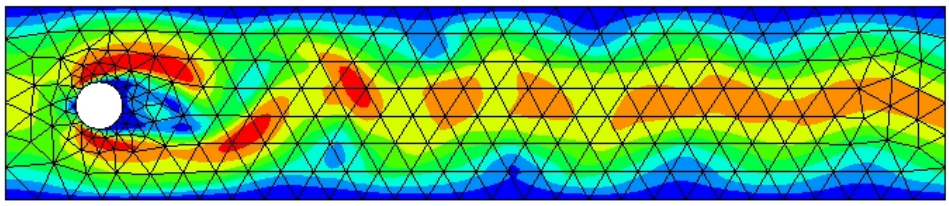
Numerics of PDEs

(summer term 2022)



Lecture Notes

Markus Faustmann, Joachim Schöberl

Institute of Analysis und Scientific Computing
TU Wien
https://www.asc.tuwien.ac.at/faustmann/
https://www.asc.tuwien.ac.at/schoeberl/

Contents

Chapter 0

Introduction

The aim of this lecture is to give an overview of common numerical methods for partial differential equations. The lecture is especially designed for the master studies Computational Science and Engineering and Interdisciplinary Mathematics.

The main topics covered are different types of finite element method. Moreover finite difference methods and combinations of both methods (e.g. for time dependent equations) are introduced.

These lecture notes are based on the assumption that the reader is familiar with concepts of higher mathematics such as

- vector spaces, integration and differentiation in more variables;
- line and surface integration and integral theorems;
- ordinary differential equations, partial differential equations.

For an overview of these topics, we refer to the lecture notes for the course *Applied Mathematics Foundations* by Markus Faustmann.

This is version 1 of the lecture notes, written during the summer term 2022 and is based on the *Numerical Methods for Partial Differential Equations* lecture notes of Joachim Schöberl.

Moreover, the lecture notes took some material from *Lecture notes on Numerical Analysis of Partial Differential Equations* by Douglas Arnold and the lecture notes on *Numerik von PDEs: stationäre Probleme* by Jens Markus Melenk.

Chapter 1

PDEs and a one dimensional model problem

1.1 Partial differential equations

Partial differential equations (PDEs) are a generalization of ordinary differential equations (ODEs) in the sense that a PDE is an equation describing the relation between a function and its derivatives, but several input arguments are allowed. Therefore, a PDE includes also partial derivatives (which explains the name PDE).
In general, we are looking at equations

$$F(u, \nabla u, \nabla^2 u, \dots) = f(x).$$

We call a PDE **linear**, if the function F only depends linearly on u and all partial derivatives of u. If $f = 0$ the PDE is called **homogeneous**, and the highest appearing derivative defines the **order** of the equation.

Studying general PDEs is very hard, but lots of physical problems are described with PDEs of second order, which means that only derivatives up to second order appear. For those, some classifications and results are known and presented in the following.

1.1.1 Elliptic, hyperbolic and parabolic PDEs

In this lecture, we mainly consider **linear PDEs of second order**, i.e., the highest appearing partial derivatives are of order 2, and F depends only linearly on u and its derivatives. To that end, let $u : \mathbb{R}^n \to \mathbb{R}$ and define

$$\sum_{i,j}^{n} a_{ij}(x) \frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{n} b_i(x) \frac{\partial u}{\partial x_i} + c(x)u(x) = f(x).$$

Here, the right-hand side f, the (symmetric) matrix valued function $A(x) = (a_{ij}(x))_{i,j=1}^n$, the vector valued function $b(x) = (b_i)_{i=1}^n$ and the (scalar) coefficient function $c(x)$ are given.

The classification of the PDE will only depend on the symmetric matrix $A(x) = (a_{ij}(x))_{ij}$, or more precisely on the eigenvalues of A. We call the PDE

- **elliptic** at the point x, if all eigenvalues of $A(x)$ fulfill $\lambda_i > 0$ for all $i = 1, \dots, n$ or all eigenvalues fulfill $\lambda_i < 0$ for all $i = 1, \dots, n$, i.e., all eigenvalues should have the same sign.
- **parabolic** at x, if there exists a zero eigenvalue $\lambda_j = 0$ of $A(x)$ and all other eigenvalues have the same sign.
- **hyperbolic** at x, if one eigenvalue of $A(x)$ has a different sign than the others. I.e., there is a $\lambda_j > 0$ and all other eigenvalues satisfy $\lambda_k < 0$ for $k \neq j$ (or the other way round).

These cases behave very differently and we note that the type of the PDE may also change for different points x. We will establish numerical results for the individual cases.
In the following, we mention the most famous PDEs of each category.

Example.

- The **Poisson equation** (elliptic): Find $u = u(x)$ (with $x \in \mathbb{R}^n$) such that

  $$-\Delta u = f$$

  for a given force f.
- The **heat equation** (parabolic): Find $u = u(x, t)$ (with $x \in \mathbb{R}^n$ and usually $t \in \mathbb{R}^+$) such that

  $$\partial_t u - \Delta u = f$$

  for a given force f.
- The **wave equation** (hyperbolic): Find $u = u(x, t)$ (with $x \in \mathbb{R}^n$ and $t \in \mathbb{R}$) such that

  $$\partial_t^2 u - \Delta u = f$$

  for a given force f.

Most of the time, PDEs are defined on a bounded region $\Omega \subset \mathbb{R}^n$ (or $\Omega \subset \mathbb{R}^{n+1}$ for time dependent problems) with $n = 1, 2, 3$. In order to obtain unique solutions, additional values have to be prescribed. This is usually done by specifying boundary conditions on $\partial\Omega$. The specific choice of the boundary condition depends on the physical model. Common types of boundary conditions are

1. **Dirichlet-conditions:** Prescribe the values at the boundary, e.g.,

   $$u = u_D(x) \qquad x \in \partial\Omega,$$

   with a given function u_D. Oftentimes, u_D is a constant, e.g., $u_D = 0$.
2. **Neumann-conditions:** Prescribe the normal flux at the boundary, e.g.,

   $$-\nabla u \cdot n = u_N \qquad x \in \partial\Omega,$$

   where n denotes the normal vector to the boundary curve/surface and u_N is a given function. Oftentimes, u_N is a constant, e.g., $u_N = 0$.

3. **Robin-conditions:** Mix the conditions from above, e.g.,

$$-\nabla u \cdot n - \alpha u = u_R \qquad x \in \partial\Omega,$$

where $\alpha \in \mathbb{R}$ is with a given constant and the function u_R is given.

Boundary conditions can also vary on parts of the boundary. To that end, we write $\partial\Omega = \Gamma_D \cup \Gamma_N \cup \Gamma_R$, with the three non-overlapping (and possibly empty) parts Γ_D (where Dirichlet conditions are imposed), Γ_N (where Neumann conditions are imposed), and Γ_R (where Robin conditions are imposed). However, exactly one boundary condition must be specified on each part of the boundary.

1.1.2 Some famous PDEs

In this subsection, we give a short description of three famous PDEs in physics.

The Schrödinger equation

The complex-valued **Schrödinger equation** describes the time evolution of the state function of a quantum mechanical system and is one of the basic equations of quantum mechanics. It reads as

$$i\hbar u_t = -H(u).$$

Here, $\hbar$ is the reduced Planck constant and H is the Hamilton operator for the system. The most famous example for H leads to the non-relativistic Schrödinger equation for the wave function of a single point particle in a potential V

$$i\hbar u_t = -\frac{\hbar}{2m}\Delta u + V(x,t)u,$$

where m is the particle mass.
Although it looks very similar to the heat equation, it is not a parabolic equation (due to the complex prefactor). In fact, its solutions can behave like waves (although it is not a hyperbolic equation either).

The Navier-Stokes equations

The Navier-Stokes equations are used to model the dynamics of a viscous fluid. Let u be the velocity of the fluid and ν the kinematic viscosity (constant). Assume that the fluid is homogeneous (constant density) and incompressible. Then, the equations

$$\begin{aligned} u_t + (u \cdot \nabla)u - \nu\Delta u &= f, \\ \operatorname{div} u &= 0, \end{aligned}$$

where u is vector-valued, are called the **incompressible Navier-Stokes equations**. Here, we are dealing with a non-linear PDE of second order.
Solution theory for the Navier-Stokes equations in 3D is one of the most famous open problems in mathematics (in 2D one can fairly easily show that the equations have a unique solution).

The Navier-Stokes equations are used in many application, such as modeling the flow in a pipe, blood in a vessel or ocean currents. Additionally, they are right now the state of the art in weather

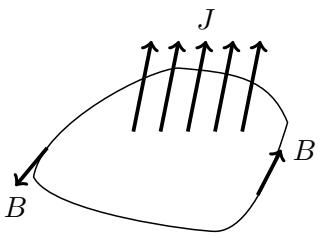
simulations, used to simulate the air flow around an object (like a wing in planes or race cars) or even used in video games to realistically simulate water flow.

Maxwell's equations

A very famous PDE with many applications in physics and electrical engineering are Maxwell's equations that describe the interaction of electric and magnetic fields.
We consider a steady current, i.e.,

$$\operatorname{curl} B - \mu_0 J = 0,$$

where B is the magnetic field density, J is the current density and μ_0 is the magnetic constant. For simplicity, we set $\mu_0 = 1$ in the following.



