H^{1/2}(\Gamma))^3$.

3.10.4 Second order elliptic PDE with $c \geq c_0 > 0$

The previous discussions give sense to the variational formulation: Seek $u \in H^1_{1,D,0}(\Omega)$ such that

$$\int_{\Omega} \langle \sigma \nabla u, \nabla v \rangle + c uv \, dx + \int_{\Gamma_R} \beta uv \, dS = \int_{\Omega} f v \, dx + \int_{\Gamma_N \cup \Gamma_R} hv \, dS \quad \forall v \in H^1_{1,D,0}(\Omega).$$

(3.23)
• For $g \in H^{1/2}(\Gamma_D)$ an extension $u_g \in H^1(\Omega)$ exists, and homogeneous or inhomogeneous Dirichlet trace is prescribed in the spaces $H^{1}_{\Gamma_D,0}(\Omega)$ and $H^{1}_{\Gamma_D,g}(\Omega) = H^{1}_{\Gamma_D,0}(\Omega) + u_g$ are well-defined.

• For $h \in H^{-1/2}(\Gamma_N \cup \Gamma_R)$ the respective linear form is the duality pairing with $H^{-1/2}(\Gamma_N \cup \Gamma_R)$ and hence continuous.

• For $f \in L^2(\Omega)$ the respective linear form is continuous as $L^2(\Omega) \subseteq (H^1_{\Gamma_D,0}(\Omega))'$.

In case of $c \geq c_0 > 0$ the bilinear form is $H^1(\Omega)$-elliptic (and so also for subspaces like $H^1_{\Gamma_D,0}(\Omega)$) with an ellipticity constant

$$\gamma = \min(\inf_{x \in \Omega} \sigma(x), \inf_{x \in \Omega} c(x)),$$

and well-posedness of (3.23) follow from Lax-Milgram’s lemma with

$$\|u\|_{H^1(\Omega)} \leq \frac{1}{\gamma} \left( \|f\|_{L^2(\Omega)} + \|g\|_{H^{1/2}(\Gamma_D)} + \|h\|_{H^{-1/2}(\Gamma_N \cup \Gamma_R)} \right).$$

\[\Box\] What is in case of $c = 0$? We have “ellipticity” for the $H^1(\Omega)$-seminorm, but the control of the $L^2(\Omega)$-norm is missing.

$$b(u, u) \geq \gamma_1 |u|_{H^1(\Omega)}^2 \geq \gamma_2 \left( |u|_{H^1(\Omega)}^2 + \|u\|_{L^2(\Omega)}^2 \right)$$

### 3.10.5 The inequalities of Poincaré and Friedrich

**Lemma 3.64 (The Friedrich inequality).** Let $\Omega$ be a open bounded domain and $|\Gamma_D| > 0$. Then, for all $u \in H^1_{\Gamma_D,0}(\Omega)$ there exists a constant $C_F(\Omega, \Gamma_D)$ such that

$$\|u\|_{L^2(\Omega)}^2 \leq C_F(\Omega, \Gamma_D) \text{diam}(\Omega)^2 \|\text{grad} u\|_{L^2(\Omega)}^2.$$ 

\[\Box\] Well-posedness in case of Dirichlet boundary conditions (at least at part of the boundary) and $c \geq 0$.

For pure Neumann boundary conditions and $c = 0$, that is

$$\text{div} \, j = - \text{div} \, \sigma \, \text{grad} u = f \quad \text{in} \, \Omega,$$

$$j \cdot n = \sigma \partial_n u = h \quad \text{on} \, \partial \Omega,$$

the solution can only be fixed up to an additive constant. Using exactly the constant function as test function in the associate variational formulation we get a **compatibility conditions**

$$\int_{\Omega} f \, dx = \int_{\partial \Omega} h \, dS,$$
the data of the problem has to fulfill. One can see this also by applying Gauss’ theorem applied to \( j \).

The constant can be fixed by demanding a vanishing mean value of \( u \), i.e., we switch to the subspace

\[
H^1_*(\Omega) := \{ v \in H^1(\Omega) : \int_{\Omega} v \, dx = 0 \}
\]

and we can apply

**Lemma 3.65** (The Poincaré inequality). Let \( \Omega \) be a open bounded domain. Then, for all \( u \in H^1(\Omega) \) there exists a constant \( C_P(\Omega) \) such that

\[
\| u - u_\Omega \|_{L^2(\Omega)}^2 \leq C_P(\Omega) \text{diam}(\Omega)^2 \| \text{grad} \, u \|_{L^2(\Omega)}^2,
\]

where \( u_\Omega = \frac{1}{|\Omega|} \int_{\Omega} u(x) \, dx \) is the mean of \( u \) in \( \Omega \).

**Remark 3.66.** One is tempted to ensure uniqueness of solutions of (3.33) by demanding \( u(x_0) = 0 \) for some \( x_0 \in \Omega \). However, this approach is flawed (fehlerhaft), because the mapping \( u \mapsto u(x_0) \) is unbounded on \( H^1(\Omega) \) (there are unbounded functions in \( H^1(\Omega) \) for \( d > 1 \)).

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.

In case of Robin boundary conditions the constant is not in the kernel of the formulation, as adding a constant would harm the Robin boundary condition. This is addressed by the following lemma.

**Lemma 3.67.** Let \( f : H^1(\Omega) \to \mathbb{R} \) be a continuous functional which fixes constants and for which we have \( |f(\lambda v)| = |\lambda|^2 |f(v)| \) for \( \lambda \in \mathbb{R} \), i.e., a constant function \( c \) vanishes if \( f(c) = 0 \). Then, it exists a constant \( C = C(f, \Omega) > 0 \) such that for all \( u \in H^1(\Omega) \) we have

\[
\| u \|_{L^2(\Omega)}^2 \leq C \left( \| \text{grad} \, u \|_{L^2(\Omega)}^2 + |f(u)| \right).
\]

**Proof.** Assume that the assertion of the lemma was false. Then, we can find a sequence \( \{ u_n \}_{n=1}^\infty \), \( u_n \in H^1(\Omega) \) such that for all \( n \in \mathbb{N} \)

\[
\left\{ \| \text{grad} \, u_n \|_{L^2(\Omega)}^2 + |f(u_n)| \right\} \leq \frac{1}{n} \| u_n \|_{L^2(\Omega)}^2 \leq \frac{1}{n} \| u_n \|_{H^1(\Omega)}^2.
\]

As the functionals on the left and right side are quadratic, we can consider instead the associated normalised sequence with \( \| u_n \|_{H^1(\Omega)} = 1 \).
As the sequence \( u_n \) is bounded, it exists a subsequence weakly convergent in \( H^1(\Omega) \) [10, Appendix A.3.8], i.e., for \( k \to \infty \)

\[ u_{n_k} \rightharpoonup u \text{ in } H^1(\Omega) \quad \text{which means} \quad (u_{n_k}, v)_{H^1(\Omega)} \to (u, v)_{H^1(\Omega)} \quad \forall v \in H^1(\Omega), \]

where the limit is denoted by \( u \).

By continuity of the functional \( \|\text{grad} v\|_{L^2(\Omega)}^2 + |f(v)| : H^1(\Omega) \to \mathbb{R} \) we can take the limit on both sides of (3.34), and it holds

\[ \|\text{grad} u\|_{L^2(\Omega)}^2 + |f(u)| = 0. \tag{3.35} \]

This implies that \( \text{grad} u \equiv 0 \) and \( u \) is a constant. Since \( f(u) = 0 \) the constant is zero, and \( u \equiv 0 \).

For bounded domains \( H^1(\Omega) \) is compactly embedded in \( L^2(\Omega) \), which mean that for any weakly convergent sequence in \( H^1(\Omega) \) there exists a subsequence strongly convergent in \( L^2(\Omega) \) (Rellich-Kondrachov theorem, [11, Chapter 2]). Denoting the subsequence again \( u_{n_k} \) it holds for \( k \to \infty \)

\[ \|u_{n_k}\|_{L^2(\Omega)} \to \|u\|_{L^2(\Omega)} = 0. \]

This mean that the subsequence \( u_{n_k} \), we have \( \|u_{n_k}\|_{H^1(\Omega)} = 1, \|\text{grad} u_{n_k}\|_{L^2(\Omega)} \to 0 \) and \( \|u_{n_k}\|_{L^2(\Omega)} \to 0 \), which is not possible. \( \square \)

So, also for the case \( c = 0 \) the bilinear form of the variational problem (3.23) is \( H^1_{1, D, 0}(\Omega) \)-elliptic.

### 3.11 Discrete variational formulations

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 \( b \in L(V \times V, \mathbb{R}) \) we pursue a Galerkin discretization of (LVP). Its gist (wesentliche) is to replace \( V \) in (LVP) by finite dimensional subspaces. The most general approach relies on two subspaces of \( V \):

\[
W_n \subset V : \quad \text{“trial space”,} \quad \dim W_n = N \\
V_n \subset V : \quad \text{“test space”,} \quad \dim V_n = N, \quad N \in \mathbb{N}.
\]

**Notation 3.68.** 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.
Figure 3.6: b-Orthogonality of the error $e_n = u - u_n$ with respect to $V_n$, if $b$ is an inner product.

Given the two spaces $W_n$ and $V_n$ and some $f \in V'$ the \textbf{discrete variational problem} corresponding to (LVP) reads: seek $u_n \in W_n$ such that
\[ b(u_n, v_n) = \langle f, v_n \rangle_{V', V} \quad \forall v_n \in V_n. \tag{DVP} \]

This most general approach, where $W_n \neq V_n$ is admitted, is often referred to as \textbf{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 \textbf{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
\[ b(u - u_n, v_n) = 0 \quad \forall v_n \in V_n. \tag{3.36} \]

Abusing terminology, this property is called \textbf{Galerkin orthogonality}, though the term orthogonality is only appropriate, if $b$ is an inner product on $V$. Sloppily speaking, the \textbf{discretization error} $e_n := u - u_n$ is “orthogonal” to the test space $V_n$, see Fig. 3.6.

\textbf{Theorem 3.69.} Let $V$ be a Banach space and $b \in L(V \times V, \mathbb{R})$ satisfy the inf-sup conditions (IS1) and (IS2) from Thm. 3.46. Further, assume that
\[ \exists \gamma_n > 0 : \inf_{w_n \in W_n} \sup_{v_n \in V_n \backslash \{0\}} \frac{|b(w_n, v_n)|}{\|v_n\|_V \|w_n\|_V} \geq \gamma_n. \tag{DIS} \]

Then, for every $f \in V' \subset V_n'$ the discrete variational problem (DVP) has a unique solution $u_n$ that satisfies the stability estimate
\[ \|u_n\|_V \leq \frac{1}{\gamma_n} \|f\|_{V'} = \frac{1}{\gamma_n} \sup_{v_n \in V_n} \frac{|f(v_n)|}{\|v_n\|_V}, \tag{3.37} \]

and the \textbf{quasi-optimality} estimate
\[ \|u - u_n\|_V \leq \left(1 + \frac{\|b\|_{V \times V \rightarrow \mathbb{R}}}{\gamma_n}\right) \inf_{w_n \in W_n} \|u - w_n\|_V, \tag{3.38} \]

where $u \in V$ solves (LVP).
Proof. Following 9 in the proof of Thm. 3.46 it is clear that (DIS) implies the uniqueness of \( u_n \), and similarly to 9 the stability estimate (3.37) follows. Since \( \dim V_n = \dim W_n \), in the finite dimensional setting this implies existence of \( u_n \) (cf. [10, Lemma A.9]).

It remains to show (3.38). For any \( w_n \in W_n \) we can estimate using the triangle inequality and (3.37):

\[
\| u - u_n \|_V \leq \| u - w_n \|_V + \| w_n - u_n \|_V \\
\leq \| u - w_n \|_V + \frac{1}{\gamma_n} \sup_{v_n \in V \setminus \{0\}} \frac{|b(w_n - u + u - u_n, v_n)|}{\| v_n \|_V} \\
\leq \| u - w_n \|_V + \frac{1}{\gamma_n} \sup_{v_n \in V \setminus \{0\}} \frac{|b(w_n, v_n)|}{\| v_n \|_V} + \frac{1}{\gamma_n} \sup_{v_n \in V \setminus \{0\}} \frac{|b(u - u_n, v_n)|}{\| v_n \|_V}.
\]

The last term vanishes due to Galerkin orthogonality (3.36) and with the continuity of \( b \) we obtain (3.37).

Remark 3.70. One can not conclude (DIS) from (IS1) and (IS2) because the supremum is taken over a much smaller set.

Remark 3.71. It goes without saying that for a \( V \)-elliptic bilinear form \( b \), cf. Def. 3.25, the assumptions of Thm. 3.69 are trivially satisfied. Moreover, we can choose \( \gamma_n \) equal to the ellipticity constant \( \gamma_e \) in this case.

Thm. 3.69 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 \( f \).

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. 3.69, the mere fact that the norm of \( V = H \) arises from an inner product, permits us to get a stronger a-priori error estimate [12].

Theorem 3.72. If \( V \) is a Hilbert space we obtain the sharper a-priori error estimate

\[
\| u - u_n \|_V \leq \frac{\| b \|_{V \times V \rightarrow \mathbb{R}}}{\gamma_n} \inf_{v_n \in V_n} \| u - v_n \|_V,
\]

(3.39)

if the assumptions of Thm. 3.69 are satisfied.

Theorem 3.73 (Céa’s Lemma). If the bilinear form \( b \) is \( V \)-elliptic the estimate

\[
\| u - u_n \|_V \leq \frac{\| b \|_{V \times V \rightarrow \mathbb{R}}}{\gamma_e} \inf_{v_n \in V_n} \| u - v_n \|_V,
\]

(3.40)

holds.

55
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 \( V_n \), i.e., the element in \( V_N \) nearest to the exact solution \( u \) in the energy norm.

**Corollary 3.74.** 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 orthogonality (3.36) and the Cauchy-Schwarz inequality

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

for any \( v_n \in V_n \). \( \square \)

**Remark 3.75.** 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 \( \gamma_n \) is bounded away from zero uniformly in \( n \). This will guarantee asymptotic quasi-optimality of the Galerkin solution: the estimate (3.37) will hold with a constant independent of \( n \). Notice that the norm of \( b \) that also enters (3.38) 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. 3.69. The “Galerkin orthogonality” (3.36) suggests that we examine the so-called **Galerkin projection** \( P_n : V \mapsto W_n \) defined by

\[
b(P_n w, v_n) = b(w, v_n) \quad \forall v_n \in V_n .
\]

**Proposition 3.76.** Under the assumption of Thm. 3.69, the equation (3.41) defines a continuous projection \( P_n : V \mapsto V \) with norm \( \| P_n \|_{V \mapsto V} \leq \gamma_n^{-1} \| b \|_{V \times V \mapsto \mathbb{R}} \).

**Proof.** As a consequence of Thm. 3.69, \( P_n \) is well defined. Its linearity is straightforward and the norm bound can be inferred from (3.37). Also \( P_n^2 = P_n \) is immediate from the definition. \( \square \)

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

\[
u_n = P_n u .
\]

**Definition 3.77.** 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 \( \operatorname{Ker}(P) \perp \operatorname{Range}(P) \).
Proposition 3.78. If \( b \in L(V \times V, \mathbb{R}) \) is an inner product on \( V \) and the assumptions of Thm. 3.69 are satisfied, then the Galerkin projection associated with \( b \) is an orthogonal projection with respect to the inner product \( b \).

Proof. According to Def. 3.77 and the previous proposition, we only have to check that
\[
\text{Ker}(P_n) \perp W_n = \text{Range}(P_n).
\]
As \( P_n \) is a projector there is a decomposition of \( V \) in a direct sum [10, Lemma A.38]
\[
V = \text{Range}(P_n) \oplus \text{Range}(Id - P_n),
\]
\( i.e. \), \( \text{Range}(P_n) \cap \text{Range}(Id - P_n) = \{0\} \). By Galerkin orthogonality (3.36) and (3.42) this direct sum is orthogonal. As \( \text{Ker}(P_n) = \text{Range}(Id - P_n) \)
\[
v \in \text{Ker}(P_n) \iff P_n v = 0 \iff (Id - P_n) v = v \iff v \in \text{Range}(Id - P_n) .
\]
the proposition holds. \( \square \)

3.12 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. 3.11. The trick to convert (DVP) into a problem that can be solved on a computer is to introduce ordered bases
\[
\mathcal{B}_V := \{p_1^N, \ldots, p_N^N\} \text{ of } V_n, \quad \mathcal{B}_W := \{q_1^N, \ldots, q_N^N\} \text{ of } W_n, \quad N := \dim V_n = \dim W_n .
\]
Remember that a basis of a finite dimensional vector space is a maximal set of linearly independent vectors. By indexing the basis vectors with consecutive integers we indicate that the order of the basis vectors will matter.

Lemma 3.79. 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
\[
B \mu = \varphi \quad \text{(LSE)}
\]
with
\[
B := \left( b(q_j^k, p_k^N) \right)_{j,k=1}^N \in \mathbb{R}^{N,N} , \quad \varphi := \left( \langle f, p_k^N \rangle_{V' \times V} \right)_{k=1}^N \in \mathbb{R}^N ,
\]
has a unique solution \( \mu = (\mu_k)_k^N \in \mathbb{R}^N \).
Then

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

**Proof.** Due to the basis property we can set

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

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

\[ b(\sum_{k=1}^{N} \mu_k q_n^k, \sum_{j=1}^{N} \nu_j p_n^j) = \left\langle f, \sum_{j=1}^{N} \nu_j p_n^j \right\rangle_{V' \times V} \]

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

\[ \sum_{j=1}^{N} \sum_{k=1}^{N} \mu_k \nu_j b(q_n^k, p_n^j) = \sum_{j=1}^{N} \nu_j \left\langle f, p_n^j \right\rangle_{V' \times V}. \quad (3.45) \]

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

\[ \sum_{k=1}^{N} \mu_k b(q_n^k, p_n^l) = \left\langle f, p_n^l \right\rangle_{V' \times V}, \quad l = 1, \ldots, N. \quad (3.46) \]

As the special test vectors span all of \( \mathbb{R}^N \) and thanks to the basis property, we conclude that (3.45) and (3.46) are equivalent. On the other hand, (3.46) corresponds to (LSE), as is clear by recalling the rules of matrix-vector multiplication.

