Numerics of Partial Differential Equations

Lecture in winter term 2015/2016 at TU Berlin/Germany
based on lectures notes from Prof. Dr. Ralf Hiptmair

Dr. Kersten Schmidt

October 18, 2017

Contents

1 Introduction

1.1 Terminology and classification of PDEs

A differential eqation 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 $\mathbf{z} = (z_1, z_2, ..., z_n) \in \mathbb{R}^n$, then there occur partial derivatives

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

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

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

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

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

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

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

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

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 $\mathbf{A}(\mathbf{z}) = \{a_{ij}(\mathbf{z})\}_{i,j=1}^n$ is symmetric.

PDEs are classified into

- In **elliptic equations** an incident at a point $\mathbf{z}$ influences all point in their neighbourhood.
- In **parabolic equations** there is in one direction from a point $\mathbf{z}$ for which the influence is only for larger values. This is often the time direction, where only later times are influenced.
- In **hyberbolic equations** there are areas of directions around a point $\mathbf{z}$ which are influenced.

The principal part (Hauptteil) of the PDE

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

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* $\mathbf{A}(\mathbf{z})$*.*

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

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

3. *The equation is said to be* hyberbolic *at* $\mathbf{z} \in \mathcal{O}$ *if* $\mathbf{A}(\mathbf{z})$ *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* hyberbolic in a set $\mathcal{O}$ of points $\mathbf{z}$*, if it has the property for all* $\mathbf{z} \in \mathcal{O}$*.*

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 convervation of energy which correspond in absence of work to the conservation of temperature

$$\frac{\partial u}{\partial t}(\mathbf{x}, t) + \operatorname{div} \mathbf{j}(\mathbf{x}, t) = f(\mathbf{x}, t) \tag{1.2}$$

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

and as constitutive equation Fourier's law

$$\mathbf{j}(\mathbf{x}, t) = -\mathbf{C}(\mathbf{x})\,\mathbf{grad}\, u(\mathbf{x}, t), \tag{1.3}$$

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

$$\begin{aligned} \text{homogen} &: \mathbf{C}(\mathbf{x}) = \mathbf{C}, \\ \text{isotrop} &: \mathbf{C}(\mathbf{x}) = \alpha(\mathbf{x}) \cdot \mathbf{1}, \end{aligned}$$

and inhomogen or anisotrop otherwise, $\alpha(\mathbf{x})$ is the conductivity.

Inserting (1.3) into (1.2) we obtain the model described by a second order partial differential equation:

$$\frac{\partial u}{\partial t} - \operatorname{div} \mathbf{C} \operatorname{\mathbf{grad}} u = f.$$

The heat equation is parabolic if $\mathbf{C}$ is positive definite. For a unique solution we have to complete the system by initial conditions

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}),$$

and boundary conditions. We may describe the temperature on the boundary of the domain of interest Ω

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

or the heat flux

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

Here, $\mathbf{n}(\mathbf{x})$ 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

$$\mathbf{j} = -\mathbf{C}(\mathbf{x}) \operatorname{\mathbf{grad}} u \,, \tag{FL}$$

$$\operatorname{div} \mathbf{j} = f, \tag{EL}$$

which has to be equipped with the above boundary conditions.

The same system arises in

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

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

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

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

A similar elliptic system is

$$\mathbf{j} = -\mathbf{C}(\mathbf{x})\ \mathbf{grad}\, u\ , \tag{FL}$$

$$\operatorname{div}\mathbf{j} = f - c(\mathbf{x})u, \tag{EL}$$

which appear in

- Molecular diffusion:

$u \to$ concentration		$[u] = 1\frac{\text{mol}}{\text{m}^3}$
$\mathbf{j} \to$ flux		$[\mathbf{j}] = 1\frac{\text{mol}}{\text{m}^2\text{s}}$
$f \to$ production/consumption rate		$[f] = 1\frac{\text{mol}}{\text{m}^3\text{s}}$

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

1.2.3 Wave equation

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

$$m(\mathbf{x})\frac{\partial^2 u}{\partial t^2}(\mathbf{x},t) = \operatorname{div}\boldsymbol{\sigma}(\mathbf{x},t) + f(\mathbf{x},t),$$

and Hook's law as constitutive equation

$$\boldsymbol{\sigma}(\mathbf{x},t) = \mathbf{C}(\mathbf{x})\,\mathbf{grad}\, u(\mathbf{x},t).$$

Here, we have

u	$\to$ vertical displacement	$[u] = 1\text{m}$
$\boldsymbol{\sigma}$	$\to$ stress vector	$[\boldsymbol{\sigma}] = 1\text{J}$
f	$\to$ external force	$[f] = 1\text{N}$
$\mathbf{grad}\, u$	$\to$ (local) deformation of the membran	
$\frac{\partial u}{\partial t}$	$\to$ (local) speed of the membran	
$\frac{\partial^2 u}{\partial t^2}$	$\to$ (local) acceleration	

and the density of mass $m(\mathbf{x})$ ($[m] = 1\frac{\text{kg}}{m^2}$).

The second-order PDE is

$$m(\mathbf{x})\frac{\partial^2 u}{\partial t^2}(\mathbf{x},t) = \operatorname{div}(\mathbf{C}(\mathbf{x})\,\mathbf{grad}\, u) + f, \tag{1.4}$$

and for homogeneous isotrope material with $\mathbf{C} = \mathbf{1}$ we have

$$m(\mathbf{x})\frac{\partial^2 u}{\partial t^2}(\mathbf{x},t) = \Delta u + f,$$

which is hyperbolic.

We have to complete the PDE with initial conditions

$$u(\mathbf{x},0) = u_0(\mathbf{x}),$$

and boundary conditions. We may describe the displacement on the boundary of the membran

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

or the boundary stress

$$\boldsymbol{\sigma}(\mathbf{x},t)\cdot\mathbf{n}(\mathbf{x}) = h(\mathbf{x},t) \text{ for } \mathbf{x} \in \partial\Omega.$$

Equation	Modelled system	Classification
Advection equation	Transport of particles in a flow.	Hyberbolic
Navier-Stokes equations	Flow of fluids.	Hyberbolic
Euler	Flow of incompressible fluids.	Hyberbolic
Shallow water equations	Flow in shallow domains (earth atmosphere).	Hyberbolic
Stokes equations	Static limit of flow of viscous flows.	Elliptic
Elasticity equations	Stresses in elastic bodies.	Elliptic
Maxwell's equations	Electromagnetic fields.	Hyberbolic
Time-harmonic Maxwell's equations	Electromagnetic fields.	Elliptic
Eddy current modell	Electromagnetic fields in low frequency.	Elliptic
Schrödinger's equation	Electron structure of matters.	Elliptic
Black-Scholes equation	Prices of stocks.	Parabolic

Table 1.1: Some examples for systems modelled by PDEs.

1.2.4 Time-harmonic wave-equation

Let $f = \mathrm{Re}(\widehat{f}\mathrm{e}^{-\mathrm{i}\omega t})$ as well as $g = \mathrm{Re}(\widehat{g}\,\mathrm{e}^{-\mathrm{i}\omega t})$, $h = \mathrm{Re}(\widehat{h}\,\mathrm{e}^{-\mathrm{i}\omega t})$ for some (angular) frequency $\omega \in \mathbb{R}^+$. With the independance of the coefficient function of t we can write $u = \mathrm{Re}(\widehat{u}\mathrm{e}^{-\mathrm{i}\omega t})$ and insertion into (1.4) leads to the time-harmonic wave-equation

$$-\operatorname{div}(\mathbf{C}(\mathbf{x})\,\mathbf{grad}\,\widehat{u}) - \omega^2 m(\mathbf{x})\widehat{u} = \widehat{f}, \tag{1.5}$$

which is a linear elliptic PDE of second order. The boundary condition transfer to $\widehat{u}$ where g and h are replaced by $\widehat{g}$ and $\widehat{h}$.

Remark 1.3. *Linear parabolic or hyberbolic 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 $\widetilde{f} = f + \delta f$ with $\|\delta f\| \ll 1$ (very small). Then $\widetilde{u}$ is the solution of the system with data $\widetilde{f}$ and $\delta u = \widetilde{u} - u$ the solution of the system with data δf. 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*

$$\begin{aligned} \Delta u &= 0 \\ u(x_1, 0) &= \frac{1}{n} \sin(nx_1) \\ \partial_2 u(x_1, 0) &= 0, \end{aligned}$$

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 $x_2 > 0$*.*

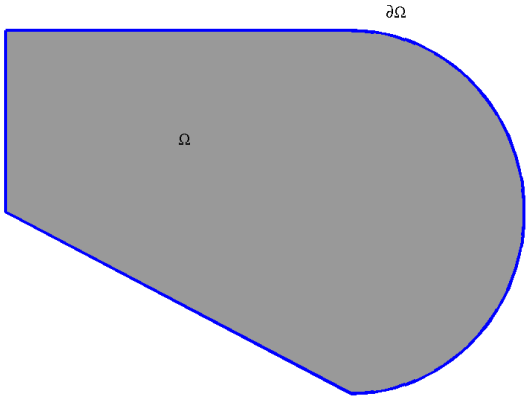
2 Finite differences scheme for second-order elliptic PDEs

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 $f \in C(\Omega)$, $g \in C(\partial\Omega)$, $0 \leq c \in C(\Omega)$

$$Lu = -\Delta u + cu = f \quad \text{in } \Omega, \tag{2.1a}$$

$$u = g \quad \text{on } \partial\Omega. \tag{2.1b}$$



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 Ω, 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 Ω, then the function u is constant.*

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

$$u(\widetilde{\mathbf{x}}) = \sup_{\mathbf{x}\in\Omega} u(\mathbf{x}) > \sup_{\mathbf{x}\in\partial\Omega} u(\mathbf{x}). \tag{2.2}$$

With

$$0 > f(\widetilde{\mathbf{x}}) \overset{(2.1a)}{=} (Lu)(\widetilde{\mathbf{x}}) = -\Delta u(\mathbf{x}) = -\sum_{i=1}^{d} \frac{\partial^2 u}{\partial x_i^2}(\widetilde{\mathbf{x}})$$

which is a contradiction to (2.2), as for a maximum point for all $1 \leq i \leq d$

$$\frac{\partial^2 u}{\partial x_i^2}(\widetilde{\mathbf{x}}) \leq 0.$$

Now, let $f \leq 0$, and $\widetilde{\mathbf{x}}$ 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

$$w = u + \beta \sum_{i=1}^{d} (x_i - \widetilde{x}_i)^2$$

attains its maximum at an interior point $\mathbf{x}_0$. 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, *i. e.*, (2.2) does not hold and u attains its maximum on the boundary.

Let the maximum is denoted by $M = \max_{\mathbf{x}\in\partial\Omega}$. There might be still points $\widetilde{\mathbf{x}}$ in the interior with $u(\widetilde{\mathbf{x}}) = M$. Assume that u is not constant in Ω. So there exists a ball $B \subset \Omega$ of positive radius R and mid-point $\mathbf{x}_0$ with $\sup_{\mathbf{x}\in B_2} u(\mathbf{x}) < M$ for any proper subset $B_2 \subsetneq B$, but with a boundary point $\widetilde{\mathbf{x}} \in \partial B$ for which $u(\widetilde{\mathbf{x}}) = M$. Let $G := B\backslash\overline{B(\mathbf{x}_0, R/2)}$ the open bounded ring domain with outer boundary $\widetilde{\Gamma}$ and inner boundary Γ. Let furthermore

$$v(\mathbf{x}) := \mathrm{e}^{-\lambda|\mathbf{x}-\mathbf{x}_0|^2} > 0$$

with $\lambda > 0$ large enough such that

$$Lv = \lambda(d - \lambda|\mathbf{x} - \mathbf{x}_0|)\mathrm{e}^{-\lambda|\mathbf{x}-\mathbf{x}_0|^2} < 0$$

for all $\mathbf{x} \in G$. Note, that

$$v(\mathbf{x}) = 0 \quad \text{ on } \widetilde{\Gamma}, \tag{2.3}$$

and so

$$w_\varepsilon := u + \varepsilon v$$

coincides with u on $\widetilde{\Gamma}$. Since $\sup_{\mathbf{x}\in\Gamma} u(\mathbf{x}) < M$ by assumption we can choose a $\varepsilon > 0$ so small such that

$$w_\varepsilon(\mathbf{x}) - w_\varepsilon(\widetilde{\mathbf{x}}) = w_\varepsilon(\mathbf{x}) - M < 0 \quad \text{ in } G. \tag{2.4}$$

Since $Lw_\varepsilon = f + \varepsilon Lv < 0$ the maximum of w_ε in $\overline{G}$ is on ∂G, which is by (2.3) and (2.4) only attained in $\widetilde{\mathbf{x}}$. Hence,

$$\mathbf{grad}\, w_\varepsilon(\widetilde{\mathbf{x}}) = \mathbf{0}$$

and as $\mathbf{grad}\, v(\widetilde{\mathbf{x}}) \neq \mathbf{0}$ it holds

$$\mathbf{grad}\, u(\widetilde{\mathbf{x}}) \neq \mathbf{0},$$

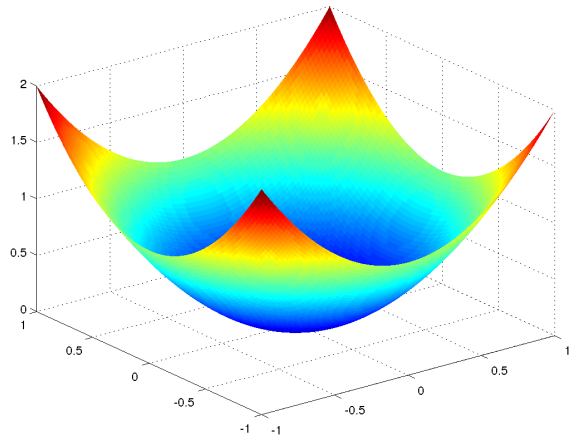
i. e., $\widetilde{\mathbf{x}}$ is not a maximum point which contradicts the assumptions. So there exists no interior maximum point if u is not constant.

For $f \geq 0$ we consider $\widetilde{u} := -u$ 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 appropiate boundary condition. Since* $f \leq 0$ *in* Ω *the solution* u *attains its maximum on the boundary* $\partial\Omega$*. This is indeed true.*



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* Ω *"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* Ω*. 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* Ω*, 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*

$$\begin{aligned} f_u &\leq f_v \quad \text{in } \Omega \\ u &\leq v \quad \text{on } \partial\Omega. \end{aligned}$$

Then, $u \leq v$ *in* Ω*.*

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 $\overline{\Omega}$. □

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. □

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 Ω.*

Proof. See [1]. □

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

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

with the remainder

$$R_n(u; x, h) = \frac{1}{n!}\int_x^{x\pm h} (x-t)^n u^{(n+1)}(t)\,\mathrm{d}t = \frac{(\pm h)^{n+1}}{(n+1)!}u^{(n+1)}(\xi) \quad \text{for some } \xi \in I.$$

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

$$(\Delta u)(\mathbf{x}) = \underbrace{\frac{u(x_1+h,x_2)+u(x_1-h,x_2)+u(x_1,x_2+h)+u(x_1,x_2-h)-4u(x_1,x_2)}{h^2}}_{=:\,\Lambda u} + O(h^2).$$

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 $u \in C^3([-h,h] \times [-h,h] + \mathbf{x})$

$$\begin{aligned}(\Delta u)(\mathbf{x}) = \quad & \frac{2}{h_L+h_R}\left(\frac{u(x_1+h_R,x_2)-u(x_1,x_2)}{h_R} - \frac{u(x_1,x_2)-u(x_1-h_L,x_2)}{h_L}\right) \\ & + \frac{2}{h_B+h_T}\left(\frac{u(x_1,x_2+h_T)-u(x_1,x_2)}{h_T} - \frac{u(x_1,x_2)-u(x_1,x_2-h_B)}{h_B}\right) + O(h).\end{aligned}$$

We call the difference quotient $\Lambda^* u$, which generalises Λu (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

$$\mathcal{T}_h := \{\mathbf{x} = (jh, ih) : \mathbf{x} \in \Omega\}, \tag{2.5}$$

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 $\overline{\mathcal{T}}_h := \mathcal{T}_h \cup \mathcal{G}_h$ (see Fig. 2.1), and $|\overline{\mathcal{T}}_h|$ 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 $v_h \in \overline{\mathcal{T}}_h$.

We ask the finite difference approximant u_h to u to fulfill

$$(L_h u_h)(\mathbf{x}) = f(\mathbf{x}) \quad \text{for all } \mathbf{x} \in \mathcal{T}_h, \tag{2.6a}$$
$$u_h(\mathbf{x}) = g(\mathbf{x}) \quad \text{for all } \mathbf{x} \in \mathcal{G}_h, \tag{2.6b}$$

with

$$(L_h u_h)(\mathbf{x}) := -\Lambda^* u_h(\mathbf{x}) + c u_h(\mathbf{x}),$$

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

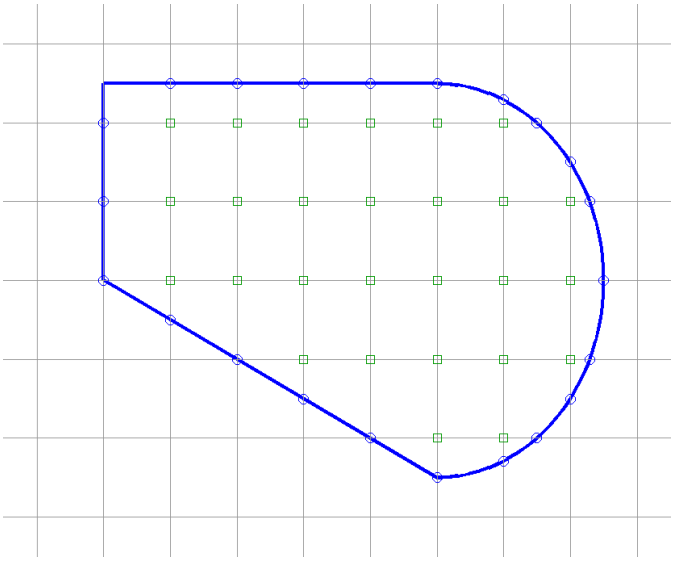


Figure 2.1: Grid $\mathcal{T}_h$ for a computational domain Ω and boundary grid $\mathcal{G}_h$.

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 $\mathbf{x} \in \mathcal{T}_h$, and $\widetilde{\mathbf{x}} \in \mathcal{T}_h$ such that

$$u_h(\widetilde{\mathbf{x}}) = \max_{\mathbf{x} \in \overline{\mathcal{T}}_h} u_h(\mathbf{x})$$

and $u_h(\widetilde{\mathbf{x}}) \geq 0$.

As consequence of the fact that $\widetilde{\mathbf{x}}$ is a maximum and an interior point it is

$$\begin{aligned}
\frac{2}{h_L + h_R} \left(\frac{u_h(\widetilde{x}_1 + h_R, \widetilde{x}_2) - u_h(\widetilde{x}_1, \widetilde{x}_2)}{h_R} - \frac{u_h(\widetilde{x}_1, \widetilde{x}_2) - u_h(\widetilde{x}_1 - h_L, \widetilde{x}_2)}{h_L} \right) \leq 0, \\
\frac{2}{h_B + h_T} \left(\frac{u_h(\widetilde{x}_1, \widetilde{x}_2 + h_T) - u_h(\widetilde{x}_1, \widetilde{x}_2)}{h_T} - \frac{u_h(\widetilde{x}_1, \widetilde{x}_2) - u_h(\widetilde{x}_1, \widetilde{x}_2 - h_B)}{h_B} \right) \leq 0,
\end{aligned}$$

and so

$$0 \geq (L_h u_h)(\widetilde{\mathbf{x}}) = -(\Lambda^* u_h)(\widetilde{\mathbf{x}}) + c(\widetilde{\mathbf{x}}) u_h(\widetilde{\mathbf{x}}) \geq c(\widetilde{\mathbf{x}}) u_h(\widetilde{\mathbf{x}}) \geq 0$$

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

$$\begin{aligned}
&c(\widetilde{\mathbf{x}}) = 0 \text{ and } u_h(\mathbf{x}) = u_h(\widetilde{\mathbf{x}}) \qquad \text{or} \\
&c(\widetilde{\mathbf{x}}) > 0 \text{ and } u_h(\mathbf{x}) = u_h(\widetilde{\mathbf{x}}) = 0
\end{aligned}$$

for all points $\mathbf{x}$ in the stencil of $\widetilde{\mathbf{x}}$.

If $\widetilde{\mathbf{x}} + (h_R, 0)^\top \in \mathcal{G}_T$ the statement of the lemma is proved, otherwise we repeat the previous steps for $\widetilde{\mathbf{x}} + (h_R, 0)^\top$ instead of $\widetilde{\mathbf{x}}$ until the (right) boundary of the domain is reached, and the lemma holds.

For $f \geq 0$ we consider $\widetilde{u}_h := -u_h$ and apply the previous steps. □

Corollary 2.11. *The only solution of* (2.6) *with* $f \equiv 0$ *on* $\mathcal{T}_h$ *and* $g \equiv 0$ *on* $\mathcal{G}_h$ *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*

$$\begin{aligned} |f_1(\mathbf{x})| &\leq f_2(\mathbf{x}) \quad \forall \mathbf{x} \in \mathcal{T}_h, \\ |g_1(\mathbf{x})| &\leq g_2(\mathbf{x}) \quad \forall \mathbf{x} \in \mathcal{G}_h. \end{aligned}$$

Then,

$$|u_{h,1}(\mathbf{x})| \leq u_{h,2}(\mathbf{x}) \quad \forall \mathbf{x} \in \overline{\mathcal{T}}_h.$$

Proof. Let $v_h := u_{h,1} - u_{h,2}$, then

$$\begin{aligned} L_h v_h = f_1 - f_2 &\leq 0 \quad \text{on } \mathcal{T}_h, \\ v_h = g_1 - g_2 &\leq 0 \quad \text{on } \mathcal{G}_h. \end{aligned}$$

If there is a non-negative maximum, it is attained on $\mathcal{G}_h$, and so $v_h(\mathbf{x}) \leq 0$ for all $\mathbf{x} \in \overline{\mathcal{T}}_h$. Otherwise, the maximum is negative and so even $v_h(\mathbf{x}) < 0$ for all $\mathbf{x} \in \overline{\mathcal{T}}_h$.

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

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 ℓ_∞ norm

$$\|v_h\|_\infty := \max_{\mathbf{x} \in \overline{\mathcal{T}}_h} |v_h(\mathbf{x})|$$

or the weighted ℓ_2 norm defined by

$$\|v_h\|_{2,h} := \frac{1}{\sqrt{|\mathcal{T}_h|}} \Big(\sum_{\mathbf{x} \in \overline{\mathcal{T}}_h} |v_h(\mathbf{x})|^2 \Big)^{1/2} \sim h \|v_h\|_2.$$

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 $\overline{\mathcal{T}}_{h_1}$ and $\overline{\mathcal{T}}_{h_2}$ 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 $\mathcal{T}_h$.

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

$$\|v_h\|_{2,h} \leq \|v_h\|_\infty. \tag{2.7}$$

Let $v \in C^0(\overline{\Omega})$. Then, we call the grid function $(R_h v) \in \overline{\mathcal{T}}_h$ its restriction to $\overline{\mathcal{T}}_h$.

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*

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

with $\|\cdot\|$ some norm on the mesh $\mathcal{T}_h$. 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 $\mathcal{T}_h$ the consistency error can be bounded as*

$$\begin{aligned} \|f - L_h u\|_\infty &\leq C_\infty h, \\ \|f - L_h u\|_{2,h} &\leq C_2\, h^{1+\ell/2}. \end{aligned}$$

Proof. Since for any $\mathbf{x} \in \mathcal{T}_h$

$$f - L_h u = -\Delta u + \Lambda^* u,$$

we have to bound the error of the finite difference approximation of the Laplacian.

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

For $u \in C^4(\Omega)$ we have on $O(h^{-2})$ points with the difference quotient Λ an point-wise error of $O(h^2)$ and on $O(h^{-1})$ points with the difference quotient Λ^* a point-wise error of $O(h)$ and so

$$\|\Delta u - \Lambda^* u\|_{2,h}^2 \leq O(h^2) \cdot \big(O(h^{-2})O(h^4) + O(h^{-1})O(h^2)\big) = O(h^3).$$

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. □

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_∞-norm)**.** *There exists a constant $C > 0$ only depending on the size of Ω such that for the solution u_h of* (2.6) *it holds*

$$\|u_h\|_\infty \leq \frac{\operatorname{diam}(\Omega)^2}{4} \|f\|_\infty + \|g\|_\infty.$$

Proof. Without loss of generality let $\mathbf{0} \in \Omega$ and let $R = \operatorname{diam}(\Omega)$. Then, for any constants $\alpha, \beta > 0$ the grid function

$$v_h = \alpha(R^2 - x_1^2 - x_2^2) + \beta \tag{2.8}$$

is positive for all $\mathbf{x} \in \overline{\mathcal{T}}_h$. Inserting (2.8) into Λ^* we have

$$\begin{aligned}
(\Lambda^* v_h)(\mathbf{x}) &= -\Lambda^*(x_1^2 + x_2^2) \\
&= -\alpha\Bigg(\frac{2}{h_L + h_R} \left(\frac{(x_1 + h_R)^2 - x_1^2}{h_R} - \frac{x_1^2 - (x_1 - h_L)^2}{h_L} \right) \\
&\qquad + \frac{2}{h_T + h_B} \left(\frac{(x_2 + h_T)^2 - x_2^2}{h_T} - \frac{x_2^2 - (x_2 - h_B)^2}{h_B} \right) \Bigg) \\
&= -4\,\alpha < 0.
\end{aligned}$$

Thus, v_h is by Corollary 2.11 the unique solution of

$$\begin{aligned}
(L_h v_h)(\mathbf{x}) &= \alpha\left(4 + c(\mathbf{x})(R^2 - x_1^2 - x_2^2)\right) + \beta =: \widetilde{f} > 0 && \text{on } \mathcal{T}_h, \\
v_h(\mathbf{x}) &= \alpha(R^2 - x_1^2 - x_2^2) + \beta =: \widetilde{g} > 0 && \text{on } \mathcal{G}_h.
\end{aligned}$$

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

$$\begin{aligned}
|f(\mathbf{x})| &\le \widetilde{f}(\mathbf{x}) \quad \forall \mathbf{x} \in \mathcal{T}_h, \\
|g(\mathbf{x})| &\le \widetilde{g}(\mathbf{x}) \quad \forall \mathbf{x} \in \mathcal{G}_h.
\end{aligned}$$

By Lemma 2.13 we have the inequality

$$|u_h(\mathbf{x})| \le v_h(\mathbf{x}) \quad \forall \mathbf{x} \in \overline{\mathcal{T}}_h,$$

and so

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

□

Lemma 2.17 (Stability of the FDM in the weighted ℓ_2-norm)**.** *Let $g \equiv 0$ and $\mathcal{T}_h$ a uniform grid. There exists a constant $C > 0$ only depending on the size of Ω such that for the solution u_h of* (2.6) *it holds*

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

Proof. As both grid functions u_h and f have same length it suffices to show the lemma directly for the ℓ_2 norm.

Since $u_h \equiv 0$ on $\mathcal{G}_h$ by (2.6b), it remains the linear system of equations (2.6a) with $N = |\mathcal{T}_h|$ unknowns (without values on the boundary) and N equations:

$$\mathbf{A}u_h = f,$$

where we give to the points in $\mathcal{T}_h$ an order. The matrix $\mathbf{A}$ is symmetric (for uniform grid $\mathcal{T}_h$), and so its eigenvalues are real.

The eigenvalues of $\mathbf{A}$ 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 $\mathcal{G}_h$ the grid function v_h is unique solution of

$$\begin{aligned}(L_{h,\lambda}v_h)(\mathbf{x}) := ((L_h - \lambda)v_h)(\mathbf{x}) = 0 \quad &\text{for all } \mathbf{x} \in \mathcal{T}_h,\\ v_h(\mathbf{x}) = 0 \quad &\text{for all } \mathbf{x} \in \mathcal{G}_h,\end{aligned}$$

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

If the inverse of $\mathbf{A}$ exists, it holds (without absolute value)

$$\|\mathbf{A}^{-1}\|_2 = \lambda_{\max}(\mathbf{A}^{-1}) = \max_{1\leq k\leq N} \lambda_k(\mathbf{A}^{-1}).$$

For the corresponding eigenvectors $v_{h,k}$ it is

$$\mathbf{A}^{-1}v_{h,k} = \lambda_k(\mathbf{A}^{-1})v_{h,k} \quad \Leftrightarrow \lambda_k^{-1}(\mathbf{A}^{-1})v_{h,k} = \mathbf{A}v_{h,k},$$

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

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

where we assume an order of the eigenvalues of $\mathbf{A}$.

The discrete eigenvalues $\lambda_{h,k}$ are upper bounds to the eigenvalues λ_k of the continuous problem in Ω which can be proved by the minimax principle for variational problems using piecewise-linear eigenfunctions as shown for example by Kuttler [2].

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 λ_1 in the domain Ω, which can be bounded from below by the lowest continuous eigenvalue in the square domain with side lengths $\operatorname{diam}(\Omega)$, which is larger than $2\pi^2/(\operatorname{diam}(\Omega))^2$ (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*

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

which is consequence of Lemma 2.16 and (2.7).

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

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

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*

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

Proof. We have the systems

$$\begin{aligned} (L_h u_h)(\mathbf{x}) &= f(\mathbf{x}) \text{ for all } \mathbf{x} \in \mathcal{T}_h & (L_h u)(\mathbf{x}) &= (L_h u)(\mathbf{x}) \text{ for all } \mathbf{x} \in \mathcal{T}_h \\ u_h(\mathbf{x}) &= g(\mathbf{x}) \text{ for all } \mathbf{x} \in \mathcal{G}_h & u(\mathbf{x}) &= g(\mathbf{x}) \quad \text{for all } \mathbf{x} \in \mathcal{G}_h, \end{aligned}$$

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

$$\begin{aligned} (L_h(u_h - u))(\mathbf{x}) &= (f - L_h u)(\mathbf{x}) \text{ for all } \mathbf{x} \in \mathcal{T}_h \\ (u_h - u)(\mathbf{x}) &= 0 \text{ for all } \mathbf{x} \in \mathcal{G}_h. \end{aligned}$$

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

$$\begin{aligned} \|u_h - u\|_\infty &\le C(\Omega)\|f - L_h u\|_\infty \le C\, h, \\ \|u_h - u\|_{2,h} &\le C(\Omega)\|f - L_h u\|_{2,h} \le C\, h^{1+\ell/2}. \end{aligned}$$

□

2.3.3 Implementational issues

The representation of (2.6) with separation of finite differences and boundary approximation is advantageously 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 $\mathcal{T}_h$, this is by ommiting (2.6b) and by moving the known values of u_h on $\mathcal{G}_h$ to the right hand side of (2.6a).