Using Ampere's law together with Stokes theorem gives the equation

$$\operatorname{curl} B - \mu_0 J = 0.$$

Similarly, one can use Stokes theorem on Faraday's law (relating the change of a magnetic field over time to the change of the electric field in space) to derive

$$\operatorname{curl} E = -\frac{\partial B}{\partial t}.$$

Gauß law for magnetic and electric fields additionally implies

$$\begin{aligned} \operatorname{div} B &= 0 \\ \operatorname{div} E &= \rho/\varepsilon, \end{aligned}$$

where ε is called the permittivity and ρ is the charge density. Finally, the material laws

$$B = \mu H, \qquad J = \sigma E,$$

with the parameters μ being the permeability and σ being the electric conductivity and H being the magnetic field intensity, allow the reduction of variables to the system of PDEs

$$\begin{aligned} \operatorname{curl} E &= -\mu \partial_t H, \\ \operatorname{curl} H &= \sigma E. \end{aligned}$$

Taking the $\operatorname{curl} \mu^{-1}$ of the first equation and the partial derivative ∂_t of the second equation and combining both gives a linear second order PDE

$$\operatorname{curl} \mu^{-1} \operatorname{curl} E = -\sigma \partial_t E,$$

where the only unknown is the electric field intensity E (a 3D-vectorfield), and which is oftentimes referred to as **Maxwell's equations**.

Oftentimes, one models time harmonic problems, i.e., the dependence in time is of the form $E(x,t) = e^{i\omega t}E(x)$ and plugging that into the equation gives the **time harmonic Maxwell equations**

$$\operatorname{curl}\mu^{-1}\operatorname{curl}E + \kappa E = 0,$$

with $\kappa = i\omega\sigma$.

1.2 A one dimensional model problem

The goal of this section is to study the simplest possible problem to demonstrate key concepts of the finite element method that will be generalized later on.

We study the Poisson equation in 1D on the interval $\Omega := (0,1)$ with homogeneous Dirichlet boundary conditions, i.e.,

$$-u''(x) = f(x) \qquad x \in (0,1) \tag{1.1}$$
$$u(0) = u(1) = 0. \tag{1.2}$$

1.2.1 Finite difference approximation

The simplest numerical approximation to differential equations can be derived by replacing the differentiation by a discrete difference quotient. This leads to a so called **finite difference method.**

We start by decomposing Ω into equidistant sub-intervals (x_i, x_{i+1}), $i = 1, \dots, N$ with $x_i = \frac{i-1}{N}$ of length $h = \frac{1}{N} = x_{i+1} - x_i$.
We set $u(x_i) =: u_i \in \mathbb{R}$ and approximate

$$u'(x_i) \simeq \frac{u(x_{i+1}) - u(x_i)}{h} = \frac{u_{i+1} - u_i}{h}$$

as well as

$$\begin{aligned} u''(x_i) &\simeq \frac{u'(x_i) - u'(x_{i-1})}{h} \simeq \frac{1}{h}\left(\frac{u(x_{i+1}) - u(x_i)}{h} - \frac{u(x_i) - u(x_{i-1})}{h}\right) \\ &= \frac{1}{h^2}\left(u_{i+1} - 2u_i + u_{i-1}\right). \end{aligned}$$

Inserting this into our 1D model problem, we obtain the equations

$$-\frac{1}{h^2}\left(u_{i+1} - 2u_i + u_{i-1}\right) = f(x_i) \qquad i = 1, \dots, N+1$$

for the unknowns u_i. As this is a linear system of $N+1$ equations for $N+1$ unknowns, we can hope for a unique computable solution.

However, we have to take the boundary conditions into account, thus the condition $u(0) = 0$ implies $u_1 = 0$ and $u(1) = 0$ implies $u_{N+1} = 0$ and consequently, the first and last equation have to be omitted, i.e., we have

$$-\frac{1}{h^2}\left(u_{i+1} - 2u_i + u_{i-1}\right) = f(x_i) \qquad i = 2, \ldots, N, \qquad u_0 = u_{N+1} = 0.$$

For different types of boundary conditions, different modifications to u_0, u_{N+1} have to be made.

Finite difference methods are widely used in engineering applications due to their advantages

- simple derivation, easy mathematical description;
- very easy implementation.

However, they have some very big drawbacks as well

- simple approach only works on very structured regions, such as squares, cubes;
- data f has to be continuous;
- solution of the linear system of equations gets expensive for large N (this is called bad conditioning).

1.2.2 Variational formulation

For solutions to PDEs (and ODEs), we thus far assumed that they are sufficiently often differentiable, i.e., solutions to the Poisson equation should be at least two times continuously differentiable. This assumption can be quite restrictive and therefore some equations do not have solutions due to this requirement.
For example, if the right-hand side f in the equations is only in $L^2(\Omega)$ (space of square integrable functions over the set Ω) but not continuous, solutions of $-u'' = f$ can not be two times differentiable.

In order to deal with this problem, the notion of solution can be changed, which leads to so called **weak solutions** that satisfy so called **variational formulations** of the PDE.
In order to derive this formulation, we multiply our model problem with a so called **test function** v in a vector space V (that is at least continuously differentiable and vanishes at the boundary of Ω) and integrate over Ω to obtain

$$\int_\Omega -u''v dx = \int_\Omega fv dx.$$

Integration by parts then gives

$$\int_\Omega u'v' dx = \int_\Omega fv dx.$$

As the test function v was arbitrary, we actually derived the following formulation: Seek u such that

$$a(u, v) := \int_\Omega u'v' dx = \int_\Omega fv dx =: l(v) \qquad \forall v \in V. \tag{1.3}$$

Here, $a(\cdot,\cdot)$ is a bilinear form and $l(\cdot)$ is a linear form (precise definitions in Section 2 below). Formulation (1.3) is called the **weak formulation** (or variational formulation) of the PDE.

It remains to specify the function space V. To that end, we make the following observations:

- The linear form $l(\cdot)$ on the right-hand side is well-defined provided fv is integrable. Using the Cauchy-Schwarz inequality, we can actually estimate

$$l(v) = \int_\Omega fv dx \leq \sqrt{\int_\Omega f^2 dx}\sqrt{\int_\Omega v^2 dx}.$$

  This shows that a requirement $f \in L^2(\Omega)$ is also sufficient.

- The bilinear form $a(\cdot,\cdot)$ on the left-hand side is well-defined, provided u, v are one time differentiable. Using the Cauchy-Schwarz inequality, we can estimate

$$a(u,v) = \int_\Omega u'v' dx \leq \sqrt{\int_\Omega (u')^2 dx}\sqrt{\int_\Omega (v')^2 dx}.$$

  Therefore, the requirement $u', v' \in L^2(\Omega)$ is also sufficient.

- In comparison with the formulation of the model problem (1.1) (which is also called strong formulation or classical formulation of the PDE), one does not require two times differentiability and continuous data!

This motivates the choice

$$V = \{v \in L^2(\Omega) \ : \ v' \in L^2(\Omega)\} =: H^1(\Omega).$$

In literature this space is called a **Sobolev** space of first order (note: its precise mathematical definition would require the notion of a weak derivative).

Exercise 1. Show that the space $H^1(\Omega)$ is a vector space. Moreover, show that

$$\|u\|^2_{H^1(\Omega)} := \|u\|^2_{L^2(\Omega)} + \|u'\|^2_{L^2(\Omega)}$$

is a norm on $H^1(\Omega)$ and

$$(u,v)_{H^1} := (u',v')_{L^2} + (u,v)_{L^2} = \int_\Omega u'v' + uv dx$$

is an inner product on $H^1(\Omega)$. □

Homogeneous Dirichlet boundary conditions are usually directly incorporated in the definition of the vector space, which leads to the space

$$V_0 = \{v \in L^2(0,1) \ : \ v' \in L^2(0,1), v(0) = v(1) = 0\} =: H^1_0(0,1).$$

Therefore, the final weak formulation reads as: Find $u \in H_0^1(0,1)$ such that

$$a(u,v) := \int_\Omega u'v'dx = \int_\Omega fvdx =: l(v) \qquad \forall v \in H_0^1(0,1). \tag{1.4}$$

By construction, we have that classical solutions are always weak solutions. However, the converse statement does not always hold. However, if a weak solution is sufficiently often continuously differentiable (here: two times), it is also a classical solution.

The following Lemma clarifies the reason why weak formulations are also called variational formulations.

Lemma 1.1. *The problem* (1.4) *is equivalent to the minimization problem:*

$$\textit{Find } u \in H_0^1(0,1) \textit{ such that } \mathcal{F}(u) := \int_0^1 \frac{(u')^2}{2} - fudx \to \ \textit{min} \tag{1.5}$$

Proof. By the theory on calculus of variations, we know that minimizers of the given functional have to satisfy the (weak form of the) Euler-Lagrange equations. For the given functional these are given as

$$-u'' + f = 0.$$

Conversely, let u solve (1.4). Let w be arbitrary. Then, we use $v := w - u$ as test function in (1.4) to derive

$$\mathcal{F}(u) \leq \mathcal{F}(u) + \frac{1}{2}\int_0^1 v'^2 dx = \frac{1}{2}\int_0^1 (u+v)'^2 - f(u+v) = \mathcal{F}(u+v) = \mathcal{F}(w).$$

As w was arbitrary, we have that u minimizes $\mathcal{F}(\cdot)$. ■

If f is continuous, Peanos theorem provides the existence of a classical solution, which is also a weak solution. Uniqueness of the weak solution follows from the following argument. Let u, w be weak solutions to (1.4). Then, the difference $u - w$ is a weak solution with data $f = 0$. Consequently, using $u - w$ as test-function gives

$$0 = \int_0^1 (u' - w')^2 dx.$$

As this can only hold, when $u' - w' = 0$, we obtain that $u - w = \text{const}$. However, the boundary condition $u(0) = w(0) = 0$ implies that the constant has to be zero, i.e., $u = w$.

1.2.3 Finite element approximation

Now, we are developing a numerical method for approximating the weak form (1.4). The main idea of the finite element method is to replace the (infinite dimensional!!) space $H_0^1(\Omega)$ in the weak formulation by a finite dimensional space to obtain a computable formulation. Obviously, this will only lead to an approximative solution to (1.4).