Note that in (3.43) \( j \) 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 \( B \) in (LSE) is given by \( b(q_n^k, p_n^l) \). The columns correspond to the trial functions where the rows correspond to the test functions.

\[ \begin{array}{c|c|c|c|c|c} \hline \cdot & \cdots & \cdot & \cdots & \cdot & \cdots \\
\downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
\mu_k & b(q_n^k, p_n^l) & \cdots & b(q_n^k, p_n^j) & \cdots & b(q_n^k, p_n^l) \\
\hline \end{array} \]

**Notation 3.80.** Throughout, bold greek symbols will be used for vectors in some Euclidean 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 3.81.** If and only if the bilinear form $b$ and trial/test space $W_n/V_n$ satisfy the assumptions of Thm. 3.69, then the matrix $B$ of (LSE) will be regular.

Thus, we have arrived at the final “algebraic problem” (LSE) through the two stage process outlined in Fig. 3.7. 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. 3.81 teaches that some properties of $B$ will only depend on $V_n$, too.

![Diagram](image)

**Figure 3.7:** Overview of stages involved in the complete Galerkin discretization of an abstract variational problem.

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

**Example 3.82.** 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 (3.43) will be the $N \times N$ identity matrix $I$.

**Lemma 3.83.** 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 $\mathcal{B} := \{p^1_n, \ldots, p^N_n\}$ and $\tilde{\mathcal{B}} := \{\tilde{p}^1_n, \ldots, \tilde{p}^N_n\}$ of $W_n = V_n$, for which

$$\tilde{p}^j_n = \sum_{k=1}^{N} s_{jk} p^k_n \quad \text{with} \quad S = (s_{jk})^{N}_{j,k=1} \in \mathbb{R}^{N,N} \text{ regular.}$$
Relying on these bases we convert (DVP) into two linear systems of equations \( B\mu = \varphi \) and \( \tilde{B}\tilde{\mu} = \tilde{\varphi} \), respectively.

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

\[
\tilde{B} = SBS^T, \quad \tilde{\varphi} = S\varphi, \quad \tilde{\mu} = S^{-T}\mu.
\] (3.47)

**Proof.** Using (3.43) and the linearity of \( b \) we get

\[
\tilde{b}_{lm} = b(p^m_n, p^l_n) = \sum_{k=1}^{N} s_{mk} b(p^k_n, p^j_n) s_{lj} = \sum_{k=1}^{N} \left( \sum_{j=1}^{N} s_{lj} b_{jk} \right) s_{mk} = (SBS^T)_{lm},
\]

which gives the relationship between \( B \) and \( \tilde{B} \). The other relationships are as straightforward. \( \square \)

**Remark 3.84.** The lemma reveals that all possible Galerkin matrices \( B \) 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. Its not an similarity transformation, so eigenvalues are not invariant (in case of an orthogonalisation the eigenvalues are all one).

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

\[
\kappa(M) = |M| \frac{\|M^{-1}\|}{\|M^{-1}\|},
\]

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

Remember that we are working on computers with finite precision algebra and that the spectral condition number measures the influence of small relative errors in the right hand side vector to the solution vector.

\[
\frac{|\mu - \tilde{\mu}|}{|\mu|} \leq |B| \frac{|\varphi - \tilde{\varphi}|}{|\varphi|}
\]

▷ Choice of basis of \( V_n, W_n \) influences the condition number

For the remainder of this section we assume that trial and test spaces and their respective bases agree. Any choice of basis \( \mathcal{B} := \{p^1_n, \ldots, p^N_n\} \) for \( V_n \) spawns a coefficient isomorphism \( C_n : \mathbb{R}^N \mapsto V_n \) by

\[
C_n\mu = \sum_{k=1}^{N} \mu_k p^k_n.
\]

It can be used to link the operator form of the discrete variational problem and the associated linear system of equations (LSE) w.r.t. \( \mathcal{B} \).
Definition 3.85. For a linear operator $T : V \mapsto W$ on vector real spaces $V, W$ we introduce its adjoint operator $T'$ by

$$T' : \begin{cases} W' & \mapsto V' \\ h & \mapsto \{ V \mapsto \mathbb{R} \\ v \mapsto \langle h, T v \rangle_{W' \times W} \} \end{cases}.$$ 

Lemma 3.86. We have

$$B = C'_n \circ B_n \circ C_n,$$

where $B_n : V_n \mapsto V_n'$ is the operator related to the bilinear form $b$ on $V_n$,

$$\langle B_n w_n, v_n \rangle_{V_n' \times V_n} = b(w_n, v_n) \quad \forall w_n \in W_n, v_n \in V_n,$$

and $B$ is the coefficient matrix for $b$ w.r.t. the basis $\mathcal{B}$ of $V_n$.

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

$$\langle \mu, \xi \rangle_{(\mathbb{R}^N)' \times \mathbb{R}^N} = \mu^T \xi.$$

Owing to Def. 3.85 and (3.43) we find

$$\mu^T B \xi = b(C_n \xi, C_n \mu) = \langle B_n C_n \xi, C_n \mu \rangle_{V_n' \times V_n}$$

$$= \langle C'_n B_n C_n \xi, \mu \rangle_{(\mathbb{R}^N)' \times \mathbb{R}^N} = \mu^T (C'_n B_n C_n) \xi.$$

Lemma 3.87. If $B$ is a matrix according to (3.43) based on a bilinear form $b$ and a trial/test space $V_n$ (equipped with basis $\mathcal{B}$) that satisfy the assumptions of Thm. 3.69, then

$$\kappa(B) \leq ||C_n||_{V_n \mapsto \mathbb{R}^N}^2 ||C_n^{-1}||_{V \mapsto \mathbb{R}^N}^2 \frac{||b||_{V \times V \mapsto \mathbb{R}}}{\gamma_n},$$

where $C_n$ is the coefficient isomorphism belonging to basis $\mathcal{B}$, and $\gamma_n$ is the inf-sup constant from (DIS).

Proof. By the definition of the operator norm,

$$||B_n w_n||_{V_n'} \leq \sup_{v_n \in V_n} \frac{|b(w_n, v_n)|}{||v_n||_V} \geq \gamma_n ||w_n||_V \quad \forall w_n \in V_n,$$

which means

$$\gamma_n ||B_n^{-1} f_n||_V \leq ||f_n||_{V_n'} \quad \forall f_n \in V_n' \quad \Rightarrow \quad ||B_n^{-1}||_{V_n' \mapsto V_n} \leq \gamma_n^{-1}.$$
The operator norm of $B_n$ can be bounded by

$$\|B\|_{V \rightarrow V'} = \|b\|_{V \times V \rightarrow \mathbb{R}}.$$

By Lemma 3.86 we have

$$B = C'_n \circ B_n \circ C_n, \quad B^{-1} = C^{-1} \circ B^{-1}_n \circ (C'_n)^{-1}.$$ 

Then the submultiplicativity of operator norms finish the proof. \hfill $\square$

$\triangleright$ condition number depends on the chosen basis via $C_n$, as well as on the discretisation space via $\gamma_n$ and on problem itself via $\|b\|_{V \times V \rightarrow \mathbb{R}}$.

References


4 Primal Finite Element Methods

In Section 3.11 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. 3.69.

In this chapter we only consider linear variational problems that arise from the primal weak formulation of boundary value problems as discussed in Sect. 3.5, see (3.23) and Sec. 3.10.1. These variational problems are set in Sobolev spaces and feature elliptic bilinear forms according to Def. 3.25. 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 3.71.

In this setting, the construction of $V_n$ has to address two major issues

1. In light of Thm. 3.69 $V_n$ 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. 3.12).

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.

4.1 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. 3.5.

**Definition 4.1.** A mesh $\mathcal{M}$ of $\Omega$ is a collection $\{K_i\}_{i=1}^M$, $M := \#\mathcal{M}$, of connected open subsets $K_i \subset \Omega$ such that

- the closure of each $K_i$ is the $C^\infty$-diffeomorphic image of a closed $d$-dimensional polytope (that is, the convex hull of $d + 1$ points in $\mathbb{R}^d$),

- $\bigcup_i K_i = \overline{\Omega}$ and $K_i \cap K_j = \emptyset$ if $i \neq j$, $i, j \in \{1, \ldots, M\}$.

$\#\mathcal{M}$ denotes the cardinality of a finite set
The $K_i$ are called cells of the mesh.

**Remark 4.2.** Sometimes the smoothness requirement on the diffeomorphism is relaxed and mapping that are continuous but only piecewise $C^\infty$ are admitted.

Following the terminology of Sect. 3.1, 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. See Fig. 4.1 for a list of cell types in 2 and 3 dimensions.

Meshes are a crucial building block in the design of the finite dimensional trial and test spaces used in the finite element method. Furthermore, they provide subdomains for integration to build the system matrix and vector of the right hand side.

**Definition 4.3.** For $d = 3$ the set of (topological) faces of a mesh $\mathcal{M}$ is given by

$$\mathcal{F}(\mathcal{M}) := \{\text{interior}(K_i \cap K_j), 1 \leq i < j \leq \#\mathcal{M}\} \cup \{\text{(geometric) faces } \subset \partial \Omega\},$$

the set of (topological) edges of $\mathcal{M}$ is defined as

$$\mathcal{E}(\mathcal{M}) := \{\text{interior}(F \cap F'), F, F' \in \mathcal{F}(\mathcal{M}), F \neq F'\},$$

whereas the set of (topological) nodes is

$$\mathcal{N}(\mathcal{M}) := \{E \cap E', E, E' \in \mathcal{E}(\mathcal{M}), E \neq E'\}.$$
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 $F(M)$ for the set of these (generalized) faces.

Figure 4.2: Two-dimensional mesh and the sets of edges (red) and nodes (blue)

**Remark 4.4.** The “is contained in the closure of” incidence relations $N(M) \times E(M) \rightarrow \{\text{true}, \text{false}\}$, $E(M) \times F(M) \rightarrow \{\text{true}, \text{false}\}$, etc., describe the topology of a mesh. The locations of nodes and shape of cells are features connected with the geometry of the mesh. We can define the nodes $N(E)$ of an edge, the edges $E(F)$ of a face and the faces $F(K)$ of a cell.

Already contained in the definition of a mesh is the notion of reference cells. By them we mean a finite set $\hat{K}_1, \ldots, \hat{K}_P$, $P \in \mathbb{N}$, of $d$-dimensional polytopes such that all cells of the mesh can be obtained from one of the $\hat{K}_i$ under a suitable diffeomorphism.

We recall that a mapping

$$\Phi : \mathbb{R}^d \mapsto \mathbb{R}^d, \quad \xi \mapsto F\xi + \tau, \quad F \in \mathbb{R}^{d,d} \text{ regular}, \tau \in \mathbb{R}^d$$

(AFF)

represents a bijective affine mapping of $d$-dimensional Euclidean space. An affine mapping from a triangle is a triangle, and an affine mapping of a square is a parallelogram.

**Definition 4.5.** A mesh $M$ 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 $\{M_n\}_{n \in \mathbb{N}}$ is affine equivalent, if this is true for each of its members and if the same reference polytope can be chosen for all $M_n$, $n \in \mathbb{N}$.

To map from a square to a general (straight) quadrilateral we need a bilinear mapping

$$\Phi : \mathbb{R}^2 \mapsto \mathbb{R}^2, \quad \xi \mapsto F\xi + \tau + \tau Q_1 \xi_1 \xi_2. \quad (4.1)$$

Note, that for the bilinear mapping to a non-convex quadrilateral (one angle $\geq 180^\circ$) is not a diffeomorphism.

With the so called blending techniques a mapping for curved cells for given parametrisation of curved edges can be defined:

$$x_K(\xi_1, \xi_2) = \Phi_K \xi = \begin{cases} (1 - \xi_2)x_1(\xi_1) + \xi_1 x_2(\xi_2) + \xi_2 x_3(1 - \xi_1) + (1 - \xi_1)x_4(1 - \xi_2) \\
(1 - \xi_1)(1 - \xi_2)p_0 - \xi_1(1 - \xi_2)p_1 - \xi_1\xi_2 p_2 - (1 - \xi_1)\xi_2 p_3. \end{cases}$$
Figure 4.3: Mapping for curved quadrilateral.

**Definition 4.6.** A mesh $\mathcal{M} = \{ K_i \}_{i=1}^M$ of $\Omega$ is called a triangulation, if $K_i \cap K_j$, $1 \leq i < j \leq M$, agrees with a geometric face/edge/vertex of $K_i$ or $K_j$. The cells of a triangulation are sometimes called elements.

Figure 4.4: 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 4.7.** Orienting an edge amounts to prescribing a direction. The orientation of a face for $d = 3$ can be fixed by specifying an ordering of the edges along its boundary.

If all geometric objects of a triangulation are equipped with an orientation, we will call it an oriented triangulation.

**Example 4.8.** 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.
Definition 4.9. A triangulation \( \mathcal{M} := \{K_i\}_{i=1}^M \) of \( \Omega \) is called \textit{conforming}, if \( K_i \cap 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 4.5: A triangulation that is not conforming and possesses two hanging nodes.

Definition 4.10. 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 \textit{hanging (dangling) node}.

In the case of triangulations we can distinguish special classes:

- \textbf{simplicial triangulations} that entirely consist of triangles \( (d = 2) \) or tetrahedra \( (d = 3) \), whose edges/faces might be curved, nevertheless.

- \textbf{quadrilateral} \( (d = 2) \) and \textbf{hexahedral} \( (d = 3) \) triangulations, which only comprise cells of these shapes. Curved edges or faces are admitted, again.

Corollary 4.11. Any family of simplicial triangulations is affine equivalent.

Remark 4.12. A quadrilateral triangulation need not be affine equivalent, because, for instance, there is no affine map taking a square to general trapezoid.

Exercise 4.1. Let \( \nu_1, \ldots, \nu_4 \in \mathbb{R}^3 \) stand for the coordinate vectors of the four vertices of a tetrahedron \( K \). Determine the affine mapping \( \text{(AFF)} \) that takes the “unit tetrahedron”

\[
\hat{K} := \text{convex} \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}
\]

to \( K \). When will the matrix \( F \) of this affine mapping be regular?

Exercise 4.2. Let \( \nu_1, \ldots, \nu_4 \in \mathbb{R}^2 \) denote the coordinate vectors belonging to the four vertices of a quadrilateral \( \hat{K} \) in the plane. Determine an analytic description of a simple smooth mapping \( \Phi : \mathbb{R}^2 \to \mathbb{R}^2 \) from the “reference square”

\[
\hat{K} := \text{convex} \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \right\}
\]

to \( K \). Compute the Jacobian \( D\Phi \) and its determinant.
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.

Automatic mesh generation is a challenging subject, which deals with the design of algorithms that create a mesh starting from a description of $\Omega$. Such a description can be given

- in terms of geometric primitives (ball, brick, etc.) whose unions or intersections constitute $\Omega$.
- by means of a parameterization of the faces of $\Omega$.
- through a function $f : \mathbb{R}^d \mapsto \mathbb{R}$, whose sign indicates whether a point is located inside $\Omega$ or outside.
- by a mesh covering the surface of $\Omega$ 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 \( \Omega \).

• the quadtree \( (d = 2) \) or octree \( (d = 3) \) approach, which fills \( \Omega \) with squares/cubes of different sizes supplemented by special measures for resolving the boundary.

• mapping techniques that split \( \Omega \) into sub-domains of “simple” shape (curved triangles, parallelograms, bricks), endow those with parametric grids and glue these together.

Remark 4.13. Traditional codes for the solution of boundary value problems based on the finite element method usually read the geometry from a file describing the topology 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. 4.8.

\[
\begin{align*}
\text{Parameters} \\
\text{Mesh generator} & \rightarrow \text{Finite element solver} \\
& \quad \text{(computational kernel)} \\
& \quad \rightarrow \text{Post-processor} \\
& \quad \text{(e.g. visualization)}
\end{align*}
\]

Figure 4.8: Flow of data in traditional finite element simulations

Remark 4.14. A typical file format for a mesh of a simplicial conforming triangulation of a two-dimensional polygonal domain is the following:

\[
\begin{align*}
\# \text{Two-dimensional simplicial mesh} \\
N \in \mathbb{N} & \quad \# \text{Number of nodes} \\
\xi_1 \eta_1 & \quad \# \text{Coordinates of first node} \\
\xi_2 \eta_2 & \quad \# \text{Coordinates of second node} \\
\vdots \\
\xi_N \eta_N & \quad \# \text{Coordinates of } N\text{-th node} \\
M \in \mathbb{N} & \quad \# \text{Number of triangles} \\
n_1^1 n_1^2 n_1^3 & X_1 \quad \# \text{Indices of nodes of first triangle} \\
n_2^2 n_2^3 n_2^3 & X_2 \quad \# \text{Indices of nodes of second triangle} \\
\vdots \\
n_M^1 n_M^2 n_M^3 & X_M \quad \# \text{Indices of nodes of } M\text{-th triangle}
\end{align*}
\]
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{align*}
K \in \mathbb{N} & \quad \text{# Number of edges on } \partial \Omega \\
n_1^1 & n_1^1 \quad Y_1 \quad \text{# Indices of endpoints of first edge} \\
n_2^1 & n_2^1 \quad Y_2 \quad \text{# Indices of endpoints of second edge} \\
\vdots & \\
n_K^1 & n_K^1 \quad Y_K \quad \text{# Indices of endpoints of } K\text{-th edge}
\end{align*}

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 (4.3).

Please note that the ordering of the nodes in the above file formats implies an orientation of triangles and edges.

For a comprehensive account on mesh generation see [16]. An interesting algorithm for Delaunay meshing is described in [17, 18]. Free mesh generation software is also available, just to name some, netgen, gmsh, triangle, emc2. However, the most sophisticated mesh generation tools are commercial products and their algorithmic details are classified.