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

$$\begin{pmatrix} -\frac{2h_L^{-1}}{h_L+h} & \frac{2(h_L^{-1}+h^{-1})}{h_L+h} & -\frac{2h_L^{-1}}{h_L+h} & \cdot & \cdot & & \\ \cdot & -h^{-2} & 2h^{-2} & -h^{-2} & \cdot & \cdot & \cdot \\ & & \ddots & \ddots & \ddots & & \\ \cdot & \cdot & \cdot & -h^{-2} & 2h^{-2} & -h^{-2} & \cdot \\ \cdot & \cdot & \cdot & \cdot & -\frac{2h_R^{-1}}{h_R+h} & \frac{2(h_R^{-1}+h^{-1})}{h_R+h} & -\frac{2h_R^{-1}}{h_R+h} \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 \end{pmatrix} \begin{pmatrix} u_L \\ u_1 \\ \vdots \\ u_N \\ u_R \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{N-1} \\ f_N \\ g_L \\ g_R \end{pmatrix}$$

the reduced one

$$\begin{pmatrix} \frac{2(h_L^{-1}+h^{-1})}{h_L+h} & -\frac{2h_L^{-1}}{h_L+h} & \cdot & & & \\ -h^{-2} & 2h^{-2} & -h^{-2} & \cdot & \cdot & \\ & \ddots & \ddots & \ddots & & \\ \cdot & \cdot & \cdot & -h^{-2} & 2h^{-2} & -h^{-2} \\ \cdot & \cdot & \cdot & \cdot & -\frac{2h_R^{-1}}{h_R+h} & \frac{2(h_R^{-1}+h^{-1})}{h_R+h} \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_N \end{pmatrix} = \begin{pmatrix} f_1 + g_L \frac{2h_L^{-1}}{h_L+h} \\ f_2 \\ \vdots \\ f_{N-1} \\ f_N + g_R \frac{2h_R^{-1}}{h_R+h} \end{pmatrix}.$$

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 straight forward (see last example). In two space dimensions often rectangular domains are used in practise, *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

$$\begin{aligned} I(0, j_{\min}(0)) &= 0, \\ I(i, j_{\min}(i)) &= I(i-1, j_{\max}(i-1)) + 1, \\ I(i,j) &= I(i, j_{\min}(i)) + j - j_{\min}(i). \end{aligned}$$

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 ℓ_∞ and weighted ℓ_2 norm.

Contras

- Limitation of the application to C^2 continuous solutions (of the strong problem).
- Limitations of the grid to simple geometries and continuous material function $\sigma(\mathbf{x})$ in case of a $\operatorname{div} \sigma(\mathbf{x})\, \mathbf{grad}$ 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 [3] (small support) or cubic Z-splines [4] (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

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

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

$$Z_2(x) = \begin{cases} 1 - \frac{5}{2}x^2 + \frac{3}{2}|x|^3 & \text{for} \quad |x| \le 1, \\ \frac{1}{2}(2 - |x|)^2(1 - |x|) & \text{for} \quad 1 \le |x| \le 2, \\ 0 & \text{for} \quad 2 \le |x|. \end{cases}$$

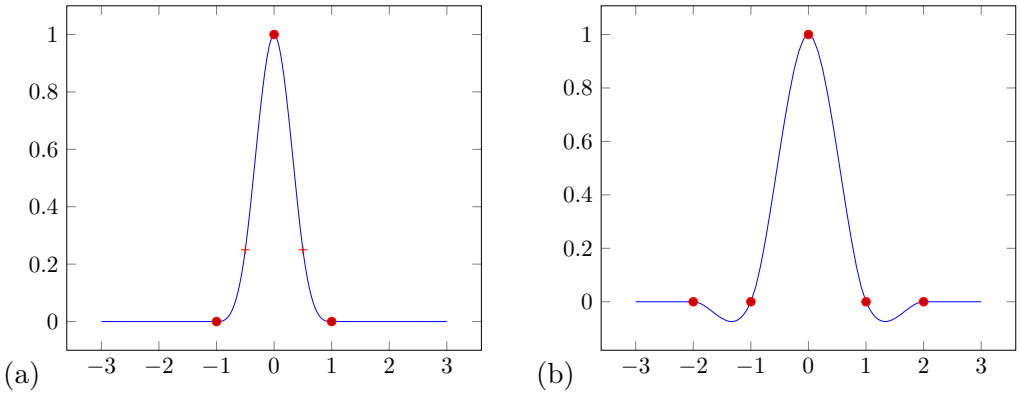


Figure 2.2: Cubic B-spline (a) and Z-spline (b) basis functions for equidistant points.

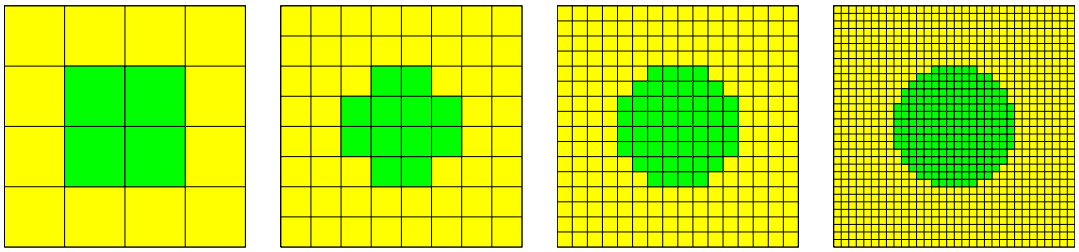


Figure 2.3: Finite difference grids with two materials, one with a circular shape.

References

[1] William McLean. *Strongly Elliptic Systems and Boundary Integral Equations*. Cambridge University Press, 2000.

[2] J. R. Kuttler and V. G. Sigillito. Eigenvalues of the Laplacian in two dimensions. *SIAM Review*, 26(2):pp. 163–193, 1984.

[3] Carl de Boor. *A Practical Guide to Splines*. Springer Verlag, New York, 2001.

[4] Julian B. Sagredo. Z-splines: moment conserving cardinal spline interpolation of compact support for arbitrarily spaced data. SAM Report 2003-10, ETH Zürich, Seminar for Applied Mathematics, Aug 2003.

3 Elliptic Boundary Value Problems

3.1 Elliptic partial differential equations of second order

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

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

where $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N \cup \overline{\Gamma}_R$, the reaction term $c \in L^\infty(\Omega)$, and the diffusion matrix $\mathbf{C} \in (L^\infty(\Omega))^{d\times d}$, and there exists constants $c_0 \geq 0$, $c_1 > 0$ such that $c \geq c_0$ and $\operatorname{ess\,inf}_{\mathbf{x}\in\Omega}(\inf_{\mathbf{v}\in\mathbb{R}^d,||v||=1}(\mathbf{C}(\mathbf{x})\mathbf{v})\cdot\mathbf{v}) \geq c_1$. The last boundary condition (3.1c) is the Robin boundary condition which models an resistant (absorbing) outer medium in $\mathbb{R}^d\backslash\overline{\Omega}$.

We consider the partial differential equations of interest on bounded, connected, and open subsets of (the affine space) $\mathbb{R}^d$, $d = 1, 2, 3$. These are called the (spatial) **domains** of related boundary value problem and will be denoted by Ω. The topological closure of Ω 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. *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

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

Definition. *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*

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

which satisfies

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

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

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

Lipschitz domains have a boundary which is "slighly" 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.

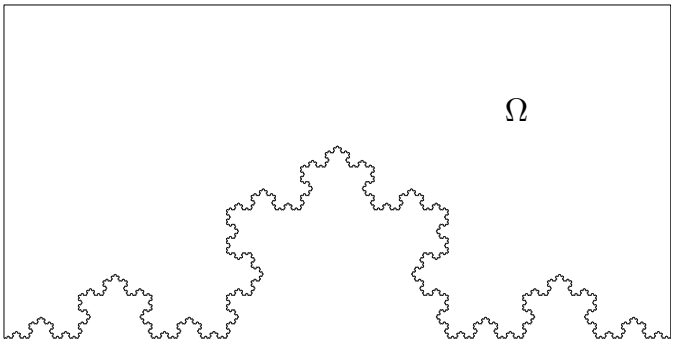


Figure 3.1: Domain whose boundary is the limit of a Koch curve is not Lipschitz (bounded domain, but boundary of infinite length).

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

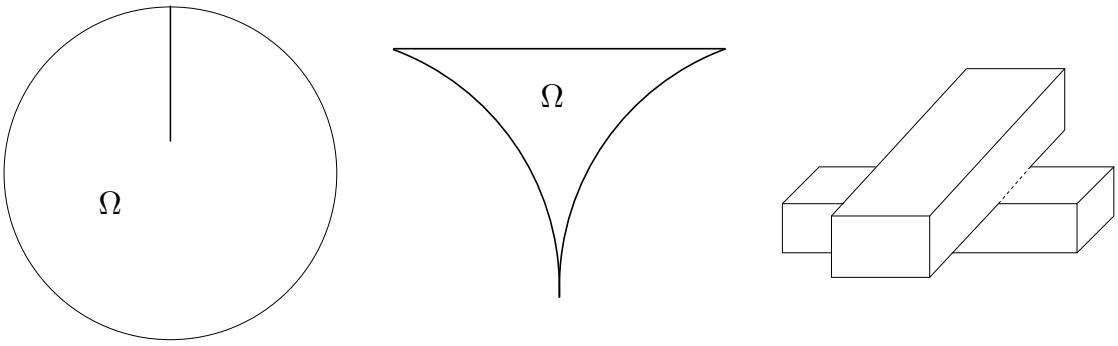


Figure 3.2: Domains that are not Lipschitz: slit domain (left), cusp domain (middle), crossing edges (right).

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

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

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. *In the case $d = 2$ a connected domain Ω is called a* ***curvilinear Lipschitz polygon****, if Ω 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*

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

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

The boundary segments are called edges, their endpoints are the vertices of Ω. A tangential direction can be defined for all points of an edge including the endpoints, which gives rise to the concept of an angle at a vertex, see Fig. 3.3. The mappings $\mathbf{\Phi}_k$ can be regarded as smooth parametrizations of the edges. An analoguous notion exists

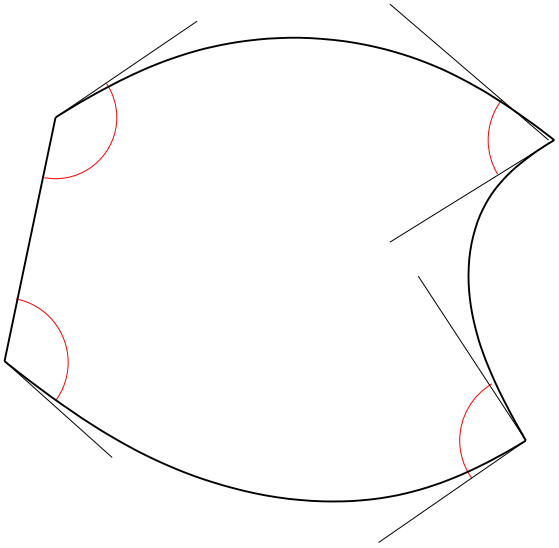


Figure 3.3: Curvilinear polygon with added angles at vertices

in three dimensions.

Definition 3.1 (Classical solution)**.** *If $u \in C^2(\Omega) \cap C(\overline{\Omega})$ satisfy* (3.1) *(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.*, (3.1) with $c = 0$, $\mathbf{C} = \mathbf{I}$. Here, we can make even a statement about the regularity of the solution.

Lemma 3.2 (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* (3.1) *with $\Gamma_D = \partial\Omega$, $c = 0$, $\mathbf{C} = \mathbf{I}$. Then $u \in C^{k+2}(\Omega)$.*

Lemma 3.3 (Continuous extension theorem)**.** *Let u be the solution of* (3.1)*. Let $\Omega' \subset \Omega$ for which $\mathbf{C} \in (C^1(\Omega'))^{d\times d}$ and $u \in H^2_{\text{loc}}(\Omega')$, and for any $D \subset \Omega'$ there exists a constant $C > 0$ such that*

$$|(\mathbf{C})_{i,j}\partial_i\partial_j u| \leq C(|u| + \|\,\mathbf{grad}\, u\|) \quad \text{a.e. in } D.$$

If furthermore $u \equiv 0$ in an open bounded subset $G \subset \Omega'$. Then, $u \equiv 0$ in Ω'.

Proof. See [5, Chap.4.3]. □

3.2 Linear differential operators

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

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

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

$$\partial^{\boldsymbol{\alpha}} := \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_n}}{\partial x_n^{\alpha_n}}$$

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

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

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

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

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

$$\mathbf{curl}\, \mathbf{f}(\mathbf{x}) := \begin{pmatrix} \dfrac{\partial f_3}{\partial x_2}(\mathbf{x}) - \dfrac{\partial f_2}{\partial x_3}(\mathbf{x}) \\ \dfrac{\partial f_1}{\partial x_3}(\mathbf{x}) - \dfrac{\partial f_3}{\partial x_1}(\mathbf{x}) \\ \dfrac{\partial f_2}{\partial x_1}(\mathbf{x}) - \dfrac{\partial f_1}{\partial x_2}(\mathbf{x}) \end{pmatrix} = \nabla \times f(\mathbf{x})\,, \quad \mathbf{x} \in \Omega \qquad \text{with } \nabla = \begin{pmatrix} \dfrac{\partial}{\partial x_1} \\ \dfrac{\partial}{\partial x_2} \\ \dfrac{\partial}{\partial x_3} \end{pmatrix},$$

where $\times$ stands for antisymmetric vector product.

Moreover, we have

$$curl \circ \mathbf{grad} = \mathbf{0}, \qquad \operatorname{div} \circ \mathbf{curl} = 0.$$

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

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

Theorem 3.4 (Gauß' theorem)**.** *If* $\mathbf{f} \in (C^1(\Omega))^d \cap (C(\overline{\Omega}))^d$*, then*

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

Proof. Please consult [6, § 15] and [7]. □

By the product rule

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

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

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

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

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

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

$$\int_\Omega \langle \mathbf{grad}\, v, \mathbf{grad}\, u \rangle + \Delta v \, u \, \mathrm{d}\mathbf{x} = \int_\Gamma \langle \mathbf{grad}\, v, \mathbf{n} \rangle \, u \, \mathrm{d}S$$

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

In three dimensions, $d = 3$, another product rule

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

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

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

Remark. *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. *Consider heat conduction in a plane wall composed of two layers of equal thickness and with heat conductivity coefficients κ_1 and κ_2. The inside of the wall is kept at temperature $u = u_1$, the outside at $u = 0$. Provided that the width and the height of the wall are much greater than its thickness, a one-dimensional model can be used. After spatial scaling it boils down to*

$$j = -\kappa(x) \frac{d}{dx} u \quad , \quad \frac{d}{dx} j = 0 \quad , \quad u(0) = 0 \, , \quad u(1) = u_1 \, , \tag{3.3}$$

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

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

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

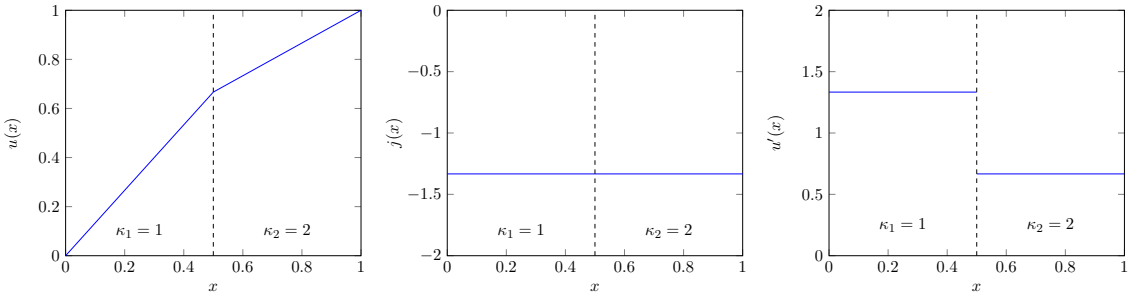


Figure 3.4: Non-classical solution $u(x)$ of Example 3.4 for $\kappa_1 = 1$ and $\kappa_2 = 2$.

Definition 3.5 (Test function space)**.** *For a non-empty open set $\Omega \subset \mathbb{R}^d$ we denote*

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

the space of test functions, which is the space of functions $C^\infty(\overline{\Omega})$ with compact support.

Lemma 3.6. *Two integrable functions f and g defined in the bounded set Ω are almost everywhere equal if and only if*

$$\int_\Omega fv \,\mathrm{d}\mathbf{x} = \int_\Omega gv \,\mathrm{d}\mathbf{x}$$

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

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

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 $\mathbf{grad}\, u$ of a function $u \in L^2(\Omega)$ will be vectorfield $\mathbf{w} \in (L^2(\Omega))^d$ with

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

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

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

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

Theorem 3.8. *If $u \in C^m(\overline{\Omega})$, then all weak derivatives of order $\leq m$ agree in $L^2(\Omega)$ with the corresponding classical derviatives.*

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

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

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

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

Lemma 3.9. *Let $\Omega \subset \mathbb{R}^d$ be bounded with Lipschitz boundary and assume a partition $\overline{\Omega} = \overline{\Omega}_1 \cup \overline{\Omega}_2$, $\Omega_1 \cap \Omega_2 = \emptyset$, where both sub-domains are supposed to have a Lipschitz boundary, too. Assume that the restriction of the function $u \in L^2(\Omega)$ to Ω_ℓ, $\ell = 1, 2$, belongs to $C^1(\Omega_\ell)$ and that $u_{|\Omega_\ell}$ can be extended to a function in $C^0(\overline{\Omega_\ell})$.*

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

$$\frac{\partial u}{\partial x_k}(\mathbf{x}) = \left\{ \begin{array}{ll} \frac{\partial}{\partial x_k} u_{|\Omega_1} & \text{if } \mathbf{x} \in \Omega_1\ , \\ \frac{\partial}{\partial x_k} u_{|\Omega_2} & \text{if } \mathbf{x} \in \Omega_2 \end{array} \right\} \in L^2(\Omega). \tag{3.4}$$

Proof. Using locally supported test functions in the definition of the weak derivative, it is clear that (3.4) supplies the only meaningful candidate for the weak derivative of u. Then we appeal to (FGF) and the fact that any crossing direction $\mathbf{n}_\Sigma$ of the interface $\Sigma := \partial\Omega_1 \cap \partial\Omega_2$ will be parallel to the exterior unit normal of one sub-domain, and anti-parallel to that of the other. Thus, we get the identity

$$\begin{aligned} \int_\Omega \langle \mathbf{grad}_{\mathrm{cl}}\, u, \mathbf{v}\rangle \,\mathrm{d}\mathbf{x} &= \int_{\Omega_1} \langle \mathbf{grad}_{\mathrm{cl}}\, u, \mathbf{v}\rangle \,\mathrm{d}\mathbf{x} + \int_{\Omega_2} \langle \mathbf{grad}_{\mathrm{cl}}\, u, \mathbf{v}\rangle \,\mathrm{d}\mathbf{x} \\ &= -\int_\Omega u \,\mathrm{div}\, \mathbf{v} \,\mathrm{d}\mathbf{x} + \int_\Sigma [u]_\Sigma \underbrace{\langle \{\mathbf{v}\}_\Sigma, \mathbf{n}_\Sigma\rangle}_{=\langle \mathbf{v}, \mathbf{n}_\Sigma\rangle} \,\mathrm{d}S + \int_\Sigma \{u\}_\Sigma \underbrace{\langle [\mathbf{v}]_\Sigma, \mathbf{n}_\Sigma\rangle}_{=0} \,\mathrm{d}S\ , \end{aligned}$$

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

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

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

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

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

Remark 3.10. *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 Ω which is in $L^2(\Omega)$, does not guarantee that it has a weak derivative in the whole domain Ω. For example a function which is piecewise $C^\infty(\overline{\Omega}_\ell)$, $\ell = 1, 2$ (notation like in Lemma 3.4), but discontinuous over Σ, does not have a weak gradient.*

Corollary 3.11. *Under the geometric assumptions of the previous lemma let $u_{|\Omega_\ell}$ belong to $C^m(\Omega_\ell)$ with possible extension to $C^{m-1}(\overline{\Omega_\ell})$. Then*

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

Lemma 3.12. *We retain the assumptions of Lemma 3.9 with the exception that u is replaced by a vectorfield $\mathbf{u} \in (L^2(\Omega))^d$ with restrictions $\mathbf{u}_{|\Omega_l} \in (C^1(\Omega_\ell))^d$ that can be extended to continuous functions on $\overline{\Omega}_\ell$, $\ell = 1, 2$.*

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

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

Example. *For the PDE*

$$-\operatorname{div} \boldsymbol{C}\,\mathbf{grad}\, u + cu = f \text{ in } \Omega, \tag{3.1a}$$

with smooth $\boldsymbol{C}$, c and f and so smooth u in subdomains Ω_1 and Ω_2 of Ω. A weak gradient $\mathbf{grad}\, u$ in Ω exists if $u \in C^0(\overline{\Omega})$ (Lemma 3.9) and a weak divergence $\operatorname{div} \boldsymbol{C}\,\mathbf{grad}\, u$ exists if the normal component of $\boldsymbol{C}\,\mathbf{grad}\, u$ (that is the co-normal derivative) is continuous, i. e., on the interface Σ between Ω_1 and Ω_2 it holds

$$\Big[(\boldsymbol{C}\,\mathbf{grad}\, u)\cdot \mathbf{n}_\Sigma\Big]_\Sigma = 0.$$

3.5 Weak formulations

3.5.1 Pure Dirichlet boundary conditions

Interpreting the div in (3.1a) as weak derivative, *i. e.*,

$$\int_\Omega \underbrace{\operatorname{div} \boldsymbol{C}\,\mathbf{grad}\, u}_{w} v \mathrm{d}\mathbf{x} = -\int_\Omega \boldsymbol{C}\,\mathbf{grad}\, u \cdot \mathbf{grad}\, v \mathrm{d}\mathbf{x},$$

we can equivalently write

$$\int_\Omega \langle \boldsymbol{C}\,\mathbf{grad}\, u, \mathbf{grad}\, v\rangle + c\,uv\,\mathrm{d}\mathbf{x} = \int_\Omega f v\,\mathrm{d}\mathbf{x} \tag{3.5}$$

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 (3.5) – have to fulfill.

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

If $g = 0$ we speak about *homogeneous Dirichlet boundary conditions*, for arbitrary g about *inhomogeneous Dirichlet boundary conditions*. For the latter case we can decompose the solution as

$$u = u_g + u_0 \qquad \text{with} \qquad u_g = g \text{ on } \partial\Omega \qquad \text{and} \qquad u_0 = 0 \text{ on } \partial\Omega,$$

with u_g given. Inserting this decomposition into (3.5) we search for u_0 satisfying

$$\begin{aligned}\int_\Omega \langle \boldsymbol{C}\,\mathbf{grad}\, u_0, \mathbf{grad}\, v\rangle + c\, u_0 v \,\mathrm{d}\mathbf{x} &= \int_\Omega f v \,\mathrm{d}\mathbf{x} - \int_\Omega \langle \boldsymbol{C}\,\mathbf{grad}\, u_g, \mathbf{grad}\, v\rangle + c\, u_g v \,\mathrm{d}\mathbf{x} \\ &= \int_\Omega (f + \operatorname{div} \boldsymbol{C}\,\mathbf{grad}\, u_g - c u_g)\, v \,\mathrm{d}\mathbf{x}.\end{aligned}$$

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

$$\int_\Omega \langle \boldsymbol{C}\,\mathbf{grad}\, u, \mathbf{grad}\, v\rangle + c\, uv \,\mathrm{d}\mathbf{x} - \int_{\partial\Omega} \langle \boldsymbol{C}\,\mathbf{grad}\, u, \mathbf{n}\rangle \; v \,\mathrm{d}S = \int_\Omega f v \,\mathrm{d}\mathbf{x}, \qquad \text{(FWP)}$$

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

$$\int_\Omega \langle \boldsymbol{C}\,\mathbf{grad}\, u, \mathbf{grad}\, v\rangle + c\, uv \,\mathrm{d}\mathbf{x} = \int_\Omega f v \,\mathrm{d}\mathbf{x} + \int_{\partial\Omega} h\, v \,\mathrm{d}S. \qquad (3.6)$$

Boundary conditions which are present in the variational formulation are called **natural**.

To see that (3.6) 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

$$\int_{\partial\Omega} \langle \boldsymbol{C}\,\mathbf{grad}\, u, \mathbf{n}\rangle \; v \,\mathrm{d}S = \int_{\partial\Omega} hv \mathrm{d}S$$

which is by Lemma 3.6 equivalent to (3.1c).

3.5.3 Robin boundary conditions

For the case of pure Robin boundary conditions ($\Gamma_R = \partial\Omega$), where Neumann boundary conditions are included with $\beta = 0$, inserting (3.1d) into (FWP) leads to

$$\int_\Omega \langle \boldsymbol{C}\,\mathbf{grad}\, u, \mathbf{grad}\, v\rangle + c\, uv \,\mathrm{d}\mathbf{x} - \int_{\partial\Omega} \underbrace{\langle \boldsymbol{C}\,\mathbf{grad}\, u, \mathbf{n}\rangle}_{-\beta u + h} \; v \,\mathrm{d}S = \int_\Omega f v \,\mathrm{d}\mathbf{x}$$

and so

$$\int_\Omega \langle \boldsymbol{C}\,\mathbf{grad}\, u, \mathbf{grad}\, v\rangle + c\, uv \,\mathrm{d}\mathbf{x} + \int_{\partial\Omega} \beta u\, v \,\mathrm{d}S = \int_\Omega f v \,\mathrm{d}\mathbf{x} + \int_{\partial\Omega} h\, v \,\mathrm{d}S \qquad \forall v \in C^\infty(\Omega) \cap C(\overline{\Omega}).$$

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

$$\mathbf{j} = -\boldsymbol{C}(\mathbf{x})\ \mathbf{grad}\, u\ , \tag{FL}$$
$$\operatorname{div}\mathbf{j} + cu = f, \tag{EL}$$

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

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

$$\begin{aligned} \int_\Omega \langle \mathbf{j}, \mathbf{q}\rangle \ \mathrm{d}\mathbf{x} &= -\int_\Omega \langle \mathbf{C}\,\mathbf{grad}\, u, \mathbf{q}\rangle \ \mathrm{d}\mathbf{x} && \forall \mathbf{q}\ , \\ -\int_\Omega \langle \mathbf{j}, \mathbf{grad}\, v\rangle + cuv \,\mathrm{d}\mathbf{x} + \int_\Gamma \langle \mathbf{j}, \mathbf{n}\rangle \ v \,\mathrm{d}S &= \int_\Omega fv \,\mathrm{d}\mathbf{x} && \forall v\ . \end{aligned} \tag{3.7}$$

Choosing $\mathbf{q} = \mathbf{grad}\, v$ 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:

$$\begin{aligned} -\int_\Omega \left\langle \boldsymbol{C}^{-1}\mathbf{j}, \mathbf{q}\right\rangle + u\ \operatorname{div}\mathbf{q}\,\mathrm{d}\mathbf{x} &= \int_{\partial\Omega} u\ \langle \mathbf{q}, \mathbf{n}\rangle \ \mathrm{d}S && \forall \mathbf{q}\ , \\ \int_\Omega \operatorname{div}\mathbf{j}\, v + cuv \,\mathrm{d}\mathbf{x} &= \int_\Omega fv \,\mathrm{d}\mathbf{x} && \forall v\ . \end{aligned} \tag{3.8}$$

Here, elimination (of u or $\mathbf{j}$) is not possible in general (see below).

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 $\mathbf{j}$ 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 $u = \beta^{-1}(h - \langle \mathbf{j}, \mathbf{n}\rangle)$, hence, they are natural.

For the notion "dual formulation" and analysis of respective finite element methods see [8].

Remark. *If $c \geq c_0 > 0$ a.e. in Ω and smooth enough and for smooth enough f we can eliminate u, since with*

$$\mathbf{grad}\, c^{-1} \operatorname{div}\mathbf{j} + \mathbf{grad}\, u = \mathbf{grad}(c^{-1} f),$$

we find

$$\mathbf{j} - \boldsymbol{C}\,\mathbf{grad}\, c^{-1} \operatorname{div}\mathbf{j} = -\boldsymbol{C}\,\mathbf{grad}(c^{-1} f).$$

Multiplying with $\boldsymbol{C}^{-1}$ from the left and with test functions $\mathbf{q}$ we obtain

$$\int_\Omega \left\langle \boldsymbol{C}^{-1}\mathbf{j}, \mathbf{q}\right\rangle - \left\langle \mathbf{grad}\, c^{-1} \operatorname{div} \mathbf{j}, \mathbf{q}\right\rangle \,\mathrm{d}\mathbf{x} = -\int_\Omega \left\langle \mathbf{grad}(c^{-1}f), \mathbf{q}\right\rangle \,\mathrm{d}\mathbf{x}$$

and so

$$\begin{aligned}\int_\Omega c^{-1} \operatorname{div}\mathbf{j} \operatorname{div}\mathbf{q} + \left\langle \boldsymbol{C}^{-1}\mathbf{j}, \mathbf{q}\right\rangle \,\mathrm{d}\mathbf{x} &= -\int_\Omega \left\langle \mathbf{grad}(c^{-1}f), \mathbf{q}\right\rangle \,\mathrm{d}\mathbf{x} + \int_{\partial\Omega} \underbrace{c^{-1}\operatorname{div}\mathbf{j}}_{c^{-1}f-u} \left\langle \mathbf{q}, \mathbf{n}\right\rangle \,\mathrm{d}S \\ &= \int_\Omega c^{-1} f \operatorname{div}\mathbf{q}\,\mathrm{d}\mathbf{x} - \int_{\partial\Omega} u \left\langle \mathbf{q}, \mathbf{n}\right\rangle \,\mathrm{d}S.\end{aligned}$$

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

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

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

Definition. *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*

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

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

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

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

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

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

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

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

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

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

Definition. *A bilinear form on a vector space V is called* **continuous** *if there exists a constant $C > 0$ such that*

$$|b(v,w)| \leq C\|v\|_V \|w\|_V \quad \forall v, w \in V.$$

In this case, we call $\|b\|_{V\times V\to\mathbb{R}} := \sup_{(v,w)\in V\times V} \frac{|b(v,w)|}{\|v\|_V \|w\|_V}$.

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

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

3.7 Definition of Sobolev spaces

3.7.1 Banach spaces

Definition. *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. *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

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

The fundamental *Cauchy-Schwarz-inequality*

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

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

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

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

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

Lemma. *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. *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

$$\int_\Omega \langle \boldsymbol{C}\,\mathbf{grad}\, u, \mathbf{grad}\, v\rangle + c\,uv\,\mathrm{d}\mathbf{x} = \int_\Omega fv\,\mathrm{d}\mathbf{x} + \int_{\partial\Omega} hv\,\mathrm{d}S \quad \forall v \in C^\infty(\Omega)\cap C(\overline{\Omega})\ . \tag{3.9}$$

The bilinear form in (3.9) is symmetric positive definite if $\boldsymbol{C}$ is symmetric positive definite with $0 < c_0 \le \langle \boldsymbol{C}(\mathbf{x})\boldsymbol{\xi},\boldsymbol{\xi}\rangle$ for all $\boldsymbol{\xi} \in \mathbb{R}^d$ and $0 < c_0 < c(\mathbf{x})$ for almost all $\mathbf{x} \in \Omega$. It is continuous if $\langle \boldsymbol{C}\boldsymbol{\xi},\boldsymbol{\eta}\rangle \le c_1 < \infty$ forall $\boldsymbol{\xi},\boldsymbol{\eta} \in \mathbb{R}^d$ and $c(\mathbf{x}) \le c_1 < \infty$ for almost all $\mathbf{x} \in \Omega$.