In order to specify this finite dimensional space, we again decompose the domain Ω into sub intervals $T_i = (x_i, x_{i+1})$, $i = 1, \ldots, N$.

We call the subintervals T_i **elements** and its collection $\mathcal{T} := \{T_i \;\; i = 1, \ldots, N\}$ a **triangulation** (the name is motivated from higher space dimensions) or **mesh**. The collection of the endpoints $\mathcal{N} = \{x_j \;:\; j = 1, \ldots, N+1\}$ is called the **nodes** of the mesh. We note that the nodes x_j do not need to be equidistant and the lengths $h_i = x_{i+1} - x_i$ are called **local mesh-widths**.

On a mesh $\mathcal{T}$, we introduce the finite element spaces $V_h \subset H^1(\Omega)$ and $V_{h,0} \subset H_0^1(\Omega)$ as

$$\begin{aligned} V_h &:= \{v \in C(\Omega) : v|_T \text{ is a polynomial of degree} \leq 1 \; \forall T \in \mathcal{T}\}, \\ V_{h,0} &:= V_h \cap H_0^1(\Omega). \end{aligned}$$

In fact, for the 1d case studied here, the space $V_{h,0}$ only contains functions in V_h satisfying $v(0) = v(1) = 0$.

Now, the finite element approximation (or FEM-solution) is defined as the function $u_h \in V_{h,0}$ satisfying

$$a(u_h, v_h) = \int_\Omega u_h' v_h' dx = \int_\Omega f v_h dx = l(v_h) \qquad \forall v_h \in V_{h,0}. \tag{1.6}$$

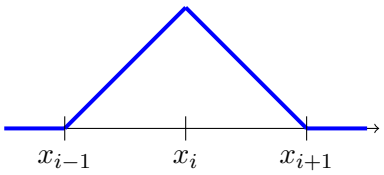
Thus, we have only replaced the space $H_0^1(0,1)$ in (1.4) by $V_{h,0}$. Existence of solutions to the FEM formulation will be discussed in Chapter 2 below, uniqueness follows from uniqueness of solutions to the problem (1.4).

As the space of linear polynomials is finite dimensional with dimension 2, we obtain that the space V_h is finite dimensional as well. In fact, a function $v_h \in V_h$ is uniquely defined by its values $v(x_j)$ in the nodes $x_j \in \mathcal{N}$. Consequently, one can obtain a so called **nodal basis** $\{\varphi_i \;:\; i = 1, \ldots, N+1\}$ of V_h by functions characterized by

$$\varphi_i(x_j) = \delta_{i,j} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases} . \tag{1.7}$$

These functions are called **hat functions** and can be written as

$$\varphi_i(x) = \begin{cases} \frac{x - x_{i-1}}{h_{i-1}}, & x \in (x_{i-1}, x_i) \\ \frac{x_{i+1} - x}{h_i}, & x \in (x_i, x_{i+1}) \\ 0, & \text{otherwise.} \end{cases}$$



Therefore, the space V_h has dimension $N+1$. For the space $V_{h,0}$ nodal values at the first and last mesh points are fixed, therefore a nodal basis consists of $\{\varphi_i \;:\; i = 2, \ldots, N\}$ and the space $V_{h,0}$ has dimension $N-1$.

With this basis available, we can represent the finite element solution $u_h \in V_{h,0}$ as

$$u_h(x) = \sum_{i=2}^{N} u_i \varphi_i(x) \tag{1.8}$$

with coefficients $u_i \in \mathbb{R}$ that are unknown. By the definition of the nodal-basis (1.7), there holds

$$u_h(x_j) = \sum_{i=2}^{N} u_i \varphi_i(x_j) = u_j.$$

In order to determine the coefficients u_j, we insert the expansion of u_h into the weak formulation and use $v = \varphi_j$ as a test function. This gives due to the linearity of $a(\cdot, \cdot)$

$$a(\sum_{i=2}^{N} u_i \varphi_i, \varphi_j) = \sum_{i=2}^{N} u_i a(\varphi_i, \varphi_j) = l(\varphi_j) \tag{1.9}$$

for all $j = 2, \ldots, N$. We define the so called **stiffness matrix** $\mathbf{A} = (A_{ij}) \in \mathbb{R}^{N-1 \times N-1}$ and the **load vector** $\mathbf{f} = (f_j) \in \mathbb{R}^{N-1}$ as

$$\begin{aligned} A_{ij} &:= \int_\Omega \varphi_j' \cdot \varphi_i' \, dx = a(\varphi_j, \varphi_i), \\ f_j &:= \int_\Omega f \varphi_j \, dx = l(\varphi_j). \end{aligned}$$

Then, with $\mathbf{u} = (u_j) \in \mathbb{R}^{N-1}$, (1.9) is equivalent to the linear system of equations

$$\mathbf{A}\mathbf{u} = \mathbf{f}.$$

One can show that the matrix $\mathbf{A}$ is invertible (here it is symmetric and positive definite, which implies invertible) and therefore, the linear system can be solved to obtain the sought coefficients as $\mathbf{u} = \mathbf{A}^{-1}\mathbf{f}$.

1.2.4 Remarks on implementation

In order to implement a FEM in a programming language, one only has to compute the stiffness matrix $\mathbf{A}$, the load vector $\mathbf{f}$ and then solve the linear system of equations. Since the solution of the linear system often can be efficiently done using existing routines (e.g., the \-Operator in MATLAB), we focus on the assembly of the linear system of equations.

Dirichlet boundary conditions: Instead of only working with the basis functions φ_i for $i = 2, \ldots, N$ one can also assemble the matrix $\widetilde{\mathbf{A}} \in \mathbb{R}^{N+1 \times N+1}$ with all basis functions φ_i for $i = 1, \ldots, N+1$ and then just delete the first and last row and column of the matrix $\widetilde{\mathbf{A}}$ to obtain the matrix $\mathbf{A}$.

Elementwise computation: Since the mesh $\mathcal{T}$ decomposes Ω, we can split the integrals in the stiffness matrix and load vector into a sum over the elements

$$\begin{aligned} A_{ij} &= \sum_{T \in \mathcal{T}} \int_T \varphi_j' \cdot \varphi_i' \, dx, \\ f_j &= \sum_{T \in \mathcal{T}} \int_T f \varphi_j \, dx. \end{aligned}$$

As the basis functions are linear polynomials on each element, we actually have that $\varphi_j' = \text{const}$ on each element. In fact, by definition of the hat functions, there each basis function is only non-zero on two elements (for φ_j the elements T_j and T_{j-1}).
This has the following consequences:

- The matrix $\mathbf{A}$ is a tridiagonal matrix. Using sparse matrix formats (MATLAB `sparse`), this reduces the storage capacity to $3(N-1)$ instead of $(N-1)^2$.
- The implementation usually is done using three loops, a outer loop over the elements, and two inner loops over the indices i, j. Using the support properties of the hat functions (i.e., the discussion on the elements where the hat functions are non-zero) theses loops do not have to run through the whole index sets.

We also note that the load vector is usually computed using formulas for numerical integration such as Gaussian quadrature.

1.2.5 The Neumann problem

We change the boundary condition in the model problem to Neumann conditions, i.e.,

$$-u''(x) = f(x) \qquad x \in (0,1) \tag{1.10}$$
$$u'(0) = u'(1) = 0. \tag{1.11}$$

In order to derive a weak formulation of the PDE, we again multiply with an (at least continuously differentiable) test function v and integrate by parts to obtain

$$\int_0^1 u'v'dx = \int_0^1 fvdx.$$

Here, the boundary terms in the integration by parts are not here due to the Neumann boundary conditions. We also note that the test function v does NOT have to satisfy any boundary condition. Consequently, the weak formulation reads as: find $u \in H^1(\Omega)$ such that

$$a(u,v) = \int_0^1 u'v'dx = \int_0^1 fvdx = l(v) \qquad \forall v \in H^1(\Omega).$$

In contrast to the Dirichlet problem (where the boundary conditions are incorporated into the vector space), the boundary conditions are "hidden" in the bilinear form.
However, this formulation may not be solvable! E.g. testing with the constant function $1 \in H^1(\Omega)$ gives

$$0 = \int_0^1 fdx,$$

which may not hold for arbitrary functions f. Moreover, provided $\int fdx = 0$, solutions may not be unique as, if u is a weak solution, $u+c$ for an arbitrary constant $c \in \mathbb{R}$ is a solution as well. In order to guarantee unique solvability, additional conditions need to be imposed. A common conditions is to only look for functions u with vanishing mean, i.e., $\int_0^1 udx = 0$.

Exercise 2. Derive a weak formulation for the 1d-model problem with Robin boundary conditions. □

1.2.6 Summary

In order to describe a finite element method for a given PDE, one needs

1. A (uniquely solvable) weak formulation posed on some Banach space V.
2. A finite dimensional subspace $V_h \subset V$, such that there is a unique weak solution for the space V_h. The space V_h usually consists of piecewise polynomials on a mesh.
3. For implementation: A basis for the finite dimensional subspace, quadrature formulas for assembly of the system matrix and a solver for the linear system of equations.

Finite element methods are widely used in structural engineering, fluid dynamics, elasticity theory, resonance problems or electrical engineering. They enjoy the advantages

- can handle arbitrary, complicated regions;
- many codes available in industry and academics (NGSolve,Comsol,Fenix,etc.);
- data f does not have to be continuous;
- allows you to "zoom in" with your computational effort (i.e., use computational power where it is really needed);
- elegant mathematical theory.

However, they have drawbacks as well

- implementation considerably harder;
- solution of the linear system of equations can get expensive for large N as well.