### 4.2 Linear finite elements on triangular meshes

#### 4.2.1 Basis functions

![Figure 4.9: FE Basis function $b_P(x)$](image-url)
Let the domain $\Omega \in \mathbb{R}^2$ be a polygon with a triangular conforming mesh $\mathcal{M}$. We define the finite element space of piecewise linear, continuous functions as

$$S^1(\Omega, \mathcal{M}) := \{ u \in C^0(\Omega) : u(x)|_K = a + bx_1 + cx_2, \ \forall K \in \mathcal{M} \}.$$  

**Proposition 4.15** (Properties of $S^1(\Omega, \mathcal{M})$).

- It holds $S^1(\Omega, \mathcal{M}) \subset H^1(\Omega)$.
- $u \in S^1(\Omega, \mathcal{M})$ is uniquely defined by the values of $u(P)$ on the nodes $P \in \mathcal{N}(\mathcal{M})$.
- $N = \dim S^1(\Omega, \mathcal{M}) = \#\mathcal{N} < \infty$.
- $S^1(\Omega, \mathcal{M}) = \text{span}\{ b_P(x) : P \in \mathcal{N}(\mathcal{M}) \}$, where the **hut functions** are defined as
  $$\varphi_P \in S^1(\Omega, \mathcal{M}), b_P(P') = \delta_{P=P'}.$$  

Let $\mathbf{b}$ the vector of the basis functions (for a fixed numbered). Then, an arbitrary FE function $v \in S^1(\Omega, \mathcal{M})$ can be written as

$$v(x) = \sum_{P \in \mathcal{N}(\mathcal{M})} v(P) b_P(x) = \mathbf{v}^\top \mathbf{b}(x),$$

so also the solution

$$u_n(x) = \sum_{P \in \mathcal{N}(\mathcal{M})} u_n(P) b_P(x) = \mathbf{\mu}^\top \mathbf{b}(x).$$

**4.2.2 Assembling of system matrix and load vector**

The support of the basis functions $b_P(x)$ is only a few triangles, i.e., the integral in the bilinear form reduces to smaller sets for pairs of basis functions.

Changing the point of view, we consider the **shape functions**, which are the restrictions of a basis function to one cell $K \in \mathcal{M}$. For $S^1(\Omega, \mathcal{M})$ these are exactly three, each for one node of $K$.

The shape functions can be defined on a single cell $K$ as

$$N_{K,P_j(K)}(x) = \hat{N}_j(\Phi_K^{-1}x).$$

(4.4)

where $\hat{N}_j$ are the **element shape functions** defined by

$$\hat{N}_0(\xi) = 1 - \xi_1 - \xi_2, \quad \hat{N}_j(\xi) = \xi_j, j = 1, 2.$$  

(4.5)

on the reference element, which is the triangle with vertices $(0,0), (1,0), (0,1)$. $P_j(K)$ is the $j$-th node of the triangle $K$, $j = 0, 1, 2$, with the coordinate $p_j$. Then, the (affine) element mapping is

$$x = \Phi_K(\xi) = p_0 + \xi_1(p_1 - p_0) + \xi_2(p_2 - p_0) = \tau + F_K(\xi),$$

72
with \( \mathbf{F} = (\mathbf{p}_1 - \mathbf{p}_0, \mathbf{p}_2 - \mathbf{p}_0) \) and \( \mathbf{\tau} = \mathbf{p}_0 \).

As the element mapping is linear, the shape functions are in fact linear as well. The basis functions result from the shape functions by glueing, i.e.,

\[
b_P(x) = \begin{cases} 
N_{K,P_j}(x) & x \in K, P \in \mathcal{N}(K), P = P_j(K), \\
0 & \text{otherwise}.
\end{cases}
\]

This can be expressed by the connectivity or \( \mathbf{T} \)-matrices

\[
\mathbf{T}_K \in \mathbb{R}^{3, N}, \quad (\mathbf{T}_K)_{ij} = \begin{cases} 
1, & P_j(K) = P_i(\mathcal{M}), \\
0, & \text{otherwise}.
\end{cases}
\]

These matrices give a relation between local number of (element) shape functions and global numbering of basis functions. They have the form

\[
\mathbf{T}_K = \begin{pmatrix}
\cdots & \cdots & 1 & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdot & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdot & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdot & \cdots & \cdots & \cdots & \cdots & \cdots 
\end{pmatrix}.
\]

For the \( \mathbf{T} \)-matrices an extremly sparse format is used. For linear finite elements on triangles we have only to store the three indices \( P_{i_0}, P_{i_1}, \) and \( P_{i_2} \) of the three nodes in the triangle.

Now, we can express a basis function in one cell \( K \) as

\[
b_{P_i}(x)|_K = \sum_{k=1}^{3} (\mathbf{T})_{kj} N_{K,P_k(K)}(x), \quad (4.6)
\]

and the system matrix is given with

\[
(\mathbf{B})_{ij} = b(b_{P_j}, b_{P_i}) = \sum_K (b_{P_j}|_K, b_{P_i}|_K)
\]

\[
= \sum_K \left( \sum_{k=1}^{3} b(\mathbf{T}_K)_{kj} N_{K,P_k(K)} \right) \sum_{\ell=1}^{3} (\mathbf{T}_K)_{\ell i} N_{K,P_\ell(K)}
\]

\[
= \sum_K \sum_{k=1}^{3} \sum_{\ell=1}^{3} (\mathbf{T}_K)_{kj} N_{K,P_k(K)} N_{K,P_\ell(K)}
\]

\[
= \sum_K (\mathbf{T}_K)_{ij} \mathbf{B}_K(\mathbf{T}_K), i
\]

as sum of “weighted” element matrices

\[
\mathbf{B} = \sum_K \mathbf{T}_K^\top \mathbf{B}_K \mathbf{T}. \quad (4.8)
\]

This mean we have to integrate only over the shape functions, and to sum up the contributions over all the cells while scattering to the right position in the system matrix.
4.2.3 Element stiffness matrix

The bilinear form in the variational problem (3.23) consists of three parts, which contribute to the system matrix. The first part

\[ a(u, v) = \int_{\Omega} \langle \sigma \nabla u, \nabla v \rangle \, dx \]

is constituting the stiffness matrix \( A \).

We transform the derivatives to \( \hat{K} \). By chain rule of differentiation and with

\[
\nabla = \left( \frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2} \right)^T, \quad \hat{\nabla} = \left( \frac{\partial}{\partial \xi_1}, \frac{\partial}{\partial \xi_2} \right)^T,
\]

we have

\[
\partial_{\xi_1} N_{K,P_j}(x(\xi)) = \partial_{x_1} N_{K,P_j}(x) \frac{\partial x_1}{\partial \xi_1} + \partial_{x_2} N_{K,P_j}(x) \frac{\partial x_2}{\partial \xi_1}
\]

\[
\partial_{\xi_2} N_{K,P_j}(x(\xi)) = \partial_{x_1} N_{K,P_j}(x) \frac{\partial x_1}{\partial \xi_2} + \partial_{x_2} N_{K,P_j}(x) \frac{\partial x_2}{\partial \xi_2}
\]

and so

\[
\hat{\nabla} N_{K,P_j}(x(\xi)) = \left( \frac{\partial \Phi_K}{\partial \xi} \right)^T \nabla N_{K,P_j}(x),
\] (4.9)

and so for triangular cells

\[
\nabla N_{K,P_j}(x) = F_K^{-\top} \hat{\nabla} N_{K,P_j}(x(\xi)) = F_K^{-\top} \hat{\nabla} \hat{N}_j(\xi).
\]

For a single cell \( K \) the entries of the element stiffness matrix \( A_K \) are so given by

\[
a_K(N_{K,P_j}, N_{K,P_i}) = \int_K \sigma(x) (\nabla N_{K,P_j}(x))^\top \nabla N_{K,P_i}(x) \, dx
\]

\[
= \int_K \sigma(\Phi_K \xi)(\hat{\nabla} \hat{N}_j(\xi))^\top F_K^{-1} F_K^{-\top} \hat{\nabla} \hat{N}_i(\xi) |F_K| \, d\xi
\]

\[
= \int_K \sigma(\Phi_K \xi)(\hat{\nabla} \hat{N}_j(\xi))^\top \text{adj}(F_K) \text{adj}(F_K)^\top \hat{\nabla} \hat{N}_i(\xi) |F_K|^{-1} \, d\xi \quad (4.10)
\]

where we use the relation of the inverse and the adjoint matrix (dt. Adjunkte)

\[
F_K^{-1} = |F_K|^{-1} \text{adj}(F_K).
\]

The adjoint matrix of a matrix \( \begin{pmatrix} a & b \\ c & d \end{pmatrix} \) is \( \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \).

Note, that the gradient of the element shape functions are constant vectors

\[
\hat{\nabla} \hat{N}_0 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \quad \hat{\nabla} \hat{N}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \hat{\nabla} \hat{N}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad (4.11)
\]
So we can simplify

\[(A_K)_{ij} = a_K(N_{K,P_j}, N_{K,P_i}) = \frac{\sigma_K}{2} (\hat{\nabla} \hat{N}_j)^\top \text{adj}(F_K) \text{adj}(F_K)^\top \hat{\nabla} \hat{N}_i |F_K|^{-1}\]

with \(\sigma_K = \frac{1}{|K|} \int_K \sigma(x) \, dx\) the average heat conduction in \(K\) (note that \(|F_K| = 2|K|\)). The latter simplifies for constant material or may be obtained by numerical quadrature for more general smooth functions.

Due to the special values of \(\hat{\nabla} \hat{N}_j\) (see (4.11)) we can even simplify [19]

\[A_K = \frac{\sigma_K}{4|K|} D_K^\top D_K\]

with a matrix \(D_K\) with coordinate differences

\[D_K = \begin{pmatrix}
y_1 - y_2 & y_2 - y_0 & y_0 - y_1 \\
x_2 - x_1 & x_0 - x_2 & x_1 - x_0
\end{pmatrix}.
\]

### 4.2.4 Element mass matrix

The mass matrix is related to the bilinear form

\[m(u, v) = \int_\Omega cuv \, dx.\]

The element mass matrix \(M_K\) for the cell \(K\) can be computed as

\[m_K(N_{K,P_j}, N_{K,P_i}) = \int_K c(x) N_{K,P_j}(x) N_{K,P_i}(x) \, dx = \int_{\hat{K}} c(\Phi_K \xi) \hat{N}_j(\xi) \hat{N}_i(\xi) |F_K| \, d\xi.\]

In case of a constant function \(c_K\) in the cell \(K\) we can write

\[(M_K)_{ij} = c_K |K| \begin{cases}
\frac{1}{9} & i = j, \\
\frac{1}{12} & i \neq j.
\end{cases}\]

### 4.2.5 Element load vector

The element load vector is related to the linear form

\[\ell_K(v) = \int_K f v \, dx.\]

For general, smooth function \(f\) we use numerical quadrature to evaluate the integrals. The simplest quadrature rule is

\[\int_K f v \, dx \approx |K| f(x_K) v(x_K)\]

where \(x_K\) is the barycenter (dt. Schwerpunkt) of the cell \(K\). As this quadrature rule is only exact for linear functions, and the shape functions are already linear, it will provide only reasonable results if \(f\) is (almost) constant in \(K\).
4.3 Higher order finite elements on curved cells

4.3.1 Linear finite elements on quadrilateral cells

The element shape functions for the linear finite elements on quadrilateral cells are

\[
\hat{N}_0(\bm{\xi}) = (1 - \xi_1)(1 - \xi_2), \\
\hat{N}_1(\bm{\xi}) = \xi_1(1 - \xi_2), \\
\hat{N}_2(\bm{\xi}) = \xi_1\xi_2, \\
\hat{N}_3(\bm{\xi}) = (1 - \xi_1)\xi_2,
\]

which span on the square \(\hat{K} = (0, 1)^2\) the space

\[\mathcal{Q}_1(\hat{K}) := \text{span}\{1, \xi_1, \xi_2, \xi_1\xi_2\}.\]

The element shape functions are linear along the edges of the square \([0, 1]^2\), and, hence, the shape functions resulting after bilinear mapping \((4.1)\) and as for triangles with \((4.4)\), are linear along the edges of \(K\) with value 1 on one of the four nodes, respectively. This allows to glue shape functions of the cells around a nodes to constitute a globally continuous basis functions.

These basis functions span the space

\[S^1(\Omega, \mathcal{M}) = \{u \in C^0(\Omega): u(x)|_K \circ \Phi_K(\bm{\xi}) \in \mathcal{Q}_1(\hat{K})\},\]

Since the inverse of the bilinear mapping \((4.1)\) is in general not polynomial, the basis functions are in general no piecewise polynomials anymore, only on each cell the image of a polynomial under \(\Phi_K\).

The matrix assembling is similar to that for triangles, whereas the \(T\)-matrices have 4 rows and the element matrices follow by \((4.10)\) and \((4.12)\) where \(F_K\) is replaced by the Jacobian matrix \(D\Phi_K(\bm{\xi})\), which is not constant anymore, but linear in \(\xi_1\) and \(x_2\).

In the formula for the stiffness matrix the inverse of the Jacobian determinant appears, which is not a polynomial.

Even for constant material functions \(\sigma(x)\) and \(c(x)\), the use of numerical quadrature rules may become interesting.

4.3.2 Numerical quadrature for quadrilateral cells

The integration is on the reference cell \([0, 1]^2\), so a tensor product of 1D quadrature rule can be applied. The quadrature rules are usually given in the interval \([-1, 1]\). So, we simply transform 1D integrals via

\[
\int_0^1 f(\xi) \, d\xi = \frac{1}{2} \int_{-1}^1 f(\frac{\xi+1}{2}) \, d\xi \approx \frac{1}{2} \sum_{j=1}^n w_j f(\frac{\xi_j+1}{2}), \quad (4.13)
\]

where \(w_j\) are the weights and \(\xi_j\) the abscissas of the quadrature rule.

Accurate quadrature rule for smooth functions are variants of the Gauss quadrature. The most well-known is the Gauss-Legendre rule for which the abscissas are the zeros.
of the $n$-th Legendre polynomial $P_n(\xi)$ and the weights are given by (see [20, page 887])

$$w_j = \frac{2}{(1 - \xi_j^2)[P'_n(\xi_j)]^2}.$$ 

The Gauss-Legendre rule is exact for polynomials of degree $2n - 1$ and the remainder (for the interval $[0, 1]$) is

$$R_n = \frac{(n!)^4}{(2n + 1)([2n]!)^3} f^{(2n)}(\xi), \quad 0 < \xi < 1.$$ 

The zeros of the Legendre polynomials are tabulated (see [20, page 921ff]) and there is an algorithm (see Matlab version \texttt{gauleg.m} on the webpage of the lecture) in the numerical recipes [21].

Only accurate for polynomials of degree $2n - 3$ is the Gauss-Lobatto rule, for which, however, both end-points are included.

For the square $[0, 1]^2$ we have the product quadrature rule

$$\int_0^1 \int_0^1 f(\xi) \, d\xi_1 \, d\xi_2 = \frac{1}{4} \int_{-1}^1 \int_{-1}^1 f\left(\xi_1 + \frac{1}{2}, \xi_2 + \frac{1}{2}\right) \, d\xi_1 \, d\xi_2 \approx \frac{1}{4} \sum_{i=1}^{n_1} w_i \sum_{j=1}^{n_2} w_j f\left(\frac{\xi_i + 1}{2}, \frac{\xi_j + 1}{2}\right),$$

with $n_1$ and $n_2$ quadrature points in the two directions.

### 4.3.3 Linear finite elements on curved cells

For curved cells the mapping from reference cell, either triangle or square, is even more general than the bilinear mapping. This influences only the Jacobian matrix $D\Phi_K(\xi)$ and numerical quadrature will be essential to compute the element matrix entries.

The continuity of the basis functions is assured if for neighbouring cells $K_1$ and $K_2$ the mapping from a reference interval $[0, 1]$ to the common edge, introduced by the mappings $\Phi_{K_1}$ and $\Phi_{K_2}$, is the same. This is also true for mixed meshes of curved triangles and quadrilaterals.

Let us define the local space for triangles as

$$\mathcal{P}_1(\tilde{K}) := \text{span}\{1, \xi_1, \xi_2\}.$$ 

Then, we are now in the position to introduce a general definition for the space

$$S^1(\Omega, M) = \{ u \in H^1(\Omega) : u(x)|_K \circ \Phi_K(\xi) \in \mathcal{P}_1(\tilde{K}) \text{ if } K \text{ is a triangle or } u(x)|_K \circ \Phi_K(\xi) \in \mathcal{Q}_1(\tilde{K}) \text{ if } K \text{ is a quadrilateral} \}.$$ 

For curved triangular cells we have also to rely on numerical quadrature.
4.3.4 Numerical quadrature for triangular cells

Numerical quadrature for triangular cells are defined on the triangle with nodes \((-1, -1), (1, -1), (-1, 1),\) where integrals over \(\hat{K}\) can easily transformed to.

\[
\int_0^1 \int_0^{1-\xi_1} f(\xi) \, d\xi_2 \, d\xi_1 = \frac{1}{4} \int_{-1}^{1} \int_{-1}^{2-\xi_1} f\left(\frac{\xi_1+1}{2}, \frac{\xi_2+1}{2}\right) \, d\xi_2 \, d\xi_1 \approx \frac{1}{4} \sum_{j=1}^{n} w_j f\left(\frac{\xi_{1,j}+1}{2}, \frac{\xi_{2,j}+1}{2}\right),
\]

Let us define generalisations of \(P_1\) and \(Q_1\).

**Definition 4.16.** Given a domain \(K \subset \mathbb{R}^d, d \in \mathbb{N},\) we write

\[
P_m(K) := \text{span}\{\xi \in K \mapsto \xi^\alpha := \xi_1^{\alpha_1} \cdots \xi_d^{\alpha_d}, \alpha \in \mathbb{N}_0^d, |\alpha| \leq m\}
\]

for the vector space of \(d\)-variate polynomials of (total) degree \(m, m \in \mathbb{N}_0.\)

If \(m = (m_1, \ldots, m_d)^T \in \mathbb{N}_0^d\) we designate by

\[
Q_m(K) := \text{span}\{\xi \in K \mapsto \xi_1^{\alpha_1} \cdots \xi_d^{\alpha_d}, 0 \leq \alpha_k \leq m_k, 1 \leq k \leq d\}
\]

the space of tensor product polynomials of maximal degree \(m_k\) in the \(k\)-th coordinate direction.