We already know that **grad** has to be used in distributional sense. The concrete spaces have deliberately been omitted in (3.9), because we want to heed the guideline formulated in the context of Example on page 29 and set out from (3.9) and design the "ideal" space. It goes without saying that the investigation of (3.9) 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

$$\|v\|_e^2 := \int_\Omega \langle \boldsymbol{C}\,\mathbf{grad}\, v, \mathbf{grad}\, v\rangle + c|v|^2\,\mathrm{d}\mathbf{x} \tag{3.10}$$

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

$$H := \{v : \Omega \mapsto \mathbb{R} :\ \text{weak } \mathbf{grad}\, v \text{ exists and energy norm (3.10) of } v < \infty\}\ .$$

Definition (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:*

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

with norm

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

where $D = \operatorname{diam}(\Omega)$ 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

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

In this case H and $H^1(\Omega)$ incorporate the same functions.

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

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

equipped with the norm

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

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

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

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

denotes a semi-norm on $H^m(\Omega)$. Hence, we can write

$$\|v\|_{H^m(\Omega)} = \left(\sum_{k=0}^{m} |v|^2_{H^k(\Omega)} \right)^{\frac{1}{2}} .$$

The Sobolev spaces are a promising framework for variational problems, see [9, Thm. 3.1]:

Theorem 3.13. *The Sobolev spaces $H^m(\Omega)$, $m \in \mathbb{N}_0$, are Hilbert spaces with the inner product*

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

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

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

where the corresponding norm on this space is given by

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

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

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

An analogue of Thm. 3.13 holds true for such spaces $H(\mathsf{D}; \Omega)$.

Example. *The most important representatives of spaces covered by Def. 3.7.3 are*

$$\begin{aligned} H(\operatorname{div}; \Omega) &:= \{\mathbf{u} \in (L^2(\Omega))^d : \ \operatorname{div} \mathbf{u} \in L^2(\Omega)\} , \\ H(\mathbf{curl}; \Omega) &:= \{\mathbf{u} \in (L^2(\Omega))^3 : \ \mathbf{curl}\,\mathbf{u} \in (L^2(\Omega))^3\} , \\ H(\Delta, \Omega) &:= \{v \in H^1(\Omega) : \ \Delta u \in L^2(\Omega)\} . \end{aligned}$$

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

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

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

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

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

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

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

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

Then we can state a key result in the theory of Sobolev spaces, the famous Meyers-Serrin theorem, whose proof is way beyond the scope of these lecture notes, see [9, Thm. 3.6]:

Theorem 3.14. *The space $C^\infty(\overline{\Omega})$ is a dense subspace of $H^m(\Omega)$ for all $m \in \mathbb{N}_0$. Moreover, the space $(C^\infty(\overline{\Omega}))^d$ is a dense subspace of $H(\mathrm{div};\Omega)$ and $H(\mathbf{curl};\Omega)$ ($d = 3$ in the latter case).*

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 [10, 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. 3.14 we can give an alternative definition of the Sobolev spaces.

Corollary 3.15. *The spaces $H^m(\Omega)$, $H(\mathrm{div};\Omega)$, and $H(\mathbf{curl};\Omega)$ (for $d = 3$) can be obtained by the completion of spaces of smooth functions with respect to the corresponding norms.*

E.g. $\overline{C^\infty(\Omega)}^{\|\cdot\|_{H^m(\Omega)}} = H^m(\Omega)$.

Remark. *Thm. 3.14 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*

$$A(v) \leq B(v) \quad \textit{or} \quad A(v) = B(v) \tag{3.11}$$

claimed for all functions u of a Sobolev space and involving continuous *expressions A, B, then it suffices to prove* (3.11) *for the dense subspace of smooth functions.*

That means for the variational formulation, *e. g.*, (3.9), we can replace the space of test function by a Sobolev space:

$$\int_\Omega \langle \boldsymbol{C}\,\mathbf{grad}\, u, \mathbf{grad}\, v\rangle + c\,uv\,\mathrm{d}\mathbf{x} = \int_\Omega fv\,\mathrm{d}\mathbf{x} + \int_{\partial\Omega} hv\,\mathrm{d}S \quad \forall v \in H^m(\Omega).$$

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

$$\int_\Omega \langle \boldsymbol{C}\,\mathbf{grad}\, u, \mathbf{grad}\, v\rangle + c\,uv\,\mathrm{d}\mathbf{x} = \int_\Omega fv\,\mathrm{d}\mathbf{x} \quad \forall v \in H^1_0(\Omega).$$

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

$$\int_\Omega \langle \boldsymbol{C}\,\mathbf{grad}\, u, \mathbf{grad}\, v\rangle + c\,uv\,\mathrm{d}\mathbf{x} = \int_\Omega fv\,\mathrm{d}\mathbf{x} \quad \forall v \in H^1(\Omega)\,. \tag{3.12}$$

Evidently, its associated bilinear form

$$\mathsf{a}(u,v) = \int_\Omega \langle \boldsymbol{C}\,\mathbf{grad}\, u, \mathbf{grad}\, v\rangle + c\,uv\,\mathrm{d}\mathbf{x}$$

is symmetric positive definite. Let us denote by

$$\ell(v) = \langle \ell, v\rangle_{V'\times V} = \int_\Omega fv\,\mathrm{d}\mathbf{x}$$

the linear form on the right hand side of (3.12).

Lemma 3.16. *The solution u of (3.12) with $V = H^1(\Omega)$ can be characterised by*

$$u = \arg\min_{v\in V} J(v) \quad \textit{with} \quad J(v) = \tfrac{1}{2}\mathsf{a}(v,v) - \langle \ell, v\rangle_{V'\times V}\,.$$

Proof. If u denotes the solution of (3.12), a simple calculation shows

$$\begin{aligned} J(v) - J(u) &= \tfrac{1}{2}\mathsf{a}(v,v) - \underbrace{\langle f, v\rangle_{V'\times V}}_{=\mathsf{a}(u,v)} - \tfrac{1}{2}\mathsf{a}(u,u) + \underbrace{\langle f, u\rangle_{V'\times V}}_{=\mathsf{a}(u,u) \text{ as } u \text{ solves (3.12)}} \\ &= \tfrac{1}{2}\mathsf{a}(v-u, v-u) =: \tfrac{1}{2}\,\|v-u\|_a^2 \quad \forall v \in V\,. \end{aligned}$$

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 polynomial in τ) and, since u is a global minimizer

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

which is equivalent to (3.12), since any v can be chosen. □

This means that a solution of (3.12) will be a global minimizer of the **energy functional** $J(v)$. This hints at the general fact that

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

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

Example. *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* $J : H^1_{\Gamma_D,g}(\Omega) \mapsto \mathbb{R}$

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

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

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

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

$$\int_\Omega \langle \boldsymbol{C}\,\mathbf{grad}\, u, \mathbf{grad}\, v\rangle + c\,uv\,\mathrm{d}\mathbf{x} + \int_{\Gamma_R}\beta\,uv = \int_\Omega fv\,\mathrm{d}\mathbf{x} + \int_{\Gamma_N\cup\Gamma_R} hv\,\mathrm{d}S \quad \forall v \in H^1_{\Gamma_D,0}(\Omega)\,.$$

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 (**??**) emerges as the link between the PDE and the minimization problem, see Fig. 3.5.

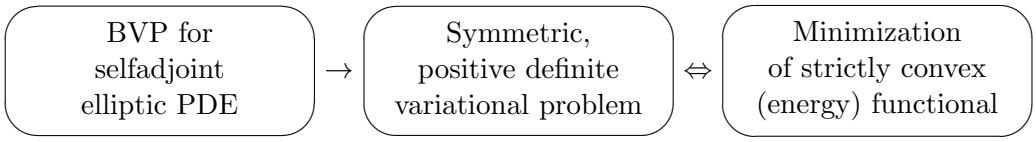


Figure 3.5: Relationship between minimization problems, variational problems, and elliptic boundary value problems

3.9 Theory of variational formulations

3.9.1 The Riesz representation theorem

Theorem 3.17 (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*

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

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.

$$\sqrt{\gamma}\|u\|_H \leq \sqrt{(u,u)_a} = \sup_{v\in H} \frac{|\ell(v)|}{\sqrt{(v,v)_a}} = \sup_{v\in H} \frac{|\ell(v)|}{\|v\|_H} \frac{\|v\|_H}{\sqrt{(v,v)_a}} \leq \|\ell\|_{H'} \frac{1}{\sqrt{\gamma}}.$$

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

$$\|v\|_{V''} = \sup_{g\in V'\setminus\{0\}} \frac{|v(g)|}{\|g\|_{V'}} = \sup_{g\in V'\setminus\{0\}} \frac{|\langle v, g\rangle_{V''\times V'}|}{\|g\|_{V'}}. \tag{3.13}$$

Obviously $V \subset V''$ and there exists a continuous embedding $J : V \to V''$ defined by

$$J(v)(g) =: v''(g) = \langle v'', g\rangle_{V''\times V'} = \langle g, v\rangle_{V'\times V} = g(v), \tag{*}$$

which preserves norms, *i. e.*, $\|v''\|_{V''} = \|v\|_V$, and so V and V'' are isomorphic. If (*) is an isomorphism, *i. e.*, J is bijective, then V is then V is called **reflexive** and we write $V \simeq V''$.

Corollary. *All Hilbert spaces are reflexive.*

Proof. As 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. □

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

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

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

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

Theorem 3.18. *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*

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

with $\gamma_s > 0$ independent of f.

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

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

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

Proof.

❶ (ii) ⇒ (i): Injectivity (Uniqueness)
Let $u_1, u_2 \in U$ be two solutions of (LVP) for the same $f \in V'$. Then $\mathsf{b}(u_1 - u_2, v) = 0$ for all $v \in V$, and from (IS1) we immediately infer $u_1 = u_2$.

❷ (ii) ⇒ (i): Closedness of the range
We define the operator $B : U \to V', \mathsf{b}(w,v) = \langle Bw, v\rangle_{V'\times V}\, \forall v \in V$ that is associated to the bilinear form b. To prove existence of solutions of (LVP) we define the following subspace of V', which correspond to B:

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

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

$$\forall k \in \mathbb{N} : \quad \exists w_k \in U : \quad b(w_k, v) = \langle g_k, v\rangle_{V'\times V} \quad \forall v \in V . \tag{3.14}$$

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

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

Hence, $\{w_k\}_{k=1}^{\infty}$ is a Cauchy-sequence, too, and will converge to some $w \in U$. The continuity of b and of the duality pairing makes it possible to pass to the limit $k \to \infty$ on both sides of (3.14), which is

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

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'.

❸ (ii) ⇒ (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 [11, Satz 4.1], confirms the existence of $v_0 \in V'' \simeq V$ (V reflexive!) such that

$$\forall g \in V_b' \quad 0 = v_0(g) = g(v_0) = \langle g, v_0\rangle_{V'\times V} \; .$$

By definition of V_b' this means $\mathsf{b}(w, v_0) = 0$ for all $w \in U$ and contradicts (IS2).

❹ (ii) ⇒ (i): Stability estimate

$$\|u\|_U \overset{\text{(IS1)}}{\leq} \frac{1}{\gamma_s} \sup_{v\in V\setminus\{0\}} \frac{|\mathsf{b}(u,v)|}{\|v\|_V} = \frac{1}{\gamma_s} \sup_{v\in V\setminus\{0\}} \frac{|\,\langle f, v\rangle_{V'\times V}\,|}{\|v\|_V} = \|f\|_{V'} .$$

❺ (i) ⇒ (ii):
Fix some $w \in V$ and denote by $g_w \in V'$ the continuous functional $v \mapsto \mathsf{b}(u,v)$, *i. e.*, w is the unique solution of

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

from which we conclude (IS1) by (ISS)

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

Now, fix some $v \in V \setminus \{0\}$. Then,

$$0 < \|v\|_{V''} = \sup_{g\in V'\setminus\{0\}}\frac{\big|\langle v,g\rangle_{V''\times V'}\big|}{\|g\|_{V'}} \quad \text{and so} \quad 0 < \sup_{g\in V'\setminus\{0\}}\big|\langle v,g\rangle_{V''\times V'}\big|$$

By the reflexivity of V we have

$$0 < \sup_{g\in V'\setminus\{0\}}\big|\langle g,v\rangle_{V'\times V}\big| = \sup_{g\in V'\setminus\{0\}}|\mathsf{b}(w,v)|,$$

where w is the unique solution of $\mathsf{b}(w,v) = \langle g,v\rangle_{V'\times V}$, $\forall v \in V$, and (IS2) follows.

□

Let us introduce sesquilinear forms $\mathsf{b} : V \times V \to \mathbb{C}$ with the properties

$$\mathsf{b}(\lambda u + \eta w, v) = \lambda\,\mathsf{b}(u,v) + \eta\,\mathsf{b}(w,v),$$
$$\mathsf{b}(u, \lambda v + \eta w) = \overline{\lambda}\,\mathsf{b}(u,v) + \overline{\eta}\,\mathsf{b}(u,w).$$

For example, if $u, v : \Omega \to \mathbb{C}$, then $\mathsf{b}(u,v) = \int_\Omega u\overline{v}\,\mathrm{d}\mathbf{x}$ is a sesquilinearform.

Definition (Ellipticity of sesquilinear forms)**.** *A sesquilinear form $\mathsf{b}(\cdot,\cdot)$ is called V-elliptic, if there exists a $\gamma > 0$ and a $\sigma \in \mathbb{C}$ with $|\sigma| = 1$, such that for all $v \in V$*

$$\mathrm{Re}(\sigma\mathsf{b}(v,v) \ge \gamma\|v\|_V^2.$$

Example. *Let $\mathsf{b}(u,v) = \int_\Omega \langle \mathbf{grad}\,u, \mathbf{grad}\,v\rangle + \mathrm{i}uv\,\mathrm{d}\mathbf{x}$. Then, $\mathsf{b}(v,v) = \|\,\mathbf{grad}\,v\|^2_{L^2(\Omega)} + \mathrm{i}\|v\|^2_{L^2(\Omega)}$. And so we obtain $H^1(\Omega)$-ellipticity,*

$$\mathrm{Re}\left(\mathrm{e}^{-\mathrm{i}\frac{\pi}{4}}\mathsf{b}(v,v)\right) \ge \tfrac{\sqrt{2}}{2}\|v\|^2_{H^1(\Omega)}.$$

Lemma 3.19 (Lax-Milgram)**.** *Let V be a reflexive Banach space. Let the bilinear form $\mathsf{b} : V \times V \to \mathbb{R}$ or sesquilinear form $\mathsf{b} : V \times V \to \mathbb{C}$ be V-elliptic. Then, the variational problem* (LVP) *has for any $f \in V'$ a unique solution $u \in V$ with*

$$\|u\|_V \le \frac{1}{\gamma}\|f\|_{V'}.$$

Proof. The V-ellipticity of the sesquilinear form b implies for any $v \in V$

$$\gamma \|v\|_V^2 \leq \mathrm{Re}(\sigma \mathsf{b}(v,v)) \leq |\sigma \mathsf{b}(v,v)| = |\mathsf{b}(v,v)|, \tag{3.15}$$

which holds directly for the bilinear form b. Then, (IS1) hols as

$$\sup_{v \in V \setminus \{0\}} \frac{|\mathsf{b}(w,v)|}{\|v\|_V} \geq \frac{|\mathsf{b}(w,w)|}{\|w\|_V} \overset{(3.15)}{\geq} \gamma \|w\|_V.$$

Furthermore, for any $v \in V$

$$\sup_{w \in V \setminus \{0\}} |\mathsf{b}(w,v)| \geq |\mathsf{b}(v,v)| \geq \gamma \|v\|_V^2 > 0.$$

and (IS2) holds. Applying Thm. 3.18 the proof is complete. □

3.10 Wellposedness for variational problem of second order elliptic PDEs

In Sec. 3.8 we have derived a variational problem for the second order BVP (3.1).

Strictly speaking, the associated variational problem does not match the definition (LVP) of a linear variational problem, because the unknown is sought in an affine space $H^1_{\Gamma_D,g}(\Omega)$ rather than a vector space. Yet, it 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 Γ_D, and plugging $u := u_g + u_0$ into the variational formulation, where, now, the **offset** $u_0 \in H^1_{\Gamma_D,0}(\Omega)$ 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 Γ_D in $H^1_{\Gamma_D,0}(\Omega)$?

3.10.1 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

$$(\mathsf{R}_m u)(\mathbf{x}) = u(\mathbf{x}) \quad \forall \mathbf{x} \in \partial\Omega,\ \forall u \in C^\infty(\overline{\Omega})\ .$$

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 $H^m(\Omega) \mapsto L^2(\partial\Omega)$).

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

$$\exists \gamma_t > 0: \quad \left\| u_{|\partial\Omega} \right\|_{L^2(\partial\Omega)} \leq \gamma_t \left\| u \right\|_{L^2(\Omega)} \quad \forall u \in C^\infty(\overline{\Omega}) \tag{3.16}$$

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

$$v(\mathbf{x}) := \begin{cases} 0 & \text{if } \varepsilon \le x_1 \le 1, 0 \le x_2 \le 1 \ , \\ 1 - \frac{x_1}{\varepsilon} & \text{if } 0 \le x_1 \le \varepsilon, 0 \le x_2 \le 1 \ . \end{cases}$$

Then we can compute

$$1 = \int_0^1 |v(0, x_2)|^2 \, \mathrm{d}x_2 \le \left\| v_{|\partial\Omega} \right\|^2_{L^2(\partial\Omega)} \ ,$$

and

$$\|v\|^2_{L^2(\Omega)} = \int_0^1 \left(1 - \tfrac{x_1}{\varepsilon}\right) \mathrm{d}x_1 \overset{y=\xi_1/h}{=} \varepsilon \int_0^1 (1-y)^2 \, dy = \tfrac{\varepsilon}{3} \ .$$

If (3.16) *were true, there would exists a constant* $\gamma_t > 0$ *such that* $1 \le \frac{1}{3}\gamma_t \varepsilon$*. For* $\varepsilon \to 0$ *we obtain a contradiction.*

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

Theorem 3.20. *The trace operator* R_1 *is continous from* $H^1(\Omega)$ *into* $L^2(\partial\Omega)$*, that is,*

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

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

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

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

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

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

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

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

Obviously, the spaces $H^1_\Gamma(\Omega)$ are closed subspaces of $H^1(\Omega)$. Another important density result holds true, see [9, Thm. 3.7].

Theorem 3.21. *The functions in* $C^\infty(\overline{\Omega})$ *whose support does not intersect* Γ *form a dense subspace of* $H^m_\Gamma(\Omega)$*,* $m \in \mathbb{N}$*. In particular,* $C^\infty_0(\Omega)$ *is a dense subspace of* $H^m_0(\Omega)$*.*

By Thm. 3.20 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 Γ temporarily stand for a connected component of the boundary of Ω. We start by introducing a norm

$$\|v\|_{H^{1/2}(\Gamma)} = \inf\{ \|w\|_{H^1(\Omega)} : \ w \in C^\infty(\overline{\Omega}), \, w_{|\Gamma} = v \} \tag{3.17}$$

on the space of restrictions of smooth functions to Γ. It is highly desirable that this norm is *intrinsic* to Γ, that is, switching to another domain $\widetilde{\Omega}$, for which Γ is also a connected component of the boundary, and using (3.17) produces an *equivalent* norm.

Definition. *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 [9, § 3].

Theorem 3.22. *The space* $H^{1/2}(\Gamma)$ *is a Hilbert space and can be equipped with the (equivalent)* ***Sobolev-Slobodeckij-norm***

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

Theorem 3.23. *The trace operator* $\mathsf{R}_1 : H^1(\Omega) \mapsto H^{1/2}(\Gamma)$ *is continuous and surjective and has a bounded right inverse* $\mathsf{F}_1 : H^{1/2}(\Gamma) \mapsto H^1(\Omega)$, *ie.,* $\mathsf{R}_1 \circ \mathsf{F}_1 = Id$.

3.10.2 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 $H^1(\Omega) \to \mathbb{R}$

$$v \mapsto \int_\Omega f v \mathrm{d}\mathbf{x}$$

is bounded in $(H^1(\Omega))'$. The same holds for subspaces $H^1_{\Gamma_D,0}(\Omega)$ or $H^1_0(\Omega)$.

It is easy to verify that

$$\|u\|_{H^{-1}(\Omega)} := \sup_{v \in H^1_0(\Omega)\setminus\{0\}} \frac{\int_\Omega uv \,\mathrm{d}\mathbf{x}}{\|v\|_{H^1(\Omega)}} \tag{3.18}$$

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

Definition. *The completion of* $L^2(\Omega)$ *with respect to the norm given by* (3.18) *is called* $H^{-1}(\Omega)$.

Lemma 3.24. *The space* $H^{-1}(\Omega)$ *is a Hilbert space, which is isometrically isomorphic to* $(H^1_0(\Omega))'$.

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

$$H^1_0(\Omega) \subset L^2(\Omega) \subset (H^1_0(\Omega))' = H^{-1}(\Omega) \,.$$

Sometimes, such an arrangement is called a Gelfand triple.

Notation. *Often the integral* $\int_\Omega \ldots \mathrm{d}\mathbf{x}$ *is used to denote the duality pairing of* $H^{-1}(\Omega)$ *and* $H^1_0(\Omega)$.

The same considerations that above targeted $H^1_0(\Omega)$ can be applied to $H^{1/2}(\Gamma)$ on a surface without boundary. This will yield the Hilbert space $H^{-1/2}(\Gamma)$, which contains $L^2(\Gamma)$ and is (isometrically isomorphic to the) dual to $H^{1/2}(\Gamma)$. As before, the integral $\int_\Gamma \ldots \mathrm{d}S$ is often used to indicate the corresponding duality pairing.

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

Lemma 3.25. *The normal components trace* $\mathsf{R}_\mathbf{n}$ *for* $\mathbf{u} \in (C^\infty(\overline{\Omega}))^d$*, defined by* $\mathsf{R}_\mathbf{n}\,\mathbf{u}(\mathbf{x}) := \langle \mathbf{u}(\mathbf{x}), \mathbf{n}(\mathbf{x})\rangle$ *for all* $\mathbf{x} \in \partial\Omega$*, can be extended to a continuous and surjective operator* $\mathsf{R}_\mathbf{n} : H(\mathrm{div};\Omega) \mapsto H^{-1/2}(\partial\Omega)$*.*

Proof. Pick some $\mathbf{u} \in (C^\infty(\overline{\Omega}))^d$. By (FGF), the Cauchy-Schwarz inequality in $L^2(\Omega)$ and Thm. 3.23 we find

$$\begin{aligned}\|\langle \mathbf{u},\mathbf{n}\rangle\|_{H^{-1/2}(\Gamma)} &= \sup_{v\in H^{1/2}(\Gamma)\setminus\{0\}} \frac{\int_\Gamma \langle \mathbf{u},\mathbf{n}\rangle\, v\,\mathrm{d}S}{\|v\|_{H^{1/2}(\Gamma)}} = \sup_{v\in H^{1/2}(\Gamma)\setminus\{0\}} \frac{\int_\Gamma \langle \mathbf{u},\mathbf{n}\rangle\, R_1(F_1 v)\,\mathrm{d}S}{\|v\|_{H^{1/2}(\Gamma)}}\\ &= \sup_{v\in H^{1/2}(\Gamma)\setminus\{0\}} \frac{1}{\|v\|_{H^{1/2}(\Gamma)}} \int_\Omega \langle \mathbf{u}, \mathbf{grad}\,\mathsf{F}_1\, v\rangle + \mathrm{div}\,\mathbf{u}\;\mathsf{F}_1\, v\,\mathrm{d}\mathbf{x}\\ &\le \frac{\|\mathsf{F}_1\, v\|_{H^1(\Omega)}}{\|v\|_{H^{1/2}(\Gamma)}}\;\|\mathbf{u}\|_{H(\mathrm{div};\Omega)} \le \|\mathsf{F}_1\|_{H^{1/2}(\Gamma)\mapsto H^1(\Omega)}\;\|\mathbf{u}\|_{H(\mathrm{div};\Omega)}\;.\end{aligned}$$

As $(C^\infty(\overline{\Omega}))^d$ is a dense subspace of $H(\mathrm{div};\Omega)$ this shows the continuity of $\mathsf{R}_\mathbf{n}$ asserted in the Lemma.

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

$$\int_\Omega \langle \mathbf{grad}\, w, \mathbf{grad}\, v\rangle + wv\,\mathrm{d}\mathbf{x} = \int_\Gamma h\;\mathsf{R}_1\, v\,\mathrm{d}S \quad \forall v \in H^1(\Omega)\;.$$

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 Sect. 3.9 the variational problem (**??**) has a unique solution in $H^1(\Omega)$.

Testing with $v \in C_0^\infty(\Omega)$ and recalling the definition of weak derivatives, see Def. 3.7, we find that

$$-\,\mathrm{div}\,\mathbf{grad}\, w + w = 0 \quad \text{in } L^2(\Omega)\;.$$

Note that div has to be understood as differential operator in the sense of distributions (see Sec. 3.4). This shows $\mathrm{div}(\mathbf{grad}\, u) \in L^2(\Omega)$, which allows to apply (FGF):

$$\int_\Omega \underbrace{\left(-\,\mathrm{div}\,\mathbf{grad}\, w + w\right)}_{=0}\, v\,\mathrm{d}\mathbf{x} + \int_\Gamma \langle \mathbf{grad}\, w, \mathbf{n}\rangle\;\mathsf{R}_1\, v\,\mathrm{d}S = \int_\Gamma h\;\mathsf{R}_1\, v\,\mathrm{d}S$$

for all $v \in C^\infty(\Omega)$. By a density argument, this amounts to $h = \langle \mathbf{grad}\, u, \mathbf{n}\rangle$ in $H^{-1/2}(\Gamma)$, and $\mathsf{R}_\mathbf{n}$ is surjective. □

As a consequence, $\mathsf{R}_\mathbf{n}\,\mathbf{grad}\, u$ is well-defined (in $H^{-1/2}(\Gamma)$) if $\mathbf{grad}\, u \in H(\mathrm{div};\Omega)$.

Moreover, the trace theorems for $H^1(\Omega)$ and $H(\mathrm{div};\Omega)$ and the density of the space of smooth functions in the respective Sobolev spaces enable us to extend (FGF), and this integration by parts formula is seen to hold for all $\mathbf{f} \in H(\mathrm{div};\Omega)$ and $u \in H^1(\Omega)$:

$$\int_\Omega \langle \mathbf{f}, \mathbf{grad}\, u\rangle + \mathrm{div}\,\mathbf{f}\, u\,\mathrm{d}\mathbf{x} = \int_\Gamma \langle \mathbf{f},\mathbf{n}\rangle\; u\mathrm{d}S \quad \forall\, \mathbf{f} \in H(\mathrm{div};\Omega), u \in H^1(\Omega)\;. \qquad \text{(FGF)}$$

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

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

$$\int_\Omega \langle \boldsymbol{C}\,\mathbf{grad}\,u, \mathbf{grad}\,v\rangle + c\,uv\,\mathrm{d}\mathbf{x} + \int_{\Gamma_R} \beta\,(\mathsf{R}_1\,u)(\mathsf{R}_1\,v)\mathrm{d}S \\ = \int_\Omega fv\,\mathrm{d}\mathbf{x} + \int_{\Gamma_N\cup\Gamma_R} h(\mathsf{R}_1\,v)\,\mathrm{d}S \quad \forall v \in H^1_{\Gamma_D,0}(\Omega)\ .$$

- For $g \in H^{1/2}(\Gamma_D)$ an extension $u_g \in H^1(\Omega)$ exists, and with $H^1_{\Gamma_D,g}(\Omega) = H^1_{\Gamma_D,0}(\Omega) + u_g$ we search for $u_0 = u - u_g \in H^1_{\Gamma_D,0}(\Omega)$, the Sobolev space with homogeneous Dirichlet traces.
- 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 $L^2(\Omega) \subset (H^1_{\Gamma_D,0}(\Omega))'$.

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

$$\gamma = \min(c_0, \inf_{\mathbf{x}\in\Omega} \min \sigma(\boldsymbol{C}(\mathbf{x})),$$

and well-posedness follow from Lax-Milgram's lemma with

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

▷ 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.

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

3.10.4 The inequalities of Poincaré and Friedrich

Lemma 3.26 (The Friedrich inequality)**.** *Let Ω 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*

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

▷ Well-posedness in case of Dirichlet boundary conditions (at least at part of the boundary) and $c \geq 0$ (γ_1 is the infimum of the minimal eigenvalue of $\boldsymbol{C}$ and $\alpha \in (0,1)$):

$$\mathsf{b}(v,v) \geq \gamma_1 |v|^2_{H^1(\Omega)} = \gamma_1(1-\alpha)|v|^2_{H^1(\Omega)} + \gamma_1\alpha|v|^2_{H^1(\Omega)} \\ \geq \gamma_1 \min(1-\alpha, \tfrac{\alpha}{C_F\,\mathrm{diam}(\Omega)^2})\|v\|^2_{H^1(\Omega)} \geq \gamma\|v\|^2_{H^1(\Omega)}.$$

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

$$\mathrm{div}\,\mathbf{j} = -\,\mathrm{div}\,\sigma\,\mathbf{grad}\,u = f \quad \text{in } \Omega, \\ \mathbf{j}\cdot\mathbf{n} = \sigma\partial_n u = h \quad \text{on } \partial\Omega,$$

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**

$$\int_\Omega f \,\mathrm{d}\mathbf{x} = \int_{\partial\Omega} h \,\mathrm{d}S,$$

the data of the problem has to fulfill. One can see this also by applying Gauss' theorem applied to $\mathbf{j}$.

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

$$H^1_*(\Omega) := \{v \in H^1(\Omega) : \int_\Omega v \,\mathrm{d}\mathbf{x} = 0\}$$

and we can apply

Lemma 3.27 (The Poincaré inequality)**.** *Let Ω be a open bounded domain. Then, for all $u \in H^1(\Omega)$ there exists a constant $C_P(\Omega)$ such that*

$$\|u - u_\Omega\|^2_{L^2(\Omega)} \leq C_P(\Omega) \operatorname{diam}(\Omega)^2 \|\operatorname{\mathbf{grad}} u\|^2_{L^2(\Omega)},$$

where $u_\Omega = \frac{1}{|\Omega|}\int_\Omega u(\mathbf{x})\mathrm{d}\mathbf{x}$ is the mean of u in Ω.