Chapter 2

Abstract framework for FEM

In this chapter we develop the abstract framework for variational problems.

2.1 Basic properties

In the following, we will work with vector spaces, Banach spaces and Hilbert spaces. We recall that Banach spaces are complete normed spaces (a norm generalizes lengths) and Hilbert spaces are complete inner product spaces (an inner-product generalizes the dot-product in $\mathbb{R}^n$). For the precise definitions of vector, Banach, Hilbert spaces, norms and inner products, we refer to the lecture notes on *Applied Mathematics Foundations*.

We also recall the best approximation (or closest point projection) property in Hilbert spaces:

Lemma 2.1. *Let S be a nonempty closed convex subset of the Hilbert space $(V, (\cdot,\cdot)_V)$. Let $u \in V$. Then, there exists a unique closest point $u_0 \in S$:*

$$\|u - u_0\|_V \leq \|u - v\|_V \qquad \forall\, v \in S.$$

Here, $\|\cdot\|_V$ denotes the induced norm from the inner product $(\cdot,\cdot)_V$.

We also recall a very important example of a Hilbert space, the space $L^2(\Omega)$.

Example. Let $\Omega \subset \mathbb{R}^n$. The set of square integrable functions on Ω, i.e., all functions satisfying

$$\|f\|_{L^2}^2 := \int_\Omega |f(x)|^2\, dx < \infty$$

is denoted by $L^2(\Omega)$ and is a Hilbert space with the inner product

$$(f,g)_{L^2} := \int_\Omega f(x)g(x)dx.$$

As it is true for any Hilbert space, there holds the Cauchy-Schwarz inequality

$$|(f,g)_{L^2}| \leq \|f\|_{L^2}\|g\|_{L^2}.$$

■

In order to define an abstract framework for variational equations, we precisely define linear and bilinear forms.

Definition 2.2. *Let V be a vector space. A* **bilinear form** *$A(\cdot,\cdot)$ on V is a mapping $A : V \times V \to \mathbb{R}$ which is linear in u and in v.*
A bilinear form is called **symmetric** *if $A(u,v) = A(v,u)$ for all $u, v \in V$.*

Example.

1. Let $A \in \mathbb{R}^{n\times n}$ be a (symmetric) matrix. Then, the mapping

$$A : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}, \quad (u,v) \mapsto u^T A v$$

   defines a (symmetric) bilinear form on $\mathbb{R}^n$.

2. Let $V = L^2(0,1)$. Then, the mapping

$$A(u,v) = \int_0^1 uv\,dx$$

   defines an inner product on $L^2(0,1)$.

3. Inner products defined on a real vector space are by definition bilinear forms. ■

Definition 2.3. *Let $(V, \|\cdot\|)$ be a Banach space. A* **functional** *or a* **linear form** *$l(\cdot)$ on V is a linear mapping $l(\cdot) : V \to \mathbb{R}$.*
The canonical norm for linear forms is the **dual norm**

$$\|l\|_{V^*} := \sup_{0\neq v\in V} \frac{l(v)}{\|v\|}.$$

A linear form l is called **bounded**, *if its norm is finite. The vector space of all bounded linear forms on V is called the* **dual space** *V^*.*

Example.

1. An example for a bounded linear form is $l(\cdot) : L^2(\Omega) \to \mathbb{R} : v \to \int_\Omega v\,dx$.

2. In $\mathbb{R}^n$ (column vectors), the multiplication with a row vector $w = (w_1, \ldots, w_n)$ from the left, i.e.,

$$l(\cdot) : \mathbb{R}^n \to \mathbb{R} : v \mapsto w \cdot v = \sum_{i=1}^n w_i v_i$$

   is a bounded linear form. In fact, this also gives a complete characterization of the dual space V^* as the space of all row vectors with n entries.

■

2.2 Solvability of variational formulations

Let $(V, (\cdot,\cdot)_V)$ be a Hilbert space. In this subsection, we present conditions for bilinear forms that guarantee unique solvability of abstract equations of the form: Find $u \in V$ such that

$$A(u,v) = l(v) \qquad \forall v \in V, \tag{2.1}$$

where $A(\cdot,\cdot)$ is a bilinear form and $l(\cdot)$ is a bounded linear form.

2.2.1 Inner products

Let $(V, (\cdot,\cdot)_V)$ be a Hilbert space and $u \in V$. Then, we can define the related continuous linear functional $l_u(\cdot) \in V^*$ by

$$l_u(v) := (u,v)_V \qquad \forall\, v \in V.$$

The opposite is also true:

Theorem 2.4. Riesz Representation Theorem. *Let $(V, (\cdot,\cdot)_V)$ be a Hilbert space. Any bounded linear form $l(\cdot)$ on V can be represented uniquely as*

$$l(v) = (u_l, v)_V \tag{2.2}$$

for some $u_l \in V$. Furthermore, we have $\|l\|_{V^*} = \|u_l\|_V$.

An application of the Riesz representation theorem is the existence of unique solutions to (2.1) provided $A(\cdot,\cdot)$ is an inner-product on V. This follows directly from using $(V, A(\cdot,\cdot))$ as Hilbert space in the Riesz Representation theorem.

Example. The bilinear form

$$A(u,v) := \int_0^1 u'v' + uv dx$$

is by Exercise 1 an inner-product on the Sobolev space $H^1(0,1)$. Moreover, since

$$\int_0^1 fv dx \leq \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)}\|v\|_{H^1(\Omega)},$$

we have that $l(v) := \int_0^1 fv dx$ is a bounded linear form. Consequently, we obtain that the variational formulation

$$A(u,v) = l(v) \quad \forall v \in H^1(\Omega)$$

has a unique solution. In fact, this problem is the weak formulation of the equation $-u'' + u = f$ with homogeneous Neumann boundary conditions. ■

2.2.2 Coercive variational problems

In the following, we want to weaken the conditions on the bilinear form a bit, but still obtain a uniquely solvable equation. This leads to the following definitions.

Definition 2.5. *A bilinear form $A(\cdot,\cdot) : V \times V \to \mathbb{R}$ is called* **coercive** *(or elliptic), if there is a constant $\alpha_1 \in \mathbb{R}$ such that*

$$A(u,u) \geq \alpha_1 \|u\|_V^2 \qquad \forall\, u \in V. \tag{2.3}$$

$A(\cdot,\cdot)$ is called **continuous***, if there is a constant $\alpha_2 \in \mathbb{R}$ such that*

$$A(u,v) \leq \alpha_2 \|u\|_V \, \|v\|_V \qquad \forall\, u, v \in V. \tag{2.4}$$

Note that we do NOT require symmetry!

Example. The previous example does fit into this framework as well, since

$$A(u,v) := \int_0^1 u'v' + uv dx = (u,v)_{H^1}.$$

Consequently, we have coercivity with $\alpha_1 = 1$. Continuity follows directly from the Cauchy-Schwarz inequality with $\alpha_2 = 1$. ■

Instead of the linear form $l(\cdot)$, we will often write $l \in V^*$.

We can also reformulate the variational problem as an operator equation: Defining the operator $A : u \in V \to A(u,\cdot) \in V^*$ (which is linear and bounded), the variational problem is equivalent to the equation

$$Au = f \qquad (\text{in } V^*).$$

Before we can show unique solvability of the variational formulation, we need a famous result from real analysis, Banach's fixed point theorem.

Theorem 2.6 (Banach's contraction mapping theorem). *Given a Banach space V and a mapping $T : V \to V$, satisfying the Lipschitz condition*

$$\|T(v_1) - T(v_2)\| \leq L \, \|v_1 - v_2\| \qquad \forall\, v_1, v_2 \in V$$

for a fixed $L \in [0,1)$. Then there exists a unique $u \in V$ such that

$$u = T(u),$$

i.e. the mapping T has a unique fixed point u.

The following theorem, called the Lax-Milgram lemma, provides a unique solution to continuous and coercive variational problems. As it is the key step in our abstract analysis, we include its proof here.

Theorem 2.7 (Lax-Milgram). *Given a Hilbert space V, a coercive and continuous bilinear form $A(\cdot,\cdot)$, and a continuous linear form $f(\cdot)$. Then there exists a unique $u \in V$ solving*

$$A(u,v) = f(v) \qquad \forall\, v \in V.$$

There holds

$$\|u\|_V \leq \alpha_1^{-1} \|f\|_{V^*} \tag{2.5}$$

Proof: Start from the operator equation $Au = f$. Let $J_V : V^* \to V$ be the so called Riesz isomorphism defined by

$$(J_V g, v)_V = g(v) \qquad \forall\, v \in V,\ \forall\, g \in V^*.$$

Then, the operator equation is equivalent to

$$J_V A u = J_V f \qquad (\text{in } V),$$

and to the fixed point equation (with some $0 \neq \tau \in \mathbb{R}$ chosen below)

$$u = u - \tau J_V (Au - f). \tag{2.6}$$

We will verify that

$$T(v) := v - \tau J_V (Av - f)$$

is a contraction mapping, i.e., $\|T(v_1) - T(v_2)\|_V \leq L \|v_1 - v_2\|_V$ with some Lipschitz constant $L \in [0,1)$. Let $v_1, v_2 \in V$, and set $v = v_1 - v_2$. Then,

$$\begin{aligned}
\|T(v_1) - T(v_2)\|_V^2 &= \|\{v_1 - \tau J_V(Av_1 - f)\} - \{v_2 - \tau J_V(Av_2 - f)\}\|_V^2 \\
&= \|v - \tau J_V A v\|_V^2 \\
&= \|v\|_V^2 - 2\tau (J_V A v, v)_V + \tau^2 \|J_V A v\|_V^2 \\
&= \|v\|_V^2 - 2\tau \langle Av, v \rangle + \tau^2 \|Av\|_{V^*}^2 \\
&= \|v\|_V^2 - 2\tau A(v,v) + \tau^2 \|Av\|_{V^*}^2 \\
&\leq \|v\|_V^2 - 2\tau\alpha_1 \|v\|_V^2 + \tau^2 \alpha_2^2 \|v\|_V^2 \\
&= (1 - 2\tau\alpha_1 + \tau^2 \alpha_2^2) \|v_1 - v_2\|_V^2
\end{aligned}$$