There are Gauss quadrature rules for triangles which are exact for polynomials of maximal total degree 1, 2, \ldots, 5 and which use only 1, 3, 4, 6 and 7 points, respectively. See e.g., page 141 in Solin [22].

To get higher accuracies in a systematic matter there are

- representation of the integral over a reference square and using a tensor-product quadrature rule,
- approximated Fekete points, where points lying on the three edges correspond to Gauss-Lobatto points, and which are tabulated,

to name just the best known.

Quadrature rules for triangles can be obtained from quadrature rules on a square using the Duffy transformation, see Fig. 4.10. When \(K = \text{conv}(\{0, 1\}, \{0, 1\})\) and \(\hat{K} = [0, 1]^2,\) then

\[
\int_K f(\xi_1, \xi_2) \, d\xi = \int_{\hat{K}} f(\xi_1(1-\hat{\xi}_2), \xi_2) (1-\hat{\xi}_2) \, d\hat{\xi}.
\]

If \(f \in P_m(K),\) then the integrand on the right hand side will belong to \(Q_{m,m+1}(\hat{K}).\)

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. |

**Remark 4.17.** An alternative to numerical quadrature is polynomial interpolation of coefficients, source functions and (inverse of) Jacobian matrix followed by analytical evaluation of the localised integrals.
4.3.5 Higher order finite elements

The linear finite elements on triangles and quadrilaterals are a special case for finite elements on high orders. The basis functions have been identified to nodes $P$ of the mesh and are linear in local coordinates along each edge.

To obtain higher order finite elements we enrich the space of element shape functions by element shape functions

- identified to a particular edge $E$ on which they have polynomial behaviour and which vanish along all other edges, and so also along the end-points of $E$

- identified to a single cell $K$ and vanishing on all edges $E$. These are called interior or bubble functions.

The former can be written for $\hat{K} = [0, 1]^2$ as

$$
\hat{N}_{e,1,j}(\xi) = (1 - \xi_2) P_j(\xi_1),
$$
lower edge,

$$
\hat{N}_{e,2,j}(\xi) = \xi_1 P_j(\xi_2),
$$
right edge,

$$
\hat{N}_{e,3,j}(\xi) = \xi_2 P_j(\xi_1),
$$
upper edge,

$$
\hat{N}_{e,4,j}(\xi) = (1 - \xi_1) P_j(\xi_2),
$$
left edge,

and the latter as

$$
\hat{N}_{c,i,j}(\xi) = P_{c,j}(\xi_1, \xi_2),
$$

where $P_j(\xi)$ are polynomials vanishing at $\xi \in \{0, 1\}$ and $P_{c,j}(\xi)$ are polynomials vanishing at $\partial \hat{K}$. Note, that the choice of the basis influences the condition number of the resulting matrices.
Remember that we are interested in an one-to-one relation between shape functions in neighbouring cells. This can be achieved by choosing for one polynomial $P$ and $P_j(\xi) = P(\pm \xi)$ dependent if the $\xi_1$ or $\xi_2$-direction along the edge coincides or does not coincide with the global orientation of the edge.

In the following we will discuss the use of a family of shape functions based on integrated Legendre polynomials, which are either symmetric or anti-symmetric and allows therefore for fast matrix assembling and a simple matching.

4.3.6 Integrated Legendre polynomials as basis in quadrilaterals

The Legendre polynomials are defined in $[-1, 1]$ and orthogonal w.r.t. the $L^2$-inner product. We can define them by

$$L_0(\xi) = 1, \quad L_1(\xi) = \xi, \quad (j + 1)L_{j+1}(\xi) = (2j + 1)\xi L_j(\xi) - jL_{j-1}(\xi), \quad j > 1,$$

and

$$\int_{-1}^{1} L_j(\xi)L_i(\xi) \, d\xi = \delta_{ij} \frac{2}{2j + 1}.$$ 

The integrated Legendre polynomials are

$$\hat{L}_0(\xi) = -1, \quad \hat{L}_1(\xi) = \xi, \quad \hat{L}_j(\xi) = \int_{-1}^{\xi} L_{j-1}(t) \, dt = \frac{1}{2j - 1} (L_j(\xi) - L_{j-2}(\xi)).$$

The 1D element shape functions are then defined in $[0, 1]$ as

$$\hat{N}_0(\xi) = 1 - \xi, \quad \hat{N}_1(\xi) = \xi,$$

$$\hat{N}_j(\xi) = \sqrt{\frac{(2j - 1)}{2}} \hat{L}_j(2\xi - 1) = \frac{1}{\sqrt{2(2j - 1)}} (L_j(2\xi - 1) - L_{j-2}(2\xi - 1)).$$

As in $\hat{N}_{2j}$ for $j \geq 1$ only even degree monomials are present, and in $\hat{N}_{2j+1}$ only odd degree monomials, we have for $j \geq 1$

$$\hat{N}_{2j}(\xi) = \hat{N}_{2j}(-\xi), \quad \hat{N}_{2j+1}(\xi) = -\hat{N}_{2j+1}(-\xi).$$

Then, the 2D element shape functions for $Q_p([0, 1]^2)$, $p \geq 1$ are given by

$$\hat{N}_{(p+1)i+j}(\xi) = \hat{N}_j(\xi_i)\hat{N}_i(\xi_2), \quad 0 \leq i, j \leq p.$$ 

For the element shape functions identified to one of the four vertices it is $i, j \leq 1$, for those related to one of the edges it is either $i \leq 1$ or $j \leq 1$, and for $i, j \geq 2$ the element shape functions are interior.

The basis in the reference element is **hierarchic** meaning that for increasing polynomial degree only new functions appear and the former remain.

In case of a constant function $\sigma$ and a parallelogram cell the element stiffness and mass matrices have $O(p)$ non-zero entries. Otherwise, the basis with integrated Legendre polynomials leads to moderately increasing condition numbers for increasing maximal polynomial degree $p$.

In the $\mathbf{T}$-matrix there is exactly one entry with value 1
Figure 4.11: The first 1D element shape functions based on integrated Legendre polynomials.

- for each element shape functions identified to a node,
- for that identified to an edge if the function along the edge is symmetric, and
- for that identified to a cell.

If the function is anti-symmetric along the edge we have to relate the local $\xi_1$ or $\xi_2$-direction to the (global) orientation of the edge. If they coincide the entry in the $T$-matrix is 1 otherwise $-1$. The latter is equivalent to mirror the element shape function w.r.t. the perpendicular bisector of the edge (dt. Mittelsenkrechte). This means that the global orientation of the edge, which is known to the two neighbouring cells, is the judge deciding the direction of the shape functions.

### 4.3.7 Integrated Legendre polynomials as basis in triangles

A similar basis for triangles can be defined using the

**Definition 4.18** (Barycentric coordinates). Given $d + 1$ points $\mathbf{p}_0, \ldots, \mathbf{p}_d \in \mathbb{R}^d$ that do not lie in a hyperplane the **barycentric coordinates** $\lambda_1 = \lambda_1(\mathbf{x}), \ldots, \lambda_{d+1} = \lambda_{d+1}(\mathbf{x}) \in \mathbb{R}$ of $\mathbf{x} \in \mathbb{R}^d$ are uniquely defined by

$$\lambda_1 + \cdots + \lambda_{d+1} = 1, \quad \lambda_1 \mathbf{p}_0 + \cdots + \lambda_{d} \mathbf{p}_d = \mathbf{x}.$$  

The barycentric coordinates for a point $\mathbf{x}$ can be obtained by solving

$$
\begin{pmatrix}
\mathbf{p}_{0,1} & \cdots & \mathbf{p}_{d,1} \\
\vdots & \ddots & \vdots \\
\mathbf{p}_{1,d} & \cdots & \mathbf{p}_{d,d} \\
1 & \cdots & 1
\end{pmatrix}
\begin{pmatrix}
\lambda_1 \\
\vdots \\
\lambda_d \\
\lambda_{d+1}
\end{pmatrix}
=
\begin{pmatrix}
x_1 \\
\vdots \\
x_d \\
1
\end{pmatrix},
$$

which shows their uniqueness and existence, if the points $\mathbf{p}_j$ are not complanar. The convex hull of $\mathbf{p}_0, \ldots, \mathbf{p}_d$ can be described by

$$\text{convex}\{\mathbf{p}_0, \ldots, \mathbf{p}_d\} = \{\mathbf{x} \in \mathbb{R}^d, \ 0 \leq \lambda_i(\mathbf{x}) \leq 1, \ 1 \leq i \leq d + 1\}.$$
For the reference triangle \( \hat{K} \) we have
\[
\lambda_1 = 1 - \xi_1 - \xi_2, \quad \lambda_2 = \xi_1, \quad \lambda_3 = \xi_2,
\]
and the three element shape functions identified to the nodes \( P_0(K), P_1(K), P_2(K) \) are given by (4.5) or
\[
\hat{N}_0(\lambda) = \lambda_1, \quad \hat{N}_1(\lambda) = \lambda_2, \quad \hat{N}_2(\lambda) = \lambda_3.
\]
They vanish on the opposite edge where \( \lambda_j = 0 \).

The element shape functions identified with an edge opposite to the node \( P_j \) have to vanish on the two neighbouring edges where \( \lambda_{j-1} = 0, \lambda_{j+1} = 0 \) (meant modulus 3). So we can write
\[
\hat{N}_{e,1,j}(\lambda) = \lambda_2 \lambda_3 P_j(\lambda_1) P_j(\xi_1), \quad \text{edge opposite to node } P_0,
\]
\[
\hat{N}_{e,2,j}(\lambda) = \lambda_1 \lambda_3 P_j(\lambda_2) P_j(\xi_1), \quad \text{edge opposite to node } P_1,
\]
\[
\hat{N}_{e,3,j}(\lambda) = \lambda_1 \lambda_2 P_j(\lambda_3) P_j(\xi_1), \quad \text{edge opposite to node } P_2.
\]
The interior element shape functions vanish on all three edges and can be written as
\[
\hat{N}_{c,i,j}(\lambda) = \lambda_1 \lambda_2 \lambda_3 P_{i,j}(\lambda).
\]
The functions \( P_j(\lambda) \) can be chosen as integrated Legendre polynomials \( \hat{L}_0, \ldots, \hat{L}_{p-2} \), and \( P_{i,j}(\lambda) \) as tensor product of integrated Legendre polynomials in \( \lambda_1 \) and \( \lambda_2 \) (\( \lambda_3 \) is dependent) with maximal total polynomial degree \( p - 3 \).

The choice of integrated Legendre polynomials is less suggesting as for quadrilaterals as the integration domain is not of tensor-product form.

The following result will be important to compute element matrices with constant coefficients:

**Lemma 4.19.** For any non-degenerate triangle \( K \) and \( \beta_1, \beta_2, \beta_3 \in \mathbb{N} \),
\[
\int_K \lambda_1^{\beta_1} \lambda_2^{\beta_2} \lambda_3^{\beta_3} \, dx = 2|K| \cdot \frac{\beta_1! \beta_2! \beta_3!}{(\beta_1 + \beta_2 + \beta_3 + 2)!}.
\]

**Proof.** The first step amounts to a transformation of \( K \) to the “unit triangle” \( \hat{K} := \text{convex } \{ (0,0), (1,0), (1,1) \} \), which leads to
\[
\int_K \lambda_1^{\beta_1} \lambda_2^{\beta_2} \lambda_3^{\beta_3} \, dx = 2|K| \int_0^1 \int_0^{1-\xi_2} \xi_1^{\beta_1} \xi_2^{\beta_2} (1 - \xi_1 - \xi_2)^{\beta_3} \, d\xi_1 d\xi_2 =: 2|K| \cdot I.
\]

Note, that the barycentric coordinates evaluated on \( \hat{K} \) are the barycentric coordinates on the reference cell, which are \( \xi_1, \xi_2, 1 - \xi_1 - \xi_2 \). Transforming \( \hat{K} \) to \((0,1)^2\) by the
Duffy-transformation $\xi_1 = \hat{\xi}_1$, $\xi_2 = \hat{\xi}_1 (1 - \hat{\xi}_2)$ (see Fig. 4.10) we get

$$I = \int_0^1 \int_0^1 \xi_1^{\beta_1} (1 - \xi_2)^{\beta_1} \xi_2^{\beta_2} (1 - \xi_2 - \hat{\xi}_1 (1 - \hat{\xi}_2))^{\beta_3} (1 - \hat{\xi}_2) d\xi_1 d\xi_2$$

$$= \int_0^1 \xi_1^{\beta_1} (1 - \xi_1)^{\beta_3} d\xi_1 \int_0^1 \xi_2^{\beta_2} (1 - \xi_2)^{\beta_1 + \beta_3 + 1} d\xi_2 = B(\beta_1 + 1, \beta_3 + 1) B(\beta_2 + 1, \beta_1 + \beta_3 + 2)$$

where we used Euler’s beta functions

$$B(\alpha, \beta) := \int_0^1 t^{\alpha-1} (1 - t)^{\beta-1} dt, \quad 0 < \alpha, \beta < \infty.$$ 

Using the known formula $\Gamma(\alpha + \beta) B(\alpha, \beta) = \Gamma(\alpha) \Gamma(\beta)$, $\Gamma$ the Gamma function, we end up with

$$I = \frac{\Gamma(\beta_1 + 1) \Gamma(\beta_3 + 1)}{\Gamma(\beta_1 + \beta_3 + 2)} \cdot \frac{\Gamma(\beta_2 + 1) \Gamma(\beta_1 + \beta_3 + 2)}{\Gamma(\beta_1 + \beta_2 + \beta_3 + 3)}.$$ 

Then, $\Gamma(n) = (n - 1)!$ finishes the proof.

### 4.4 Conforming finite element basis on non-conforming meshes

In case of non-conforming meshes hanging nodes and hanging edges appear that will not carry basis functions. The associated shape functions do not contribute by a one-to-one relation to a basis function. However, they are needed to represent hut functions or basis functions identified to an edge on a parents edge.

**Example 4.20** (T-Matrices for an irregular mesh).

*Three cells with each three shape functions and four global basis functions. The hanging node is marked with $\circ$.*

\[
T_K = \begin{pmatrix}
1 & 2 & 3 & 4 \\
1 & 1/2 & 1/2 & \\
2 & & & 1 \\
3 & & & \\
\end{pmatrix}
\]

\[
T_L = \begin{pmatrix}
1 & 2 & 3 & 4 \\
1 & 1 & . & . \\
2 & & & 1 \\
3 & 1/2 & 1/2 & . \\
\end{pmatrix}
\]
Let the non-conforming mesh be generated by a refinement procedure from a conforming mesh. The interior basis functions vanishing on all edges follow as for conforming meshes. Only non-hanging nodes and non-hanging edges carry basis functions. For the integration routine we have to represent the basis function by shape functions on the cells of the mesh. For the cells in the support of such a basis function we can find parent cells such that on each edge in the interior of the support there are only two cells adjacent. On this conforming level we can construct the representation by $\mathbf{T}$ matrices as introduced in the former sections.

![Subdivision variants on the reference square.](image)

Figure 4.12: Subdivision variants on the reference square.

To represent the basis functions on each cell on the smallest level we can introduce a relation between the shape functions on different level. So let $K$ a parent cell of a cell $K'$. Then, we can represent each shape function defined on $K$ in the cell $K'$ as

$$N_{K,i}(\mathbf{x})|_{K'} = \sum_j (\mathbf{S}_{K'K})_{ji} N_{K',j}(\mathbf{x}),$$

where a $\mathbf{S}$ matrix is involved. The shape functions $N_{K,i}$ are defined as element shape function $\hat{N}_i$ on the reference element $\hat{K}$ and then transformed with the element map $\Phi_K$ to $K$. As $K'$ is a part of $K$ we can write a representation between $\hat{K}$ and $\hat{K}' = \Phi^{-1}_K K'$, where $\mathbf{x} \in \hat{K}'$

$$\hat{N}_i(\Phi_K^{-1} \mathbf{x}) = \sum_j (\mathbf{S}_{K'K})_{ji} \hat{N}_j(\Phi_K^{-1} \mathbf{x})$$

or

$$\hat{N}_i(\Phi_{K'} \mathbf{\xi}) = (\mathbf{S}_{K'K})_{ji} \hat{N}_j(\mathbf{\xi})$$

Obviously, we can also write $\mathbf{S}_{\hat{K}'\hat{K}}$ instead of $\mathbf{S}_{K'K}$. For each refinement variant we have such an $\mathbf{S}$-matrix. It simplifies when restricting to the those only with division ratio $\frac{1}{2}$, where we have that of Fig. 4.12 for quadrilateral cells.
(a) The even element shape function $\hat{N}_2(\xi)$.  

(b) $\hat{N}_2(\xi)$ evaluated on $(0, \frac{1}{2})$ mapped to $(0, 1)$.  

(c) $-\sqrt{6} \hat{N}_1(\xi)$ and $\frac{1}{4} \hat{N}_2(\xi)$. 

Figure 4.13: The representation of $\hat{N}_2(\xi)$ evaluated in $[0, \frac{1}{2}]$ by element shape functions.

**S-matrices in 1D**  The reference interval $\hat{I} = [0, 1]$ is subdivided in the left part $\hat{I}' = [0, \frac{1}{2}]$ and the right one $\hat{I}^* = [\frac{1}{2}, 1]$. So for the left part we have $\Phi_{\hat{K}'} = \frac{\xi}{2}$.