Hence, similarly we obtain for all $v \in H^1(\Omega)$

$$\mathsf{b}(v,v) \geq \gamma_1 |v|^2_{H^1(\Omega)} \geq \gamma_1 \min(1-\alpha, \tfrac{\alpha}{C_P(\Omega)\operatorname{diam}(\Omega)^2}) \|v\|^2_{H^1(\Omega)} \geq \gamma \|v\|^2_{H^1(\Omega)}.$$

Remark. *One is tempted to ensure uniqueness of solutions of (??) by demanding $u(\mathbf{x}_0) = 0$ for some $\mathbf{x}_0 \in \Omega$. However, this approach is flawed (fehlerhaft), because the mapping $u \mapsto u(\mathbf{x}_0)$ 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 3.28. *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*

$$\|u\|^2_{L^2(\Omega)} \leq C \left(\|\operatorname{\mathbf{grad}} u\|^2_{L^2(\Omega)} + |f(u)| \right).$$

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 $n \in \mathbb{N}$

$$\left\{ \left| \|\mathbf{grad}\, u_n\|^2_{L^2(\Omega)} + |f(u_n)| \right\} \leq \frac{1}{n}\|u_n\|^2_{L^2(\Omega)} \leq \frac{1}{n}\|u_n\|^2_{H^1(\Omega)}. \qquad (*)$$

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 $H^1(\Omega)$ [12, 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 $\|\mathbf{grad}\, v\|^2_{L^2(\Omega)} + |f(v)| : H^1(\Omega) \to \mathbb{R}$ we can take the limit on both sides of (*), and it holds

$$\|\mathbf{grad}\, u\|^2_{L^2(\Omega)} + |f(u)| = 0.$$

This implies that $\mathbf{grad}\, u \equiv 0$ 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, [13, 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 $\|u_{n_k}\|_{H^1(\Omega)} = 1$, $\|\mathbf{grad}\, u_{n_k}\|_{L^2(\Omega)} \to 0$ and $\|u_{n_k}\|_{L^2(\Omega)} \to 0$, which is not possible. □

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

$$\begin{aligned} \mathsf{b}(v, v) &\geq \gamma_1 |v|^2_{H^1(\Omega)} + \underbrace{\beta_0}_{=\inf_{\mathbf{x}\in\Gamma_R}(\beta)} \|\,\mathsf{R}_1\, v\|^2_{L^2(\Gamma_R)} \\ &\geq \gamma_1 \min(1-\alpha, \tfrac{\alpha\beta_0}{C}) \|v\|^2_{H^1(\Omega)} \geq \gamma \|v\|^2_{H^1(\Omega)}, \end{aligned}$$

since the quadratic functional $\|\,\mathsf{R}_1\, v\|^2_{L^2(\Gamma_R)}$ fixes constants:

$$\|\,\mathsf{R}_1\, 1\|^2_{L^2(\Gamma_R)} = \int_{\Gamma_R} 1 \,\mathrm{d}S = |\Gamma_R| > 0.$$

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 $\mathsf{b} \in L(V \times V, \mathbb{R})$ 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:

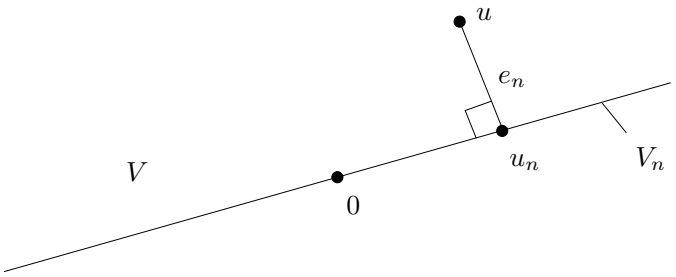


Figure 3.6: b-Orthogonality of the error $e_n = u - u_n$ with respect to V_n, if b is an inner product.

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

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

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

$$\mathsf{b}(u_n, v_n) = \ell(v_n) \quad \forall v_n \in V_n \, . \tag{DVP}$$

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

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

$$\mathsf{b}(u - u_n, v_n) = 0 \quad \forall v_n \in V_n \, . \tag{3.19}$$

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

Theorem 3.29. *Let V be a Banach space and $\mathsf{b} \in L(V \times V, \mathbb{R})$ satisfy the inf-sup conditions* (IS1) *and* (IS2) *from Thm. 3.18. Further, assume that*

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

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*

$$\|u_n\|_V \leq \frac{1}{\gamma_n} \|f\|_{V_n'} = \frac{1}{\gamma_n} \sup_{v_n \in V_n} \frac{|f(v_n)|}{\|v_n\|_V} \, , \tag{3.20}$$

*and the **quasi-optimality** estimate*

$$\|u - u_n\|_V \leq \left(1 + \frac{\|\mathsf{b}\|_{V \times V \mapsto \mathbb{R}}}{\gamma_n}\right) \inf_{w_n \in W_n} \|u - w_n\|_V \, , \tag{3.21}$$

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

Proof. Following ❶ in the proof of Thm. 3.18 it is clear that (DIS) implies the uniqueness of u_n, and similarly to ❹ the stability estimate (3.20) follows. Since $N = \dim V_n = \dim W_n$, in the finite dimensional setting this implies existence of u_n (*cf.* [12, Lemma A.9]): for the operator B_n defined by $\langle B_n w_n, v_n\rangle_{V'\times V} = \mathsf{b}(w_n, v_n)$ for all $w_n \in W_n$ and all $v_n \in V_n$ it holds

$$N = \dim(\mathrm{Ker}(B_n)) + \dim(\mathrm{Range}(B_n)).$$

It remains to show (3.21). For any $w_n \in W_n$ we can estimate using the triangle inequality and (3.20):

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

The last term vanishes due to Galerkin orthogonality (3.19) and with the continuity of b we obtain (3.20). □

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

Remark. *It goes without saying that for a V-elliptic bilinear form* b*the assumptions of Thm. 3.29 are trivially satisfied. Moreover, we can choose γ_n equal to the ellipticity constant γ_e in this case.*

Theorem 3.29 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. 3.29, the mere fact that the norm of $V = H$ arises from an inner product, permits us to get a stronger a-priori error estimate [14].

Theorem 3.30. *If V is a Hilbert space we obtain the sharper a-priori error estimate*

$$\|u-u_n\|_V \le \frac{\|\mathsf{b}\|_{V\times V\mapsto\mathbb{R}}}{\gamma_n}\inf_{v_n\in V_n}\|u-v_n\|_V \ , \tag{3.22}$$

if the assumptions of Thm. 3.29 are satisfied.

Theorem 3.31 (Céas's Lemma)**.** *If the bilinear form* b *is V-elliptic the estimate*

$$\|u-u_n\|_V \le \frac{\|\mathsf{b}\|_{V\times V\mapsto\mathbb{R}}}{\gamma_e}\inf_{v_n\in V_n}\|u-v_n\|_V \ , \tag{3.23}$$

holds.

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 3.32. *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 (3.19) and the Cauchy-Schwarz inequality

$$\begin{aligned} \|u - u_n\|_b^2 = \mathsf{b}(u - u_n, u - u_n) &= \mathsf{b}(u - u_n, u - v_n) + \underbrace{\mathsf{b}(u - u_n, v_n - u_n)}_{=0 \text{ by Galerkin orthog.}} \\ &\leq \|u - u_n\|_b \, \|u - v_n\|_b \ , \end{aligned}$$

for any $v_n \in V_n$. □

Remark. *Many of our efforts will target* ***asymptotic a-priori estimates*** *that involve sequences* $\{V_n\}_{n=1}^{\infty}$*,* $\{W_n\}_{n=1}^{\infty}$ *of test and trial spaces. Then it will be the principal objective to ensure that the constant* γ_n *is bounded away from zero uniformly in* n*. This will guarantee* ***asymptotic quasi-optimality*** *of the Galerkin solution: the estimate* (3.20) *will hold with a constant independent of* n*. Notice that the norm of* b *that also enters* (3.21) *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. 3.29. The "Galerkin orthogonality" (3.19) suggests that we examine the so-called **Galerkin projection** $\mathsf{P}_n : V \mapsto W_n$ defined by

$$\mathsf{b}(\mathsf{P}_n w, v_n) = \mathsf{b}(w, v_n) \quad \forall v_n \in V_n \ . \tag{3.24}$$

Proposition 3.33. *Under the assumption of Thm. 3.29, the equation* (3.24) *defines a continuous projection* $\mathsf{P}_n : V \mapsto V$ *with norm* $\|\mathsf{P}_n\|_{V \mapsto V} \leq \gamma_n^{-1} \, \|\mathsf{b}\|_{V \times V \mapsto \mathbb{R}}$*.*

Proof. As a consequence of Thm. 3.29, P_n is well defined. Its linearity is straightforward and the norm bound can be inferred from (3.20). To see this let $w \in V$ be fixed and $w_n \in W_n$ be solution of $\mathsf{b}(w_n, v_n) = \mathsf{b}(w, v_n)$ for all $v_n \in V_n$. Hence, using (3.20) and the Cauchy-Schwarz inequality we find

$$\|w_n\|_V = \|\mathsf{P}_n w\|_V \leq \frac{1}{\gamma_n} \sup_{v_n \in V_n \setminus \{0\}} \frac{|\mathsf{b}(w, v_n)|}{\|v_n\|_V} \leq \frac{1}{\gamma_n} \|\mathsf{b}\|_{V \times V \to \mathbb{R}} \frac{\|w\|_V \|v_n\|_V}{\|v_n\|_V},$$

and so

$$\|\mathsf{P}_n\|_{V \to V} = \sup_{w \in V \setminus \{0\}} \frac{\|\mathsf{P}_n w\|_V}{\|w\|_V} \leq \frac{1}{\gamma_n} \|\mathsf{b}\|_{V \times V \to \mathbb{R}}.$$

Also $\mathsf{P}_n^2 = \mathsf{P}_n$ immediate from the definition since for all $v_n \in V_n$ we have

$$\mathsf{b}(\underbrace{\mathsf{P}_n(\mathsf{P}_n w)}_{\in W_n}, v_n) \overset{(3.24)}{=} \mathsf{b}(\underbrace{P_n w_n}_{\in W_n}, v_n) \overset{(3.24)}{=} \mathsf{b}(w, v_n)$$

and the statement of the proposition is proved. □

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

$$u_n = \mathsf{P}_n u \,. \tag{3.25}$$

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

Proposition 3.34. *If $\mathsf{b} \in L(V \times V, \mathbb{R})$ is an inner product on V, $V_n = W_n$ and the assumptions of Thm. 3.29 are satisfied, then the Galerkin projection associated with b is an orthogonal projection with respect to the inner product b.*

Proof. According to the definition of the orthogonal projection and the previous proposition, we only have to check that $\mathrm{Ker}(\mathsf{P}_n) \perp \mathrm{Range}(\mathsf{P}_n)$. As P_n is a projector there is a decomposition of V in a direct sum [12, Lemma A.38]

$$V = \mathrm{Range}(\mathsf{P}_n) \oplus \mathrm{Range}(Id - \mathsf{P}_n),$$

i. e., $\mathrm{Range}(\mathsf{P}_n) \cap \mathrm{Range}(Id - \mathsf{P}_n) = \{0\}$. By Galerkin orthogonality (3.19) and (3.25) this direct sum is orthogonal:

$$\mathsf{b}(\underbrace{u - u_n}_{(Id-\mathsf{P}_n)u}, \underbrace{u_n}_{\mathsf{P}_n u \in V_n = W_n}) = 0 \quad \Leftrightarrow \quad \mathsf{b}(\mathsf{P}_n u, (Id - \mathsf{P}_n)u) = 0.$$

As $\mathrm{Ker}(\mathsf{P}_n) = \mathrm{Range}(Id - \mathsf{P}_n)$

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

and the proposition holds. □

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**

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

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

Lemma 3.35. *The following is equivalent:*

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

(ii) The linear system of equations

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

with

$$\mathbf{B} := \left(\mathsf{b}(q_n^k, p_n^j)\right)_{j,k=1}^{N} \in \mathbb{R}^{N,N} , \tag{3.26}$$

$$\boldsymbol{\varphi} := \left(\left\langle f, p_n^k \right\rangle_{V'\times V}\right)_{k=1}^{N} \in \mathbb{R}^{N} , \tag{3.27}$$

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

Then

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

Proof. Due to the basis property we can set

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

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

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

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

$$\sum_{j=1}^{N}\sum_{k=1}^{N} \mu_k \nu_j b(q_n^k, p_n^j) = \sum_{j=1}^{N} \nu_j \left\langle f, p_n^j \right\rangle_{V'\times V} . \tag{3.28}$$

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

$$\sum_{k=1}^{N} \mu_k \mathsf{b}(q_n^k, p_n^l) = \left\langle f, p_n^l \right\rangle_{V'\times V}, \quad l = 1, \dots, N . \tag{3.29}$$

As the special test vectors span all of $\mathbb{R}^N$ and thanks to the basis property, we conclude that (3.28) and (3.29) are equivalent. On the other hand, (3.29) corresponds to (LSE), as is clear by recalling the rules of matrix×vector multiplication. □

Note that in (3.26) j is the row index, whereas k is the column index. Consequently, the element in the j-th row and k-th column of the matrix $\mathbf{B}$ in (LSE) is given by

$\mathsf{b}(q_n^k, p_n^j)$. The columns correspond to the trial functions where the rows correspond to the test functions.

$$
\begin{array}{c}
\\
j \rightarrow
\end{array}
\begin{array}{c}
\begin{array}{ccccc} & & k & & \end{array} \\
\downarrow \\
\begin{pmatrix}
\cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & * & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot
\end{pmatrix}
\end{array}
$$

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

Corollary 3.36. *If and only if the bilinear form b and trial/test space W_n/V_n satisfy the assumptions of Thm. 3.29, then the matrix $\mathbf{B}$ of* (LSE) *will be regular.*

Thus, we have arrived at the final "algebraic problem" (LSE) through the two stage process outlined in Fig. 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. 3.36 teaches that some properties of $\mathbf{B}$ will only depend on V_n, too.

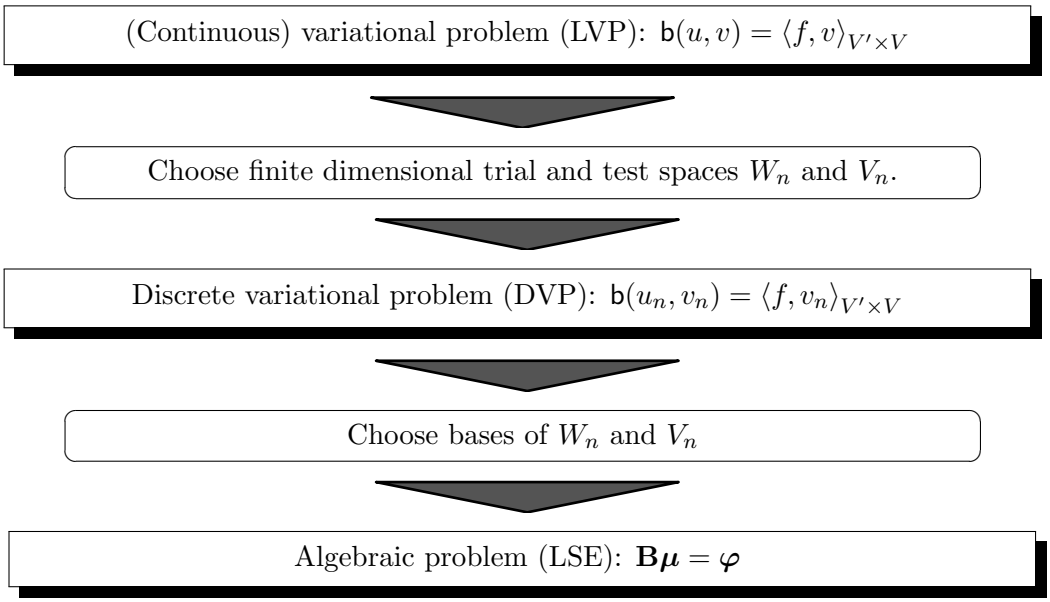


Figure 3.7: Overview of stages involved in the complete Galerkin discretization of an abstract variational problem.

References

[5] William McLean. *Strongly Elliptic Systems and Boundary Integral Equations*. Cambridge University Press, 2000.

[6] O. Forster. *Analysis 3. Integralrechung im $\mathbb{R}^n$ mit Anwendungen*. Vieweg–Verlag, Wiesbaden, 3rd edition, 1984.

[7] H. König. Ein einfacher Beweis des Integralsatzes von Gauss. *Jahresber. Dtsch. Math.-Ver.*, 66:119–138, 1964.

[8] F. Brezzi and M. Fortin. *Mixed and hybrid finite element methods*. Springer, 1991.

[9] J. Wloka. *Partial differential equations*. Cambridge University Press, Cambridge, UK, 1987.

[10] F. Hirzebruch and W. Scharlau. *Einführung in die Funktionalanalysis*, volume 296 of *BI Hochschultaschenbücher*. Bibliographisches Institut, Mannheim, 1971.

[11] H. W. Alt. *Lineare Funktionalanalysis*. Springer-Verlag, Heidelberg, 1985.

[12] Pavel Šolín. *Partial Differential Equations and the Finite Element Method*. Wiley-Interscience, Hoboken, USA, 2006.

[13] Dietrich Braess. *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics*. Cambridge University Press, 3th edition, 2007.

[14] J. Xu and L. Zikatanov. Some oberservations on Babuška and Brezzi theories. Report AM222, PennState Department of Mathematics, College Park, PA, September 2000. To appear in Numer. Math.

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. 3.29.

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 (**??**) and Sec. **??**). These variational problems are set in Sobolev spaces and feature elliptic bilinear forms according to Def. **??**. 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 **??**.

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

1. In light of Thm. 3.29 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. **??**.

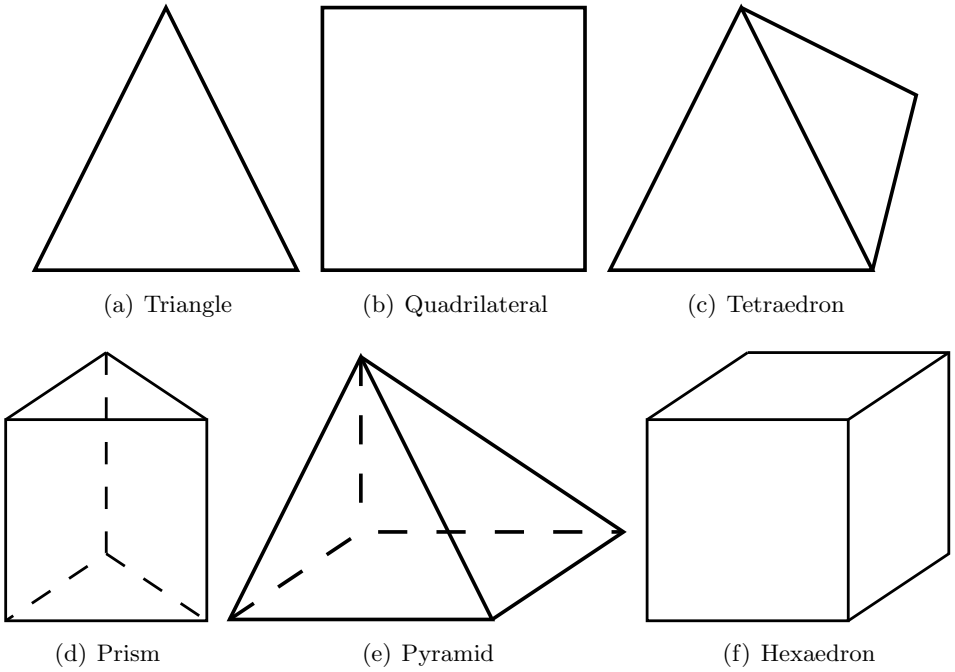
Definition 4.1. *A **mesh** $\mathcal{M}$ of Ω is a collection $\{K_i\}_{i=1}^M$, $M := \sharp\mathcal{M}$[1], of connected* open *subsets $K_i \subset \Omega$ such that*

- *the closure of each K_i is the C^∞-diffeomorphic image of a closed d-dimensional polytope (that is, the convex hull of $n \geq d+1$ points in $\mathbb{R}^d$),*
- $\bigcup_i \overline{K}_i = \overline{\Omega}$ *and* $K_i \cap K_j = \emptyset$ *if* $i \neq j$, $i, j \in \{1, \ldots, M\}$.

*The K_i are called **cells** of the mesh.*

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

[1] $\sharp$ denotes the cardinality of a finite set



(a) Triangle (b) Quadrilateral (c) Tetraedron

(d) Prism (e) Pyramid (f) Hexaedron

Figure 4.1: Cell types in 2D and 3D.

Following the terminology of Sect. **??**, 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*

$$\mathcal{F}(\mathcal{M}) := \left\{\text{interior}(\overline{K}_i \cap \overline{K}_j),\ 1 \le i < j \le \sharp\mathcal{M}\right\} \cup \{\textit{(geometric) faces } \subset \partial\Omega\} \ ,$$

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

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

*whereas the **set of (topological) nodes** is*

$$\mathcal{N}(\mathcal{M}) := \left\{\overline{E} \cap \overline{E'},\ E, E' \in \mathcal{E}(\mathcal{M}), E \neq E'\right\} \ .$$

*Similarly, we can define sets of edges and nodes for $d = 2$, and the set of nodes for $d = 1$. Often the term **face** is used for components of a mesh of dimension $d - 1$, that is, for faces in three dimensions, edges in two dimensions, and nodes in one dimension. We still write $\mathcal{F}(\mathcal{M})$ for the set of these (generalized) faces.*

Remark 4.4. *The "is contained in the closure of" **incidence relations** $\mathcal{N}(\mathcal{M}) \times \mathcal{E}(\mathcal{M}) \mapsto \{\text{true}, \text{false}\}$, $\mathcal{E}(\mathcal{M}) \times \mathcal{F}(\mathcal{M}) \mapsto \{\text{true}, \text{false}\}$, etc., describe the **topology** of a*

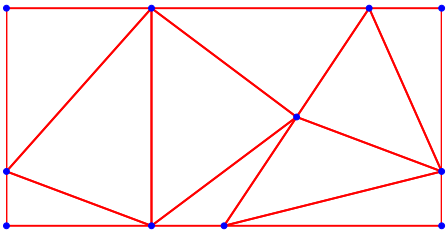


Figure 4.2: Two-dimensional mesh and the sets of edges (red) and nodes (blue)

mesh. The locations of nodes and shape of cells are features connected wity the ***geometry*** *of the mesh. We can define the nodes $\mathcal{N}(E)$ of an edge, the edges $\mathcal{E}(F)$ of a face and the faces $\mathcal{F}(K)$ of a cell.*

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

We recall that a mapping

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

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

Definition 4.5. *A mesh $\mathcal{M}$ of $\Omega \subset \mathbb{R}^d$ is called* ***affine equivalent****, if all its cells arise as affine images of a single d-dimensional (reference) polytope.*

A family of meshes $\{\mathcal{M}_n\}_{n\in\mathbb{N}}$ is affine equivalent, if this is true for each of its members and if the same reference polytope can be chosen for all $\mathcal{M}_n$, $n \in \mathbb{N}$.

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

$$\boldsymbol{\Phi} : \mathbb{R}^2 \mapsto \mathbb{R}^2, \quad \boldsymbol{\xi} \mapsto \mathbf{F}\boldsymbol{\xi} + \boldsymbol{\tau} + \boldsymbol{\tau}_Q \xi_1 \xi_2. \tag{4.1}$$

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

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

$$\begin{aligned} \boldsymbol{x}_K(\xi_1, \xi_2) = \boldsymbol{\Phi}_K \boldsymbol{\xi} = \quad & (1-\xi_2)\,\mathbf{x}_1(\xi_1) + \xi_1\,\mathbf{x}_2(\xi_2) + \xi_2\,\mathbf{x}_3(1-\xi_1) + (1-\xi_1)\,\mathbf{x}_4(1-\xi_2) \\ & - (1-\xi_1)(1-\xi_2)\,\boldsymbol{p}_0 - \xi_1(1-\xi_2)\,\boldsymbol{p}_1 - \xi_1\xi_2\,\boldsymbol{p}_2 - (1-\xi_1)\xi_2\,\boldsymbol{p}_3. \end{aligned}$$

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

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.*

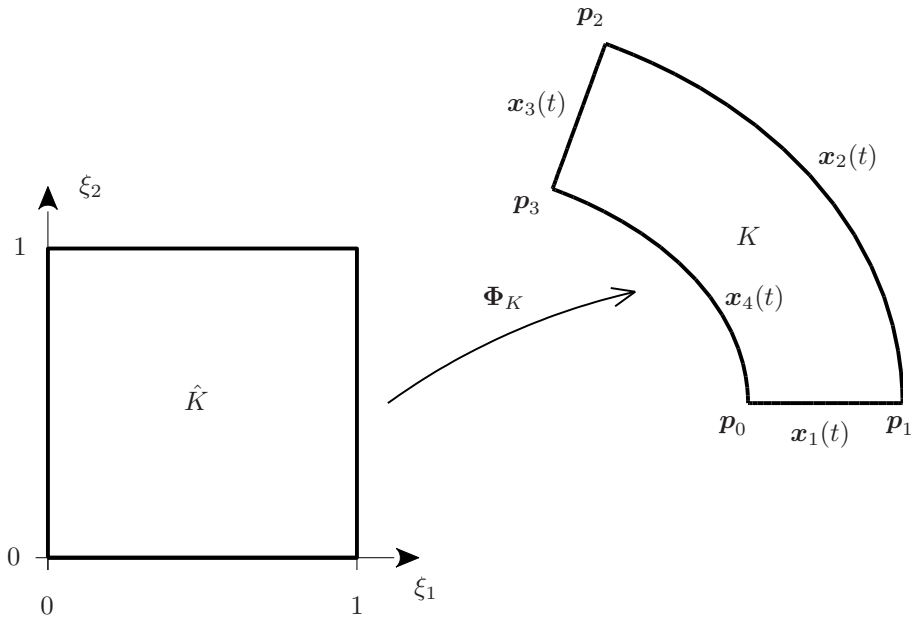


Figure 4.3: Mapping for curved quadrilateral.

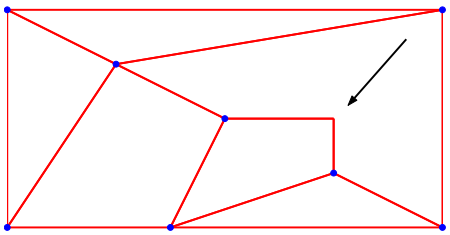


Figure 4.4: A mesh that is not a triangulation. The arrow points at the culprit.

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 Ω is called **conforming**, if $\overline{K}_i \cap \overline{K}_j$, $1 \leq i < j \leq M$, is a (geometric) face of both K_i and K_j.*

Unless clearly stated otherwise we will tacitly assume that all meshes that will be used for the construction of finite element spaces in the remainder of this chapter are conforming.

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 **hanging (dangling) node**.*

In the case of triangulations we can distinguish special classes:

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

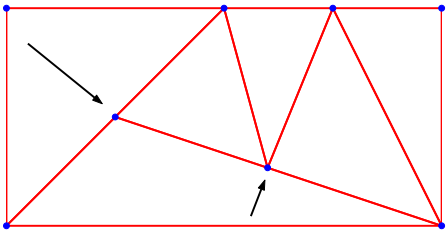


Figure 4.5: A triangulation that is not conforming and possesses two hanging nodes.

- **quadrilateral** ($d = 2$) and **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 quadrilaterial triangulation need not be affine equivalent, because, for instance, there is no affine map taking a square to general trapezoid.*

Exercise 4.1. *Let $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_4 \in \mathbb{R}^3$ stand for the coordinate vectors of the four vertices of a tetrahedron K. Determine the affine mapping* (AFF) *that takes the "unit tetrahedron"*

$$\widehat{K} := \text{convex}\left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$$

to K. When will the matrix $\mathbf{F}$ of this affine mapping be regular?

Exercise 4.2. *Let $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_4 \in \mathbb{R}^2$ denote the coordinate vectors belonging to the four vertices of a quadrilateral K in the plane. Determine an analytic description of a simple smooth mapping $\boldsymbol{\Phi} : \mathbb{R}^2 \mapsto \mathbb{R}^2$ from the "reference square"*

$$\widehat{K} := \text{convex}\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix} \right\}$$

to K. Compute the Jacobian $D\boldsymbol{\Phi}$ and its determinant.

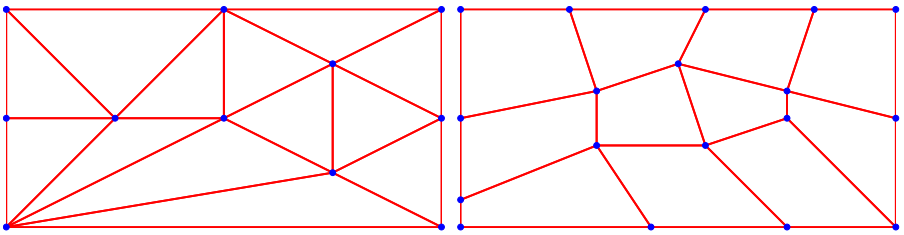


Figure 4.6: Examples of triangular and quadrilateral meshes in two dimensions

The term grid is often used as a synonym for triangulation, but we will reserve it for meshes with a locally **translation invariant** structure. These can be **tensor product grids**, that is meshes whose cells are quadrilaterals ($d = 2$) or hexahedra ($d = 3$) with parallel sides.

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

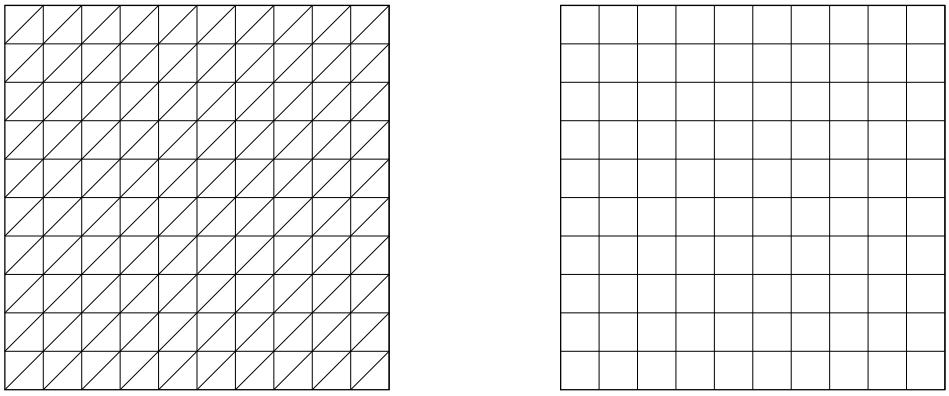


Figure 4.7: Example of triangular and quadrilateral grids in two dimensions

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

Various strategies can be employed for automatic grid generation:

- advancing front method that build cells starting from the boundary.
- Delaunay refinement techniques that can create a mesh starting from a mesh for $\partial\Omega$ or a "cloud" of points covering Ω.
- the quadtree ($d = 2$) or octree ($d = 3$) approach, which fills Ω with squares/cubes of different sizes supplemented by special measures for resolving the boundary.
- mapping techniques that split Ω into sub-domains of "simple" shape (curved triangles, parallelograms, bricks), endow those with parametric grids and glue these together.