Now, we choose $\tau = \alpha_1/\alpha_2^2$, and obtain a Lipschitz constant

$$L^2 = 1 - \alpha_1^2/\alpha_2^2 \in [0,1).$$

Banach's contraction mapping theorem state that (2.6) has a unique fixed point. Finally, we obtain the bound (2.5) from

$$\|u\|_V^2 \leq \alpha_1^{-1} A(u,u) = \alpha_1^{-1} f(u) \leq \alpha_1^{-1} \|f\|_{V^*} \|u\|_V,$$

and dividing by $\|u\|_V$. □

2.2.3 Approximation of coercive variational problems

Now, let V_h be a closed subspace of V. Replacing V by V_h leads to a so called **Galerkin method** of finding $u_h \in V_h$ such that

$$A(u_h, v_h) = f(v_h) \qquad \forall\, v_h \in V_h. \tag{2.7}$$

This variational problem is again uniquely solvable by Lax-Milgram, since, $(V_h, \|.\|_V)$ is a Hilbert space, and continuity and coercivity on V_h are inherited from the original problem on V.

A fundamental property is so called Galerkin orthogonality.

Lemma 2.8. *Let V be a Hilbert space and u be the solution to the variational problem $A(u, v) = f(v)$ for all $v \in V$. Let $V_h \subset V$ be closed subspace and u_h be the solution to $A(u_h, v_h) = f(v_h)$ for all $v_h \in V_h$. Then, there holds the so called* **Galerkin orthogonality**

$$A(u - u_h, v_h) = 0 \qquad \forall v_h \in V_h. \tag{2.8}$$

Proof. Using the linearity of A and the assumption that $V_h \subset V$, we can use v_h as test function in both variational formulations to compute

$$A(u - u_h, v_h) = A(u, v_h) - A(u_h, v_h) = f(v_h) - f(v_h) = 0 \qquad \forall\, v_h \in V_h,$$

which is the stated orthogonality. ■

The next theorem says, that the solution defined by the Galerkin method is, up to a constant factor, as good as the best possible approximation in the space V_h.

Theorem 2.9 (Cea). *Let the assumptions of Lemma 2.8 hold and $A(\cdot, \cdot)$ be continuous and coercive. Then,*

$$\|u - u_h\|_V \leq \alpha_2/\alpha_1 \inf_{v_h \in V_h} \|u - v_h\|_V,$$

i.e., the approximation error of the Galerkin method is quasi optimal.

Proof. Let $v_h \in V_h$ be arbitrary. Using Galerkin orthogonality, we compute

$$\begin{aligned} \|u - u_h\|_V^2 &\leq \alpha_1^{-1} A(u - u_h, u - u_h) \\ &= \alpha_1^{-1} A(u - u_h, u - v_h) + \alpha_1^{-1} A(u - u_h, \underbrace{v_h - u_h}_{\in V_h}) \\ &= \alpha_1^{-1} A(u - u_h, u - v_h) \\ &\leq \alpha_2/\alpha_1 \, \|u - u_h\|_V \|u - v_h\|_V. \end{aligned}$$

Divide one factor $\|u - u_h\|_V$. Since $v_h \in V_h$ was arbitrary, the estimation holds true also for the infimum in V_h. ■

Chapter 3

Sobolev spaces and weak formulation of Poisson problem

In the abstract framework of the previous setting, we used a Hilbert space V and a (finite dimensional) subspace V_h. In this section, we want to introduce spaces commonly employed for elliptic PDEs.

Throughout this section Ω is an open subset of $\mathbb{R}^d$ and is either bounded or unbounded. We recall the notation

$$C^k(\overline{\Omega}) := \{u : \overline{\Omega} \to \mathbb{R} \ : \ u \text{ is k-times continuously differentiable}\}$$

for differentiable functions of order $k \in \mathbb{N} \cup \{\infty\}$.

3.1 Sobolev spaces

In the following, we introduce the concept of generalized derivatives and give a precise description of Sobolev spaces in arbitrary space dimensions.

3.1.1 Generalized derivatives

Let $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}_0^d$ be a multi-index, $|\alpha| = \sum \alpha_i$, and define the classical differential operator for functions in $C^\infty(\Omega)$

$$D^\alpha = \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \cdots \left(\frac{\partial}{\partial x_n}\right)^{\alpha_d}.$$

Example. For $d = 2$ and $\alpha = (1, 0)$, we have $D^\alpha = \frac{\partial}{\partial x}$. For $\alpha = (1, 1)$, we have $D^\alpha = \frac{\partial}{\partial x}\frac{\partial}{\partial y}$.
For a function $u \in C(\Omega)$, the **support** is defined as

$$\operatorname{supp}\{u\} := \overline{\{x \in \Omega : u(x) \neq 0\}}.$$

This is a compact set if and only if it is bounded. We say u has **compact support in** Ω, if $\operatorname{supp} u \subset \Omega$. If Ω is a bounded domain, then u has compact support in Ω if and only if u vanishes in a neighborhood of $\partial\Omega$.

The space of smooth functions with compact support is denoted as

$$\mathcal{D}(\Omega) := C_0^\infty(\Omega) := \{u \in C^\infty(\Omega) : \ u \text{ has compact support in } \Omega\}. \tag{3.1}$$

For a smooth function $u \in C^{|\alpha|}(\Omega)$, there holds the formula of integration by parts

$$\int_\Omega D^\alpha u \varphi \, dx = (-1)^{|\alpha|} \int_\Omega u D^\alpha \varphi \, dx \qquad \forall \, \varphi \in \mathcal{D}(\Omega). \tag{3.2}$$

Definition 3.1. *We denote the dual space (i.e. the space of all bounded linear forms) of $\mathcal{D}(\Omega)$ by $\mathcal{D}'(\Omega)$ and call elements in $\mathcal{D}'(\Omega)$* **distributions**.

Example.

1. The $L^2(\Omega)$-inner product with a fixed function u in $C(\Omega)$ defines the linear functional on $\mathcal{D}(\Omega)$
$$u(\varphi) := \langle u, \varphi \rangle_{\mathcal{D}' \times \mathcal{D}} := \int_\Omega u \varphi \, dx, \quad \varphi \in \mathcal{D}(\Omega).$$
In this sense, every continuous function generates a distribution.

2. The Dirac delta defined as
$$\delta_0(\varphi) := \int_\Omega \varphi(x) \delta_0(x) dx = \varphi(0), \quad \varphi \in \mathcal{D}(\Omega)$$
is a distribution (this justifies the name Dirac δ-distribution used in Applied Mathematics Foundations).

The formula (3.2) is valid for functions $u \in C^{|\alpha|}$. The strong regularity is needed only on the left hand side. Thus, we use the less demanding right hand side to extend the definition of differentiation for distributions:

Definition 3.2. *For $u \in \mathcal{D}'$, we define $g \in \mathcal{D}'$ to be the* **generalized derivative** *$D^\alpha_g u$ of u by*

$$g(\varphi) = \langle g, \varphi \rangle_{\mathcal{D}' \times \mathcal{D}} = (-1)^{|\alpha|} u(D^\alpha \varphi) \qquad \forall \, \varphi \in \mathcal{D}.$$

If $u \in C^{|\alpha|}$, then D^α_g coincides with D^α. Moreover, by this definition, distributions always have generalized derivatives of arbitrary order!!

The function space of **locally integrable** functions on Ω is called

$$L^1_{loc}(\Omega) = \{u : u_K \in L^1(K) \; \forall \text{ compact } K \subset \Omega\}.$$

It contains functions which can behave very badly near $\partial\Omega$. E.g., $e^{e^{1/x}}$ is in $L^1_{loc}(0,1)$. If Ω is unbounded, then the constant function 1 is in $L^1_{loc}(\Omega)$, but not in $L^1(\Omega)$.

Definition 3.3. *For $u \in L_1^{loc}(\Omega)$, we call g the* **weak derivative** *$D^\alpha_w u$, if $g \in L^1_{loc}(\Omega)$ satisfies*

$$\int_\Omega g(x) \varphi(x) \, dx = (-1)^{|\alpha|} \int_\Omega u(x) D^\alpha \varphi(x) \, dx \qquad \forall \, \varphi \in \mathcal{D}.$$

The weak derivative is more general than the classical derivative, but more restrictive than the generalized derivative.

Example. Let $\Omega = (-1, 1)$ and

$$u(x) = \begin{cases} 1 + x & x \leq 0 \\ 1 - x & x > 0 \end{cases} .$$

Then,

$$g(x) = \begin{cases} 1 & x \leq 0 \\ -1 & x > 0 \end{cases}$$

is the first generalized derivative D_g^1 of u, which is also a weak derivative. The second generalized derivative h is

$$\langle h, \varphi \rangle = -2\varphi(0) \qquad \forall\, \varphi \in \mathcal{D},$$

i.e., -2 times the Dirac δ-distribution. As this is not a function, it can not be in $L^1_{loc}(\Omega)$ and therefore it is not a weak derivative. ■

In the following, we will focus on weak derivatives. Unless it is essential we will skip the sub-scripts w and g.

3.1.2 Definition of Sobolev spaces

With the use of the weak derivative, we can now precisely define Sobolev spaces.