**Example 4.21.** If we take for instance the 1D element shape function (see Fig. 4.13) $\hat{N}_2(\xi) = \sqrt{6} \xi (\xi - 1)$ and we search $(S_{\hat{I}'\hat{I}})_{2,j}$ such that for $\xi \in \hat{I}$

$$\hat{N}_2 \left( \frac{\xi}{2} \right) = \sum_{j=0}^{2} (S_{\hat{I}'\hat{I}})_{2,j} \hat{N}_j(\xi).$$

We find

$$\hat{N}_2 \left( \frac{\xi}{2} \right) = \frac{\sqrt{6}}{4} \xi (\xi - 2) = -\frac{\sqrt{6}}{4} \hat{N}_1(\xi) + \frac{1}{4} \hat{N}_2(\xi),$$

and so

$$(S_{\hat{I}'\hat{I}})_{2,-} = \begin{pmatrix} 0 & -\frac{\sqrt{6}}{4} & \frac{1}{4} \end{pmatrix}.$$

In general the 1D S matrices can be obtained by evaluating the element shape functions on to $[0, 1]$ transformed Chebychev points $\xi_{\ell}$ and on the points $\Phi_{\hat{K}'} \xi_{\ell}$, and solving a (small) linear system. Note, that in case of a hierarchic basis $\{\hat{N}_j\}_j$ the 1D S matrix for a maximal polynomial degree $p$ is just a part of the 1D S for maximal polynomial degrees $q > p$. So the S matrices have just to be computed once for the maximal polynomial degree in the space.

**S-matrices in 2D**  For element shape functions with (functional) tensor product structure, i. e.,

$$\hat{N}_{kt} = \hat{N}_k \otimes \hat{N}_t.$$
For an anisotropic subdivision the 2D $S$ matrices are a tensor product like
\[ S_{\tilde{K} \tilde{K}} = S_{\tilde{P} \tilde{P}} \otimes I \]
where $I$ is the identity matrix.

For subdivision in both direction we have a tensor product like
\[ S_{\tilde{K} \tilde{K}'} = S_{\tilde{P} \tilde{P}} \otimes S_{\tilde{P} \tilde{P}}. \]

### 4.5 Local and global degrees of freedom

Let us collect the shape functions $N_{K,i}$ of a cell $K$ in the local trial space $\Pi_K$ which has a particular dimension. In the space $\Pi_K$ we have the from the reference element transformed polynomials.

Then we call a local degrees of freedom a linear functional $(C^\infty(K))^l \to \mathbb{R}$, if a set $\Sigma_K$ of local degrees of freedom provide a basis for the dual space $(\Pi_K)'$, i.e.
\[ \sharp \Sigma_K = \dim \Pi_K \text{ and } \forall v \in \Pi_K : l(v) = 0 \forall l \in \Sigma_K \Rightarrow v = 0. \]

The property that the local degrees of freedom fix the function in the local trial space is called unisolvence.

We can always choose the basis of $\Sigma_K$ such that
\[ l^K_m(N_{K,j}) = \delta_{mj}, \quad m,j \in \{1, \ldots, \dim \Pi_K\}. \tag{4.15} \]

**Example 4.22** (Linear finite elements). The local degrees of freedom can be chosen as point evaluations on the nodes
\[ l^K_j(v) := v(p_j(K)), \tag{4.16} \]
which are well-defined for $v \in L^\infty(K)$.

**Example 4.23** (Higher order finite elements on quadrilaterals). The local degrees of freedom of the four nodal shape functions can be chosen as (4.16). Then, for bounded functions $v$
\[ v_c(x) := v(x) - \sum_{j=1}^4 l^K_j(v)N_{K,j}(x) \]
is zero on all four nodes.

Note, that the integrated Legendre polynomials vanish on both end points for $i \geq 2$, i.e., $\hat{L}_i(\pm 1) = 0$, and so for any $j \in \mathbb{N}$
\[ \int_{-1}^1 \hat{L}_i'(\xi)\hat{L}_j'(\xi)d\xi = -\int_{-1}^1 \hat{L}_i(\xi)\hat{L}_j''(\xi)d\xi. \]
On the other hand

\[ \int_{-1}^{1} \hat{L}'(\xi) \hat{L}_j'(\xi) d\xi = \int_{-1}^{1} L_{i-1}(\xi)L_{j-1}(\xi) d\xi = \delta_{ij} \frac{2}{2j-1}. \]

So

\[ -2 \int_{0}^{1} \hat{N}_i(\xi) \hat{N}'_j(\xi) d\xi = -2 \sqrt{2j-1} \sqrt{2i-1} \int_{0}^{1} \hat{L}_i(2\xi - 1) \hat{L}_j''(2\xi - 1) d\xi \]

\[ = \frac{1}{2} \sqrt{(2j-1)(2i-1)} \int_{-1}^{1} \hat{L}_i(\xi) \hat{L}_j'(\xi) d\xi = \delta_{ij}. \]

To fix the trace on each edge \( E_i(K) \), \( i = 1, 2, 3, 4 \) in case that \( v_e \) would be a shape functions identified to this edge, we introduce for \( j = 0, \ldots, p-2 \)

\[ l_{e,i,j}^K(v) = -2 \int_{0}^{1} \frac{\hat{v}_e(\xi)}{v_e(\Phi_{E_i}(K)\xi)} \hat{N}'_{j+2}(\xi) d\xi \]

\[ = -2 \int_{E_i(K)} v_e(x) \hat{N}'_{j+2}(\Phi_{E_i}(K)x) |D\Phi_{E_i}(K)(\Phi_{E_i}^{-1}(K)x)|^{-1} dS(x), \quad (4.17) \]

where \( \Phi_{E_i}(K) \) is the mapping from the reference interval \([0, 1]\) to \( E_i(K) \) which is introduced as the mapping \( \Phi_K \) of the respective edge in \( \hat{K} \) to \( E_i(K) \) respecting the (global) orientation of \( E_i(K) \) (if the local \( \xi_1 \) or \( \xi_2 \) direction along the edge in \( \hat{K} \) is opposite to global orientation, \( \xi_1 \) or \( \xi_2 \), respectively, is replaced by \( 1 - \xi_1 \) or \( 1 - \xi_2 \)).

Then, in case \( v \in \Pi_K \) the function

\[ v_e(x) := v_e(x) - \sum_{i=1}^{4} \sum_{j=0}^{p-2} \frac{l_{e,i,j}^K(v) N_{e,i,j}(x)}{\hat{N}_{e,i,j}(\Phi_{E_i}(K)x)} \]

is vanishing on all edges of \( K \).

Similarly, we define for \( i, j = 0, \ldots, p-2 \),

\[ l_{c,i,j}^K(v) = 4 \int_{\hat{K}} \frac{\hat{v}_e(\xi)}{v_e(\Phi_{E_i}(K)\xi)} \hat{N}'_{i+2}(\xi_1) \hat{N}'_{j+2}(\xi_2) d\xi \]

\[ := 4 \int_{\hat{K}} v_e(x) \hat{N}'_{i+2}(\xi_1) \hat{N}'_{j+2}(\xi_2) |D\Phi_{K}(\Phi_K^{-1}x)|^{-1} d\Sigma(x), \]

with \( \xi = \Phi_{K}^{-1}x \).

**Definition 4.24.** A **finite element** is a triple \((K, \Pi_K, \Sigma_K)\) such that

(i) \( K \) is a cell of a mesh \( M \) of the computational domain \( \Omega \subset \mathbb{R}^d \).

(ii) \( \Pi_K \subset (C^\infty(\hat{K}))^l \) is the trial space with \( \dim \Pi_K < \infty \).
(iii) $\Sigma_K$ is a set of local degrees of freedom.

A finite element is called \textbf{V-conforming}, if

(iv) for any face $F$ of $K$ the degrees of freedom localized on $F$ uniquely determine the natural trace $R u|_F$ of a $u \in \Pi_K$ onto $F$.

**Definition 4.25.** Let $K$ be a cell of a mesh $\mathcal{M}$ and $F$ be a face/edge/node of the mesh that is contained in $K$. Given a local trial space $\Pi_K$, $\Pi_K \subset (C^\infty(K))^l$, and a set $\Sigma_K$ of local degrees of freedom, a linear functional $l \in \Sigma_K$ is called \textbf{localized/supported on $F$} or \textbf{associated with $F$}, if

$$l(v) = 0 \quad \forall v \in (C^\infty(K))^l, \quad \text{supp}(v) \cap F = \emptyset.$$ 

**Notation 4.26.** 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”.

To obtain $V$-conforming basis functions the local degrees of freedom localised on a face/edge/node on the cells sharing face/edge/node has to be matched to \textbf{global degrees of freedoms}. We have unisolvence, i.e., the global degrees of freedom determine the basis functions.

**Remark 4.27.** 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 $V$.

The construction of finite element spaces can be started from local/global shape functions or equivalently via the approach of degrees of freedom (“dual view”, see Fig. 4.14).

![Diagram](image.png)

**Figure 4.14:** Duality of degrees of freedom and shape functions

The concepts of degrees of freedom may be helpful to obtain finite element spaces which are $C^m$ continuous, $m > 0$, or where only the tangential components or normal component of vector fields are continuous ($H(\text{curl}, \Omega)$ or $H(\text{div}, \Omega)$ conforming.

**References**


5 Basic Finite Element Theory

5.1 Discretisation error is bounded by the interpolation error

If the discrete variational problem is well-posed (see Theorem 3.69) the discretisation error is bounded “quasi optimally” by the best approximation error:

\[
\| u - u_n \|_V \leq \left( 1 + \frac{\| b \|_{V \times V \to \mathbb{R}}} {\gamma_n} \right) \inf_{w_n \in W_n} \| u - w_n \|_V ,
\]

Following the usual way one estimates the best approximation error by the error of the interpolation of the solution \( u \) to the finite trial space \( W_n \)

\[
\inf_{w_n \in W_n} \| u - w_n \|_V \leq \| u - I_h u \|_V ,
\]

with \( I_h : V \to W_n \) or at least \( I_h : V_S \to W_n \), \( V_S \subset V \) is some interpolation operator. The latter interpolation operators can be used if the solution is in fact in a smaller subspace \( V_S \) of \( V \).

So, its all to define some interpolation of \( u \) in the discrete space \( W_n \) and to estimate its error.

5.2 The Bramble-Hilbert lemma

A generalisation of Poincaré’s inequality (Lemma 3.65) and a best approximation of polynomials in Sobolev spaces similarly to the point wise statement of Taylor’s theorem is the

**Lemma 5.1** (Bramble-Hilbert lemma). If \( \Omega \subset \mathbb{R}^d \) is a bounded Lipschitz-domain and \( m \in \mathbb{N} \), then

\[
\exists \gamma = \gamma(m, \Omega) > 0 : \inf_{p \in \mathcal{P}_{m-1}(\Omega)} \| v - p \|_{H^m(\Omega)} \leq \gamma |v|_{H^m(\Omega)} \quad \forall v \in H^m(\Omega) .
\]

**Proof.** The proof can be found in [23]. \( \square \)

In other words, the norm on the quotient space \( H^m(\Omega)/\mathcal{P}_{m-1}(\Omega) \) is equivalent to the seminorm \( |\cdot|_{H^m(\Omega)} \).

Note, that for \( \ell > m \) we have

\[
\inf_{p \in \mathcal{P}_{\ell-1}(\Omega)} \| v - p \|_{H^m(\Omega)} \leq \inf_{p \in \mathcal{P}_{m-1}(\Omega)} \| v - p \|_{H^m(\Omega)} \leq \gamma |v|_{H^m(\Omega)} .
\]

There is also a version of the Bramble-Hilbert lemma for tensor-product polynomials [24]:

90
Lemma 5.2. If \( \Omega \subset \mathbb{R}^d \) is a bounded Lipschitz-domain and \( m \in \mathbb{N} \), then
\[
\exists \gamma = \gamma(m, \Omega) > 0 : \inf_{p \in \mathcal{Q}_{m-1}(\Omega)} \| v - p \|_{H^m(\Omega)} \leq \gamma \left( \sum_{i=1}^{d} \left\| \frac{\partial^m v}{\partial \xi_i^m} \right\|_{L^2(\Omega)}^2 \right)^{1/2} \forall v \in H^m(\Omega).
\]

5.3 The interpolation operator of Raviart-Thomas

Let the finite element space consist of \( V \)-conforming finite elements.

The interpolation operator of Raviart-Thomas is defined as
\[(l_n v)(x) := \sum_{j=1}^{N} l_j(v) b_j(x)\]
where \( l_j \) are the global degrees of freedom. As the global degrees of freedom result by matching of local degrees of freedom we have for the restriction on one cell \( K \)
\[(l_n v)(x) := \sum_{j=1}^{\dim \Pi_K} l^K_j(v) N_{K,j}(x), \quad x \in K.\]

5.4 The interpolation error estimates on simplices

We need three properties of \( l_n \) on simplices:

- The shape functions are polynomials on \( K \), \( \Pi_K = \mathcal{P}_m(K) \).
- The interpolation operator \( l_n \) preserves polynomials, i.e., for \( p \in \mathcal{P}_m(K) \)
\[l_n p = l_n( \sum_{i=1}^{\dim \Pi_K} \alpha_i N_{K,i} ) = \sum_{j=1}^{\dim \Pi_K} \sum_{i=1}^{\dim \Pi_K} \alpha_i l_j(N_{K,i}) N_{K,j} = \sum_{j=1}^{\dim \Pi_K} \alpha_j N_{K,j} = p.\]
- The interpolation operator \( l_n \) is continuous for functions in \( H^t(K) \) for some \( t \in \mathbb{R} \).

Since the local degrees of freedom for linear finite elements as well as those of higher orders involve point evaluations we have \( t = 1 \) for \( d = 1 \) and \( t = 2 \) for \( d = 2 \). We have the following theorem [25, Thm. 6.2].

Theorem 5.3. If and only if \( d/2 < m \), \( m \in \mathbb{N} \), then \( H^m(\Omega) \) is continuously embedded in \( C^0(\Omega) \).
With the Bramble-Hilbert-Lemma, the triangle inequality and \( t \leq m+1 \) we can estimate
\[
\|u - I_n v\|_{H^1(K)} \leq \inf_{P_m(K)} \| (u - p) - I_n(u - p)\|_{H^1(K)}
\]
\[
\leq \inf_{P_m(K)} \left( \| u - p\|_{H^1(K)} + \| I_n(u - p)\|_{H^1(K)} \right)
\]
\[
\leq (1 + C(m,K)) \inf_{P_m(K)} \| u - p\|_{H^1(K)}
\]
\[
\leq \begin{cases} 
(1 + C(m,K)) \inf_{P_{t-1}(K)} \| u - p\|_{H^1(K)} & \leq \gamma(t,K)(1 + C(m,K)) |u|_{H^1(K)}, \\
(1 + C(m,K)) \inf_{P_m(K)} \| u - p\|_{H^{m+1}(K)} & \leq \gamma(m,K)(1 + C(m,K)) |u|_{H^{m+1}(K)}.
\end{cases}
\]

\[\triangleright\] How does the constant in the estimate depend on \( K \), especially the size of \( K \) ?

To answer this question we consider the interpolation operator on the usual reference element, which has a fixed size and angles.

**Interpolation operator on the reference element** We have defined the shape functions \( N_{K,j} \) as from a reference element transformed element shape functions \( \hat{N}_j(\xi) = \hat{N}_j(\xi) \).

Equivalently, we can define local degrees of freedom on \( \hat{K} \) with
\[
\hat{l}_j(\hat{v}) = l^K_j(v)
\]
where \( \hat{v}(\xi) := v(\Phi_K(\xi)) \) is the pullback of \( v \).

Then, we can define an interpolation operator on \( \hat{K} \) as
\[
(\hat{I} \hat{v})(\xi) := \sum_{j=1}^{\dim \Pi_K} \hat{l}_j(\hat{v}) \hat{N}_j(\xi), \tag{5.2}
\]
which equalise the transformed interpolation operator
\[
\hat{I}_n \hat{v}(\xi) := (I_n v)(x(\xi)) = \sum_{j=1}^{\dim \Pi_K} l^K_j(v) N_{K,j}(x(\xi)) = \sum_{j=1}^{\dim \Pi_K} \hat{l}_j(\hat{v}) \hat{N}_j(\xi) = (\hat{I} \hat{v})(\xi). \tag{5.3}
\]

**Estimate on the reference element** Repeating the steps, we went on \( K \), on the reference element \( \hat{K} \) we obtain estimates for the interpolation error for \( \hat{I} \). Let us measure the error in the \( H^r \)-norm with \( r \leq t \leq m+1 \) (\( t \geq 1 \) for \( d = 1 \) and \( t \geq 2 \) for \( d = 2,3 \)).

\[
\|\hat{u} - \hat{I} \hat{u}\|_{H^r(\hat{K})} = \inf_{p \in P_m(\hat{K})} \| (\hat{u} - p) - \hat{I}(\hat{u} - p)\|_{H^r(\hat{K})}
\]
\[
\leq (1 + C) \inf_{p \in P_m(\hat{K})} \| \hat{u} - p\|_{H^1(\hat{K})} \leq \gamma(m)(1 + C)|\hat{u}|_{H^1(\hat{K})}.
\]
Transformation techniques  Let us transform Sobolev norms between the reference triangle $\tilde{K}$ and the triangle $K$ and vice-versa.

**Lemma 5.4.** If $\Phi_K : \tilde{K} \to K$ is an affine mapping $\xi \mapsto F_K \xi + \tau$, then, for all $m \in \mathbb{N}_0$,

$$|\hat{u}|_{H^m(\tilde{K})} \leq \binom{m + d}{d} d^m \|F_K\|^m |\det(F_K)|^{-1/2} |u|_{H^m(K)},$$

$$|u|_{H^m(K)} \leq \binom{m + d}{d} d^m \|F_K^{-1}\|^m |\det(F_K)|^{1/2} |\hat{u}|_{H^m(\tilde{K})},$$

with $\|F\|_K$ denoting the matrix norm of $F_K$ associated with the Euclidean vector norm.