Remark 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 topolgy and geometry of the underlying mesh. Then an approximate solution is computed and written to file in order to be read by post-processing tools like visualization software, see Fig. 4.8.*

Remark 4.14. *A typical file format for a mesh of a simplicial conforming triangulation*

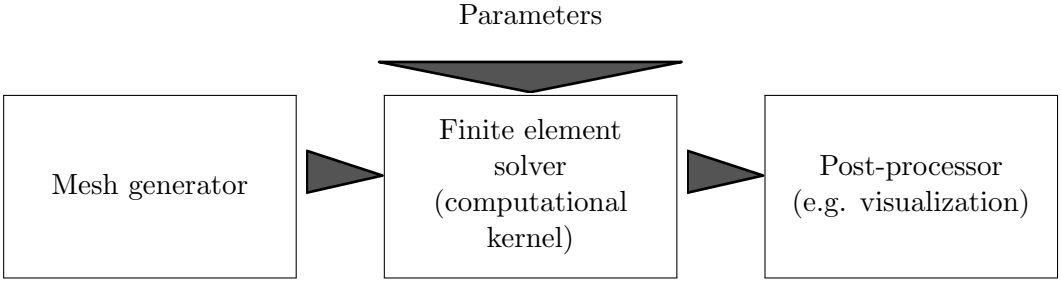


Figure 4.8: Flow of data in traditional finite element simulations

of a two-dimensional polygonal domain is the following:

$$
\begin{array}{lll}
\# \texttt{Two-dimensional simplicial mesh} & & \\
N \in \mathbb{N} & & \# \texttt{Number of nodes} \\
\xi_1 \ \eta_1 & & \# \texttt{Coordinates of first node} \\
\xi_2 \ \eta_2 & & \# \texttt{Coordinates of second node} \\
\vdots & & \\
\xi_N \ \eta_N & & \# \texttt{Coordinates of } N\texttt{-th node} \\
M \in \mathbb{N} & & \# \texttt{Number of triangles} \\
n_1^1 \ n_2^1 \ n_3^1 & X_1 & \# \texttt{Indices of nodes of first triangle} \\
n_1^2 \ n_2^2 \ n_3^2 & X_2 & \# \texttt{Indices of nodes of second triangle} \\
\vdots & & \\
n_1^M \ n_2^M \ n_3^M & X_M & \# \texttt{Indices of nodes of } M\texttt{-th triangle}
\end{array}
\tag{4.2}
$$

Here, X_i, $i = 1, \ldots, M$, is an additional piece of information that may, for instance, describe what kind of material properties prevail in triangle $\#i$. In this case X_i may be an integer index into a look-up table of material properties or the actual value of a coefficient function inside the triangle.

Additional information about edges located on $\partial\Omega$ may be provided in the following form:

$$
\begin{array}{lll}
K \in \mathbb{N} & & \# \texttt{Number of edges on } \partial\Omega \\
n_1^1 \ n_2^1 & Y_1 & \# \texttt{Indices of endpoints of first edge} \\
n_1^2 \ n_2^2 & Y_2 & \# \texttt{Indices of endpoints of second edge} \\
\vdots & & \\
n_1^K \ n_2^K & Y_K & \# \texttt{Indices of endpoints of } K\texttt{-th edge}
\end{array}
\tag{4.3}
$$

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 [15]. An interesting algorithm for Delaunay meshing is described in [16, 17]. 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

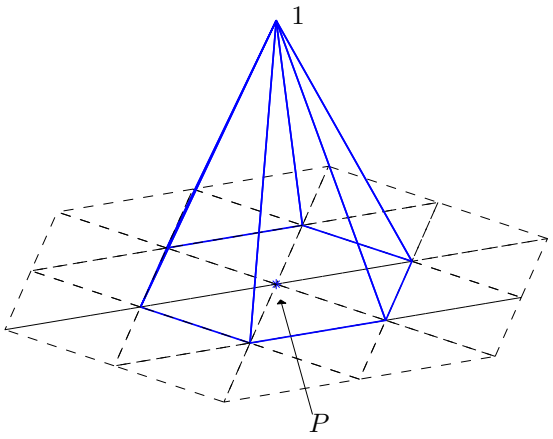


Figure 4.9: FE Basis function $b_P(\mathbf{x})$.

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

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

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* $P \in \mathcal{N}(\mathcal{M})$.
- $N = \dim S^1(\Omega, \mathcal{M}) = \sharp\mathcal{N} < \infty$.
- $S^1(\Omega, \mathcal{M}) = \operatorname{span}\{b_P(\mathbf{x}) : P \in \mathcal{N}(\mathcal{M})\}$, *where the* ***hut functions*** *are defined as*

$$b_P \in S^1(\Omega, \mathcal{M}), b_P(P') = \delta_{P=P'}.$$

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

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

so also the solution

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

4.2.2 Assembling of system matrix and load vector

The support of the basis functions $b_P(\mathbf{x})$ 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

$$N_{K,P_j(K)}(\mathbf{x}) = \widehat{N}_j(\Phi_K^{-1}\mathbf{x}). \tag{4.4}$$

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

$$\widehat{N}_0(\boldsymbol{\xi}) = 1 - \xi_1 - \xi_2, \qquad \widehat{N}_j(\boldsymbol{\xi}) = \xi_j, j = 1, 2. \tag{4.5}$$

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 $\mathbf{p}_j$. Then, the (affine) element mapping is

$$\mathbf{x} = \Phi_K(\boldsymbol{\xi}) = \mathbf{p}_0 + \xi_1(\mathbf{p}_1 - \mathbf{p}_0) + \xi_2(\mathbf{p}_2 - \mathbf{p}_0) = \boldsymbol{\tau} + \mathbf{F}_K\boldsymbol{\xi},$$

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

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.*,

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

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

$$\mathbf{T}_K = \begin{pmatrix} \cdot & \dots & \cdot & 1 & \cdot & \dots & \cdot & \cdot & \cdot & \dots & \cdot & \cdot & \cdot & \dots & \cdot \\ \cdot & \dots & \cdot & \cdot & \cdot & \dots & \cdot & 1 & \cdot & \dots & \cdot & \cdot & \cdot & \dots & \cdot \\ \cdot & \dots & \cdot & \cdot & \cdot & \dots & \cdot & \cdot & \cdot & \dots & \cdot & 1 & \cdot & \dots & \cdot \end{pmatrix}.$$

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

$$b_{P_i}(\mathbf{x})|_K = \sum_{k=1}^{3} (\mathbf{T})_{ki} N_{K,P_k(K)}(\mathbf{x}), \tag{4.6}$$

and the system matrix is given with

$$\begin{aligned}(\mathbf{B})_{ij} &= \mathsf{b}(b_{P_j}, b_{P_i}) = \sum_K (b_{P_j}|_K, b_{P_i}|_K) \\ &= \sum_K (\sum_{k=1}^{3} \mathsf{b}(\mathbf{T}_K)_{kj} N_{K,P_k(K)}, \sum_{\ell=1}^{3} (\mathbf{T}_K)_{\ell i} N_{K,P_\ell(K)}) \\ &= \sum_K \sum_{k=1}^{3} \sum_{\ell=1}^{3} (\mathbf{T}_K)_{kj} (\mathbf{T}_K)_{\ell i} \mathsf{b}(N_{K,P_k(K)}, N_{K,P_\ell(K)}) \\ &= \sum_K (\mathbf{T}_K)_{\cdot,j}^\top \mathbf{B}_K (\mathbf{T}_K)_{\cdot,i}\end{aligned} \tag{4.7}$$

as sum of "weighted" element matrices

$$\mathbf{B} = \sum_K \mathbf{T}_K^\top \mathbf{B}_K \mathbf{T}. \tag{4.8}$$

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 (**??**) consists of three parts, which contribute to the system matrix. The first part

$$\mathsf{a}(u, v) = \int_\Omega \langle \sigma \, \mathbf{grad}\, u, \mathbf{grad}\, v \rangle \, \mathrm{d}\mathbf{x}$$

is constituting the **stiffness matrix A**.

We transform the derivatives to $\widehat{K}$. By chain rule of differentiation and with

$$\nabla = \left(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2} \right)^\top, \quad \widehat{\nabla} = \left(\frac{\partial}{\partial \xi_1}, \frac{\partial}{\partial \xi_2} \right)^\top,$$

we have

$$\begin{aligned}\partial_{\xi_1} N_{K,P_j}(\mathbf{x}(\boldsymbol{\xi})) &= \partial_{x_1} N_{K,P_j}(\mathbf{x}) \frac{\partial x_1}{\partial \xi_1} + \partial_{x_2} N_{K,P_j}(\mathbf{x}) \frac{\partial x_2}{\partial \xi_1} \\ \partial_{\xi_2} N_{K,P_j}(\mathbf{x}(\boldsymbol{\xi})) &= \partial_{x_1} N_{K,P_j}(\mathbf{x}) \frac{\partial x_1}{\partial \xi_2} + \partial_{x_2} N_{K,P_j}(\mathbf{x}) \frac{\partial x_2}{\partial \xi_2}\end{aligned}$$

and so

$$\widehat{\nabla} N_{K,P_j}(\mathbf{x}(\boldsymbol{\xi})) = \underbrace{\left(\frac{\partial \boldsymbol{\Phi}_K}{\partial \boldsymbol{\xi}} \right)^\top}_{D\boldsymbol{\Phi}_K^\top} \nabla N_{K,P_j}(\mathbf{x}), \tag{4.9}$$

and so for triangular cells

$$\nabla N_{K,P_j}(\mathbf{x}) = \mathbf{F}_K^{-\top} \widehat{\nabla} N_{K,P_j}(\mathbf{x}(\boldsymbol{\xi})) = \mathbf{F}_K^{-\top} \widehat{\nabla} \widehat{N}_j(\boldsymbol{\xi}).$$

For a single cell K the entries of the element stiffness matrix $\mathbf{A}_K$ are so given by

$$\begin{aligned}\mathsf{a}_K(N_{K,P_j}, N_{K,P_i}) &= \int_K \sigma(\mathbf{x})(\nabla N_{K,P_j}(\mathbf{x}))^\top \nabla N_{K,P_i}(\mathbf{x})\mathrm{d}\mathbf{x} \\ &= \int_{\widehat{K}} \sigma(\mathbf{\Phi}_K\boldsymbol{\xi})(\widehat{\nabla}\widehat{N}_j(\boldsymbol{\xi}))^\top \mathbf{F}_K^{-1}\mathbf{F}_K^{-\top}\widehat{\nabla}\widehat{N}_i(\boldsymbol{\xi})|\mathbf{F}_K|\,\mathrm{d}\boldsymbol{\xi} \\ &= \int_{\widehat{K}} \sigma(\mathbf{\Phi}_K\boldsymbol{\xi})(\widehat{\nabla}\widehat{N}_j(\boldsymbol{\xi}))^\top \mathrm{adj}(\mathbf{F}_K)\mathrm{adj}(\mathbf{F}_K)^\top\widehat{\nabla}\widehat{N}_i(\boldsymbol{\xi})|\mathbf{F}_K|^{-1}\,\mathrm{d}\boldsymbol{\xi} \quad (4.10)\end{aligned}$$

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

$$\mathbf{F}_K^{-1} = |\mathbf{F}_K|^{-1}\mathrm{adj}(\mathbf{F}_K).$$

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

$$\widehat{\nabla}\widehat{N}_0 = \begin{pmatrix}-1\\-1\end{pmatrix}, \qquad \widehat{\nabla}\widehat{N}_1 = \begin{pmatrix}1\\0\end{pmatrix}, \qquad \widehat{\nabla}\widehat{N}_2 = \begin{pmatrix}0\\1\end{pmatrix}, \qquad (4.11)$$

So we can simplify

$$(\mathbf{A}_K)_{ij} = \mathsf{a}_K(N_{K,P_j}, N_{K,P_i}) = \frac{\sigma_K}{2}(\widehat{\nabla}\widehat{N}_j)^\top \mathrm{adj}(\mathbf{F}_K)\mathrm{adj}(\mathbf{F}_K)^\top\widehat{\nabla}\widehat{N}_i|\mathbf{F}_K|^{-1}$$

with $\sigma_K = \frac{1}{|K|}\int_K \sigma(\mathbf{x})\mathrm{d}\mathbf{x}$ the average heat conduction in K (note that $|\mathbf{F}_K| = 2|K|$). The latter simplies for constant material or may be obtained by numerical quadrature for more general smooth functions.

Due to the special values of $\widehat{\nabla}\widehat{N}_j$ (see (4.11)) we can even simplify [18]

$$\mathbf{A}_K = \frac{\sigma_K}{4|K|}\mathbf{D}_K^\top\mathbf{D}_K$$

with a matrix $\mathbf{D}_K$ with coordinate differences

$$\mathbf{D}_K = \begin{pmatrix} y_1 - y_2 & y_2 - y_0 & y_0 - y_1 \\ x_2 - x_1 & x_0 - x_2 & x_1 - x_0 \end{pmatrix}.$$

4.2.4 Element mass matrix

The mass matrix is related to the bilinear form

$$\mathsf{m}(u, v) = \int_\Omega c\, uv\,\mathrm{d}\mathbf{x}.$$

The element mass matrix $\mathbf{M}_K$ for the cell K can be computed as

$$\mathsf{m}_K(N_{K,P_j}, N_{K,P_i}) = \int_K c(\mathbf{x})\, N_{K,P_j}(\mathbf{x})N_{K,P_i}(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int_{\widehat{K}} c(\mathbf{\Phi}_K\boldsymbol{\xi})\widehat{N}_j(\boldsymbol{\xi})\widehat{N}_i(\boldsymbol{\xi})|F_K|\,\mathrm{d}\boldsymbol{\xi}. \qquad (4.12)$$

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

$$(\mathbf{M}_K)_{ij} = c_K|K|\begin{cases}\frac{1}{6} & i = j, \\ \frac{1}{12} & i \neq j.\end{cases}$$

4.2.5 Element load vector

The element load vector is related to the linear form

$$\ell_K(v) = \int_K f v \mathrm{d}\mathbf{x}.$$

For general, smooth function f we use numerical quadrature to evaluate the integrals. The simpliest quadrature rule is

$$\int_K f v \mathrm{d}\mathbf{x} \approx |K| f(\mathbf{x}_K) v(\mathbf{x}_K)$$

where $\mathbf{x}_K$ 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

$$\widehat{N}_0(\boldsymbol{\xi}) = (1-\xi_1)(1-\xi_2), \qquad \widehat{N}_1(\boldsymbol{\xi}) = \xi_1(1-\xi_2),$$
$$\widehat{N}_2(\boldsymbol{\xi}) = \xi_1\xi_2, \qquad \widehat{N}_3(\boldsymbol{\xi}) = (1-\xi_1)\xi_2,$$

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

$$\mathcal{Q}_1(\widehat{K}) := \mathrm{span}\{1, \xi_1, \xi_2, \xi_1\xi_2\}.$$

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

$$S^1(\Omega, \mathcal{M}) = \{u \in C^0(\Omega) : u(\mathbf{x})|_K \circ \boldsymbol{\Phi}_K(\boldsymbol{\xi}) \in \mathcal{Q}_1(\widehat{K})\},$$

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 $\boldsymbol{\Phi}_K$.

The matrix assembling is similar to that for triangles, whereas the $\mathbf{T}$-matrices have 4 rows and the element matrices follow by (4.10) and (4.12) where $\mathbf{F}_K$ is replaced by the Jacobian matrix $D\Phi_K(\boldsymbol{\xi})$, which is not constant anymore, but linear in ξ_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 $\sigma(\mathbf{x})$ and $c(\mathbf{x})$, 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 intervall $[-1,1]$. So, we simply transform 1D integrals via

$$\int_0^1 f(\xi)\,\mathrm{d}\xi = \frac{1}{2}\int_{-1}^1 f(\tfrac{\xi+1}{2})\,\mathrm{d}\xi \approx \frac{1}{2}\sum_{j=1}^{n} w_j f(\tfrac{\xi_j+1}{2}), \tag{4.13}$$

where w_j are the weights and ξ_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 [19, 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 [19, page 921ff]) and there is an algorithm (see Matlab version `gauleg.m` on the webpage of the lecture) in the numerical recipes [20].

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

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

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 $D\boldsymbol{\Phi}_K(\boldsymbol{\xi})$ 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 $\boldsymbol{\Phi}_{K_1}$ and $\boldsymbol{\Phi}_{K_2}$, is the same. This is also true for mixed meshes of curved triangles and quadrilaterals.

Let us define the local space for triangles as

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

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

$$\begin{aligned} S^1(\Omega,\mathcal{M}) = \{\, u\in H^1(\Omega) : u(\mathbf{x})|_K \circ \boldsymbol{\Phi}_K(\boldsymbol{\xi}) \in \mathcal{P}_1(\widehat{K}) \text{ if } K \text{ is a triangle or} \\ u(\mathbf{x})|_K \circ \boldsymbol{\Phi}_K(\boldsymbol{\xi}) \in \mathcal{Q}_1(\widehat{K}) \text{ if } K \text{ is a quadrilateral}\,\}. \end{aligned}$$

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 $\widehat{K}$ can easily transformed to.

$$\int_0^1 \int_0^{1-\xi_1} f(\boldsymbol{\xi})\,\mathrm{d}\xi_2\,\mathrm{d}\xi_1 = \frac{1}{4}\int_{-1}^{1}\int_{-1}^{2-\xi_1} f(\tfrac{\xi_1+1}{2},\tfrac{\xi_2+1}{2})\,\mathrm{d}\xi_2\,\mathrm{d}\xi_1 \approx \frac{1}{4}\sum_{j=1}^{n} w_j f(\tfrac{\xi_{1,j}+1}{2},\tfrac{\xi_{2,j}+1}{2}),$$

Let us define generalisations of $\mathcal{P}_1$ and $\mathcal{Q}_1$.

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

$$\mathcal{P}_m(K) := \operatorname{span}\{\boldsymbol{\xi} \in K \mapsto \boldsymbol{\xi}^{\boldsymbol{\alpha}} := \xi_1^{\alpha_1} \cdot \cdots \cdot \xi_d^{\alpha_d}, \boldsymbol{\alpha} \in \mathbb{N}_0^d,\, |\boldsymbol{\alpha}| \le m\}$$

*for the vector space of d-**variate polynomials** of (total) degree m, $m \in \mathbb{N}_0$.*

If $\mathbf{m} = (m_1,\ldots,m_d)^T \in \mathbb{N}_0^d$ we designate by

$$\mathcal{Q}_{\mathbf{m}}(K) := \operatorname{span}\{\boldsymbol{\xi} \in K \mapsto \xi_1^{\alpha_1} \cdot \cdots \cdot \xi_d^{\alpha_d},\, 0 \le \alpha_k \le m_k,\, 1 \le k \le d\}$$

*the space of **tensor product polynomials** of maximal degree m_k in the k-th coordinate direction.*

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 [21].

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 $K = \operatorname{convex}\{\binom{0}{0},\binom{1}{0},\binom{1}{1}\}$ and $\widehat{K} =]0,1[^2$, then

$$\int_K f(\xi_1,\xi_2)\,\mathrm{d}\mathbf{x} = \int_{\widehat{K}} f(\widehat{\xi}_1(1-\widehat{\xi}_2),\widehat{\xi}_2)\,(1-\widehat{\xi}_2)\,\mathrm{d}\widehat{\boldsymbol{\xi}}\,.$$

If $f \in \mathcal{P}_m(K)$, then the integrand on the right hand side will belong to $\mathcal{Q}_{m,m+1}(\widehat{K})$.

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.*

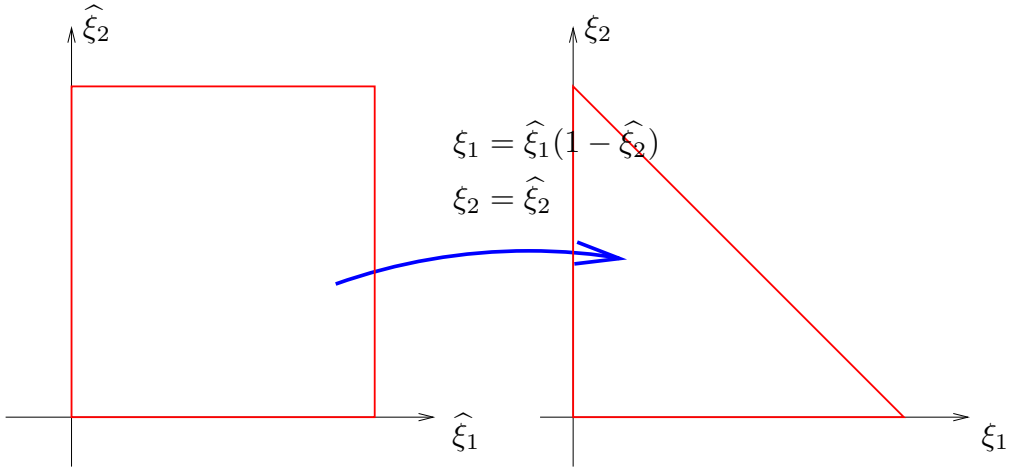


Figure 4.10: "Duffy transformation" of a square into a triangle.

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 $\widehat{K} = [0,1]^2$ as

$$\begin{aligned}
\widehat{N}_{e,1,j}(\boldsymbol{\xi}) &= (1-\xi_2)\, P_j(\xi_1), && \text{lower edge,}\\
\widehat{N}_{e,2,j}(\boldsymbol{\xi}) &= \xi_1\, P_j(\xi_2), && \text{right edge,}\\
\widehat{N}_{e,3,j}(\boldsymbol{\xi}) &= \xi_2\, P_j(\xi_1), && \text{upper edge,}\\
\widehat{N}_{e,4,j}(\boldsymbol{\xi}) &= (1-\xi_1)\, P_j(\xi_2), && \text{left edge,}
\end{aligned}$$

and the latter as

$$\widehat{N}_{c,i,j}(\boldsymbol{\xi}) = P_{i,j}(\xi_1, \xi_2),$$

where $P_j(\xi)$ are polynomials vanishing at $\xi \in \{0,1\}$ and $P_{i,j}(\boldsymbol{\xi})$ are polynomials vanishing at $\partial\widehat{K}$. 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 ξ_1 or ξ_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

$$L_0(\xi) = 1, \qquad L_1(\xi) = \xi \qquad (j+1)L_{j+1}(\xi) = (2j+1)\xi L_j(\xi) - jL_{j-1}(\xi), \quad j > 1,$$

and

$$\int_{-1}^{1} L_j(\xi)L_i(\xi)\,\mathrm{d}\xi = \delta_{ij}\frac{2}{2j+1}.$$

The integrated Legendre polynomials are

$$\widehat{L}_0(\xi) = -1, \qquad \widehat{L}_1(\xi) = \xi, \qquad \widehat{L}_j(\xi) = \int_{-1}^{\xi} L_{j-1}(t)\,\mathrm{d}t = \frac{1}{2j-1}\left(L_j(\xi) - L_{j-2}(\xi)\right).$$

The 1D element shape functions are then defined in $[0,1]$ as

$$\begin{aligned}
&\widehat{N}_0(\xi) = 1-\xi, \quad \widehat{N}_1(\xi) = \xi, \\
&\widehat{N}_j(\xi) = \sqrt{\frac{(2j-1)}{2}}\widehat{L}_j(2\xi-1) = \frac{1}{\sqrt{2(2j-1)}}\left(L_j(2\xi-1) - L_{j-2}(2\xi-1)\right).
\end{aligned}$$

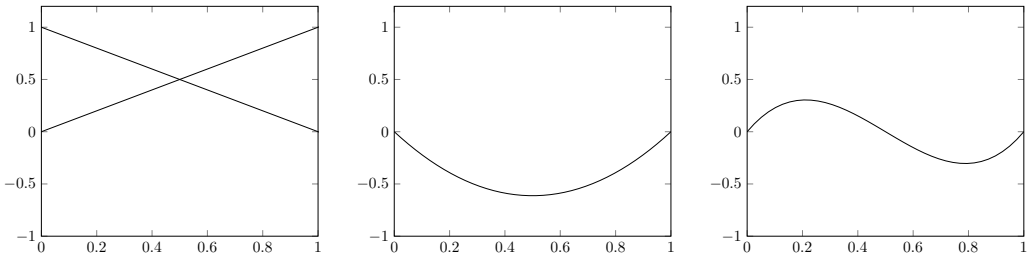
As in $\widehat{N}_{2j}$ for $j \geq 1$ only even degree monomials are present, and in $\widehat{N}_{2j+1}$ only odd degree monomials, we have for $j \geq 1$

$$\widehat{N}_{2j}(\xi) = \widehat{N}_{2j}(-\xi), \qquad\qquad \widehat{N}_{2j+1}(\xi) = -\widehat{N}_{2j+1}(-\xi).$$

Then, the 2D element shape functions for $\mathcal{Q}_p([0,1]^2)$, $p \geq 1$ are given by

$$\widehat{N}_{(p+1)i+j}(\boldsymbol{\xi}) = \widehat{N}_j(\xi_1)\widehat{N}_i(\xi_2), \quad 0 \leq i,j \leq p.$$

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.



(a) The linear element shape function $\widehat{N}_0(\xi)$, $\widehat{N}_1(\xi)$.

(b) The even element shape function $\widehat{N}_2(\xi)$.

(c) The odd element shape function $\widehat{N}_3(\xi)$.

Figure 4.11: The first 1D element shape functions based on integrated Legendre polynomials.

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 σ 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

- 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 ξ_1 or ξ_2-direction to the (global) orientation of the edge. If they coincide the entry in the $\mathbf{T}$-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*

$$\lambda_1 + \cdots + \lambda_{d+1} = 1 \quad , \quad \lambda_1 \mathbf{p}_0 + \cdots + \lambda_d \mathbf{p}_d = \mathbf{x} \, .$$

The barycentric coordinates for a point $\mathbf{x}$ can be obtained by solving

$$\begin{pmatrix} p_{0,1} & \cdots & p_{d,1} \\ \vdots & & \vdots \\ p_{1,d} & \cdots & p_{d,d} \\ 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_d \\ \lambda_{d+1} \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_d \\ 1 \end{pmatrix} \, , \tag{4.14}$$

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

$$\text{convex}\{\mathbf{p}^0, \ldots, \mathbf{p}^d\} = \{\mathbf{x} \in \mathbb{R}^d, \; 0 \le \lambda_i(\mathbf{x}) \le 1, \; 1 \le i \le d+1\} \, .$$

For the reference triangle $\widehat{K}$ we have

$$\lambda_1 = 1 - \xi_1 - \xi_2, \qquad \lambda_2 = \xi_1, \qquad \lambda_3 = \xi_2,$$

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

$$\widehat{N}_0(\boldsymbol{\lambda}) = \lambda_1, \qquad \widehat{N}_1(\boldsymbol{\lambda}) = \lambda_2, \qquad \widehat{N}_2(\boldsymbol{\lambda}) = \lambda_3.$$

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

$$\begin{aligned}
\widehat{N}_{e,1,j}(\boldsymbol{\lambda}) &= \lambda_2\lambda_3 P_j(\lambda_2)\,, && \text{edge opposite to node } P_0,\\
\widehat{N}_{e,2,j}(\boldsymbol{\lambda}) &= \lambda_1\lambda_3 P_j(\lambda_3)\,, && \text{edge opposite to node } P_1,\\
\widehat{N}_{e,3,j}(\boldsymbol{\lambda}) &= \lambda_1\lambda_2 P_j(\lambda_1)\,, && \text{edge opposite to node } P_2.
\end{aligned}$$

The interior element shape functions vanish on all three edges and can be written as

$$\widehat{N}_{c,i,j}(\boldsymbol{\lambda}) = \lambda_1\lambda_2\lambda_3 P_{i,j}(\boldsymbol{\lambda}).$$

The functions $P_j(\lambda)$ can be chosen as integrated Legendre polynomials $\widehat{L}_0, \dots \widehat{L}_{p-2}$, and $P_{i,j}(\boldsymbol{\lambda})$ as tensor product of integrated Legendre polynomials in λ_1 and λ_2 (λ_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 $\beta_1, \beta_2, \beta_3 \in \mathbb{N}$,*

$$\int_K \lambda_1^{\beta_1}\lambda_2^{\beta_2}\lambda_3^{\beta_3}\,\mathrm{d}\mathbf{x} = 2|K| \cdot \frac{\beta_1!\,\beta_2!\,\beta_3!}{(\beta_1+\beta_2+\beta_3+2)!}\,.$$

Proof. The first step amounts to a transformation of K to the "unit triangle" $\widehat{K} := \text{convex}\left\{\binom{0}{0}, \binom{1}{0}, \binom{0}{1}\right\}$, which leads to

$$\int_K \lambda_1^{\beta_1}\lambda_2^{\beta_2}\lambda_3^{\beta_3}\,\mathrm{d}\mathbf{x} = 2|K| \int_0^1 \int_0^{1-\xi_2} \xi_1^{\beta_1}\xi_2^{\beta_2}(1-\xi_1-\xi_2)^{\beta_3}\,\mathrm{d}\xi_1\mathrm{d}\xi_2 =: 2|K|\; I.$$

Note, that the barycentric coordinates evaluated on $\widehat{K}$ are the barycentric coordinates on the reference cell, which are ξ_1, ξ_2, $1-\xi_1-\xi_2$. Transforming $\widehat{K}$ to $(0,1)^2$ by the Duffy-transformation $\xi_1 = \widehat{\xi}_1$, $\xi_2 = \widehat{\xi}_1(1-\widehat{\xi}_2)$ (see Fig. 4.10) we get

$$\begin{aligned}
I &= \int_0^1 \int_0^1 \widehat{\xi}_1^{\beta_1}(1-\widehat{\xi}_2)^{\beta_1}\widehat{\xi}_2^{\beta_2}(1-\widehat{\xi}_2-\widehat{\xi}_1(1-\widehat{\xi}_2))^{\beta_3}\,(1-\widehat{\xi}_2)\mathrm{d}\widehat{\xi}_1\,\mathrm{d}\widehat{\xi}_2\\
&= \int_0^1 \widehat{\xi}_1^{\beta_1}(1-\widehat{\xi}_1)^{\beta_3}\mathrm{d}\widehat{\xi}_1 \int_0^1 \widehat{\xi}_2^{\beta_2}(1-\widehat{\xi}_2)^{\beta_1+\beta_3+1}\mathrm{d}\widehat{\xi}_2 = B(\beta_1+1,\beta_3+1)\,B(\beta_2+1,\beta_1+\beta_3+2)
\end{aligned}$$

where we used Euler's beta functions

$$B(\alpha,\beta) := \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,\mathrm{d}t\,, \quad 0<\alpha,\beta<\infty\,.$$

Using the known formula $\Gamma(\alpha+\beta)\,B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)$, Γ the Gamma function, we end up with

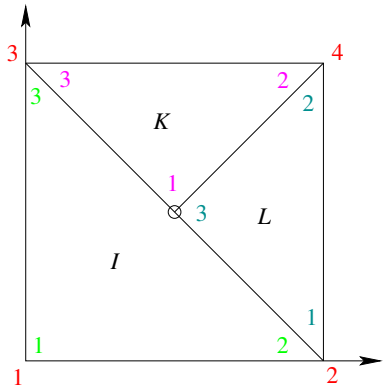
$$I = \frac{\Gamma(\beta_1+1)\Gamma(\beta_3+1)}{\Gamma(\beta_1+\beta_3+2)} \cdot \frac{\Gamma(\beta_2+1)\Gamma(\beta_1+\beta_3+2)}{\Gamma(\beta_1+\beta_2+\beta_3+3)}.$$

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.

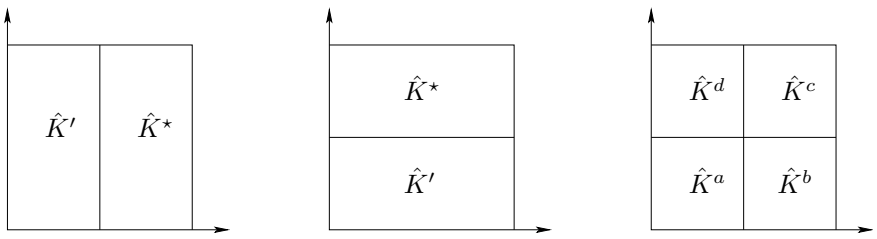
Example 4.20 (**T**-Matrices for an irregular mesh)**.**
Three cells with each three shape functions and four global basis functions. The hanging node is marked with ○.



$$\mathbf{T}_L = \begin{pmatrix} & 1 & 2 & 3 & 4 \\ 1 & \cdot & 1 & \cdot & \cdot \\ 2 & \cdot & \cdot & \cdot & 1 \\ 3 & \cdot & 1/2 & 1/2 & \cdot \end{pmatrix}$$

$$\mathbf{T}_K = \begin{pmatrix} & 1 & 2 & 3 & 4 \\ 1 & \cdot & 1/2 & 1/2 & \cdot \\ 2 & \cdot & \cdot & \cdot & 1 \\ 3 & \cdot & \cdot & 1 & \cdot \end{pmatrix}$$

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 **T** matrices as introduced in the former sections.