Definition 3.4. *For $k \in \mathbb{N}_0$, we define the* **Sobolev spaces** *via*

$$H^k(\Omega) = \{u \in L^1_{loc}(\Omega) : \|u\|_{H^k(\Omega)} < \infty\},$$

where the **Sobolev norms** *are given as (here D^α is the weak derivative)*

$$\|u\|_{H^k(\Omega)} := \left(\sum_{|\alpha| \leq k} \|D^\alpha u\|^2_{L^2(\Omega)} \right)^{1/2} .$$

The **Sobolev semi-norms** $|u|^2_{H^k(\Omega)} := \sum_{|\alpha|=k} \|D^\alpha u\|^2_{L^2(\Omega)}$ *contain only the highest order derivatives.*

In the previous chapter we have seen the importance of complete spaces. This is the case for Sobolev spaces:

Theorem 3.5. *The Sobolev space $H^k(\Omega)$ is a Banach space.*

Proof. Let v_j be a Cauchy sequence with respect to $\|\cdot\|_{H^k(\Omega)}$. This implies that $D^\alpha v_j$ is a Cauchy sequence in $L^2(\Omega)$, and thus converges to some v^α in $\|.\|_{L^2(\Omega)}$.

We verify that $D^\alpha v_j \to v^\alpha$ implies $\int_\Omega D^\alpha v_j \varphi \, dx \to \int_\Omega v^\alpha \varphi \, dx$ for all $\varphi \in \mathcal{D}$. Let K be the compact support of φ. There holds

$$\begin{aligned}\int_\Omega (D^\alpha v_j - v^\alpha)\varphi \, dx &= \int_K (D^\alpha v_j - v^\alpha)\varphi \, dx \\ &\leq \|D^\alpha v_j - v^\alpha\|_{L^1(K)} \|\varphi\|_{L^\infty(\Omega)} \\ &\leq \|D^\alpha v_j - v^\alpha\|_{L^2(K)} \|\varphi\|_{L^\infty(\Omega)} \to 0\end{aligned}$$

Finally, we compute

$$\begin{aligned}\int v^\alpha \varphi \, dx &= \lim_{j\to\infty} \int_\Omega D^\alpha v_j \varphi \, dx \\ &= \lim_{j\to\infty} (-1)^{|\alpha|} \int_\Omega v_j D^\alpha \varphi \, dx = \\ &= (-1)^\alpha \int_\Omega v D^\alpha \varphi \, dx,\end{aligned}$$

which verifies that v^α is the weak derivative of v. ■

With the inner product

$$(u, v)_{H^k} := \sum_{|\alpha| \leq k} (D^\alpha u, D^\alpha v)_{L^2(\Omega)}$$

the spaces $(H^k(\Omega), (\cdot, \cdot)_{H^k})$ are even Hilbert spaces.

3.1.3 Important theorems for Sobolev spaces

In order to obtain certain strong results for Sobolev spaces, a little bit more regularity has to be assumed for the region Ω.

Definition 3.6. *We call a bounded domain $\Omega \subset \mathbb{R}^n$ a* **Lipschitz domain**, *if its boundary $\partial\Omega$ can be locally parametrized with Lipschitz continuous functions. (More precisely, for each point $x \in \partial\Omega$ exists a neighborhood U_x in the parameter domain, such that $\gamma : U_x \to \partial\Omega$ is Lipschitz continuous.)*

The first statement shows that smooth functions are dense in Sobolev spaces (i.e., for every function in the Sobolev space there exists a sequence of smooth functions that converges to this function).

Theorem 3.7 (Meyers-Serrin). *Smooth functions in $C^\infty(\Omega)$ are dense in $H^k(\Omega)$ for every $k \in \mathbb{N}$.*

The following result is a very strong theorem that states compactness of the inclusion of Sobolev spaces of higher orders.

Theorem 3.8 (Rellich compactness theorem). *Let $k, \ell \in \mathbb{N}$ with $\ell > k$. Then, there holds $H^\ell(\Omega) \subset H^k(\Omega)$. Moreover, the inclusion mapping*

$$\iota : H^k(\Omega) \to H^\ell(\Omega),\ u \mapsto u$$

is a compact mapping, i.e., the image of each bounded sequence has a convergent subsequence.

Provided the order of differentiation is large enough Sobolev functions are actually continuous.

Theorem 3.9 (Sobolev embedding). *Let $k > d/2$. Then, $H^k(\Omega) \subset C(\Omega)$ and*

$$\|u\|_\infty \leq C\|u\|_{H^k(\Omega)}$$

for all $u \in H^k(\Omega)$.

Finally, we state the very famous Poincaré inequality, which can be used to show that the L^2-norm of the gradient is a norm on certain subspaces of $H^1(\Omega)$.

Lemma 3.10 (Poincaré inequality). *Let $u \in H^1(\Omega)$. Then,*

$$\|u\|_{L^2(\Omega)} \leq C\left(\|\nabla u\|_{L^2(\Omega)} + \left|\int_\Omega u\, dx\right|\right). \tag{3.3}$$

In particular, there holds $\|u\|_{L^2(\Omega)} \leq C\|\nabla u\|_{L^2(\Omega)}$ for all $u \in H^1(\Omega)$ with $\int_\Omega u\, dx = 0$.

Proof. We show the result by contradiction. Therefore, we assume that (3.3) does not hold. Consequently, there exists a sequence $(u_n)_{n\in\mathbb{N}}$ in $H^1(\Omega)$ such that

$$\|u_n\|_{H^1(\Omega)} > n\left(\|\nabla u_n\|_{L^2(\Omega)} + \left|\int_\Omega u_n\, dx\right|\right) \qquad \forall n \in \mathbb{N}.$$

One can actually normalize the u_n, i.e., assume $\|u_n\|_{H^1(\Omega)} = 1$. Consequently, one obtains

$$\|\nabla u_n\|_{L^2(\Omega)} + \left|\int_\Omega u_n\, dx\right| \leq \frac{1}{n} \to 0.$$

Therefore, by the Rellich compactness theorem, the bounded sequence u_n in $H^1(\Omega)$ has a subsequence $(u_{n_k})_{k\in\mathbb{N}}$ that converges in $L^2(\Omega)$. From the equation above, we infer that this subsequence also satisfies $\|\nabla u_{n_k}\|_{L^2(\Omega)} \to 0$. Consequently, the sequence $(u_{n_k})_{k\in\mathbb{N}}$ converges in $H^1(\Omega)$ and as $H^1(\Omega)$ is a complete normed space, there is a $u \in H^1(\Omega)$ such that

$$u_{n_k} \to u.$$

By $\|\nabla u_{n_k}\|_{L^2(\Omega)} \to 0$ there must hold $\|\nabla u\|_{L^2(\Omega)} = 0$ and by $\left|\int_\Omega u_{n_k}\, dx\right| \to 0$ there must hold $\left|\int_\Omega u\, dx\right| = 0$. Consequently, $u = 0$, which is a contradiction to $\|u\|_{H^1(\Omega)} = \lim \|u_{n_k}\|_{H^1(\Omega)} = 1$. ■

Replacing u by $w := u - \frac{1}{|\Omega|}\int_\Omega u\,dx \in H^1(\Omega)$ in the Poincaré inequality, one can also write (since then $\int_\Omega w\,dx = 0$)

$$\|u - \frac{1}{|\Omega|}\int_\Omega u\|_{L^2(\Omega)} \le C\|\nabla u\|_{L^2(\Omega)}.$$

The following theorem generalizes this result to higher order Sobolev spaces.

Theorem 3.11 (Bramble-Hilbert/Deny-Lions lemma). *Let $k \ge 1$ and $\mathcal{P}_k$ be the space of polynomials of maximal degree k. Let $u \in H^k(\Omega)$. Then,*

$$\inf_{p\in\mathcal{P}_{k-1}} \|u-p\|_{H^k(\Omega)} \le C\,|u|_{H^k(\Omega)}\,.$$

Proof. Essentially the same as for the Poincaré inequality, just replacing the H^1-norm by the H^k-norm and the integral mean by the projection Π onto the subspace of polynomials of degree $k-1$. ■

3.1.4 Traces of Sobolev functions

Our goal is to use Sobolev spaces in weak formulations. However, for the Dirichlet problem it is essential to be able to prescribe boundary values on $\partial\Omega$. For $u \in L^2(\Omega)$ this obviously might not be possible. However, for function $H^1(\Omega)$ this can be made formally precise.

We start in one dimension. Let $u \in C^1([0,h])$ with some $h > 0$. Then, we can bound

$$\begin{aligned}
u(0) &= \left(1-\frac{x}{h}\right)u(x)|_{x=0} = -\int_0^h \left\{\left(1-\frac{x}{h}\right)u(x)\right\}'\,dx \\
&= -\int_0^h \frac{-1}{h}u(x) + \left(1-\frac{x}{h}\right)u'(x)\,dx \\
&\le \left\|\frac{1}{h}\right\|_{L^2(0,h)}\|u\|_{L^2(0,h)} + \left\|1-\frac{x}{h}\right\|_{L^2(0,h)}\|u'\|_{L^2(0,h)} \\
&\simeq h^{-1/2}\|u\|_{L^2(0,h)} + h^{1/2}\|u'\|_{L_2(0,h)}. \qquad (3.4)
\end{aligned}$$

Next, we extend the trace operator to the whole Sobolev space H^1:

Theorem 3.12. *There is a well defined and continuous trace operator*

$$\mathrm{tr} : H^1(0,h) \to \mathbb{R}$$

whose restriction to $C^1([0,h])$ coincides with

$$u \to u(0).$$

Proof. We use that $C^1([0,h])$ is dense in $H^1(0,h)$ (by the Meyers-Serrin theorem). Let $u \in H^1(0,h)$. Take a sequence u_j in $C^1([0,h])$ with

$$u_j \to u \qquad \text{in } H^1(0,h).$$

Since by (3.4) we have

$$|u_m(0) - u_n(0)| \le C\|u_m - u_n\|_{H^1(\Omega)} \to 0,$$

the sequence $(u_j(0))_{j\in\mathbb{N}}$ is a Cauchy-sequence in $\mathbb{R}$. As $\mathbb{R}$ is complete, there is a limit $u_0 \in \mathbb{R}$ (and the limit is independent of the choice of the sequence u_j). This allows to define $\operatorname{tr} u := u_0$. ■

Now, we extend this 1D result to domains in more dimensions.