**Proof.** Without loss of generality we can assume that $u \in C^\infty(\tilde{K})$. Let $\alpha \in \mathbb{N}^d$ with $|\alpha| = m$, $m \in \mathbb{N}_0$. Remember that the Gateaux derivative in direction $\delta$ is

$$D \hat{u}(\xi)(\delta) := \lim_{\varepsilon \to 0} \frac{\hat{u}(\xi + \varepsilon \delta) - \hat{u}(\xi)}{\varepsilon}.$$ 

So, the $m$-th Gateaux-derivative $D^m : \mathbb{R}^d \times \cdots \times \mathbb{R}^d \to \mathbb{R}$ allows to express

$$\partial^\alpha \hat{u} = D^m \hat{u}(\xi)(\delta^1, \ldots, \delta^m),$$

with $\delta^1 = \cdots = \delta^{\alpha_1} = e_1$, $\delta^{\alpha_1+1} = \cdots = \delta^{\alpha_1+\alpha_2} = e_2$, etc. Here, $e_k$ designates the $k$-th unit vector in $\mathbb{R}^d$. This means for example

$$\partial_{\xi_1} \partial_{\xi_2}^2 \hat{u}(\xi) = D^3 \hat{u}(\xi)(e_1, e_2, e_2).$$

So, we deduce that

$$|\partial^\alpha \hat{u}(\xi)| \leq \|D^m \hat{u}(\xi)\| := \sup\{D^m \hat{u}(\xi)(\delta^1, \ldots, \delta^m), \delta^k \in \mathbb{R}^d, |\delta^k| = 1\},$$

(instead of the particular choice of $\delta^1, \ldots, \delta^m$ we take the supremum) which implies

$$|\hat{u}|^2_{H^m(\tilde{K})} = \sum_{|\alpha|=m} \int_{\tilde{K}} |\partial^\alpha \hat{u}(\xi)|^2 \, d\xi \leq \binom{m + d}{m} \int_{\tilde{K}} \|D^m \hat{u}(\xi)\|^2 \, d\xi. \quad (5.4)$$

The chain rule gives

$$D^m \hat{u}(\xi)(\delta^1, \ldots, \delta^m) = D^m u(\Phi_K(\xi))(F_K \delta_1, \ldots, F_K \delta_m). \quad (5.5)$$

By linearity of the derivative we have

$$\|D^m \hat{u}(\xi)\| \leq \|F_K\|^m \|D^m u(\Phi(\xi))\|.$$ 

Next, we use the transformation formula for multidimensional integrals and apply it to (5.4):

$$|\hat{u}|^2_{H^m(\tilde{K})} \leq \binom{m + d}{m} \int_{\tilde{K}} \|F_K\|^{2m} \|D^m u(\xi)\|^2 |\det(F_K)|^{-1} \, d\xi \leq \binom{m + d}{m} \|F_K\|^{2m} |\det(F_K)|^{-1} \int_{\tilde{K}} \|D^m u(\xi)\|^2 \, d\xi. \quad (5.6)$$

93
Then, observe that
\[
\|D^m u(x)\| = \sup\{D^m u(x)(\delta^1, \ldots, \delta^m), \delta^k \in \mathbb{R}^d, |\delta^k| = 1\}
\leq \sup\{\sum_{\alpha_1=1}^{d} \cdots \sum_{\alpha_m=1}^{d} |D^m u(\delta^1_{\alpha_1}e_{\alpha_1}, \ldots, \delta^m_{\alpha_m}e_{\alpha_m})|, \delta^k \in \mathbb{R}^d, |\delta^k| = 1\}
\leq \sum_{\alpha_1=1}^{d} \cdots \sum_{\alpha_m=1}^{d} |D^m u(e_{\alpha_1}, \ldots, e_{\alpha_m})| \leq d^m \max\{|\partial^\alpha u(x)|,
\]
Finally,
\[
\int_K \|D^m u(x)\|^2 dx \leq d^m \int_K \max\{|\partial^\alpha u(x)|^2 dx \leq d^{2m} \sum_{|\alpha|=m} \int_K |\partial^\alpha u(x)|^2 dx = d^{2m} \|u\|_{H^m(K)}^2.
\]

\[\square\]

**Remark 5.5.** If \(\Phi\) is a general \(C^\infty\)-diffeomorphism \(\hat{K} \rightarrow K\), then the analogue of (5.5) will involve derivatives of \(u\) from \(Du\) up to \(D^m u\) and derivatives \(D\Phi_K\) up to \(D^m \Phi_K\).

Thus, in order to estimate the Sobolev-seminorm \(|\hat{u}|_{H^m(\hat{K})}\), we have to resort to the full Sobolev norm \(\|u\|_{H^m(K)}\) and vice versa.

**Estimate on a cell** \(K\). With the transformation techniques we can relate the interpolation error on \(K\) with that of the push-backed function
\[
\|u - I_n u\|_{H^r(K)}^2 = \sum_{\ell=0}^{r} |u - I_n u|_{H^r(K)}^2
\leq \sum_{\ell=0}^{r} \left(\ell + \frac{d}{2}\right)^2 d^{2\ell} \|\Phi_K^{-1}\|^{2\ell} |\det(\Phi_K)| \left\|\hat{u} - \hat{I_n u}\right\|_{H^r(\hat{K})}^2
\leq C(r) \left(\frac{r + d}{d}\right)^2 d^{2r} \|\Phi_K^{-1}\|^{2r} |\det(\Phi_K)| \left\|\hat{u} - \hat{I_n u}\right\|_{H^r(\hat{K})}^2.
\]

Using the estimate on the reference element we get
\[
\|u - I_n u\|_{H^r(K)} \leq C(m) \left(\frac{r + d}{d}\right) d^{r} \|\Phi_K^{-1}\|^{r} |\det(\Phi_K)|^{1/2} |\hat{u}|_{H^r(\hat{K})},
\]

and transformed back to \(K\)
\[
\|u - I_n u\|_{H^r(K)} \leq C(m) \left(\frac{r + d}{d}\right) \left(\frac{t + d}{d}\right) d^{r+t} \|\Phi_K^{-1}\|^{r} \|\Phi_K\|^{t} |u|_{H^t(K)}.
\]

The estimate depends on the size and shape of the triangle through \(\Phi_K\). We will see in the following that \(\|\Phi_K\| = O(h_K)\) and \(\|\Phi_K^{-1}\| = O(h_K^{-1})\), where \(h_K\) is the cell diameter.

\[\triangleright\] Convergence in terms of \(h_K\) is only expected if the solution has higher regularity (even for \(d = 1\) where the interpolation operator is continuous in \(H^t(K)\) for \(t = 1\)).

94
Estimates of the simplicial element mapping Let us consider an affine equivalent simplicial triangulation $\mathcal{M}$, see Def. 4.5. We fix a reference simplex $\hat{K}$ and find affine mappings $\Phi_K : \hat{K} \mapsto K$, $\Phi_K(\xi) := F_K \xi + \tau_K$ for each $K \in \mathcal{M}$. In light of the general strategy outlined above, we have to establish bounds for $\|F_K\|, \|F_K^{-1}\|, |\det(F_K)|$, and $|\det(F_K)|^{-1}$ that depend on controllable geometric features of $\mathcal{M}$.

**Definition 5.6.** Given a cell $K$ of a mesh $\mathcal{M}$ we define its diameter $h_K := \sup\{|x - y|, x, y \in K\}$, and the maximum radius of an inscribed ball $r_K := \sup\{r > 0 : \exists x \in K : |x - y| < r \Rightarrow y \in K\}$. The ratio $h_K/r_K$ is called the shape regularity measure $\rho_K$ of $K$.

![Figure 5.1: Diameter $h_K$ and $r_K$ for a triangular cell](image)

**Lemma 5.7.** If $\hat{K}, K \subset \mathbb{R}^d$, $d = 2, 3$, are a generic non-degenerate simplices and $\Phi_K : \hat{K} \mapsto K$, $\Phi_K(\xi) := F_K \xi + \tau$, the associated bijective affine mapping, then

$$
\left(\frac{h_K}{h_{\hat{K}}}\right)^d \rho_{\hat{K}}^{1-d} = \frac{h_K^d}{h_{\hat{K}}^d} \leq |\det(F_K)| = \frac{|K|}{|\hat{K}|} \leq \frac{h_K^d}{h_{\hat{K}}^d} \rho_{\hat{K}}^{d-1} = \left(\frac{h_K}{h_{\hat{K}}}\right)^d \rho_{\hat{K}}^{d-1},
$$

(5.9)

$$
\|F_K\| \leq \frac{h_K}{2r_{\hat{K}}} = \frac{1}{2} \rho_{\hat{K}} \frac{h_K}{h_{\hat{K}}}, \quad \|F_K^{-1}\| \leq \frac{h_{\hat{K}}}{2r_K} = \frac{1}{2} \rho_K \frac{h_{\hat{K}}}{h_K}.
$$

(5.10)

**Proof.** The inequalities (5.9) can be concluded from the volume formula for simplices by elementary geometric considerations.

Write $\zeta \in \hat{K}$ for the center of the largest inscribed ball of $\hat{K}$. Then estimates (5.10) follow from

$$
\|F_K\| = \sup\{|F_K(\zeta) - \xi|, |\xi| = 1\} = \frac{1}{2} r_{\hat{K}}^{-1} \sup\{|F_K(\hat{K} - \zeta)|, |\hat{K} - \zeta| = 2r_{\hat{K}}\}
$$

= $\frac{1}{2} r_{\hat{K}}^{-1} \sup\{|\Phi_K(\zeta) - \Phi_K(\zeta)|, |\zeta - \zeta| = 2r_{\hat{K}}\} \leq h_K/2r_{\hat{K}},$

because both $\Phi(\xi)$ and $\Phi(\xi)$ lie inside $K$. A role reversal of $\hat{K}$ and $K$ establishes the other estimate. \qed
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 5.8.** If the smallest angle of a triangle $K$ is bounded from below by $\alpha > 0$, then

$$\sin(\alpha/2)^{-1} \leq \rho_K \leq 2 \sin(\alpha/2)^{-1}.$$ 

![Figure 5.2: Angle condition for shape regularity of a triangle](image)

**Proof.** It is immediate from Fig. 5.2 that

$$\frac{1}{2} h_K \sin(\alpha/2) \leq l \sin(\alpha/2) = r_K \leq h_K \sin(\alpha/2).$$

Lemma 5.7 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 5.9.** 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. 5.6 have been used. Moreover, the **quasi-uniformity measure** of $\mathcal{M}$ is the quantity

$$\mu_\mathcal{M} := \max \{ h_K / h_{K'}, K, K' \in \mathcal{M} \} = \max \{ h_K, K \in \mathcal{M} \} \cdot \min \{ h_K^{-1}, K \in \mathcal{M} \}.$$ 

**Remark 5.10.** 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 $\pi$. 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

\[
\hat{K} := \text{convex}\left\{\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}\right\} \text{ for } d = 2,
\]

\[
\hat{K} := \text{convex}\left\{\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}\right\} \text{ for } d = 3.
\]

**Corollary 5.11.** Let $\mathcal{M}$ be a simplicial triangulation and choose the reference simplex according to (5.11) and (5.12), respectively. Then the affine mappings $\Phi_K : \hat{K} \mapsto K$, $\Phi_K(\xi) := F_K \xi + \tau_K$, $K \in \mathcal{M}$, satisfy

\[
\frac{\rho_{\mathcal{M}}^{1-d}}{\mu_{\mathcal{M}}^d} h_{\mathcal{M}}^d \leq |\det(F_K)| \leq h_{\mathcal{M}}^d, \quad \|F_K\| \leq h_{\mathcal{M}}, \quad \|F_K^{-1}\| \leq \rho_{\mathcal{M}} \mu_{\mathcal{M}}^{-1} h_{\mathcal{M}}^{-1}.
\]

**Interpolation error estimate in the mesh** The interpolation error can be decomposed into the contributions from the all triangles of $\mathcal{M}$, using (5.8) and Corollary 5.11 and summing the errors from each triangle we get

**Theorem 5.12.** Let $I_n$ stand for the finite element interpolation operator belonging to the finite element space $S_m(\mathcal{M})$ on a simplicial mesh $\mathcal{M}$. Then, for $2 \leq t \leq m+1$, $0 \leq r \leq t$

\[
\exists \gamma = \gamma(t, r, m, \rho_{\mathcal{M}}, \mu_{\mathcal{M}}) : \|u - I_n u\|_{H^r(\Omega)} \leq \gamma h_{\mathcal{M}}^{t-r} |u|_{H^t(\Omega)} \quad \forall u \in H^t(\Omega).
\]

### 5.5 A-priori error estimates for finite elements

Now, we are in the position to conclude in an a-priori error estimate for the discretisation error of the finite elements solution $u_h$. Applying (3.38), (5.1) and Theorem 5.12 we end up with

\[
\|u - u_h\|_{H^1(\Omega)} \leq \gamma h_{\mathcal{M}}^{t-1} |u|_{H^t(\Omega)} \quad \text{for } 2 \leq t \leq m+1 \text{ and } u \in H^t(\Omega),
\]

where $\gamma = \gamma(\Omega, \gamma_n, \|b\|, \rho_{\mathcal{M}}, \mu_{\mathcal{M}})$.

Let us discuss this a priori finite element discretization error estimate:

1. The estimate (5.13) 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 (3.1) 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 $\Omega$ we can expect $u \in H^{r+2}(\Omega)$.

**Example 5.13.** For $d = 1$ we have $u \in H^{r+2}(\Omega)$, if $f \in H^r(\Omega)$. 

97
2. The bound from (5.13) 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 $\Omega$. They are assumed to be uniformly shape-regular, that is,

$$\exists \gamma > 0 : \rho_{\mathcal{M}_n} < \gamma \quad \forall n \in \mathbb{N}.$$ 

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 (5.13) can be expressed by

$$\|u - u_n\|_{H^1(\Omega)} = O(h_{\mathcal{M}_n}^{t-1}) \quad \text{for} \quad n \to \infty. \quad (5.14)$$

If (5.14) 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 5.14.** With considerable extra effort, more sophisticated best approximation estimates can be derived: for $m,t \geq 1$ we have

$$\inf_{v_n \in S_m(\mathcal{M})} \|u - v_n\|_{H^1(\Omega)} \leq \gamma(\rho_{\mathcal{M}},\mu_{\mathcal{M}}) \left( \frac{h_{\mathcal{M}}}{m} \right)^{\min\{m+1,t\}-1} \|u\|_{H^t(\Omega)}. \quad (5.15)$$

This paves the way for a-priori error estimates for the $p$-version of $H^1$-conforming elements.

### 5.6 Duality techniques

Let us deal with the variational problem

$$b(u,v) := \int_\Omega \langle A \nabla u, \nabla v \rangle \, dx = \int_\Omega fv \, dx \quad \forall v \in H^1_0(\Omega). \quad (5.16)$$

and its Galerkin discretization based on the finite element space $S_m(\mathcal{M})$ 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 Sec. 3.9 (inf-sup-conditions, quasi-optimality) and will rely on particular techniques for elliptic boundary value problems.

**Assumption 5.15.** We assume that (5.16) is 2-regular, that is, all $u \in H^1_0(\Omega)$ with $-\text{div}(A \nabla u) \in L^2(\Omega)$ satisfy

$$u \in H^2(\Omega) \quad \text{and} \quad \|u\|_{H^2(\Omega)} \leq \gamma \|\text{div}(A \nabla u)\|_{L^2(\Omega)},$$

with a constant $\gamma = \gamma(A,\Omega) > 0$ independent of $u$. 

98
We write \( u_n \) for the unique solution of the discrete variational problem

\[
    u_n \in S_m(M) \cap H^1_0(\Omega) : \quad b(u_n, v_n) = \int_\Omega f v_n \, dx \quad \forall v_n \in S_m(M).
\]

Write \( u \in H^1_0(\Omega) \) for the exact solution of (5.16) and \( e_h := u - u_n \in H^1_0(\Omega) \) for the discretization error. From Sect. 3.11 we recall the Galerkin orthogonality (3.36)

\[
    b(e_h, v_n) = 0 \quad \forall v_n \in S_m(M).
\]

The solution \( w \in H^1_0(\Omega) \) of the dual linear variational problem

\[
    w \in H^1_0(\Omega) : \quad b(w, v) = \int_\Omega e_h v \, dx \quad \forall v \in H^1_0(\Omega),
\]

(5.17) will be a solution of the elliptic boundary value problem

\[
    -\text{div}(A \text{grad} w) = e_h \text{ in } \Omega, \quad w = 0 \text{ on } \Gamma.
\]

Since \( e_h \in L^2(\Omega) \), by Assumption 5.15 we know

\[
    w \in H^2(\Omega), \quad \|w\|_{H^2(\Omega)} \leq \gamma \|e_h\|_{L^2(\Omega)},
\]

(5.18) with \( \gamma = \gamma(\Omega, A) > 0 \).

Next, we plug \( v = e_h \) into (5.17) and arrive at

\[
    \|e_h\|^2_{L^2(\Omega)} = b(w, e_h) = \inf_{v_n \in S_m(M)} b(w - v_n, e_h),
\]

where Galerkin orthogonality came into play. We may now plug in \( v_n := I_n w \), where \( I_n \) is the finite element interpolation operator for \( S_m(M) \). Then we can use the continuity of \( b \) in \( H^1(\Omega) \) and the interpolation error estimate of Thm. 5.12 for \( r = 1 \) and \( t = 2 \):

\[
    \|e_h\|^2_{L^2(\Omega)} \leq b(w - I_n w, e_h) \leq \gamma \|w - I_n w\|_{H^1(\Omega)} \cdot \|e_h\|_{H^1(\Omega)} \leq \gamma h_M \|w\|_{H^2(\Omega)} \cdot \|e_h\|_{H^1(\Omega)}.
\]

Here, the final constant \( \gamma \) will depend on \( A, m, \rho_M, \) and \( \mu_M \), but not on \( u \) or \( u_n \).

Eventually, we resort to the 2-regularity in the form of estimate (5.18) and cancel one power of \( \|e_h\|_{L^2(\Omega)} \).

This technique is known as duality technique, because it relies on the dual variational problem (5.17). Sometimes the term “Aubin-Nitsche trick” can be found. Summing up we have proved the following result:

**Theorem 5.16.** Assuming 2-regularity according to Assumption 5.15, we obtain

\[
    \|u - u_n\|_{L^2(\Omega)} \leq \gamma h_M \|u - u_n\|_{H^1(\Omega)},
\]

where the constant \( \gamma > 0 \) depends on \( \Omega, A, m, \rho_M, \mu_M \).
Remark 5.17. Thm. 5.16 tells us that under suitable assumptions in the h-version of finite elements we can gain another power of $h_M$ 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 $h_M \to 0$.