(a) Vertical. (b) Horizontal. (c) Into four childs.

Figure 4.12: Subdivison 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 **S** matrix is involved. The shape functions $N_{K,i}$ are defined as element shape function $\widehat{N}_i$ on the reference element $\widehat{K}$ and then transformed with the element map $\mathbf{\Phi}_K$

to K. As K' is a part of K we can write a representation between $\widehat{K}$ and $\widehat{K}' = \mathbf{\Phi}_K^{-1} K'$, where $\mathbf{x} \in \widehat{K}'$

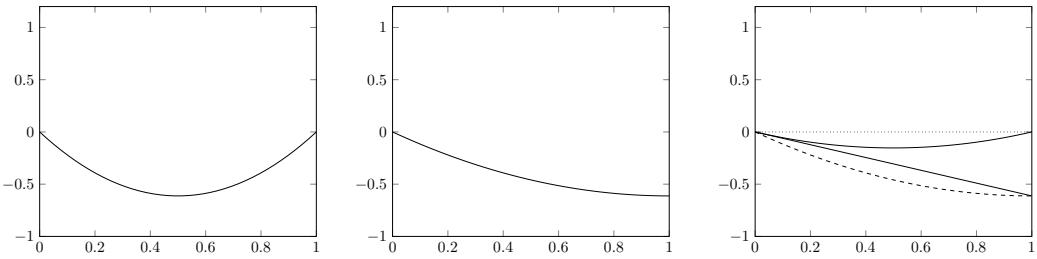
$$\widehat{N}_i(\underbrace{\mathbf{\Phi}_K^{-1}\mathbf{x}}_{\in \widehat{K}'}) = \sum_j (\mathbf{S}_{K'K})_{ji} \widehat{N}_j(\underbrace{\mathbf{\Phi}_{K'}^{-1}\mathbf{x}}_{\in \widehat{K}})$$

or

$$\widehat{N}_i(\underbrace{\mathbf{\Phi}_{\widehat{K}'}\boldsymbol{\xi}}_{\in \widehat{K}'}) = (\mathbf{S}_{K'K})_{ji} \widehat{N}_j(\underbrace{\boldsymbol{\xi}}_{\in \widehat{K}})$$

Obviously, we can also write $\mathbf{S}_{\widehat{K}'\widehat{K}}$ 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.

S-matrices in 1D The reference interval $\widehat{I} = [0, 1]$ is subdivided in the left part $\widehat{I}' = [0, \frac{1}{2}]$ and the right one $\widehat{I}^\star = [\frac{1}{2}, 1]$. So for the left part we have $\mathbf{\Phi}_{\widehat{K}'} = \frac{\xi}{2}$.



(a) The even element shape function $\widehat{N}_2(\xi)$. (b) $\widehat{N}_2(\xi)$ evaluated on $(0, \frac{1}{2})$ mapped to $(0, 1)$. (c) $-\frac{\sqrt{6}}{4}\widehat{N}_1(\xi)$ and $\frac{1}{4}\widehat{N}_2(\xi)$.

Figure 4.13: The representation of $\widehat{N}_2(\xi)$ evaluated in $[0, \frac{1}{2}]$ by element shape functions.

Example 4.21. *If we take for instance the 1D element shape function (see Fig. 4.13)*

$$\widehat{N}_2(\xi) = \sqrt{6}\xi(\xi - 1)$$

and we search $(\mathbf{S}_{\widehat{I}'\widehat{I}})_{2,j}$ *such that for* $\xi \in \widehat{I}$

$$\widehat{N}_2\left(\frac{\xi}{2}\right) = \sum_{j=0}^{2} (\mathbf{S}_{\widehat{I}'\widehat{I}})_{2,j} \widehat{N}_j(\xi).$$

We find

$$\widehat{N}_2\left(\frac{\xi}{2}\right) = \frac{\sqrt{6}}{4}\xi(\xi - 2) = -\frac{\sqrt{6}}{4}\underbrace{\widehat{N}_1(\xi)}_{\xi} + \frac{1}{4}\underbrace{\widehat{N}_2(\xi)}_{\sqrt{6}\xi(\xi-1)},$$

and so

$$(\mathbf{S}_{\widehat{I}'\widehat{I}})_{2,\cdot} = \begin{pmatrix} 0 & -\frac{\sqrt{6}}{4} & \frac{1}{4}. \end{pmatrix}$$

In general the 1D **S** matrices can be obtained by evaluating the element shape functions on to $[0,1]$ transformed Chebychev points ξ_ℓ and on the points $\mathbf{\Phi}_{\widehat{K}'}\xi_\ell$, and solving a (small) linear system. Note, that in case of a hierarchic basis $\{\widehat{N}_j\}_j$ 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.*,

$$\widehat{N}_{k\ell} = \widehat{N}_k \otimes \widehat{N}_\ell.$$

For an anisotropic subdivision the 2D **S** matrices are a tensor product like

$$\mathbf{S}_{\widehat{K}'\widehat{K}} = \mathbf{S}_{\widehat{I}'\widehat{I}} \otimes \mathbf{I}$$

where **I** is the identity matrix.

For subdivision in both direction we have a tensor product like

$$\mathbf{S}_{\widehat{K}^d\widehat{K}} = \mathbf{S}_{\widehat{I}'\widehat{I}} \otimes \mathbf{S}_{\widehat{I}^\star\widehat{I}}.$$

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** Π_K which has a particular dimension. In the space Π_K we have the polynomials transformed from the reference element.

Then we call a **local degrees of freedom** a linear functional $(C^\infty(\overline{K}))^l \mapsto \mathbb{R}$, if a set Σ_K of local degrees of freedom provide a basis for the dual space $(\Pi_K)'$, *i. e.*

$$\sharp\Sigma_K = \dim \Pi_K \quad \text{and} \quad \forall v \in \Pi_K : \quad l(v) = 0 \; \forall l \in \Sigma_K \quad \Rightarrow \quad v = 0 \; .$$

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 Σ_K such that

$$l_m^K(N_{K,j}) = \delta_{mj} \; , \quad m, j \in \{1, \ldots, \dim \Pi_K\} \; . \tag{4.15}$$

Example 4.22 (Linear finite elements)**.** *The local degrees of freedom can be chosen as point evaluations on the nodes*

$$l_j^K(v) := v(\mathbf{p}_j(K)), \tag{4.16}$$

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*

$$v_e(\mathbf{x}) := v(\mathbf{x}) - \sum_{j=1}^{4} l_j^K(v) N_{K,j}(\mathbf{x})$$

is zero on all four nodes.

Note, that the integrated Legendre polynomials vanish on both end points for $i \geq 2$, i. e., $\widehat{L}_i(\pm 1) = 0$, and so for any $j \in \mathbb{N}$

$$\int_{-1}^{1} \widehat{L}_i'(\xi)\widehat{L}_j'(\xi)\mathrm{d}\xi = -\int_{-1}^{1} \widehat{L}_i(\xi)\widehat{L}_j''(\xi)\mathrm{d}\xi.$$

On the other hand

$$\int_{-1}^{1} \widehat{L}_i'(\xi)\widehat{L}_j'(\xi)\mathrm{d}\xi = \int_{-1}^{1} L_{i-1}(\xi)L_{j-1}(\xi)\mathrm{d}\xi = \delta_{ij}\frac{2}{2j-1}.$$

So

$$\begin{aligned}-2\int_0^1 \widehat{N}_i(\xi)\widehat{N}_j''(\xi)\mathrm{d}\xi &= -2\sqrt{\frac{2j-1}{2}}\sqrt{\frac{2i-1}{2}}\int_0^1 \widehat{L}_i(2\xi-1)\widehat{L}_j''(2\xi-1)\mathrm{d}\xi \\ &= \frac{1}{2}\sqrt{(2j-1)(2i-1)}\int_{-1}^{1} \widehat{L}_i'(\xi)\widehat{L}_j'(\xi)\mathrm{d}\xi = \delta_{ij}.\end{aligned}$$

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 $j = 0, \ldots, p-2$

$$\begin{aligned}l_{e,i,j}^K(v) &= -2\int_0^1 \underbrace{\widehat{v}_e(\xi)}_{v_e(\mathbf{\Phi}_{E_i(K)}\xi)} \widehat{N}_{j+2}''(\xi)\,\mathrm{d}\xi \\ &= -2\int\limits_{E_i(K)} v_e(\mathbf{x})\,\widehat{N}_{j+2}''(\mathbf{\Phi}_{E_i(K)}^{-1}\mathbf{x})\,|D\mathbf{\Phi}_{E_i(K)}(\mathbf{\Phi}_{E_i(K)}^{-1}\mathbf{x})|^{-1}\,\mathrm{d}S(\mathbf{x}), \qquad (4.17)\end{aligned}$$

where $\mathbf{\Phi}_{E_i(K)}$ is the mapping from the reference interval $[0,1]$ to $E_i(K)$ which is introduced as the mapping $\mathbf{\Phi}_K$ of the respective edge in $\widehat{K}$ to $E_i(K)$ respecting the (global) orientation of $E_i(K)$ (if the local ξ_1 or ξ_2 direction along the edge in $\widehat{K}$ is opposite to global orientation, ξ_1 or ξ_2, respectively, is replaced by $1-\xi_1$ or $1-\xi_2$).

Then, in case $v \in \Pi_K$ the function

$$v_c(\mathbf{x}) := v_e(\mathbf{x}) - \sum_{i=1}^{4}\sum_{j=0}^{p-2} l_{e,i,j}^K(v)\underbrace{N_{e,i,j}(\mathbf{x})}_{\widehat{N}_{e,i,j}(\mathbf{\Phi}_{E_i(K)}^{-1}\mathbf{x})}$$

is vanishing on all edges of K.

Similarily, we define for $i, j = 0, \ldots, p-2$,

$$\begin{aligned}l_{c,i,j}^K(v) &= 4\int_{\widehat{K}} \underbrace{\widehat{v}_c(\boldsymbol{\xi})}_{v_c(\mathbf{x})} \widehat{N}_{i+2}''(\xi_1)\widehat{N}_{j+2}''(\xi_2)\mathrm{d}\boldsymbol{\xi} \\ &:= 4\int_K v_c(\mathbf{x})\,\widehat{N}_{i+2}''(\xi_1)\widehat{N}_{j+2}''(\xi_2)\,|D\mathbf{\Phi}_K(\mathbf{\Phi}_K^{-1}\mathbf{x})|^{-1}\mathrm{d}\mathbf{x},\end{aligned}$$

with $\boldsymbol{\xi} = \mathbf{\Phi}_K^{-1}\mathbf{x}$.

Definition 4.24. *A **finite element** is a triple (K, Π_K, Σ_K) such that*

(i) K *is a cell of a mesh* $\mathcal{M}$ *of the computational domain* $\Omega \subset \mathbb{R}^d$.

(ii) $\Pi_K \subset (C^\infty(\overline{K}))^l$ *is the trial space with* $\dim \Pi_K < \infty$.

(iii) Σ_K *is a set of local degrees of freedom.*

A finite element is called V***-conforming****, if*

(iv) for any face F *of* K *the degrees of freedom localized on* F *uniquely determine the natural trace* $\mathsf{R}\, u_{|F}$ *of a* $u \in \Pi_K$ *onto* F.

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* $\overline{K}$. *Given a local trial space* Π_K, $\Pi_K \subset (C^\infty(\overline{K}))^l$, *and a set* Σ_K *of local degrees of freedom, a linear functional* $l \in \Sigma_K$ *is called* ***localized/supported on*** F *or* ***associated with*** F, *if*

$$l(v) = 0 \quad \forall v \in (C^\infty(\overline{K}))^l,\ \mathrm{supp}(v) \cap F = \emptyset\ .$$

Notation 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 **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).

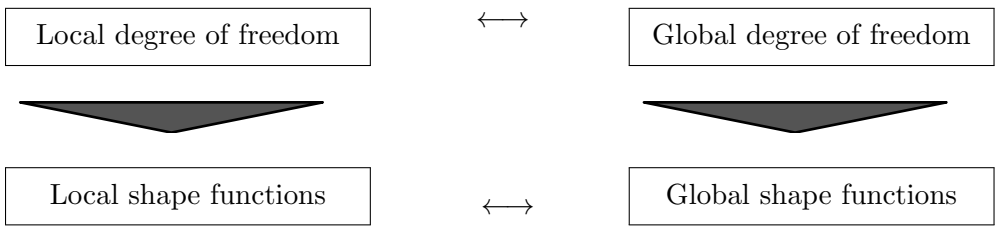


Figure 4.14: Duality of degrees of freedom and shape functions

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 ($H(\mathbf{curl}, \Omega)$ or $H(\mathrm{div}, \Omega)$ conforming.

References

[15] P.J. Frey and P.-L. George. *Mesh generation. Application to finite elements.* Hermes Science Publishing, Oxford, UK, 2000.

[16] J. Ruppert. A Delaunay refinement algorithm for quality 2-dimensional mesh generation. *J. Algorithms*, 18(3):548–585, 1995.

[17] J.R. Shewchuk. Triangle: Engineering a 2D Quality Mesh Generator and Delaunay Triangulator. In Ming C. Lin and Dinesh Manocha, editors, *Applied Computational Geometry: Towards Geometric Engineering*, volume 1148 of *Lecture Notes in Computer Science*, pages 203–222. Springer-Verlag, May 1996.

[18] Christoph Schwab. Einführung in die Numerik partieller Differentialgleichungen: Stationäre Probleme. Vorlesungsmanuskript, WS 2000/01.

[19] Milton Abramowitz and Irene A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth dover printing, tenth gpo printing edition, 1964.

[20] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, et al. *Numerical recipes*, volume 3. Cambridge university press Cambridge, 2007.

[21] Pavel Šolín. *Partial Differential Equations and the Finite Element Method*. Wiley-Interscience, Hoboken, USA, 2006.

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 3.29) the discretisation eror is bounded "quasi optimally" by the best approximation error:

$$\|u - u_n\|_V \leq \left(1 + \frac{\|\mathsf{b}\|_{V\times V\mapsto\mathbb{R}}}{\gamma_n}\right) \inf_{w_n\in W_n} \|u - w_n\|_V \ , \tag{3.21}$$

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

$$\inf_{w_n\in W_n} \|u - w_n\|_V \leq \|u - \mathsf{I}_h\, u\|_V \,, \tag{5.1}$$

with $\mathsf{I}_h : V \to W_n$ or at least $\mathsf{I}_h : V_S \to W_n$, $V_S \subset V$ 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 3.27) 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 Lipshitz-domain and $m \in \mathbb{N}$, then*

$$\exists\gamma = \gamma(m,\Omega) > 0 : \qquad \inf_{p\in\mathcal{P}_{m-1}(\Omega)} \|v - p\|_{H^m(\Omega)} \leq \ \gamma\, |v|_{H^m(\Omega)} \quad \forall v \in H^m(\Omega) \ .$$

Proof. The proof can be found in [22]. □

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 [23]:

Lemma 5.2. *If $\Omega \subset \mathbb{R}^d$ is a bounded Lipshitz-domain and $m \in \mathbb{N}$, then*

$$\exists\gamma = \gamma(m,\Omega) > 0 : \qquad \inf_{p\in\mathcal{Q}_{m-1}(\Omega)} \|v - p\|_{H^m(\Omega)} \leq \gamma \left(\sum_{i=1}^{d} \left\|\frac{\partial^m v}{\partial\xi_i^m}\right\|^2_{L^2(\Omega)}\right)^{1/2} \quad \forall v \in H^m(\Omega) \ .$$

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

$$(\mathsf{I}_n\, v)(\mathbf{x}) := \sum_{j=1}^{N} l_j(v) b_j(\mathbf{x})$$

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

$$(\mathsf{I}_n\, v)(\mathbf{x}) := \sum_{j=1}^{\dim \Pi_K} l_j^K(v) N_{K,j}(\mathbf{x}), \quad \mathbf{x} \in K.$$

5.4 The interpolation error estimates on simplices

We need three properties of I_n on simplices:

- The shape functions are polynomials on K, $\Pi_K = \mathcal{P}_m(K)$.
- The interpolation operator I_n preserves polynomials, *i. e.*, for $p \in \mathcal{P}_m(K)$

$$\mathsf{I}_n\, p = \mathsf{I}_n(\sum_{i=1}^{\dim \Pi_K} \alpha_i N_{K,i}) = \sum_{j=1}^{\dim \Pi_K} \sum_{i=1}^{\dim \Pi_K} \alpha_i \underbrace{l_j(N_{K,i})}_{\delta_{ij}} N_{K,j} = \sum_{j=1}^{\dim \Pi_K} \alpha_j N_{K,j} = p.$$

- The interpolation operator I_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 [24, 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 $C^0(\overline{\Omega})$.*

With the Bramble-Hilbert-Lemma, the triangle inequality and $t \leq m+1$ we can estimate

$$\begin{aligned}
\|u - \mathsf{I}_n\, v\|_{H^1(K)} &\leq \inf_{\mathcal{P}_m(K)} \|(u-p) - \mathsf{I}_n(u-p)\|_{H^1(K)} \\
&\leq \inf_{\mathcal{P}_m(K)} \left(\|u-p\|_{H^1(K)} + \|\,\mathsf{I}_n(u-p)\|_{H^1(K)}\right) \\
&\leq (1 + C(m,K)) \inf_{\mathcal{P}_m(K)} \|u-p\|_{H^t(K)} \\
&\leq \begin{cases} (1 + C(m,K)) \inf_{\mathcal{P}_{t-1}(K)} \|u-p\|_{H^t(K)} & \leq \gamma(t,K)(1 + C(m,K))\ |u|_{H^t(K)}, \\ (1 + C(m,K)) \inf_{\mathcal{P}_m(K)} \|u-p\|_{H^{m+1}(K)} & \leq \gamma(m,K)(1 + C(m,K))\ |u|_{H^{m+1}(K)}. \end{cases}
\end{aligned}$$

▷ 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

$$N_{K,j}(\mathbf{\Phi}_K(\boldsymbol{\xi})) = \widehat{N}_j(\boldsymbol{\xi}).$$

Equivalently, we can define local degrees of freedom on $\widehat{K}$ with

$$\widehat{l}_j(\widehat{v}) = l_j^K(v)$$

where $\widehat{v}(\boldsymbol{\xi}) := v(\mathbf{\Phi}_K(\boldsymbol{\xi}))$ is the **pullback** of v.

Then, we can define an interpolation operator on $\widehat{K}$ as

$$(\widehat{I}\,\widehat{v})(\boldsymbol{\xi}) := \sum_{j=1}^{\dim \Pi_K} \widehat{l}_j(\widehat{v})\widehat{N}_j(\boldsymbol{\xi}), \tag{5.2}$$

which equalise the transformed interpolation operator

$$\widehat{\mathsf{I}_n\, v}(\boldsymbol{\xi}) := (\mathsf{I}_n\, v)(\mathbf{x}(\boldsymbol{\xi})) = \sum_{j=1}^{\dim \Pi_K} l_j^K(v) N_{K,j}(\mathbf{x}(\boldsymbol{\xi})) = \sum_{j=1}^{\dim \Pi_K} \widehat{l}(\widehat{v})\widehat{N}_j(\boldsymbol{\xi}) = (\widehat{I}\,\widehat{v})(\boldsymbol{\xi}). \tag{5.3}$$

Estimate on the reference element Repeating the steps, we went on K, on the reference element $\widehat{K}$ we obtain estimates for the interpolation error for $\widehat{I}$. Let us measure the error in the H^r-norm with $r \le t \le m+1$ ($t \ge 1$ for $d = 1$ and $t \ge 2$ for $d = 2, 3$).

$$\begin{aligned}\left\|\widehat{u} - \widehat{I}\widehat{u}\right\|_{H^r(\widehat{K})} &= \inf_{p\in\mathcal{P}_m(\widehat{K})} \left\|(\widehat{u}-p) - \widehat{I}(\widehat{u}-p)\right\|_{H^r(\widehat{K})} \\ &\le (1+C) \inf_{p\in\mathcal{P}_m(\widehat{K})} \|\widehat{u}-p\|_{H^t(\widehat{K})} \le \gamma(m)(1+C)|\widehat{u}|_{H^t(\widehat{K})}.\end{aligned}$$

Transformation techniques Let us transform Sobolev norms between the reference triangle $\widehat{K}$ and the triangle K and vice-versa.

Lemma 5.4. *If $\mathbf{\Phi}_K : \widehat{K} \mapsto K$ is an affine mapping $\boldsymbol{\xi} \mapsto \mathbf{F}_K\boldsymbol{\xi} + \boldsymbol{\tau}$, then, for all $m \in \mathbb{N}_0$,*

$$\begin{aligned}|\widehat{u}|_{H^m(\widehat{K})} &\le \binom{m+d-1}{d}^{1/2} d^m \left\|\mathbf{F}_K\right\|^m |\det(\mathbf{F}_K)|^{-1/2} |u|_{H^m(K)} \quad \forall u \in H^m(K)\ , \\ |u|_{H^m(K)} &\le \binom{m+d-1}{d}^{1/2} d^m \left\|\mathbf{F}_K^{-1}\right\|^m |\det(\mathbf{F}_K)|^{1/2} |\widehat{u}|_{H^m(\widehat{K})} \quad \forall u \in H^m(\widehat{K})\ .\end{aligned}$$

with $\|\mathbf{F}\|_K$ denoting the matrix norm of $\mathbf{F}_K$ associated with the Euclidean vector norm.

Proof. Without loss of generality we can assume that $u \in C^\infty(\overline{K})$. Let $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $|\boldsymbol{\alpha}| = m$, $m \in \mathbb{N}_0$. Remember that the Gateaux derivative in direction $\boldsymbol{\delta}$ is

$$D\widehat{u}(\boldsymbol{\xi})(\boldsymbol{\delta}) := \lim_{\varepsilon\to 0} \frac{\widehat{u}(\boldsymbol{\xi}+\varepsilon\boldsymbol{\delta})}{\varepsilon}.$$

So, the m-th Gateaux-derivative $D^m : \mathbb{R}^d \times \cdots \times \mathbb{R}^d \mapsto \mathbb{R}$ allows to express

$$\partial^{\boldsymbol{\alpha}}\widehat{u} = D^m\widehat{u}(\boldsymbol{\xi})(\boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^m)\ ,$$

with $\boldsymbol{\delta}^1 = \cdots = \boldsymbol{\delta}^{\alpha_1} = \mathbf{e}_1$, $\boldsymbol{\delta}^{\alpha_1+1} = \cdots = \boldsymbol{\delta}^{\alpha_1+\alpha_2} = \mathbf{e}_2$, etc. Here, $\mathbf{e}_k$ designates the k-th unit vector in $\mathbb{R}^d$. This means for example

$$\partial_{\xi_1}\partial^2_{\xi_2}\widehat{u}(\boldsymbol{\xi}) = D^3\widehat{u}(\boldsymbol{\xi})(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_2).$$

So, we deduce that

$$|\partial^{\boldsymbol{\alpha}}\widehat{u}(\boldsymbol{\xi})| \leq \|D^m\widehat{u}(\boldsymbol{\xi})\| := \sup\{D^m\widehat{u}(\boldsymbol{\xi})(\boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^m),\, \boldsymbol{\delta}^k \in \mathbb{R}^d, \left|\boldsymbol{\delta}^k\right| = 1\}\,,$$

(instead of the particular choice of $\boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^m$ we take the supremum) which implies

$$|\widehat{u}|^2_{H^m(\widehat{K})} = \sum_{|\boldsymbol{\alpha}|=m} \int_{\widehat{K}} |\partial^{\boldsymbol{\alpha}}\widehat{u}(\boldsymbol{\xi})|^2 \,\mathrm{d}\boldsymbol{\xi} \leq \binom{m+d-1}{m} \int_{\widehat{K}} \|D^m\widehat{u}(\boldsymbol{\xi})\|^2 \,\mathrm{d}\boldsymbol{\xi}\,. \tag{5.4}$$

The chain rule gives

$$D^m\widehat{u}(\boldsymbol{\xi})(\boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^m) = D^m u(\boldsymbol{\Phi}_K(\boldsymbol{\xi}))(\mathbf{F}_K\boldsymbol{\delta}_1, \ldots, \mathbf{F}_K\boldsymbol{\delta}_m)\,. \tag{5.5}$$

By linearity of the derivative we have

$$\|D^m\widehat{u}(\boldsymbol{\xi})\| \leq \|\mathbf{F}_K\|^m \, \|D^m u(\boldsymbol{\Phi}(\boldsymbol{\xi}))\|\,.$$

Next, we use the transformation formula for multidimensional integrals and apply it to (5.4):

$$\begin{aligned} |\widehat{u}|^2_{H^m(\widehat{K})} &\leq \binom{m+d-1}{m} \int_K \|\mathbf{F}_K\|^{2m} \|D^m u(\mathbf{x})\|^2 \,|\det(\mathbf{F}_K)|^{-1} \,\mathrm{d}\mathbf{x} \\ &\leq \binom{m+d-1}{m} \|\mathbf{F}_K\|^{2m} \,|\det(\mathbf{F}_K)|^{-1} \int_K \|D^m u(\mathbf{x})\|^2 \,\mathrm{d}\mathbf{x}\,. \end{aligned} \tag{5.6}$$

Then, observe that

$$\begin{aligned} \|D^m u(\mathbf{x})\| &= \sup\{D^m u(\mathbf{x})(\boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^m),\, \boldsymbol{\delta}^k \in \mathbb{R}^d, \left|\boldsymbol{\delta}^k\right| = 1\} \\ &\leq \sup\{\sum_{\alpha_1=1}^d \cdots \sum_{\alpha_m=1}^d |D^m u(\delta^1_{\alpha_1}\mathbf{e}_{\alpha_1}, \ldots, \delta^m_{\alpha_m}\mathbf{e}_{\alpha_m})|,\, \boldsymbol{\delta}^k \in \mathbb{R}^d, \left|\boldsymbol{\delta}^k\right| = 1\} \\ &\leq \sum_{\alpha_1=1}^d \cdots \sum_{\alpha_m=1}^d |D^m u(\mathbf{e}_{\alpha_1}, \ldots, \mathbf{e}_{\alpha_m})| \;\leq\; d^m \;\max\{|\partial^{\boldsymbol{\alpha}} u(\mathbf{x})|, \end{aligned}$$

Finally,

$$\int_K \|D^m u(\mathbf{x})\|^2 \mathrm{d}\mathbf{x} \leq d^m \int_K \max\{|\partial^{\boldsymbol{\alpha}} u(\mathbf{x})|^2 \mathrm{d}\mathbf{x} \leq d^{2m} \sum_{|\alpha|=m} \int_K |\partial^{\boldsymbol{\alpha}} u(\mathbf{x})|^2 \mathrm{d}\mathbf{x} = d^{2m}|u|^2_{H^m(K)}.$$

□

Remark 5.5. *If $\boldsymbol{\Phi}$ is a general C^∞-diffeomorphism $\widehat{K} \mapsto K$, then the analogue of* (5.5) *will involve derivatives of u from Du up to $D^m u$ and derivatives $D\boldsymbol{\Phi}_K$ up to $D^m\boldsymbol{\Phi}_K$. Thus, in order to estimate the Sobolev-seminorm $|\widehat{u}|_{H^m(\widehat{K})}$, we have to resort to the full Sobolev norm $\|u\|_{H^m(K)}$ and vice versa.*

Estimate on a cell K With the transformation techniques we can relate the interpolation error on K with the that of the push-backed function

$$\begin{aligned}\left\|u-\mathsf{I}_n\,u\right\|^2_{H^r(K)} &= \sum_{\ell=0}^{r}|u-\mathsf{I}_n\,u|^2_{H^\ell(K)}\\ &\le \sum_{\ell=0}^{r}\binom{\ell+d-1}{d}d^{2\ell}\left\|\mathbf{F}_K^{-1}\right\|^{2\ell}|\det(\mathbf{F}_K)|\,|\widehat{u}-\underbrace{\widehat{\mathsf{I}_n\,u}}_{\widehat{I}\widehat{u}}|^2_{H^\ell(\widehat{K})}\\ &\le C(r)\binom{r+d-1}{d}d^{2r}\left\|\mathbf{F}_K^{-1}\right\|^{2r}|\det(\mathbf{F}_K)|\left\|\widehat{u}-\widehat{I}\widehat{u}\right\|^2_{H^r(\widehat{K})}. \end{aligned} \tag{5.7}$$

Using the estimate on the reference element we get

$$\left\|u-\mathsf{I}_n\,u\right\|_{H^r(K)} \le C(m)\binom{r+d-1}{d}^{1/2}d^r\left\|\mathbf{F}_K^{-1}\right\|^r|\det(\mathbf{F}_K)|^{1/2}|\widehat{u}|_{H^t(\widehat{K})},$$

and transformed back to K

$$\left\|u-\mathsf{I}_n\,u\right\|_{H^r(K)} \le \underbrace{C(m)\binom{r+d-1}{d}^{1/2}\binom{t+d-1}{d}^{1/2}d^{r+t}}_{C(m,d)}\left\|\mathbf{F}_K^{-1}\right\|^r\left\|\mathbf{F}_K\right\|^t|u|_{H^t(K)}. \tag{5.8}$$

The estimate depends on the size and shape of the triangle through $\mathbf{F}_K$. We will see in the following that $\|\mathbf{F}_K\| = O(h_K)$ and $\|\mathbf{F}_K^{-1}\| = O(h_K^{-1})$, where h_K is the cell diameter.

▷ 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 $t=1$).