Theorem 3.13. *Let Ω be a Lipschitz domain. There exists a well defined and continuous operator*

$$\operatorname{tr} : H^1(\Omega) \to L^2(\partial\Omega),$$

which coincides with $u|_{\partial\Omega}$ for $u \in C^1(\overline{\Omega})$ and satisfies

$$\|\operatorname{tr} u\|_{L^2(\partial\Omega)} \le C\|u\|_{H^1(\Omega)}.$$

Proof. We show a special case of $\Omega \subset \mathbb{R}^2$ being a convex polygon with $0 \in \Omega$. Moreover, let $B_\rho(0) \subset \Omega$ be a ball of radius ρ contained in Ω. By the divergence theorem, we have for $v \in C^\infty(\overline{\Omega})$

$$\int_{\partial\Omega} v^2\, x \cdot n ds = \int_\Omega \operatorname{div}(v^2 x) dx$$

for all $v \in C^1(\overline{\Omega})$. Denote the outer normal vector to the line segment Γ_j by n_j. Then,

$$x \cdot n_j = |x| \cos\alpha \ge \operatorname{dist}(0, \Gamma_j) \ge \rho.$$

Therefore,

$$\int_{\partial\Omega} v^2\, x \cdot n ds = \sum_j \int_{\Gamma_j} v^2 x \cdot n_j ds \ge \sum_j \int_{\Gamma_j} v^2 \rho ds = \rho \|v\|^2_{L^2(\partial\Omega)}.$$

On the other hand, we may compute

$$\begin{aligned}\int_\Omega \operatorname{div}(v^2 x) dx &= 2\int_\Omega v^2 dx + 2\int_\Omega x \cdot \nabla v v dx \le 2\|v\|^2_{L^2(\Omega)} + 2\operatorname{diam}(\Omega)\int_\Omega |\nabla v|\,|v|\,dx \\ &\le 2\|v\|^2_{L^2(\Omega)} + 2\operatorname{diam}(\Omega)\|v\|_{L^2(\Omega)}\|\nabla v\|_{L^2(\Omega)} \\ &\le C\|v\|_{L^2(\Omega)}\|v\|_{H^1(\Omega)}.\end{aligned}$$

Combining both inequalities, we arrive at

$$\rho\|v\|^2_{L^2(\partial\Omega)} \le C\|v\|_{L^2(\Omega)}\|v\|_{H^1(\Omega)} \le C\|v\|^2_{H^1(\Omega)},$$

which shows continuity of the trace operator mapping from $C^\infty(\overline{\Omega})$ to $L^2(\partial\Omega)$. By density of $C^\infty(\overline{\Omega})$ in $H^1(\Omega)$ this result can be extended to $H^1(\Omega)$. ■

By the previous theorem, every function in $H^1(\Omega)$ has a trace in $L^2(\partial\Omega)$. However, not every $g \in L^2(\partial\Omega)$ is a trace of some $u \in H^1(\Omega)$.
We introduce a stronger space, such that the trace operator is still continuous, and bijective. In fact, we define it as the range of the trace operator

$$H^{1/2}(\partial\Omega) := \{\text{tr } u : u \in H^1(\Omega)\}$$

with the norm

$$\|\text{tr } u\|_{H^{1/2}(\partial\Omega)} = \inf_{\substack{v \in H^1(\Omega) \\ \text{tr } u = \text{tr } v}} \|v\|_{H^1(\Omega)}. \tag{3.5}$$

This is indeed a norm on $H^{1/2}(\partial\Omega)$.

Lemma 3.14. *The space $(H^{1/2}(\partial\Omega), \|.\|_{H^{1/2}(\partial\Omega)})$ is a Banach space. Moreover, for all $g \in H^{1/2}(\partial\Omega)$ there exists an $u \in H^1(\Omega)$ such that $\text{tr } u = g$ and $\|u\|_{H^1(\Omega)} = \|g\|_{H^{1/2}(\partial\Omega)}$.*

Now, we can define an appropriate subspace with zero boundary conditions.

Definition 3.15. *By means of the trace operator, we can define the space*

$$H_0^1(\Omega) = \{u \in H^1(\Omega) : \text{tr } u = 0\}$$

which is a closed subspace of $H^1(\Omega)$. If one requires only zero boundary values on a subset $\Gamma_D \subset \partial\Omega$ (of positive measure), one may define the space

$$H_D^1(\Omega) := \{u \in H^1(\Omega) : \text{tr } u = 0 \text{ on } \Gamma_D\}.$$

We note that constant functions like $u = 1$ belong to $H^1(\Omega)$ for any bounded $\Omega \subset \mathbb{R}^n$, but do not belong to $H_0^1(\Omega)$ by construction of the trace operator.

Integration by parts

For any $u \in C^1(\overline{\Omega})$ and any vector field $\varphi \in [C^1(\overline{\Omega})]^n$, we introduced integration by parts in Applied Mathematics Foundations using the divergence theorem. By definition of the trace operator and density of C^1 in $H^1(\Omega)$ (Meyers-Serrin theorem), this can be extended to functions $u \in H^1(\Omega)$ and vector fields $\varphi \in [H^1(\Omega)]^n$, i.e., there holds

$$\int_\Omega \nabla u \cdot \varphi \, dx = -\int_\Omega u \, \text{div } \varphi \, dx + \int_{\partial\Omega} \text{tr } u \, \varphi \cdot n \, dx.$$

The same can be done for Green's identities, i.e., for $u \in H^2(\Omega)$ and $v \in H^1(\Omega)$, there holds

$$-\int_\Omega \Delta u v \, dx = \int_\Omega \nabla u \cdot \nabla v \, dx - \int_{\partial\Omega} \nabla u \cdot n \, \text{tr } v \, dx.$$

Sobolev spaces over sub-domains

Let Ω consist of M Lipschitz-continuous sub-domains Ω_i such that

- $\overline{\Omega} = \cup_{i=1}^{M} \overline{\Omega}_i$
- $\Omega_i \cap \Omega_j = \emptyset \quad$ if $i \neq j$

The interfaces are $\gamma_{ij} = \overline{\Omega}_i \cap \overline{\Omega}_j$. The outer normal vector of Ω_i is n_i.

Theorem 3.16. *Let $u \in L^2(\Omega)$ such that*

- *$u_i := u|_{\Omega_i}$ is in $H^1(\Omega_i)$, and $g_i = \nabla u_i$ is its weak gradient*
- *the traces on common interfaces coincide:*

$$\mathrm{tr}_{\gamma_{ij}}\, u_i = \mathrm{tr}_{\gamma_{ij}}\, u_j$$

Then, u belongs to $H^1(\Omega)$. Its weak gradient $g = \nabla u$ fulfills $g|_{\Omega_i} = g_i$.

Proof: We have to verify that $g \in L^2(\Omega)^n$, defined by $g|_{\Omega_i} = g_i$, is the weak gradient of u, i.e.,

$$\int_\Omega g \cdot \varphi \, dx = -\int_\Omega u \, \mathrm{div}\, \varphi \, dx \qquad \forall\, \varphi \in [C_0^\infty(\Omega)]^n$$

We are using Green's formula on the sub-domains

$$\begin{aligned}
\int_\Omega g \cdot \varphi \, dx &= \sum_{i=1}^{M} \int_{\Omega_i} g_i \cdot \varphi \, dx = \sum_{i=1}^{M} \int_{\Omega_i} \nabla u_i \cdot \varphi \, dx \\
&= \sum_{i=1}^{M} -\int_{\Omega_i} u_i \, \mathrm{div}\, \varphi \, dx + \int_{\partial\Omega_i} \mathrm{tr}\, u_i \, \varphi \cdot n_i \, ds \\
&= -\int_\Omega u \, \mathrm{div}\, \varphi \, dx + \sum_{\gamma_{ij}} \int_{\gamma_{ij}} \left\{ \mathrm{tr}_{\gamma_{ij}}\, u_i \, \varphi \cdot n_i + \mathrm{tr}_{\gamma_{ij}}\, u_j \, \varphi \cdot n_j \right\} ds \\
&= -\int_\Omega u \, \mathrm{div}\, \varphi \, dx.
\end{aligned}$$

We have used that $\varphi = 0$ on $\partial\Omega$, and $n_i = -n_j$ on γ_{ij}. □

Applications of this theorem are (conforming nodal) finite element spaces of the next section. The partitioning Ω_i is the mesh. On each sub-domain, i.e., on each element T, the functions are polynomials and thus in $H^1(T)$. The finite element functions are constructed to be continuous, i.e., the traces match on the interfaces. Thus, the finite element space is a sub-space of H^1.

Finally, we present a variant of the Poincaré inequality for functions with zero trace.

Theorem 3.17 (Poincaré-Friedrichs inequality). *Let $\Gamma_D \subset \partial\Omega$ with positive measure $|\Gamma_D|$. Then,*

$$\|v\|_{L^2(\Omega)} \leq C \|\nabla v\|_{L^2(\Omega)}$$

for all $v \in H^1_D(\Omega)$.