What remains to be settled is whether Assumption 5.15 is reasonable. This is part of elliptic regularity theory. In particular, we have the following result [26]

Theorem 5.18. If the computational domain $\Omega \subset \mathbb{R}^d$ is convex or has $C^1$-boundary and $A \in C^1(\Omega)$, then the elliptic boundary value problem belonging to (5.16) is 2-regular.

5.7 Estimates for quadrature errors

As explained in Sect. 4.3.2 and 4.3.4, usually the finite element discretization of (3.10) or (3.12) 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.

5.7.1 Abstract estimates

We consider a linear variational problem (LVP) on a Banach space $V$

$$u \in V : \quad b(u,v) = (f,v)_{V' \times V}, \quad \forall v \in V,$$

with bilinear form $b \in L(V \times V, \mathbb{R})$ satisfying the inf-sup conditions (IS1), (IS2) and $f \in V'$, see Sect. 3.9. Existence and uniqueness of a solution $u \in V$ are guaranteed by Thm. 3.46.

Based on $V_n \subset V$, dim($V_n$) < $\infty$, we arrive at the discrete variational problem (DVP), see Sect. 3.6.

$$u_n \in V_n : \quad b(u_n,v_n) = (f,v_n)_{V' \times V}, \quad \forall v_n \in V_n.$$
We assume the discrete inf-sup condition (DIS) to be satisfied, which implies existence and uniqueness of $u_n$.

From an abstract point of view the application of numerical quadrature and an inexact boundary approximation in a finite element context means that the discrete variational problem will suffer a perturbation

$$\tilde{u}_n \in V_n : \quad \tilde{b}(\tilde{u}_n, v_n) = \langle \tilde{f}, v_n \rangle_{V' \times V}, \quad \forall v_n \in V_n,$$

with a bilinear form $\tilde{b} \in L(V_n \times V_n, \mathbb{R})$ and $\tilde{f} \in V'_n$. The perturbation destroys Galerkin orthogonality and leads to extra terms in the discretization error estimate of Cor. 3.74.

**Theorem 5.19** (First Strang’s lemma). Beside the assumptions on $b$ and $\tilde{b}$ stated above we demand that $\tilde{b}$ satisfies (DIS) with constant $\gamma_n$. Then, (5.19) will have a unique solution $\tilde{u}_n \in W_n$, which satisfies the a-priori error estimate

$$\|u - \tilde{u}_n\|_V \leq \gamma \left( \inf_{w_n \in W_n} \left( \|u - w_n\|_V + \sup_{v_n \in V_n} \frac{|b(w_n, v_n) - \tilde{b}(w_n, v_n)|}{\|v_n\|_V} \right) + \sup_{v_n \in V_n} \frac{|\langle f, v_n \rangle_{V' \times V'} - \langle \tilde{f}, v_n \rangle_{V' \times V'}|}{\|v_n\|_V} \right),$$

with $\gamma = \gamma(\|b\|, \gamma_n) > 0$

**Proof.** Similarly to the proof of quasi-optimality in Theorem 3.69 we use triangle inequality and (DIS) to estimate

$$\|u - \tilde{u}_n\|_V \leq \|u - w_n\|_V + \|w_n - \tilde{u}_n\|_V$$

$$\leq \|u - w_n\|_V + \frac{1}{\gamma_n} \sup_{v_n \in V_n \setminus \{0\}} \left| \tilde{b}(w_n - \tilde{u}_n, v_n) \right|.$$

With

$$\tilde{b}(w_n - \tilde{u}_n, v_n) = (b(u, v_n) - \tilde{b}(\tilde{u}_n, v_n)) + (\tilde{b}(w_n, v_n) - b(w_n, v_n)) + b(w_n - u, v_n)$$

$$= (\langle f, v_n \rangle_{V' \times V'} - \langle \tilde{f}, v_n \rangle_{V' \times V'}) + (\tilde{b}(w_n, v_n) - b(w_n, w_n)) + b(w_n - u, v_n),$$

the continuity of $b$ and as $w_n \in W_N$ has been arbitrary we conclude in the statement of the lemma.

The two terms

$$\sup_{v_n \in V_n} \frac{|b(w_n, v_n) - \tilde{b}(w_n, v_n)|}{\|v_n\|_V}, \quad \sup_{v_n \in V_n} \frac{|\langle f, v_n \rangle_{V' \times V'} - \langle \tilde{f}, v_n \rangle_{V' \times V'}|}{\|v_n\|_V},$$

are called **consistency (error) terms**. They have to be tackled, when we aim to gauge the impact of numerical quadrature or inexact boundary representation quantitatively.
Remark 5.20. Note, that ellipticity of \( \tilde{b} \)
\[
\tilde{b}(v_n, v_n) \geq \gamma_1 \| v \|_V
\]
for some positive constant \( \gamma_1 \) implies (DIS). This property for the perturbed bilinear form is called \textbf{h-ellipticity}. In the h-version of finite elements (mesh refinement) we want \( \gamma_1 \) to be independent of the meshwidth (“uniform h-ellipticity”).

Remark 5.21. If \( \tilde{b} \) is still an continuous bilinear form on \( V \times V \), then the estimate of the theorem can be simplified in the following way:
\[
\| u - \hat{u}_n \|_V \leq \| u - w_n \|_V + \| w_n - \hat{u}_n \|_V
\]
\[
\leq \| u - w_n \|_V + \frac{1}{\gamma_n} \sup_{v_n \in V_n \setminus \{0\}} \left| \frac{\tilde{b}(w_n - u + \hat{u}_n, v_n)}{\| v_n \|_V} \right|
\]
\[
\leq \left( 1 + \frac{\| \tilde{b} \|}{\gamma_n} \right) \| u - w_n \|_V + |R(u)|
\]

with the residual term
\[
R(u) = \sup_{v_n \in V_n} \frac{\tilde{b}(u, v_n) - \tilde{f}(v_n)}{\| v_n \|_V}.
\]

5.7.2 Uniform h-ellipticity

Let us consider the variational problem
\[
b(u, v) := \int_\Omega \langle A \grad u, \grad v \rangle \, dx = \int_\Omega f v \, dx \quad \forall v \in H^1_0(\Omega). \quad (5.16)
\]
discretized by means of finite elements of uniform polynomial degree \( m \) on a simplicial triangulation \( M \) of a polygonal/polyhedral computational domain \( \Omega \).

Applying local quadrature rules of the form
\[
\int_\Omega f(x) \, dx \approx \sum_{K \in M} |K| \sum_{l=1}^{P_K} \omega^K_l f(\pi^K_l) \quad \text{(NUQ)}
\]
the perturbed bilinear form for \( u_n, v_n \in S_{m,0}(M) \) reads
\[
\tilde{b}(u_n, v_n) := \sum_{K \in M} |K| \sum_{l=1}^{P_K} \omega^K_l \langle A(\pi^K_l) \grad u_n(\pi^K_l), \grad v_n(\pi^K_l) \rangle \quad \text{(5.20)}
\]

For the analysis we must rely on a certain smoothness of the coefficient function \( A \):

**Assumption 5.22.** The restriction of the coefficient function \( A : \Omega \rightarrow \mathbb{R}^{d,d} \) to any cell \( K \in M \) belongs to \( C^m(K)^{d,d} \) and can be extended to a function \( \in C^m(\overline{K})^{d,d} \).
Lemma 5.23. Let $A$ satisfy
$$
\gamma |\mu|^2 \leq \mu^T A(x) \mu \quad \forall \mu \in \mathbb{R}^d \text{ and almost all } x \in \Omega,
$$
with $\gamma > 0$ and Assumption 5.22, and let the local quadrature weights $\omega^K_l$ be positive. If the local quadrature rules are exact for polynomials up to degree $2m - 2$, then
$$
\tilde{b}(v_n, v_n) \geq \gamma |v_n|_{H^1(\Omega)}^2 \quad \forall v_n \in S_m(\mathcal{M}).
$$

Proof. Since $A$ is uniformly positive definite and the quadrature weights are positive
$$
\tilde{b}(v_n, v_n) \geq \sum_{K \in \mathcal{M}} |K| \sum_{l=1}^{P_K} \omega^K_l |\nabla v_n(\pi^K_l)|^2 \geq \gamma \sum_{K \in \mathcal{M}} |K| \sum_{l=1}^{P_K} \omega^K_l |\nabla v_n(\pi^K_l)|^2
\frac{1}{2} = \gamma |v_n|_{H^1(\Omega)}^2,
$$
because on each $K \in \mathcal{M}$ we know $\nabla v_n \in P_{m-1}(K)^d$ so that the numerical quadrature of $|\nabla v_n|^2$ is exact.

References


103
6 Adaptive Finite Elements

In this chapter we only consider the primal variational formulation of a second order elliptic boundary value problem

$$-\Delta u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{or} \quad \langle \nabla u, n \rangle = 0 \quad \text{on } \Gamma.$$  (6.1)

in a bounded polygon $\Omega \subset \mathbb{R}^2$ with Lipschitz boundary $\Gamma$.

**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 variational problem (LVP), we could just choose $V_n := \text{span}\{u\}$ 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**).

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 $C^\infty$ functions), then for the solution $u$ of (6.1) it holds

$$f \in H^k(\Omega) \implies u \in H^{k+2}(\Omega) \quad \text{for } k \in \mathbb{N}_0,$$

and

$$\forall k \in \mathbb{N}_0, \exists \gamma = \gamma(\Omega, k) : \|u\|_{H^{k+2}(\Omega)} \leq \gamma(\Omega, k) \|f\|_{H^k(\Omega)} \quad \forall f \in H^k(\Omega).$$

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. The solution gets singular meaning that some (higher) derivative is not square integrable.

![Figure 6.1: Polygon $\Omega$ and notation for the corners.](image)

**Theorem 6.4.** Let $\Omega \subset \mathbb{R}^2$ be a polygon with $J$ corners $\pi_j$. Denote the polar coordinates in the corner $\pi_j$ by $(r_j, \theta_j)$ and the inner angle at the corner $\pi_j$ by $\omega_j$ as in Figure 6.1. Additionally, let $f \in H^{-1+s}(\Omega)$ with $s \geq 1$ integer and $s \neq \lambda_{jk}$, where the $\lambda_{jk}$ are given by the singular exponents

$$\lambda_{jk} = \frac{k\pi}{\omega_j} \quad \text{for } k \in \mathbb{N}. \quad (6.2)$$

Then, we have the following decomposition [27] 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

---

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.
from a smooth boundary according to Thm. 6.3) and finitely many so-called **singular functions** \( s_{jk}(r, \theta) \):

\[
u = u^0 + \sum_{j=1}^{J} \psi(r_j) \sum_{\lambda_{jk} < s} \alpha_{jk} s_{jk}(r_j, \theta_j).
\]

Here, \( u^0 \in H^{1+s}(\Omega) \) and \( \psi \) is a \( C^\infty \) cut off function (\( \psi \equiv 1 \) in a neighborhood of 0).

The singular functions \( s_{jk} \) are explicitly given [27, Sect. 4.2] by

\[
\begin{align*}
\lambda_{jk} \text{ non-integer:} & \quad s_{jk}(r, \theta) = r^{\lambda_{jk}} \sin(\lambda_{jk} \theta), \\
\lambda_{jk} \in \mathbb{N}, \omega \notin \{\pi, 2\pi\}: & \quad s_{jk}(r, \theta) = r^{\lambda_{jk}} \ln r \sin(\lambda_{jk} \theta)
\end{align*}
\]

Note, that for \( \lambda_{jk} \in \mathbb{N}, \omega \notin \{\pi, 2\pi\} \) the singular function \( s_{jk}(r, \theta) \) is not harmonic \((\Delta s_{jk} \neq 0)\) and does not fulfill the Dirichlet boundary condition for \( \theta = \omega \). The former is cured by adding the smooth function \( r^{\lambda_{jk}} \theta \cos(\lambda_{jk} \theta) \) and the latter by adding the harmonic polynomial [28] \( \mathcal{H}_{\lambda_{jk}}(x, y) \) of degree \( \lambda_{jk} \) satisfying the boundary data.

For the homogeneous Neumann problem in (6.1), sin has to be replaced by cos and vice-versa.

**Remark 6.5.** The coefficients \( \alpha_{jk} \) 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

\[
- \text{div}(A \grad u) = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \Gamma_D, \quad \langle A \grad u, n \rangle = 0 \quad \text{on } \Gamma_N.
\]

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 \( S \) with a tip at the singular point.

![Figure 6.2: Corner \( \pi_j \) with changing boundary conditions and the infinite sector \( S_\omega \).](image)
Example 6.7. Consider $-\Delta u = f$ in $\Omega$ with mixed boundary conditions at $\pi_j$. Let $\pi_j \in \partial \Omega$ be a boundary point where the type of the boundary conditions changes from Dirichlet to Neumann (cf. Figure 6.2).

In the infinite sector

$$S_\omega = \{(r, \theta) : 0 < r < \infty, \ 0 < \theta < \omega\},$$

we are looking for non-trivial solutions of the homogeneous problem

$$\Delta s = 0 \text{ in } S_\omega, \quad \frac{\partial s}{\partial n}\bigg|_{\theta=0} = 0, \quad s\bigg|_{\theta=\omega} = 0$$

of the form $s(r, \theta) = r^\lambda \Theta(\theta)$. Using $s = r^\lambda \Theta(\theta)$, it follows in $S_\omega$:

$$0 = \Delta s = r^{\lambda-2}(\Theta'' + \lambda^2 \Theta) \text{ for } r > 0,$$

i.e. the pairs $(\lambda, \Theta(\theta))$ are eigenpairs of a Sturm-Liouville problem

$$L\Theta = \Theta'' + \lambda^2 \Theta = 0 \text{ in } (0, \omega), \quad \Theta'(0) = 0, \quad \Theta(\omega) = 0.$$

One recalculates that the eigenpairs are explicitly given by

$$\lambda_k = (k - 1/2) \frac{\pi}{\omega}, \quad \Theta_k(\theta) = \cos(\lambda_k \theta), \quad k = 1, 2, 3, \ldots.$$

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.

![Figure 6.3: Cracked panel.](image)

Example 6.8. 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 6.3). 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 6.9. 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 \( \Omega \). The singular behaviour of \( u \) is restricted to the corners \( \pi_j \).

Remark 6.10. 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 \( (\partial^\alpha s_{j1})(r_j, \theta_j) \sim r_j^{\lambda_{j1}-|\alpha|} \) for \( r_j \to 0 \) that the derivative \( \partial^\alpha \) of the singular functions \( s_{jk} \) for \( |\alpha| = 2 \) is not square integrable since \( \lambda_{j1} - |\alpha| < -1 \), i.e. for \( |\alpha| = 2 \) we have

\[ ||(\partial^\alpha s_{j1})(r_j, \theta_j)||^2 \sim r_j^{-2+\epsilon} \notin L^1(\Omega). \]

The shift theorem Thm. 6.3 does no longer hold.

References. Corner and edge singularities for solutions of elliptic problems are discussed in [29, 30, 27].

6.2 Convergence of finite element solutions

Let \( u_n \in S_m(M_n) \) 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 \( M_n \). Temporarily, we will allow \( d \in \{1, 2, 3\} \).

Let \( \{M_n\}_{n=1}^\infty \) denote a uniformly shape-regular and quasi-uniform family of triangulations of the polygon \( \Omega \) such that \( h_n := h_{M_n} \to 0 \) as \( n \to \infty \). From Sect. 5.5 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 \( \{M_n\}_{n=1}^\infty \) of simplicial meshes. In the case of finite elements of polynomial degree \( m \) we have the crude estimates

\[ N_n := \dim(S_m(M_n)) \leq \binom{d+m}{d} \cdot \sharp M_n \Rightarrow \sharp M_n \approx h^{-d} M_n, \]

with constants depending on shape-regularity 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 $\gamma$ depends on $\Omega, A$ and the bounds for $\rho_{\mathcal{M}_n}, \mu_{\mathcal{M}_n}$. This reveals an algebraic asymptotic convergence rate of the h-version of 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 $f \in H^k(\Omega)$:

$$u = u^0 + u_{\text{sing}}$$

with a smooth part $u^0 \in H^k(\Omega), k \geq 2$, and with a singular part $u_{\text{sing}}$, which is a (finite!) sum of singular functions $s(r, \theta)$, which have the explicit form

$$s(r, \theta) = r^\lambda \Theta(\theta),$$

with piecewise smooth $\Theta$, where $0 < \lambda < k - 1$ (we assume here that log $r$ terms are absent). The singular functions (6.7) are only poorly approximated by finite element functions on sequences of quasi-uniform meshes. For the singular functions $s(r, \theta)$ as in (6.7) and with $(r, \theta)$ denoting polar coordinates at a vertex of $\Omega$ the (optimal) error estimate

$$\min_{v_n \in S_m(\mathcal{M}_n)} \| s - v_n \|_{H^1(\Omega)} \leq \gamma h_n \min(m, \lambda) \leq \gamma N_n^{-\min(m, \lambda)/2}$$

holds, where again $N_n := \dim S_m(\mathcal{M}_n) = O(h_n^{-2})$ 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

$$\| u - u_n \|_{H^1(\Omega)} \leq \gamma h_n \min(m, \lambda^*) \leq \gamma N_n^{-\min(m, \lambda^*)/2},$$

where $\lambda^* = \min\{\lambda_{jk} : j = 1, \ldots, J, k = 1, 2, \ldots\}$, 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 6.11.** If the exact solution $u$ is very smooth, that is, $t \gg 1$, raising the polynomial degree $m$ is preferable (p-version), because we have $N_n \approx m^d h_n^{-d}$ and, thus the estimate (see Remark 5.14)

$$\inf_{v_n \in S_m(\mathcal{M})} \| u - v_n \|_{H^1(\Omega)} \leq \gamma (\rho_{\mathcal{M}}, \mu_{\mathcal{M}}) \left( \frac{h_n}{m} \right)^{\min\{m+1, t\}} \| u \|_{H^1(\Omega)},$$

(5.15)

gives asymptotically for $t \leq m \to \infty$

$$\| u - u_n \|_{H^1(\Omega)} \leq \gamma N_n^{t-1/d}. $$

(6.9)
For p-FEM the regularity of the solution determine the rate of convergence, where for h-FEM the rate is mainly determined by the polynomial degree. 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 6.12.** If the exact solution is analytic in $\Omega$, that is, it is $C^\infty$ and can be expanded into a locally convergent power series in each point of $\Omega$, then the p-version yields an exponential asymptotic convergence rate

$$\|u - u_n\|_{H^1(\Omega)} \leq \gamma \exp(-\gamma' N_n^\beta),$$

with $\gamma, \gamma', \beta > 0$ only depending on problem parameters and the fixed triangulation, but independent of the polynomial degree $m$ of the finite elements.