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 $\widehat{K}$ and find affine mappings $\mathbf{\Phi}_K : \widehat{K} \mapsto K$, $\mathbf{\Phi}_K(\boldsymbol{\xi}) := \mathbf{F}_K\boldsymbol{\xi} + \boldsymbol{\tau}_K$ for each $K \in \mathcal{M}$. In light of the general strategy outlined above, we have to establish bounds for $\|\mathbf{F}\|_K$, $\left\|\mathbf{F}_K^{-1}\right\|$, $|\det(\mathbf{F}_K)|$, and $|\det(\mathbf{F}_K)|^{-1}$ 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***

$$h_K := \sup\{|\mathbf{x}-\mathbf{y}|\,,\ \mathbf{x},\mathbf{y}\in K\}\,,$$

and the maximum radius of an inscribed ball

$$r_K := \sup\{r>0:\ \exists\mathbf{x}\in K:\ |\mathbf{x}-\mathbf{y}|<r\ \Rightarrow\ \mathbf{y}\in K\}\,.$$

The ratio h_K/r_K is called the ***shape regularity measure*** *ρ_K of K.*

Lemma 5.7. *If $\widehat{K}, K \subset \mathbb{R}^d$, $d=2,3$, are a generic non-degenerate simplices and $\mathbf{\Phi}_K : \widehat{K} \mapsto K$, $\mathbf{\Phi}_K(\boldsymbol{\xi}) := \mathbf{F}_K\boldsymbol{\xi} + \boldsymbol{\tau}$, the associated bijective affine mapping, then*

$$\left(\frac{h_K}{h_{\widehat{K}}}\right)^d \rho_K^{1-d} = \frac{h_K r_K^{d-1}}{h_{\widehat{K}}^d} \le |\det(\mathbf{F}_K)| = \frac{|K|}{|\widehat{K}|} \le \frac{h_K^d}{h_{\widehat{K}} r_{\widehat{K}}^{d-1}} = \left(\frac{h_K}{h_{\widehat{K}}}\right)^d \rho_{\widehat{K}}^{d-1}\,, \tag{5.9}$$

$$\|\mathbf{F}_K\| \le \frac{h_K}{2r_{\widehat{K}}} = \tfrac{1}{2}\rho_{\widehat{K}}\frac{h_K}{h_{\widehat{K}}} \quad , \quad \left\|\mathbf{F}_K^{-1}\right\| \le \frac{h_{\widehat{K}}}{2r_K} = \tfrac{1}{2}\rho_K\frac{h_{\widehat{K}}}{h_K}\,. \tag{5.10}$$

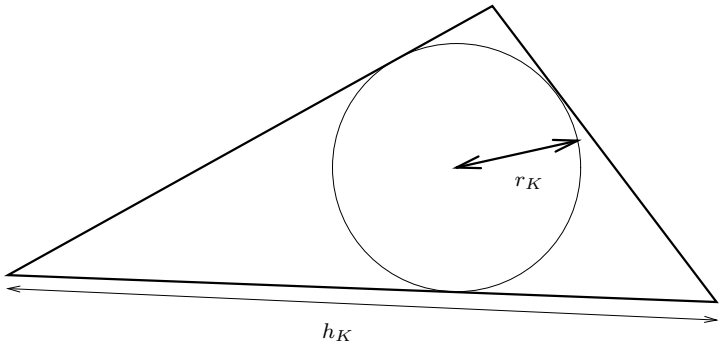


Figure 5.1: Diameter h_K and r_K for a triangular cell

Proof. The inequalities (5.9) can be concluded from the volume formula for simplices by elementary geometric considerations.

Write $\boldsymbol{\zeta} \in \widehat{K}$ for the center of the largest inscribed ball of $\widehat{K}$. Then estimates (5.10) follow from

$$\begin{aligned}\left\|\mathbf{F}_K\right\| &= \sup\{|\mathbf{F}_K\boldsymbol{\xi}|,\ |\boldsymbol{\xi}|=1\} = \tfrac{1}{2}r_{\widehat{K}}^{-1}\sup\{|\mathbf{F}_K(\boldsymbol{\xi}-\boldsymbol{\zeta})|,\ |\boldsymbol{\xi}-\boldsymbol{\zeta}|=2r_{\widehat{K}}\} \\ &= \tfrac{1}{2}r_{\widehat{K}}^{-1}\sup\{|\boldsymbol{\Phi}_K(\boldsymbol{\xi})-\boldsymbol{\Phi}_K(\boldsymbol{\zeta})|,\ |\boldsymbol{\xi}-\boldsymbol{\zeta}|=2r_{\widehat{K}}\} \le h_K/2r_{\widehat{K}}\ ,\end{aligned}$$

because both $\boldsymbol{\Phi}(\boldsymbol{\xi})$ and $\boldsymbol{\Phi}(\boldsymbol{\zeta})$ lie inside K. A role reversal of $\widehat{K}$ and K establishes the other estimate. □

The shape regularity measure of a simplex can be calculated from bounds for the smallest and largest angles enclosed by edge/face normals. We give the result for two dimensions:

Lemma 5.8. *If the smallest angle of a triangle K is bounded from below by $\alpha > 0$, then*

$$\sin(\alpha/2)^{-1} \le \rho_K \le 2\sin(\alpha/2)^{-1}\ .$$

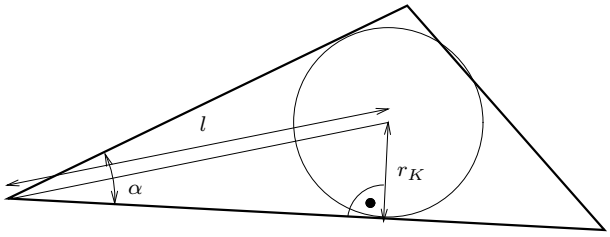


Figure 5.2: Angle condition for shape regularity of a triangle

Proof. It is immediate from Fig. 5.2 that

$$1/2 h_K \sin(\alpha/2) \le l\sin(\alpha/2) = r_K \le h_K \sin(\alpha/2)\ .$$

□

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*

$$\mu_{\mathcal{M}} := \max\{h_K/h_{K'},\ K, K' \in \mathcal{M}\}\ = \max\{h_K,\ K \in \mathcal{M}\} \cdot \max\{h_K^{-1},\ K \in \mathcal{M}\}.$$

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 π. Hence, it is not unreasonable to assume good shape regularity of simplicial meshes that are used for finite element computations.*

The choice of reference simplices is arbitrary. So we may just opt for

$$\widehat{K} := \text{convex}\left\{\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}\right\} \quad \text{for } d = 2\ , \tag{5.11}$$

$$\widehat{K} := \text{convex}\left\{\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}0\\1\\0\end{pmatrix}, \begin{pmatrix}0\\0\\1\end{pmatrix}\right\} \quad \text{for } d = 3\ . \tag{5.12}$$

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 $\mathbf{\Phi}_K : \widehat{K} \mapsto K$, $\mathbf{\Phi}_K(\boldsymbol{\xi}) := \mathbf{F}_K\boldsymbol{\xi} + \boldsymbol{\tau}_K$, $K \in \mathcal{M}$, satisfy*

$$\frac{\rho_{\mathcal{M}}^{1-d}}{\mu_{\mathcal{M}}^d}\, h_{\mathcal{M}}^d \le |\det(\mathbf{F}_K)| \le h_{\mathcal{M}}^d\ , \quad \|\mathbf{F}_K\| \le h_{\mathcal{M}}\ , \quad \left\|\mathbf{F}_K^{-1}\right\| \le \rho_{\mathcal{M}}\mu_{\mathcal{M}}\, h_{\mathcal{M}}^{-1}\ .$$

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 $\mathcal{S}_m(\mathcal{M})$ on a simplicial mesh $\mathcal{M}$. Then, for $2 \le t \le m+1$, $0 \le r \le t$*

$$\exists \gamma = \gamma(t, r, m, \rho_{\mathcal{M}}, \mu_{\mathcal{M}}) : \quad \|u - \mathsf{I}_n\, u\|_{H^r(\Omega)} \le \gamma\, h_{\mathcal{M}}^{t-r}\, |u|_{H^t(\Omega)} \quad \forall u \in H^t(\Omega)\ .$$

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 (3.21), (5.1) and Theorem 5.12 we end end up with

$$\|u - u_n\|_{H^1(\Omega)} \le \gamma\, h_{\mathcal{M}}^{t-1}\, |u|_{H^t(\Omega)} \quad \text{for } 2 \le t \le m+1 \ \text{ and } \ u \in H^t(\Omega)\ , \tag{5.13}$$

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 Ω we can expect $u \in H^{r+2}(\Omega)$.

   Example 5.13. *For $d = 1$ we have $u \in H^{r+2}(\Omega)$, if $f \in H^r(\Omega)$.*

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 Ω. They are assumed to be *uniformly* shape-regular, that is,

$$\exists \gamma > 0 : \quad \rho_{\mathcal{M}_n} < \gamma \quad \forall n \in \mathbb{N} .$$

   Moreover, the meshes are to become infinitely fine

$$h_{\mathcal{M}_n} \to 0 \quad \text{as} \quad n \to \infty .$$

   Then the statement of (5.13) can be expressed by

$$\|u - u_n\|_{H^1(\Omega)} = O(h_{\mathcal{M}_n}^{t-1}) \quad \text{for } n \to \infty . \tag{5.14}$$

   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*

$$\inf_{v_n \in \mathcal{S}_m(\mathcal{M})} \|u - v_n\|_{H^1(\Omega)} \leq \gamma(\rho_{\mathcal{M}}, \mu_{\mathcal{M}}) \left(\frac{h_{\mathcal{M}}}{m}\right)^{\min\{m+1,t\}-1} \|u\|_{H^t(\Omega)} . \tag{5.15}$$

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

$$\mathsf{b}(u, v) := \int_\Omega \langle \mathbf{A}\,\mathbf{grad}\, u, \mathbf{grad}\, v \rangle \, \mathrm{d}\mathbf{x} = \int_\Omega f v \, \mathrm{d}\mathbf{x} \quad \forall v \in H_0^1(\Omega) . \tag{5.16}$$

and its Galerkin discretization based on the finite element space $\mathcal{S}_m(\mathcal{M})$ on a simplicial mesh $\mathcal{M}$. Now, we aim to establish an estimate of the discretization error in the $L^2(\Omega)$-norm.

This is beyond the scope of the theory presented in 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 $u \in H_0^1(\Omega)$ with $-\operatorname{div}(\mathbf{A}\,\mathbf{grad}\, u) \in L^2(\Omega)$ satisfy*

$$u \in H^2(\Omega) \quad \textit{and} \quad \|u\|_{H^2(\Omega)} \leq \gamma \, \|\operatorname{div}(\mathbf{A}\,\mathbf{grad}\, u)\|_{L^2(\Omega)} ,$$

with a constant $\gamma = \gamma(\mathbf{A}, \Omega) > 0$ independent of u.

We write u_n for the unique solution of the discrete variational problem

$$u_n \in \mathcal{S}_m(\mathcal{M}) \cap H^1_0(\Omega): \quad \mathsf{b}(u_n, v_n) = \int_\Omega f v_n \,\mathrm{d}\mathbf{x} \quad \forall v_n \in \mathcal{S}_m(\mathcal{M}) \,.$$

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* (3.19)

$$\mathsf{b}(e_h, v_n) = 0 \quad \forall v_n \in \mathcal{S}_m(\mathcal{M}) \,.$$

The solution $w \in H^1_0(\Omega)$ of the *dual linear variational problem*

$$w \in H^1_0(\Omega): \quad \mathsf{b}(w, v) = \int_\Omega e_h v \,\mathrm{d}\mathbf{x} \quad \forall v \in H^1_0(\Omega) \,, \tag{5.17}$$

will be a solution of the the elliptic boundary value problem

$$-\operatorname{div}(\mathbf{A}\,\mathbf{grad}\, w) = e_h \text{ in } \Omega \quad , \quad w = 0 \text{ on } \Gamma \,.$$

Since $e_h \in L^2(\Omega)$, by Assumption 5.15 we know

$$w \in H^2(\Omega) \quad , \quad \|w\|_{H^2(\Omega)} \le \gamma \, \|e_h\|_{L^2(\Omega)} \,, \tag{5.18}$$

with $\gamma = \gamma(\Omega, \mathbf{A}) > 0$.

Next, we plug $v = e_h$ into (5.17) and arrive at

$$\|e_h\|^2_{L^2(\Omega)} = \mathsf{b}(w, e_h) = \inf_{v_n \in \mathcal{S}_m(\mathcal{M})} \mathsf{b}(w - v_n, e_h) \,,$$

where Galerkin orthogonality came into play. We may now plug in $v_n := \mathsf{I}_n w$, where I_n is the finite element interpolation operator for $\mathcal{S}_m(\mathcal{M})$. Then we can use the the continuity of b in $H^1(\Omega)$ and the interpolation error estimate of Thm. 5.12 for $r = 1$ and $t = 2$:

$$\|e_h\|^2_{L^2(\Omega)} \le \mathsf{b}(w - \mathsf{I}_n w, e_h) \le \gamma \, \|w - \mathsf{I}_n w\|_{H^1(\Omega)} \cdot \|e_h\|_{H^1(\Omega)} \le \gamma \, h_{\mathcal{M}} \, |w|_{H^2(\Omega)} \cdot \|e_h\|_{H^1(\Omega)} \,.$$

Here, the final constant γ will depend on $\mathbf{A}$, m, $\rho_{\mathcal{M}}$, and $\mu_{\mathcal{M}}$, but not on u or u_n. Eventually, we resort to the 2-regularity in the form of estimate (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*

$$\|u - u_n\|_{L^2(\Omega)} \le \gamma \, h_{\mathcal{M}} \, \|u - u_n\|_{H^1(\Omega)} \,,$$

where the constant $\gamma > 0$ depends on $\Omega, \mathbf{A}, m, \rho_{\mathcal{M}}, \mu_{\mathcal{M}}$.

Remark 5.17. *Thm. 5.16 tells us that under suitable assumptions in the h-version of finite elements we can gain another power of $h_{\mathcal{M}}$ when measuring the discretization error in the $L^2(\Omega)$-norm. More generally, often we can expect that, sloppily speaking,*

> *the weaker the norm of the discretization error that we consider the faster it will converge to zero as $h_{\mathcal{M}} \to 0$.*

What remains to be settled is whether Assumption 5.15 is reasonable. This is part of **elliptic regularity theory**. In particular, we have the following result [25]

Theorem 5.18. *If the computational domain $\Omega \subset \mathbb{R}^d$ is convex or has C^1-boundary and $\mathbf{A} \in C^1(\overline{\Omega})$, then the elliptic boundary value problem 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 (3.5) or (**??**) 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

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

with bilinear form $\mathsf{b} \in L(V \times V, \mathbb{R})$ 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. 3.18.

Based on $V_n \subset V$, $\dim(V_n) < \infty$, we arrive at the discrete variational problem (DVP), see Sect. 3.6.

$$u_n \in V_n : \quad \mathsf{b}(u_n, v_n) = \langle f, v_n\rangle_{V'\times V} \quad \forall v_n \in V_n \ .$$

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

$$\widetilde{u}_n \in V_n : \quad \widetilde{\mathsf{b}}(\widetilde{u}_n, v_n) = \langle \widetilde{f}, v_n\rangle_{V'\times V} \quad \forall v_n \in V_n \ , \tag{5.19}$$

with a bilinear form $\widetilde{\mathsf{b}} \in L(V_n \times V_n, \mathbb{R})$ and $\widetilde{f} \in V_n'$. The perturbation destroys Galerkin orthogonality and leads to extra terms in the discretization error estimate of Cor. 3.32.

Theorem 5.19 (First Strang's lemma)**.** *Beside the assumptions on* b *and* $\widetilde{\mathsf{b}}$ *stated above we demand that* $\widetilde{\mathsf{b}}$ *satisfies* (DIS) *with constant* γ_n*. Then,* (5.19) *will have a unique solution* $\widetilde{u}_n \in W_n$*, which satisfies the a-priori error estimate*

$$\begin{aligned}\|u - \widetilde{u}_n\|_V \ \le \ \gamma \Big(\inf_{w_n\in W_n} \big(\|u - w_n\|_V + \sup_{v_n\in V_n} \frac{|\mathsf{b}(w_n,v_n) - \widetilde{\mathsf{b}}(w_n,v_n)|}{\|v_n\|_V}\big) \\ + \sup_{v_n\in V_n} \frac{|\langle f, v_n\rangle_{V\times V'} - \langle \widetilde{f}, v_n\rangle_{V\times V'}|}{\|v_n\|_V}\Big) \ ,\end{aligned}$$

with $\gamma = \gamma(\|\mathsf{b}\|, \gamma_n) > 0$

Proof. Similarly to the proof of quasi-optimality in Theorem 3.29 we use triangle inequality and (DIS) to estimate

$$\begin{aligned}\|u-\widetilde{u}_n\|_V &\le \|u-w_n\|_V + \|w_n-\widetilde{u}_n\|_V \\ &\le \|u-w_n\|_V + \frac{1}{\gamma_n}\sup_{v_n\in V_n\setminus\{0\}}\frac{\left|\widetilde{\mathsf{b}}(w_n-\widetilde{u}_n,v_n)\right|}{\|v_n\|_V}.\end{aligned}$$

With

$$\begin{aligned}\widetilde{\mathsf{b}}(w_n-\widetilde{u}_n,v_n) &= (\mathsf{b}(u,v_n)-\widetilde{\mathsf{b}}(\widetilde{u}_n,v_n)) + (\widetilde{\mathsf{b}}(w_n,v_n)-\mathsf{b}(w_n,v_n)) + \mathsf{b}(w_n-u,v_n) \\ &= (\langle f,v_n\rangle_{V\times V'} - \langle \widetilde{f},v_n\rangle_{V\times V'}) + (\widetilde{\mathsf{b}}(w_n,v_n)-\mathsf{b}(w_n,w_n)) + \mathsf{b}(w_n-u,v_n),\end{aligned}$$

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

$$\sup_{v_n\in V_n}\frac{|\mathsf{b}(w_n,v_n)-\widetilde{\mathsf{b}}(w_n,v_n)|}{\|v_n\|_V}\,, \qquad \sup_{v_n\in V_n}\frac{|\langle f,v_n\rangle_{V\times V'}-\langle \widetilde{f},v_n\rangle_{V\times V'}|}{\|v_n\|_V}\,,$$

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* $\widetilde{\mathsf{b}}$

$$\widetilde{\mathsf{b}}(v_n,v_n) \ge \gamma_1\|v\|_V$$

for some positive constant γ_1 *implies* (DIS). *This property for the pertubed bilinear form is called* ***h-ellipticity****. In the h-version of finite elements (mesh refinement) we want* γ_1 *to be indepedent of the meshwidth ("uniform h-ellipticity").*

Remark 5.21. *If* $\widetilde{\mathsf{b}}$ *is still an continuous bilinear form on* $V\times V$*, then the estimate of the theorem can be simplified in the following way:*

$$\begin{aligned}\|u-\widetilde{u}_n\|_V &\le \|u-w_n\|_V + \|w_n-\widetilde{u}\|_V \\ &\le \|u-w_n\|_V + \frac{1}{\gamma_n}\sup_{v_n\in V_n\setminus\{0\}}\frac{\left|\widetilde{\mathsf{b}}(w_n-u+u-\widetilde{u}_n,v_n)\right|}{\|v_n\|_V}, \\ &\le \left(1+\frac{\|\widetilde{\mathsf{b}}\|}{\gamma_n}\right)\|u-w_n\|_V + |R(u)|\end{aligned}$$

with the residual *term*

$$R(u) = \sup_{v_n\in V_n}\frac{\widetilde{\mathsf{b}}(u,v_n)-\widetilde{f}(v_n)}{\|v_n\|_V}\,.$$

5.7.2 Uniform h-ellipticity

Let us consider the variational problem

$$\mathsf{b}(u,v) := \int_\Omega \langle \mathbf{A}\,\mathbf{grad}\,u, \mathbf{grad}\,v\rangle\,\mathrm{d}\mathbf{x} = \int_\Omega fv\,\mathrm{d}\mathbf{x} \quad \forall v\in H^1_0(\Omega)\,. \tag{5.16}$$

discretized by means of finite elements of uniform polynomial degree m on a simplicial triangulation $\mathcal{M}$ of a polygonal/polyhedral computational domain Ω.

Applying local quadrature rules of the form

$$\int_\Omega f(\mathbf{x})\,\mathrm{d}\mathbf{x} \approx \sum_{K\in\mathcal{M}} |K| \sum_{l=1}^{P_K} \omega_l^K\, f(\boldsymbol{\pi}_l^K)\;, \tag{NUQ}$$

the perturbed bilinear form for $u_n, v_n \in \mathcal{S}_{m,0}(\mathcal{M})$ reads

$$\widetilde{\mathsf{b}}(u_n, v_n) := \sum_{K\in\mathcal{M}} |K| \sum_{l=1}^{P_K} \omega_l^K \left\langle \mathbf{A}(\boldsymbol{\pi}_l^K)\,\mathbf{grad}\, u_n(\boldsymbol{\pi}_l^K), \mathbf{grad}\, v_n(\boldsymbol{\pi}_l^K)\right\rangle\;. \tag{5.20}$$

For the analysis we must rely on a certain smoothness of the coefficient function $\mathbf{A}$:

Assumption 5.22. *The restriction of the coefficient function $\mathbf{A}:\Omega\mapsto\mathbb{R}^{d,d}$ to any cell $K\in\mathcal{M}$ belongs to $C^m(K)^{d,d}$ and can be extended to a function $\in C^m(\overline{K})^{d,d}$.*

Lemma 5.23. *Let* $\mathbf{A}$ *satisfy*

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

with $\underline{\gamma}>0$ and Assumption 5.22, and let the local quadrature weights ω_l^K be positive. If the local quadrature rules are exact for polynomials up to degree $2m-2$, then

$$\widetilde{\mathsf{b}}(v_n, v_n) \ge \underline{\gamma}\,|v_n|^2_{H^1(\Omega)} \quad \forall v_n\in\mathcal{S}_m(\mathcal{M})\;.$$

Proof. Since $\mathbf{A}$ is uniformly positive definite and the quadrature weights are positive

$$\begin{aligned}\widetilde{\mathsf{b}}(v_n, v_n) &\ge \sum_{K\in\mathcal{M}} |K| \sum_{l=1}^{P_K} \omega_l^K \underline{\gamma}|\,\mathbf{grad}\, v_n(\boldsymbol{\pi}_l^K)|^2 \ge \underline{\gamma} \sum_{K\in\mathcal{M}} |K| \sum_{l=1}^{P_K} \omega_l^K |\,\mathbf{grad}\, v_n(\boldsymbol{\pi}_l^K)|^2\\ &= \underline{\gamma}\,|v_n|^2_{H^1(\Omega)}\;,\end{aligned}$$

because on each $K\in\mathcal{M}$ we know $\mathbf{grad}\, v_n \in \mathcal{P}_{m-1}(K)^d$ so that the numerical quadrature of $|\,\mathbf{grad}\, v_n|^2$ is exact. □

References

[22] P.G. Ciarlet. *The Finite Element Method for Elliptic Problems*, volume 4 of *Studies in Mathematics and its Applications*. North-Holland, Amsterdam, 1978.

[23] K.T. Smith. Inequalities for formally positive integro-differential forms. *Bull. Am. Math. Soc.*, 67:368–370, 1961.

[24] J. Wloka. *Partial differential equations*. Cambridge University Press, Cambridge, UK, 1987.

[25] P. Grisvard. *Elliptic Problems in Nonsmooth Domains*. Pitman, Boston, 1985.

6 Adaptive Finite Elements

In this chapter we only consider the primal variational formulation of a second order elliptic boundary value problem

$$-\Delta u = f \quad \text{in } \Omega \quad , \quad u = 0 \quad \text{or} \quad \langle \mathbf{grad}\, u, \mathbf{n} \rangle = 0 \quad \text{on } \Gamma \, . \tag{6.1}$$

in a bounded polygon $\Omega \subset \mathbb{R}^2$ with Lipschitz boundary Γ.

Definition 6.1. *A Galerkin discretization of a variational problem is called* ***adaptive****, if it employs a trial space V_n that is based on non-uniform meshes or non-uniform polynomial degree of the finite elements. We distinguish*

- ***a priori*** *adapted finite element spaces, which aim to take into account known features of the exact solution.*
- ***a posteriori*** *adapted finite element spaces, whose construction relies on the data of the problem.*

The next example shows that a posteriori adaptivity can dramatically enhance accuracy:

Example 6.2. *If we knew the continuous solution $u \in V$ of the linear varitional problem* (LVP)*, we could just choose $V_n := \operatorname{span}\{u\}$ and would end up with a perfect Galerkin discretization.*

Three basic policies can be employed to achieve a good fit of the finite element space and the continuous solution:

- adjusting of the mesh $\mathcal{M}$ while keeping the type of finite elements (**h-adaptivity**).
- adjusting the local trial spaces (usually by raising/lowering the local polynomial degree) while retaining a single mesh (**p-adaptivity**).
- combining both of the above approaches (**hp-adaptivity**).

6.1 Regularity of solutions of second-order elliptic boundary value problems

If the geometry does not interfere, the solution of (6.1) is as smooth as the data f permit:

Theorem 6.3. *If $\partial\Omega$ is smooth (i. e. $\partial\Omega$ has a parameter representation with C^∞ functions), then for the solution u of (6.1) it holds*

$$f \in H^k(\Omega) \Longrightarrow u \in H^{k+2}(\Omega) \quad \textit{for } k \in \mathbb{N}_0 \, ,$$

and

$$\forall k \in \mathbb{N}_0,\ \exists \gamma = \gamma(\Omega, k) : \quad \|u\|_{H^{k+2}(\Omega)} \leq \gamma(\Omega, k)\ \|f\|_{H^k(\Omega)} \quad \forall f \in H^k(\Omega) \, .$$

Similar results hold for Neumann boundary conditions on the whole of $\partial\Omega$.

If $\partial\Omega$ has corners (as in the case of a polygonal domain), the results from the previous section do not hold any longer. The solution gets **singular** meaning that some (higher) derivative is not square integrable.

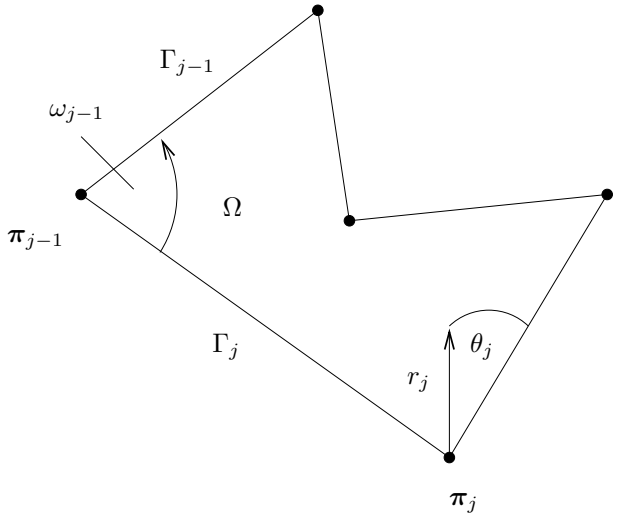


Figure 6.1: Polygon Ω and notation for the corners.

Theorem 6.4. *Let $\Omega \subset \mathbb{R}^2$ be a polygon with J corners $\boldsymbol{\pi}_j$. Denote the polar coordinates in the corner $\boldsymbol{\pi}_j$ by (r_j, θ_j) and the inner angle at the corner $\boldsymbol{\pi}_j$ by w_j as in Figure 6.1. Additionally, let $f \in H^{-1+s}(\Omega)$ with $s \geq 1$ integer[1] and $s \neq \lambda_{jk}$, where the λ_{jk} are given by the **singular exponents***

$$\lambda_{jk} = \frac{k\pi}{\omega_j} \quad \text{for } k \in \mathbb{N}\,. \tag{6.2}$$

*Then, we have the following decomposition [26] of the solution $u \in H_0^1(\Omega)$ of the **Dirichlet problem** (6.1) into a regular part (i. e. with the regularity one would expect from a smooth boundary according to Thm. 6.3) and finitely many so-called **singular functions** $s_{jk}(r, \theta)$:*

$$u = u^0 + \sum_{j=1}^{J} \psi(r_j) \sum_{\lambda_{jk}<s} \alpha_{jk} s_{jk}(r_j, \theta_j). \tag{6.3}$$

Here, $u^0 \in H^{1+s}(\Omega)$ and ψ is a C^∞ cut off function ($\psi \equiv 1$ in a neighborhood of 0). The singular functions s_{jk} are explicitly given [26, Sect. 4.2] by

$$\begin{aligned} &\lambda_{jk} \text{ non-integer:} && s_{jk}(r,\theta) = r^{\lambda_{jk}} \sin(\lambda_{jk}\theta), \\ &\lambda_{jk} \in \mathbb{N}, \omega \notin \{\pi, 2\pi\}: && s_{jk}(r,\theta) = r^{\lambda_{jk}} \ln r \sin(\lambda_{jk}\theta) \end{aligned}$$

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

[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.

is cured by adding the smooth function $r^{\lambda_{jk}}\theta\cos(\lambda_{jk}\theta))$ and the latter by adding the harmonic polynomial [27] $\mathcal{H}_{\lambda_{jk}}(x,y)$ of degree λ_{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 α_{jk} in* (6.3) *depend only on f and are called (generalised)* ***stress intensity factors****.*

Remark 6.6. *At first glance, the decomposition (6.3) appears to be very special and restricted to the problem (6.1). Yet, similar decompositions with suitable $s_{jk}(r,\theta)$ hold for all elliptic boundary value problems of the form*

$$-\operatorname{div}(\mathbf{A}\,\mathbf{grad}\,u) = f \quad \text{in } \Omega \quad , \quad u = 0 \quad \text{on } \Gamma_D \ , \ \langle \mathbf{A}\,\mathbf{grad}\,u, \mathbf{n}\rangle = 0 \quad \text{on } \Gamma_N \ .$$

Generally, $s_{jk}(r,\theta) = r^{\lambda_{jk}}\,\Theta_{jk}(\theta)$ is a non-trivial solution of the homogeneous differential equation in an infinite sector $\mathbf{S}$ with a tip at the singular point.

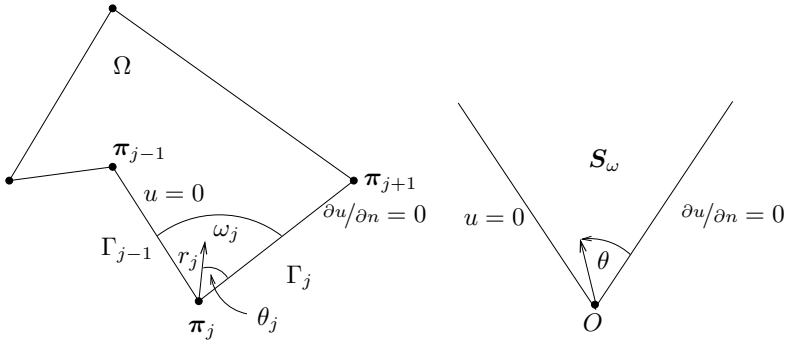


Figure 6.2: Corner $\boldsymbol{\pi}_j$ with changing boundary conditions and the infinite sector $\mathbf{S}_\omega$.

Example 6.7. *Consider $-\Delta u = f$ in Ω with mixed boundary conditions at $\boldsymbol{\pi}_j$. Let $\boldsymbol{\pi}_j \in \partial\Omega$ be a boundary point where the type of the boundary conditions changes from Dirichlet to Neumann (cf. Figure 6.2).*

In the infinite sector

$$\mathbf{S}_\omega = \{(r,\theta) : 0 < r < \infty,\ 0 < \theta < \omega\} \ ,$$

we are looking for non-trivial solutions of the homogeneous problem

$$\Delta s = 0 \text{ in } \mathbf{S}_\omega, \frac{\partial s}{\partial n}\Big|_{\theta=0} = 0, \quad s\Big|_{\theta=\omega} = 0$$

of the form $s(r,\theta) = r^\lambda\,\Theta(\theta)$. Using $s = r^\lambda\,\Theta(\theta)$, it follows in $\mathbf{S}_\omega$:

$$0 = \Delta s = r^{\lambda-2}(\Theta'' + \lambda^2\Theta) \quad \text{for } r > 0,$$

i. e. the pairs $(\lambda, \Theta(\theta))$ are ***eigenpairs of a Sturm-Liouville problem***

$$\mathcal{L}\Theta = \Theta'' + \lambda^2\Theta = 0 \text{ in } (0,\omega) \ , \quad \Theta'(0) = 0 \ , \quad \Theta(\omega) = 0.$$

One recalculates that the eigenpairs are explicitly given by

$$\lambda_k = (k - 1/2)\,\frac{\pi}{\omega} \quad , \quad \Theta_k(\theta) = \cos(\lambda_k\theta), \quad k = 1,2,3,\dots \ .$$

Note: even if $\omega = \pi$, i. e. for changing boundary conditions on a straight edge, there exists a singularity $r^{1/2}\cos(\theta/2)$ for changing boundary conditions.