### 6.3 A priori adaptivity by graded meshes

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 finite elements on quasi-uniform meshes pointed out in Sect. 6.2.

![Figure 6.4: Polygon Ω with corner $\pi_j$, subdomain $\text{conv}(\pi_j, \kappa_1, \kappa_2)$ adjacent to it and its representation by an affine map $\Phi$ from the standard triangle $\hat{K}$.](image)

One option is judicious (vernünftig) mesh refinement towards the vertices of the polygon. Consider the polygon $\Omega$ shown in Fig. 6.4. In $\Omega$, consider any vertex $\pi_j$ (In Fig. 6.4 we chose a convex corner, the approach to a re-entrant corner at $\pi_{j+1}$ is indicated in Fig. 6.5). We denote again by $(r, \theta)$ polar coordinates at vertex $\pi_j$, and by $s(r, \theta)$ a singular function as in (6.7)

$$s(r, \theta) = r^\lambda \Theta(\theta)$$

110
with a smooth $\Theta(\theta)$. The triangle $K = \text{conv}(\pi_j, \kappa_1, \kappa_2)$ denotes a neighbourhood of vertex $\pi_j$ in $\Omega$ (shown shaded in Fig. 6.4). By means of an affine map $\Phi$ the triangle $K$ is mapped onto the reference triangle $\hat{K}$ with polar coordinates $(\hat{r}, \hat{\theta})$. The singular function $s(r, \theta)$ in $\Omega$ is transformed by $\Phi$ into

$$s(\hat{r}, \hat{\theta}) = \hat{r}^\lambda \hat{\Theta}(\hat{\theta})$$

in $\hat{K}$, with the same exponent $\lambda$ but with another $C^\infty$-function $\hat{\Theta}(\hat{\theta})$:

**Example 6.13.** Let $\pi_j = (0, 0)$, $(x_1, x_2)$ stand for the coordinates in $\Omega$,

$$x_1 = r \cos \theta, \quad x_2 = r \sin \theta$$

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = F \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \hat{r} \begin{pmatrix} \cos \hat{\theta} \\ \sin \hat{\theta} \end{pmatrix}$$

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 $\xi_1, \xi_2$, once again of the form (6.7) let $F = (f_{ij})_{1 \leq i,j \leq 2}$. Then

$$r^2 = x_1^2 + x_2^2 = (f_{11} \xi_1 + f_{12} \xi_2)^2 + (f_{21} \xi_1 + f_{22} \xi_2)^2$$

$$= \hat{r}^2 \{(f_{11} \cos \hat{\theta} + f_{12} \sin \hat{\theta})^2 + (f_{21} \cos \hat{\theta} + f_{22} \sin \hat{\theta})^2\},$$

and

$$r^\lambda = \hat{r}^\lambda \{(f_{11} \cos \hat{\theta} + f_{12} \sin \hat{\theta})^2 + (f_{21} \cos \hat{\theta} + f_{22} \sin \hat{\theta})^2\}^{\frac{\lambda}{2}} = \hat{r}^\lambda \Theta_1(\hat{\theta}),$$

with a smooth (analytic) function $\Theta_1(\hat{\theta})$. Analogously, we have that $\Theta(\theta) = \hat{\Phi}_2(\hat{\theta})$ with a smooth function $\hat{\Theta}_2(\hat{\theta})$.  

111
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 \( \hat{K} \) as shown in Figure 6.6. In the case of a re-entrant corner, the reference domain consists of three triangles, see Fig. 6.5, and the ensuing considerations can be applied to each of them.

In what follows we show that by using so-called algebraically graded meshes \( M_\beta^n \) at the vertices of \( \Omega \) the optimal asymptotic behavior \( O(N^{-m/2}) \) of the best approximation error of finite elements of uniform global degree \( m \) can be retained for singular functions as well.

![Graded mesh on (0,1).](image1)

![Graded mesh on reference triangle.](image2)

Figure 6.6: Construction of a graded meshes \( M_\beta^n \) in \( \Omega = (0, 1) \) for \( \beta > 1 \) and on reference triangle \( \hat{K} \).

**Definition 6.14.** A family \( \{M_\beta^n\}_{n=1}^\infty \) of meshes of a computational domain \( \Omega \subset \mathbb{R}^2 \) is called algebraically graded with respect to \( \pi \in \overline{\Omega} \) and grading factor \( \beta \geq 1 \) if

(i) the meshes are uniformly shape-regular, and

(ii) with constants independent of \( n \) and \( h_n := h_{M_\beta^n} \),

\[
\forall K \in M_\beta^n, \ \pi \notin \overline{K} : \ h_K \approx n^{-1} \text{dist}(\pi, K)^{1-1/\beta}.
\]

We will describe the concrete construction of algebraically graded meshes \( M_\beta^n, n \in \mathbb{N} \), with grading factor \( \beta \geq 1 \) \( n \in \mathbb{N} \) on the reference domain \( \hat{K} \) with respect to the vertex \( \binom{0}{0} \), see Fig. 6.6:

**Algorithm 6.15** (Graded mesh on reference triangle). On \( \hat{K} = \text{conv}(\binom{0}{0}, \binom{1}{0}, \binom{1}{1}) \) we proceed as follows:
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 $L_j = \{ \xi \in \hat{K} : \tau_{j-1}^n < \xi_1 + \xi_2 < \tau_j^n \}$, $j = 1, \ldots, n$.

3. Equip each layer $L_j$, $j = 1, \ldots, n$ with a simplicial triangulation $\mathcal{M}_{n|L_j}^\beta$ such that
   a) their union yields a simplicial triangulation of $\hat{K}$,
   b) the shape regularity measure of $\mathcal{M}_{n|L_j}^\beta$ (see Def. 5.9) is uniformly bounded independently of $j$ and $n$,
   c) for each $K \in \mathcal{M}_{n|L_j}^\beta$ we have $h_K \approx \tau_j - \tau_{j-1}$ with constants independent of $j$ and $n$, and
   d) $\mathcal{M}_{n|L_1}^\beta$ consists of a single triangle $K^*$ adjacent to $\binom{0}{0}$.

Remark 6.16. For $\beta = 1$ the meshes $\mathcal{M}_{n}^\beta$ are quasi-uniform with meshwidth $1/n$.

Lemma 6.17. Fix $\beta > 1$. Then, with constants only depending on the bounds on the shape-regularity measure $\rho(\mathcal{M}_{n|L_j}^\beta)$ and quasi-uniformity measure $\mu(\mathcal{M}_{n|L_j}^\beta)$ we find

$$h_K \approx \frac{\beta}{n} \left( \frac{j}{n} \right)^{\beta-1} = \frac{\beta}{n} \left( \frac{j}{n} \right) \left( \frac{j}{n} \right)^{1-1/\beta} \quad \forall K \in \mathcal{M}_{n|L_1}^\beta,$$

(6.11)

and

$$\sharp \mathcal{M}_{n|L_1}^\beta \approx j.$$

(6.12)

Proof. Pick $n \in \mathbb{N}$ and $j \in \{2, \ldots, n\}$. Then, with the monotonously increasing function $f(t) = (t/n)^\beta$ and its derivative $f'(t) = \beta/n(t/n)^{\beta-1}$ we have $\tau_j - \tau_{j-1} = f(j) - f(j-1)$ which can be bounded by the mean value theorem by $f'(j)$ and $f'(j-1)$ from above and below. Thus,

$$\frac{\beta}{n} \left( \frac{j-1}{n} \right)^{\beta-1} \leq \tau_j - \tau_{j-1} \leq \frac{\beta}{n} \left( \frac{j}{n} \right)^{\beta-1}. $$

Together with $h_{K^*} = n^{-\beta}$ we conclude the first assertion of the lemma.

To confirm the second, we start with the volume formula

$$2|L_j| = \left( \frac{j}{n} \right)^{2\beta} - \left( \frac{j-1}{n} \right)^{2\beta} \approx \frac{2\beta}{n} \left( \frac{j}{n} \right)^{2\beta-1}. $$

(6.13)

As a consequence of (6.11), the area of a triangle $\subset L_j$ is

$$2|K| \approx h_K^2 \approx \frac{\beta^2}{n^2} \left( \frac{j}{n} \right)^{2\beta-2} \quad \forall K \in \mathcal{M}_{n|L_j}^\beta.$$

(6.14)

None of the constants depends on $n$ and $j$. Dividing (6.13) by (6.14) yields (6.12). □
**Corollary 6.18.** The family \( \{ M_\beta^n \}, n \in \mathbb{N}, \beta \geq 1, \) of meshes emerging from construction 6.15 is algebraically graded with respect to \((0_0)\) and grading factor \(\beta\).

**Corollary 6.19.** The algebraically graded meshes \( M_\beta^n, n \in \mathbb{N}, \) of \(\hat{K}\) constructed as above for \(\beta \geq 1\) satisfy
\[
h_n := h_{M_\beta^n} \approx \frac{\beta}{n}, \quad \sharp M_\beta^n \approx n^2,
\]
with constants independent of \(n\).

As a consequence, \(h(M_\beta^n) \to 0\) for \(n \to \infty\), if \(\beta \geq 1\). Moreover, for fixed \(m \in \mathbb{N}\), we get from Cor. 6.19 that
\[
N_n = \dim S_m(M_\beta^n) \leq \gamma \sharp M_\beta^n \leq \gamma n^2,
\]
holds with a constant independent of \(n\).

**Approximation of the regular part** As, again by Cor. 6.19, \(n^{-1} \leq \gamma N_n^{-\frac{1}{2}}\), we deduce from Thm. 5.12 that for the regular part \(u^0 \in H^{m+1}(\hat{K})\) of the decomposition (6.6) of the solution \(u \in H^1_0(\Omega)\) of \(-\Delta u = f\) holds
\[
\min_{v_n \in S_m(M_\beta^n)} \| u^0 - v_n \|_{H^1(\hat{K})} \leq \| u^0 - I_n u^0 \|_{H^1(\hat{K})} \leq \gamma h_n^m \leq \gamma N_n^{-m/2}, \tag{6.15}
\]
with \(\gamma = \gamma(m, \rho, M_\beta^n)\). Here \(I_n\) is the finite element interpolation operator for \(S_m(M_\beta^n)\).

This implies that the regular part \(u^0\) of the solution \(u\) can also be approximated on algebraically \(\beta\)-graded meshes at the optimal rate (6.4), independently of the size of \(\beta \geq 1\) (the size of the constant \(\gamma\) in the error estimates (6.15) depends of course on \(\beta\) and possibly grows strongly with \(\beta > 1\); for fixed \(\beta\) and \(n \to \infty\) the convergence rate (6.15) is optimal, however).

**Approximation of the singular part** Let us now consider the singular part of \(u\) in the decomposition (6.6). According to (6.3) the solution \(v_{\text{sing}}\) 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 6.20.** Let \(s(r, \theta) = r^{\lambda} \Theta(\theta)\) with \(\lambda > 0\) and \(\Theta \in C^\infty([0, \pi/2])\) for \((r, \theta) \in \tilde{K}\) as in Fig. 6.6. Let further
\[
\beta > \max\{m/\lambda, 1\}.
\]
Then there holds, as \(N_n = \dim S_m(M_\beta^n) \to \infty\)
\[
\min_{v_n \in S_m(M_\beta^n)} |s - v_n|_{H^1(\tilde{K})} \leq \gamma(m, \lambda, \beta) N_n^{-\frac{m}{2}},
\]
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 \( s - l_n s \). 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, \( l_n \) designates linear interpolation on \( \mathcal{M}_n^\beta \).

Let \( K^* \in \mathcal{M}_n^\beta \) denote the triangle which contains the origin \((0, 0)\) 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

\[
|s|_{H^1(K^*)}^2 \leq \int_0^{\pi/2} \int_0^{\tau_1} \left( \frac{\partial s}{\partial r}^2 + \left( \frac{1}{r} \frac{\partial s}{\partial \theta} \right)^2 \right) r \, dr \, d\theta \leq \gamma(\Theta) \int_0^{\pi/2} \int_0^{\tau_1} \left( \lambda r^{2\lambda-1} \Theta'(\theta) \right)^2 + \left( \frac{1}{r} \lambda^\beta \Theta(\theta) \right)^2 r \, dr \, d\theta.
\]

Since \( l_n(s) \) is linear on \( K^* \), we find

\[
|l_n s|_{H^1(K^*)}^2 = |K^*| \int_0^{\pi/2} \left( s(\tau_1, 0)^2 + s(0, \tau_1)^2 \right) = \gamma(\Theta) \lambda^{2\lambda} \left( \frac{1}{n} \right)^{2\beta\lambda}.
\]

Thus, a simple application of the triangle inequality yields

\[
|s - l_n s|_{H^1(K^*)} \leq \gamma(\Theta) \left( \frac{1}{n} \right)^{2\beta\lambda} \leq \gamma N_n^{-\beta\lambda} \leq \gamma N_n^{-m},
\]

since \( \beta > \lambda/m \).

Next, consider \( \hat{K} \setminus K^* \). For \( x \in \hat{K} \) 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. 5.12 (with \( r = 1, m = 1, t = 2 \)) yields

\[
|s - l_n s|_{H^1(\hat{K} \setminus K^*)}^2 = \sum_{K \in \mathcal{M}_n^\beta, \ K \neq K^*} h_K^2 \left| s \right|_{H^2(K)}^2 \leq \gamma \sum_{K \in \mathcal{M}_n^\beta, \ K \neq K^*} h_K^2 \|D^2 s\|^2 \, dx.
\]

The construction of the graded mesh implies that at any point \( \xi \in \hat{K} \setminus K^* \) it holds

\[
r = |\xi| \geq \frac{\sqrt{3}}{2} \left( \frac{j}{n} \right) \beta \quad \text{for some } j \in \mathbb{N},
\]

\[
|h(\xi)| \leq \gamma n^{-1} \left( \frac{j}{n} \right)^{1-1/\beta} \leq \gamma n^{-1} r^{1-1/\beta},
\]

\[
|D^2 s| \leq \gamma r^{\lambda-2}, \quad \text{with } \gamma > 0 \text{ independent of } n.
\]
Hence, for \( n \gg 1 \) and \( \lambda > \beta^{-1} \),
\[
\int_{\hat{K} \setminus K^*} (h(\xi))^2 |D^2 s|^2 \, d\xi \leq \gamma \int_1^n r^{-2(1-1/\beta) + 2\lambda - 4} \, dr \\
\leq \gamma n^{-2} \left[ r^{2\lambda - 2/\beta} \right] \frac{1}{\sqrt{2} n^{-\beta}} \leq \gamma n^{-2} \leq \gamma N_n^{-1},
\]
using \( \lambda > \beta^{-1} \) implies the assertion. For general \( m \geq 1 \) we get in the case of \( \lambda > m/\beta \)
\[
|s - \ln s|^2_{H^1(\hat{K} \setminus K^*)} \leq \gamma n^{-2m} \left[ r^{2\lambda - 2/\beta - 2(m-1)} \right] \frac{1}{\sqrt{2} n^{-\beta}} \\
\leq \gamma (n^{-2m} + n^{-2\beta}) \leq \gamma (n^{-2m} + n^{-2\beta} \lambda m - 1)^{-2} \\
\leq \gamma (n^{-2m} + n^{-2m-2}) \leq \gamma n^{-2m} \leq \gamma N_n^{-m}.
\]

**Remark 6.21.** From Thm. 6.20 we learn that either \( \lambda > m \) and \( \beta = 1 \) or \( \lambda = m/\beta \)
will ensure the optimal rate of convergence of the finite element solution in terms of the dimension of the finite element space.

**Corollary 6.22.** For \( \hat{K} \) as shown in Fig. 6.6 and \( k_{\text{max}} \) such that \( \lambda_{k_{\text{max}}} = \max \{ \lambda_k : \lambda_k < m \} \), we have \( u = u^0 + u_{\text{sing}} \) with \( u^0 \in H^{m+1}(\hat{K}) \) and
\[
u_{\text{sing}} = \sum_{k=1}^{k_{\text{max}}} \alpha_k r^{\lambda_k} \Phi_k(\theta),
\]
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
\[
\min_{v \in S_m(M^0)} \|u_{\text{sing}} - v\|_{H^1(\hat{K})} \leq \gamma N_n^{-m/2} \text{ with } \gamma = \gamma(p, \alpha_k, \beta).
\]

**Remark 6.23.**
1. If \( m/\lambda_1 > 1 \) we achieve with the grading factor
\[
\beta > m/\lambda_1 \quad (6.17)
\]
the same convergence rate as for smooth solutions \( u \) with a quasi uniform triangulation.

2. For fixed \( \sharp M \) we have to increase \( \beta \) (i.e. the grading must be more pronounced) if
a) \( m \) is raised and b) if the singular exponent \( \lambda_1 \) is reduced.

3. Usually the singular exponent \( \lambda_1 > 0 \) is unknown. Thm. 6.20 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 \( \lambda_1 \) 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 $\Omega$, where $\lambda_j := \pi/\omega_j > 1$. For $m > 1$ mesh refinement near vertices is also required in convex domains, in general.

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

$$f \in H^{k-2}(\Omega), \; k \geq 2.$$ 

Let further $u \in H^1_0(\Omega)$ 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

$$\inf_{v_n \in S_m(M_\beta)} |u - v_n|_{H^1(\Omega)} \leq \gamma N_n^{-m/2},$$

for $N_n = \dim S_m(M_\beta) \to \infty$.

References