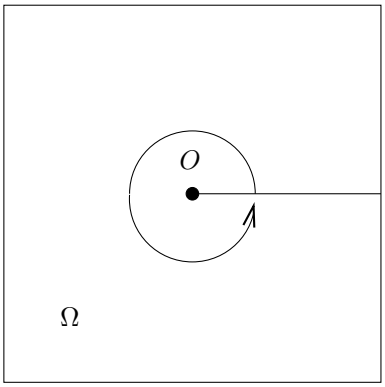


Figure 6.3: Cracked panel.

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 Ω. The singular behaviour of u is restricted to the corners $\boldsymbol{\pi}_j$.*

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 $(\partial^{\boldsymbol{\alpha}} s_{j1})(r_j,\theta_j) \sim r_j^{\lambda_{j1}-|\boldsymbol{\alpha}|}$ for $r_j \to 0$ that the derivative $\partial^{\boldsymbol{\alpha}}$ of the singular functions s_{jk} for $|\boldsymbol{\alpha}| = 2$ is not square integrable since $\lambda_{j1} - |\boldsymbol{\alpha}| < -1$, i. e. for $|\boldsymbol{\alpha}| = 2$ we have*

$$|(\partial^{\boldsymbol{\alpha}} s_{j1})(r_j,\theta_j)|^2 \sim r_j^{-2-\varepsilon} \notin L^1(\Omega).$$

The shift theorem Thm. 6.3 does no longer hold.

References. Corner and edge singularities for solutions of elliptic problems are discussed in [28, 29, 26].

6.2 Convergence of finite element solutions

Let $u_n \in \mathcal{S}_m(\mathcal{M}_n)$ stand for the Galerkin solution of (6.1) obtained by means of Lagrangian finite elements of uniform polynomial degree $m \in \mathbb{N}$ on the mesh $\mathcal{M}_n$. Temporarily, we will allow $d \in \{1,2,3\}$.

Let $\{\mathcal{M}_n\}_{n=1}^{\infty}$ denote a uniformly shape-regular and quasi-uniform family of triangulations of the polygon Ω such that $h_n := h_{\mathcal{M}_n} \to 0$ as $n \to \infty$. From Sect. 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 $\{\mathcal{M}_n\}_{n=1}^{\infty}$ of simplicial meshes. In the case of finite elements of polynomial degree m we have the crude estimates

$$N_n := \dim(\mathcal{S}_m(\mathcal{M}_n)) \leq \binom{d+m}{d} \cdot \sharp\mathcal{M}_n \quad \Rightarrow \quad \sharp\mathcal{M}_n \approx h_{\mathcal{M}_n}^{-d} ,$$

with constants depending on shape-regtularity and m. Thus, if $t \geq m+1$ we get asymptotically

$$\|u - u_n\|_{H^1(\Omega)} \leq \gamma N_n^{-m/d} . \tag{6.5}$$

The constant γ depends on $\Omega, \mathbf{A}$ and the bounds for $\rho_{\mathcal{M}_n}$, $\mu_{\mathcal{M}_n}$. This reveals an **algebraic asymptotic convergence rate** of the h-version of finite elements for second order elliptic problems.

However, even for small m the regularity $u \in H^{m+1}(\Omega)$ cannot be taken for granted. Consider $d = 2$ and remember that from Sect. 6.1 it is merely known that for $f \in H^{k-2}(\Omega)$:

$$u = u^0 + u_{\text{sing}} \tag{6.6}$$

with a smooth part $u^0 \in H^k(\Omega)$, $k \geq 2$, and with a singular part u_{sing}, which is a (finite!) sum of singular functions $s(r_i, \theta_i)$, which have the explicit form

$$s(r, \theta) = r^{\lambda}\, \Theta(\theta) , \tag{6.7}$$

with piecewise smooth Θ, where $0 < \lambda < k-1$ (we assume here that $\log r$ terms are absent). The singular functions (6.7) are only poorly approximated by finite element functions on sequences of quasi-uniform meshes. For the singular functions $s(r, \theta)$ as in (6.7) and with (r, θ) denoting polar coordinates at a vertex of Ω the (optimal) error estimate

$$\min_{v_n \in \mathcal{S}_m(\mathcal{M}_n)} \|s - v_n\|_{H^1(\Omega)} \leq \gamma h_n^{\min(m,\lambda)} \leq \gamma N_n^{-\min(m,\lambda)/2}$$

holds, where again $N_n := \dim \mathcal{S}_m(\mathcal{M}_n) = O(h_n^{-2})$ denotes the number of degrees of freedom.

For a sequence $\{\mathcal{M}_n\}_{n=1}^{\infty}$ of quasi uniform meshes one therefore observes only the suboptimal convergence rate

$$\|u - u_n\|_{H^1(\Omega)} \leq \gamma h_n^{\min(m,\lambda^*)} \leq \gamma N_n^{-\min(m,\lambda^*)/2} , \tag{6.8}$$

where $\lambda^* = \min\{\lambda_{jk} : j = 1, \dots, J,\ k = 1, 2, \dots\}$, as $h_n \to 0$ (or for $N_n \to \infty$), instead of the optimal asymptotic convergence rate (6.5) supported by the polynomial degree of the finite element space.

Since often $\lambda^* < 1$, one observes even for the simple piecewise linear ($m = 1$) elements a reduced convergence rate, and for $m > 1$ we hardly ever get the optimal asymptotic rate $O(N_n^{-m/d})$.

Remark 6.11. *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 $N_n \approx m^d h_{\mathcal{M}}^{-d}$ and, thus the estimate (see Remark 5.14)*

$$\inf_{v_n \in \mathcal{S}_m(\mathcal{M})} \|u - v_n\|_{H^1(\Omega)} \leq \gamma(\rho_{\mathcal{M}}, \mu_{\mathcal{M}}) \left(\frac{h_{\mathcal{M}}}{m}\right)^{\min\{m+1,t\}-1} \|u\|_{H^t(\Omega)} \ , \tag{5.15}$$

gives asymptotically for $t \leq m \to \infty$

$$\|u - u_n\|_{H^1(\Omega)} \leq \gamma N_n^{(t-1)/d} \ . \tag{6.9}$$

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 $\overline{\Omega}$, that is, it is C^∞ and can be expanded into a locally convergent power series in each point of Ω, then the p-version yields an* ***exponential asymptotic convergence rate***

$$\|u - u_n\|_{H^1(\Omega)} \leq \gamma \exp(-\gamma' N_n^\beta) \ , \tag{6.10}$$

with $\gamma, \gamma', \beta > 0$ only depending on problem parameters and the fixed triangulation, but independent of the polynomial degree m of the 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.

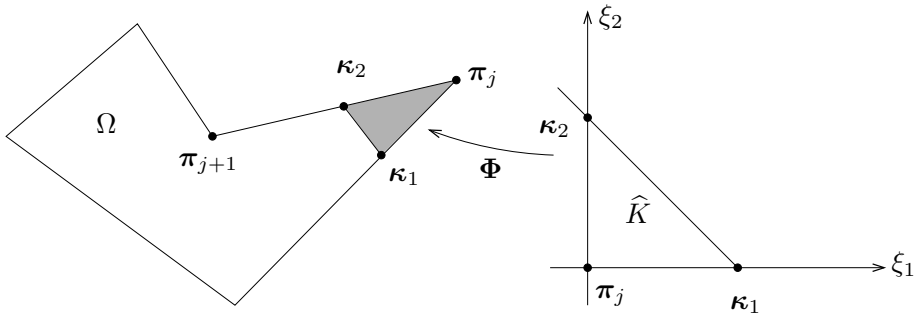


Figure 6.4: Polygon Ω with corner $\boldsymbol{\pi}_j$, subdomain $\mathrm{conv}(\boldsymbol{\pi}_j, \boldsymbol{\kappa}_1, \boldsymbol{\kappa}_2)$ adjacent to it and its representation by an affine map $\boldsymbol{\Phi}$ from the standard triangle $\widehat{K}$.

One option is **judicious** (vernünftig) **mesh refinement towards the vertices of the polygon**. Consider the polygon Ω shown in Fig. 6.4. In Ω, consider any vertex $\boldsymbol{\pi}_j$

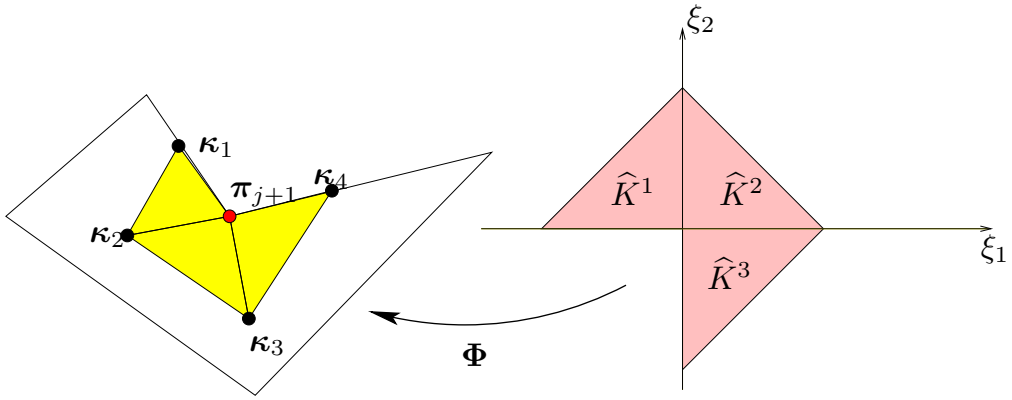


Figure 6.5: Mapping for a re-entrant corner at $\boldsymbol{\pi}_{j+1}$.

(In Fig. 6.4 we chose a convex corner, the approach to a re-entrant corner at $\boldsymbol{\pi}_{j+1}$ is indicated in Fig. 6.5). We denote again by (r, θ) polar coordinates at vertex $\boldsymbol{\pi}_j$, and by $s(r, \theta)$ a singular function as in (6.7)

$$s(r, \theta) = r^\lambda \, \Theta(\theta)$$

with a smooth $\Theta(\theta)$. The triangle $K = \mathrm{conv}(\boldsymbol{\pi}_j, \boldsymbol{\kappa}_1, \boldsymbol{\kappa}_2)$ denotes a neighbourhood of vertex $\boldsymbol{\pi}_j$ in Ω (shown shaded in Fig. 6.4). By means of an affine map $\boldsymbol{\Phi}$ the triangle K is mapped onto the reference triangle $\widehat{K}$ with polar coordinates $(\widehat{r}, \widehat{\theta})$. The singular function $s(r, \theta)$ in Ω is transformed by $\boldsymbol{\Phi}$ into

$$\widehat{s}(\widehat{r}, \widehat{\theta}) = \widehat{r}^\lambda \, \widehat{\Theta}(\widehat{\theta}) \quad \text{in } \widehat{K} \ ,$$

with the same exponent λ but with another C^∞-function $\widehat{\Theta}(\widehat{\theta})$:

Example 6.13. *Let $\boldsymbol{\pi}_j = (0, 0)$, (x_1, x_2) stand for the coordinates in Ω,*

$$x_1 = r \cos\theta \quad , \quad x_2 = r \sin\theta$$
$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \mathbf{F} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \widehat{r} \begin{pmatrix} \cos\widehat{\theta} \\ \sin\widehat{\theta} \end{pmatrix}$$

and

$$s(r, \theta) = r^\lambda \, \Theta(\cos\theta, \sin\theta)$$

denote the singular function in (6.7). To prove that $s(r, \theta)$ is, in the coordinates ξ_1, ξ_2, once again of the form (6.7) let $\mathbf{F} = (f_{ij})_{1 \le i,j \le 2}$. Then

$$\begin{aligned} r^2 = x_1^2 + x_2^2 = \ & (f_{11}\,\xi_1 + f_{12}\,\xi_2)^2 + (f_{21}\,\xi_1 + f_{22}\,\xi_2)^2 \\ = \ & \widehat{r}^2 \, \{(f_{11}\cos\widehat{\theta} + f_{12}\sin\widehat{\theta})^2 + (f_{21}\cos\widehat{\theta} + f_{22}\sin\widehat{\theta})^2\} \ , \end{aligned}$$

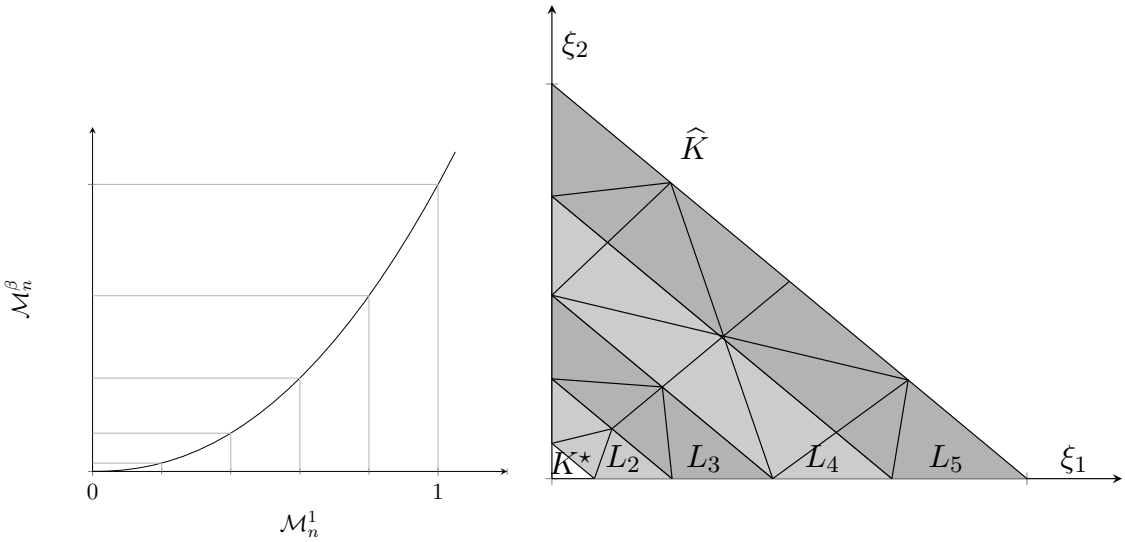
and

$$r^\lambda = \ \widehat{r}^\lambda \, \{(f_{11}\cos\widehat{\theta} + f_{12}\sin\widehat{\theta})^2 + (f_{21}\cos\widehat{\theta} + f_{22}\sin\widehat{\theta})^2\}^{\frac{\lambda}{2}} = \ \widehat{r}^\lambda \, \Theta_1(\widehat{\theta}) \ ,$$

with a smooth (analytic) function $\Theta_1(\widehat{\theta})$. Analogously, we have that $\Theta(\theta) = \widehat{\Phi}_2(\widehat{\theta})$ with a smooth function $\widehat{\Theta}_2(\widehat{\theta})$.

Due to the transformation theorem it is therefore sufficient to investigate the finite element approximation of $s(r,\theta)$ in (6.7) in the reference domain $\widehat{K}$ 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** $\mathcal{M}_n^\beta$ at the vertices of Ω the optimal asymptotic behavior $O(N_n^{-m/2})$ of the best approximation error of finite elements of uniform global degree m can be retained for singular functions as well.



(a) Graded mesh on $(0,1)$.

(b) Graded mesh on reference triangle.

Figure 6.6: Construction of a graded meshes $\mathcal{M}_n^\beta$ in $\Omega = (0,1)$ for $\beta > 1$ and on reference triangle $\widehat{K}$.

Definition 6.14. *A family $\{\mathcal{M}_n^\beta\}_{n=1}^\infty$ of meshes of a computational domain $\Omega \subset \mathbb{R}^2$ is called **algebraically graded** with respect to $\boldsymbol{\pi} \in \overline{\Omega}$ and grading factor $\beta \geq 1$ if*

(i) the meshes are uniformly shape-regular, and

(ii) with constants independent of n and $h_n := h_{\mathcal{M}_n^\beta}$,

$$\forall K \in \mathcal{M}_n^\beta,\ \boldsymbol{\pi} \notin \overline{K}: \quad h_K \approx n^{-1}\,\mathrm{dist}(\boldsymbol{\pi}, K)^{1-1/\beta}\,.$$

We will describe the concrete construction of algebraically graded meshes $\mathcal{M}_n^\beta$,$n \in \mathbb{N}$, with grading factor $\beta \geq 1$ $n \in \mathbb{N}$ on the reference domain $\widehat{K}$ with respect to the vertex $\binom{0}{0}$, see Fig. 6.6:

Algorithm 6.15 (Graded mesh on reference triangle)**.** *On $\widehat{K} = \mathrm{convex}\{\binom{0}{0}, \binom{1}{0}, \binom{0}{1}\}$ 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

$$L_j = \{\boldsymbol{\xi} \in \widehat{K} : \ \tau_{j-1}^n < \xi_1 + \xi_2 < \tau_j^n\}\,, \quad j = 1, \dots, n\,.$$

3. Equip each layer L_j, $j = 1, \dots, n$ with a simplicial triangulation $\mathcal{M}^\beta_{n|L_j}$ such that

a) their union yields a simplicial triangulation of $\widehat{K}$,

b) the shape regularity measure of $\mathcal{M}^\beta_{n|L_j}$ (see Def. 5.9) is uniformly bounded idependently of j and n,

c) for each $K \in \mathcal{M}^\beta_{n|L_j}$ we have $h_K \approx \tau_j - \tau_{j-1}$ with constants independent of j and n, and

d) $\mathcal{M}^\beta_{n|L_1}$ consists of a single triangle K^* adjacent to $\binom{0}{0}$.

Remark 6.16. *For $\beta = 1$ the meshes $\mathcal{M}^\beta_n$ 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 $\rho(\mathcal{M}^\beta_{n|L_j})$ and quasi-uniformity measure $\mu(\mathcal{M}^\beta_{n|L_j})$ we find*

$$h_K \approx \frac{\beta}{n}\left(\frac{j}{n}\right)^{\beta-1} = \frac{\beta}{n}\left(\left(\frac{j}{n}\right)^{\beta}\right)^{1-1/\beta} \qquad \forall K \in \mathcal{M}^\beta_{n|L_1}\,, \tag{6.11}$$

and

$$\sharp\mathcal{M}^\beta_{n|L_1} \approx j\,. \tag{6.12}$$

Proof. Pick $n \in \mathbb{N}$ and $j \in \{2, \dots, n\}$. 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,

$$\frac{\beta}{n}\left(\frac{j-1}{n}\right)^{\beta-1} \le \tau_j - \tau_{j-1} \le \frac{\beta}{n}\left(\frac{j}{n}\right)^{\beta-1}\,.$$

Together with $h_{K^*} = n^{-\beta}$ we conclude the first assertion of the lemma.

To confirm the second, we start with the volume formula

$$2|L_j| = \left(\frac{j}{n}\right)^{2\beta} - \left(\frac{j-1}{n}\right)^{2\beta} \approx \frac{2\beta}{n}\left(\frac{j}{n}\right)^{2\beta-1}\,. \tag{6.13}$$

As a consequence of (6.11), the area of a triangle $\subset L_j$ is

$$2|K| \approx h_K^2 \approx \frac{\beta^2}{n^2}\left(\frac{j}{n}\right)^{2\beta-2} \qquad \forall K \in \mathcal{M}^\beta_{n|L_j} \tag{6.14}$$

None of the constants depends on n and j. Dividing (6.13) by (6.14) yields (6.12). □

Corollary 6.18. *The family $\{\mathcal{M}^\beta_n\}$, $n \in \mathbb{N}$, $\beta \ge 1$, of meshes emerging from construction 6.15 is algebraically graded with respect to $\binom{0}{0}$ and grading factor β.*

Corollary 6.19. *The algebraically graded meshes $\mathcal{M}_n^\beta$, $n \in \mathbb{N}$, of $\widehat{K}$ constructed as above for $\beta \geq 1$ satisfy*

$$h_n := h_{\mathcal{M}_n^\beta} \approx {}^{\beta}\!/_{n} \quad , \quad \sharp\mathcal{M}_n^\beta \approx n^2 \, ,$$

with constants independent of n.

As a consequence, $h(\mathcal{M}_n^\beta) \to 0$ for $n \to \infty$, if $\beta \geq 1$. Moreover, for fixed $m \in \mathbb{N}$, we get from Cor. 6.19 that

$$N_n = \dim \mathcal{S}_m(\mathcal{M}_n^\beta) \leq \gamma \, \sharp\mathcal{M}_n^\beta \leq \gamma \, n^2$$

holds with a constant independent of n.

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 $u^0 \in H^{m+1}(\widehat{K})$ of the decomposition (6.6) of the solution $u \in H_0^1(\Omega)$ of $-\Delta u = f$ holds

$$\min_{v_n \in \mathcal{S}_m(\mathcal{M}_n^\beta)} \left\| u^0 - v_n \right\|_{H^1(\widehat{K})} \leq \left\| u^0 - \mathsf{I}_n \, u^0 \right\|_{H^1(\widehat{K})} \leq \gamma h_n^m \leq \gamma N_n^{-m/2} \, , \tag{6.15}$$

with $\gamma = \gamma(m, \rho_{\mathcal{M}_n^\beta})$. Here I_n is the finite element interpolation operator for $\mathcal{S}_m(\mathcal{M}_n^\beta)$.

This implies that the regular part u^0 of the solution u can also be approximated on algebraically β-graded meshes at the optimal rate (6.4), independently of the size of $\beta \geq 1$ (the size of the constant γ in the error estimates (6.15) depends of course on β and possibly grows strongly with $\beta > 1$; for fixed β and $n \to \infty$ the convergence rate (6.15) is optimal, however).

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 u_{sing} is a finite sum of terms of the form (6.7) (the treatment of terms of the form $r^\lambda |\log r| \, \Theta(\theta)$ is left to the reader as an exercise), where $\Theta(\theta) \in C^\infty([0, \omega])$ is assumed without loss of generality, and where $\lambda > 0$.

Theorem 6.20. *Let $s(r, \theta) = r^\lambda \, \Theta(\theta)$ with $\lambda > 0$ and $\Theta \in C^\infty([0, \pi/2])$ for $(r, \theta) \in \widehat{K}$ as in Fig. 6.6. Let further*

$$\beta > \max\{m/\lambda, 1\}.$$

Then there holds, as $N_n = \dim \mathcal{S}_m(\mathcal{M}_n^\beta) \to \infty$

$$\min_{v_n \in \mathcal{S}_m(\mathcal{M}_n^\beta)} |s - v_n|_{H^1(\widehat{K})} \leq \gamma(m, \lambda, \beta) \, N_n^{-\frac{m}{2}} ,$$

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 $s - \mathsf{I}_n \, s$. 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, I_n designates linear interpolation on $\mathcal{M}_n^\beta$.

Let $K^* \in \mathcal{M}_n^\beta$ denote the triangle which contains the origin $\binom{0}{0}$ in its closure. We are going to demonstrate that the contribution of this triangle to the interpolation error is negligible.

First, using polar coordinates, observe that

$$|s|^2_{H^1(K^*)} \le \int_0^{\tau_1}\int_0^{\pi/2} \left(\left(\frac{\partial s}{\partial r}\right)^2 + \left(\frac{1}{r}\frac{\partial s}{\partial \theta}\right)^2 \right) r \,\mathrm{d}\theta \,\mathrm{d}r = \int_0^{\tau_1}\int_0^{\pi/2} \left(\left(\lambda r^{\lambda-1}\Theta(\theta)\right)^2 + \left(\frac{1}{r}\, r^\lambda \Theta'(\theta)\right)^2 \right) r \,\mathrm{d}\theta \,\mathrm{d}r$$
$$\le \gamma(\Theta) \int_0^{\tau_1} \lambda^2 r^{2\lambda-1} + r^{2\lambda-1} \,\mathrm{d}r = \gamma(\Theta)\,(1+\lambda^2)\,\left(\frac{1}{n}\right)^{2\beta\lambda} .$$

Since $\mathsf{I}_n(s)$ is linear on K^*, we find

$$|\mathsf{I}_n\, s|^2_{H^1(K^*)} = |K^*|\tau_1^{-2}\left(s(\tau_1,0)^2 + s(0,\tau_1)^2\right) = {}^1\!/_2 \tau_1^{2\lambda}(\Theta(0)^2 + \Theta(\pi/2)^2)$$
$$= \gamma(\Theta)\,\tau_1^{2\lambda} = \gamma(\Theta)\,\left(\frac{1}{n}\right)^{2\beta\lambda} .$$

Thus, a simple application of the triangle inequality yields

$$|s - \mathsf{I}_n\, s|^2_{H^1(K^*)} \le \gamma(\Theta)\,\left(\frac{1}{n}\right)^{2\beta\lambda} \le \gamma N_n^{-\beta\lambda} \le \gamma N_n^{-m},$$

since $\beta > \lambda/m$.

Next, consider $\widehat{K} \setminus K^*$. For $x \in \widehat{K}$ 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

$$|s - \mathsf{I}_n\, s|^2_{H^1(\widehat{K}\setminus K^*)} = \sum_{\substack{K\in\mathcal{M}_n^\beta \\ K\neq K^*}} |s - \mathsf{I}_n\, s|^2_{H^1(K)}$$
$$\le \gamma \sum_{\substack{K\in\mathcal{M}_n^\beta \\ K\neq K^*}} h_K^2\, |s|^2_{H^2(K)} = \gamma \int_{\widehat{K}\setminus K^*} h^2\, |D^2 s|^2 \,\mathrm{d}\mathbf{x} .$$

The construction of the graded mesh implies that at any point $\boldsymbol{\xi} \in \widehat{K} \setminus K^*$ it holds

$$r = |\boldsymbol{\xi}| \ge \frac{\sqrt{2}}{2}\left(\frac{j}{n}\right)^\beta \text{ for some } j \in \mathbb{N} \qquad |h(\boldsymbol{\xi})| \le \gamma\, n^{-1}\left(\left(\frac{j}{n}\right)^\beta\right)^{1-1/\beta} \le \gamma\, n^{-1}\, r^{1-1/\beta},$$
$$|D^2 s| \le \gamma\, r^{\lambda-2} , \text{ with } \gamma > 0 \text{ independent of } n.$$

Hence, for $n \gg 1$ and $\lambda > \beta^{-1}$,

$$\int_{\widehat{K}\setminus K^\star} (h(\boldsymbol{\xi}))^2\, |D^2 s|^2 \,\mathrm{d}\boldsymbol{\xi} \le \gamma \int_{\frac{\sqrt{2}}{2}n^{-\beta}}^1 n^{-2}\, r^{2(1-1/\beta)+2\lambda-4}\, r \,\mathrm{d}r$$
$$\le \gamma\, n^{-2}\, \left[r^{2\lambda-2/\beta}\right]_{\frac{\sqrt{2}}{2}n^{-\beta}}^1 \le \gamma\, n^{-2} \le \gamma\, N_n^{-1} ,$$

using $\lambda > \beta^{-1}$ implies the assertion. For general $m \geq 1$ we get in the case of $\lambda > m/\beta$

$$\begin{aligned} |s - \mathsf{I}_n\, s|^2_{H^1(\widehat{K}\setminus K^*)} &\leq \gamma\, n^{-2m} \left| \left[r^{2\lambda - 2/\beta - 2(m-1)} \right]^1_{\frac{\sqrt{2}}{2} n^{-\beta}} \right| \\ &\leq \gamma\, (n^{-2m} + n^{-2\beta(\lambda+m-1)-2}) \leq \gamma\, (n^{-2m} + n^{-2\frac{m}{\lambda}(\lambda+m-1)-2}) \\ &\leq \gamma\, (n^{-2m} + n^{-2m-2}) \leq \gamma n^{-2m} \leq \gamma\, N_n^{-m}\ . \end{aligned}$$

□

Remark 6.21. *From Thm. 6.20 we learn that $\lambda < m$ and $\beta > m/\lambda$ will ensure the optimal rate of convergence of the finite element solution in terms of the dimension of the finite element space. If $\lambda > m$, then $u \in H^{m+1}(\widehat{K})$, and we can use the uniform mesh (for which $\beta = 1$).*

Corollary 6.22. *For $\widehat{K}$ as shown in Fig. 6.6 and $k_{\max}$ such that $\lambda_{k_{\max}} = \{\max \lambda_k : \lambda_k < m\}$, we have $u = u^0 + u_{\text{sing}}$ with $u^0 \in H^{m+1}(\widehat{K})$ and*

$$u_{\text{sing}} = \sum_{k=1}^{k_{\max}} \alpha_k\, r^{\lambda_k}\, \Phi_k(\theta)\ ,$$

where we assume that $\alpha_1 \neq 0$, and $0 < \lambda_1 \leq \lambda_2 \leq \ldots$, and $\Phi_k \in C^\infty([0,\omega])$. Then for $m \geq 1$, $\beta > \max\{1, m/\lambda_1\}$ it holds

$$\min_{v \in \mathcal{S}_m(\mathcal{M}_n^\beta)} |u_{\text{sing}} - v|_{H^1(\widehat{K})} \leq \gamma N_n^{-m/2} \quad \text{with } \gamma = \gamma(p, \alpha_k, \beta)\ . \tag{6.16}$$

Remark 6.23.

1. *If $m/\lambda_1 > 1$ we achieve with the grading factor*

$$\beta > m/\lambda_1 \tag{6.17}$$

   the same convergence rate as for smooth solutions u with a quasi uniform triangulation.

2. *For fixed $\sharp\mathcal{M}$ we have to increase β (i. e. the grading must be more pronounced) if a) m is raised and b) if the singular exponent λ_1 is reduced.*

3. *Usually the singular exponent $\lambda_1 > 0$ is unknown. Thm. 6.20 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 λ_1 is not necessary.*

4. *No refinement is required, if $\lambda_1 \geq m$. This is the case e. g. for $m = 1$ and the Laplace equation in convex polygons Ω, where $\lambda_j := \pi/\omega_j > 1$. For $m > 1$ mesh refinement near vertices is also required in convex domains, in general.*

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 $u \in H_0^1(\Omega)$ denote the solution of the homogeneous Dirichlet problem for $-\Delta u = f$. Then, for every $m \geq 1$ there exists a $\beta > 1$ such that for $k \geq m+1$ it holds

$$\inf_{v_n \in \mathcal{S}_m(\mathcal{M}_n^\beta)} |u - v_n|_{H^1(\Omega)} \leq \gamma N_n^{-m/2}\ ,$$

for $N_n = \dim \mathcal{S}_m(\mathcal{M}_n^\beta) \to \infty$.

References

[26] C. Schwab. *p- and hp-Finite Element Methods. Theory and Applications in Solid and Fluid Mechanics*. Numerical Mathematics and Scientific Computation. Clarendon Press, Oxford, 1998.

[27] M. Chamberland and D. Siegel. Polynomial solutions to dirichlet problems. *Proceedings of the american mathematical society*, 129(1):211–218, 2001.

[28] P. Grisvard. *Singularities in boundary value problems*, volume 22 of *Research Notes in Applied Mathematics*. Springer–Verlag, New York, 1992.

[29] S.A. Nazarov and B.A. Plamenevskii. *Elliptic Problems in Domains with Piecewise Smooth Boundaries*, volume 13 of *Expositions in Mathematics*. Walter de Gruyter, Berlin, 1994.