Proof. Essential identical to the proof of the Poincaré inequality. The only difference is that now the boundary condition $\operatorname{tr} u = 0$ and continuity of the trace operator allows to argue that the function $u = 0$. ■

3.2 The weak formulation of the Poisson equation

We are now able to give a precise definition of the weak formulation of the Poisson problem in $\mathbb{R}^n$. We start with the homogeneous Dirichlet problem.

3.2.1 The Dirichlet problem

Let $\Omega \subset \mathbb{R}^n$ be a bounded Lipschitz domain. For $f \in L^2(\Omega)$, we study

$$\begin{aligned} -\Delta u &= f \qquad \text{in } \Omega \\ u &= 0 \qquad \text{on } \partial\Omega. \end{aligned}$$

In order to describe the weak formulation of the Dirichlet problem, we use the vector space $V := H_0^1(\Omega)$. Multiplication with a smooth test-function in $H_0^1(\Omega)$ and integration by parts leads to the problem of finding $u \in H_0^1(\Omega)$ such that

$$A(u,v) := \int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega fv \, dx =: l(v) \qquad \forall v \in H_0^1(\Omega).$$

The following theorem provides well-posedness of the variational formulation.

Theorem 3.18. *The weak formulation of the Poisson problem*

Find $u \in H_0^1(\Omega)$ such that

$$A(u,v) = l(v) \qquad \forall\, v \in H_0^1(\Omega)$$

has a unique solution u and there holds the bound

$$\|u\|_{H^1(\Omega)} \leq C \|f\|_{L^2(\Omega)}.$$

Proof. We want to apply the Lax-Milgram lemma. In order to do so, we need to check that $A(\cdot,\cdot) : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ is a coercive, continuous bilinear form. By definition of $A(\cdot,\cdot)$ bilinearity is obvious. We check continuity by using the Cauchy-Schwarz inequality:

$$|A(u,v)| = \left| \int_\Omega \nabla u \cdot \nabla v \, dx \right| \leq \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} \leq \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)}.$$

For coercivity, we employ the Poincaré-Friedrich inequality (denoting the constant there by C_P) to derive

$$\|u\|_{H^1(\Omega)}^2 = \|u\|_{L^2(\Omega)}^2 + \|\nabla u\|_{L^2(\Omega)}^2 \leq (C_P^2 + 1) \|\nabla u\|_{L^2(\Omega)}^2 = (C_P^2 + 1) A(u,u).$$

Again, by the Cauchy-Schwarz inequality, we obtain that $l(\cdot)$ is continuous on $H_0^1(\Omega)$, i.e.,

$$|l(v)| \leq \|v\|_{L^2(\Omega)}\|f\|_{L^2(\Omega)} \leq \|v\|_{H^1(\Omega)}\|f\|_{L^2(\Omega)}.$$

Now, the statement of the theorem follows directly from the Lax-Milgram lemma. ■

3.2.2 Other boundary conditions

Let $\Omega \subset \mathbb{R}^n$ be a bounded Lipschitz domain and its boundary $\partial\Omega$ is decomposed as $\partial\Omega = \Gamma_D \cup \Gamma_N \cup \Gamma_R$ according to Dirichlet, Neumann and Robin boundary conditions. We study

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega \\ u &= u_D && \text{on } \Gamma_D \\ \nabla u \cdot n &= u_N && \text{on } \Gamma_N \\ \nabla u \cdot n + \alpha u &= u_R && \text{on } \Gamma_N \end{aligned}$$

For the given data, we assume

- $u_D \in H^{1/2}(\Gamma_D)$, $u_N \in L^2(\Gamma_N)$, $u_R \in L^2(\Gamma_R)$,
- $f \in L_2(\Omega)$, $\alpha : \Gamma_R \to \mathbb{R}^+$ is bounded.

Additionally, there should hold

(a) The Dirichlet part has positive measure $|\Gamma_D| > 0$.

Multiplying the PDE with a test-function $v \in H_D^1(\Omega)$ and integrating by parts, we obtain

$$\int_\Omega \nabla u \cdot \nabla v dx - \int_{\partial\Omega} \nabla u \cdot n \operatorname{tr} v \, ds = \int_\Omega f v \, dx.$$

For the integral over the boundary, we employ the decomposition into the subparts and the given boundary conditions to obtain

$$\begin{aligned} \int_{\partial\Omega} \nabla u \cdot n \operatorname{tr} v \, ds &= \int_{\Gamma_D} \nabla u \cdot n \operatorname{tr} v \, ds + \int_{\Gamma_N} \nabla u \cdot n \operatorname{tr} v \, ds + \int_{\Gamma_R} \nabla u \cdot n \operatorname{tr} v \, ds \\ &= 0 + \int_{\Gamma_N} u_N \operatorname{tr} v \, ds + \int_{\Gamma_R} u_R \operatorname{tr} v \, ds - \int_{\Gamma_R} \alpha \operatorname{tr} u \operatorname{tr} v \, ds. \end{aligned}$$

Therefore the bilinear form for the weak formulation reads as

$$A(u, v) = \int_\Omega \nabla u \cdot \nabla v \, dx + \int_{\Gamma_R} \alpha \operatorname{tr} u \operatorname{tr} v \, ds$$

and the linear form

$$f(v) = \int_\Omega f v \, dx + \int_{\Gamma_N} u_N \operatorname{tr} v \, ds + \int_{\Gamma_R} u_R \operatorname{tr} v \, ds.$$

Exercise 3. Show that $A(\cdot,\cdot)$ is indeed a continuous and coercive bilinear form on $H^1_D(\Omega)$ and that $f(\cdot)$ is a bounded linear form on $H^1_D(\Omega)$. □

Theorem 3.19. *The weak formulation of the Poisson problem*

Find $u \in \{H^1(\Omega) : \text{tr } u = u_D \text{ on } \Gamma_D\}$ *such that*

$$A(u,v) = f(v) \qquad \forall\, v \in H^1_D(\Omega) \tag{3.6}$$

has a unique solution.

Proof. By definition of $H^{1/2}(\Gamma_D)$ there exists a function there exists an $\tilde{u}_D \in \{H^1(\Omega) : \text{tr } u = u_D \text{ on } \Gamma_D\}$ such that

$$\text{tr } \tilde{u}_D = u_D \qquad \text{and} \qquad \|\tilde{u}_D\|_{H^1(\Omega)} \leq C\|u_D\|_{H^{1/2}(\Gamma_D)}.$$

Now, pose the problem: Find $z \in H^1_D(\Omega)$ such that

$$A(z,v) = f(v) - A(\widetilde{u}_D, v) \qquad \forall\, v \in H^1_D(\Omega).$$

The right hand side is the evaluation of the bounded linear form $f(\cdot) - A(\widetilde{u}_D, \cdot)$ on $H^1_D(\Omega)$. Due to Exercise 3, the Lax-Milgram lemma provides a unique solution z. Then, $u := \widetilde{u}_D + z$ solves (3.6). The choice of $\widetilde{u}_D$ is not unique, but, the constructed u is unique. ■

Remark. Using the so called Tartar theorem, one can also show that the bilinear form is coercive provided the Robin term has positive contribution $\int_{\Gamma_R} \alpha\, dx > 0$ in the case of a Dirichlet part with $|\Gamma_D| = 0$. ■

Chapter 4

Finite Element Method

In this chapter we introduce various finite dimensional spaces V_h that can be used in the abstract Galerkin framework of Section 2.

4.1 Finite Elements

We will consider spline spaces of piecewise polynomials based on so called **finite elements**. The basic definition (due to P.Ciarlet) of a finite element is:

Definition 4.1 (Finite element). *A finite element is a triple (T, V_T, Ψ_T), where*

1. *T is a bounded set;*
2. *V_T is a finite dimensional function space on T of dimension $N_T \in \mathbb{N}$;*
3. *$\Psi_T = \{\psi_T^1, \ldots, \psi_T^{N_T}\}$ is a set of linearly independent functionals on V_T.*

In practice, T will be a fixed interval (in 1D) or polygon (triangle, quadrilateral in 2D) or polyhedron (tetrahedron, cuboid in 3D) and V_T will be a space of polynomials on T (so called *local degrees of freedom*). The linear functionals in Ψ_T allow the definition of a local nodal basis $\{\varphi_T^1 \ldots \varphi_T^{N_T}\}$ (also called *local shape functions*) of V_T via

$$\psi_T^i(\varphi_T^j) = \delta_{ij}.$$

The local nodal interpolation operator $I_T : C^m(\overline{T}) \to V_T$ is defined by

$$I_T v := \sum_{j=1}^{N_T} \psi_T^j(v)\varphi_T^j.$$

It is a projection, i.e., for all $v \in V_T$ there holds $I_T v = v$.

Example. In the 1D-example at the beginning of the lecture, we took

1. $T = (0, 1)$;

2. $V_T = P^1(T)$ with $P^1(T)$ being the space of polynomials of maximal degree 1 on T;
3. $\Psi_T = \{\ell_1, \ell_2\}$ with $\ell_1(v) = v(0)$ and $\ell_2(v) = v(1)$.

The corresponding nodal basis consists of the functions $\{1 - x, x\}$. ■

Example. A generalization to 2D can be

1. T is a triangle with vertices $(0, 0)$, $(1, 0)$ and $(0, 1)$;
2. $V_T = P^1(T)$ with $P^1(T)$ being the space of polynomials of maximal degree 1 on T;
3. $\Psi_T = \{\ell_1, \ell_2, \ell_3\}$ with $\ell_1(v) = v(0, 0)$ and $\ell_2(v) = v(1, 0)$ and $\ell_3(v) = v(0, 1)$.

The corresponding nodal basis consists of the functions $\{1 - x - y, x, y\}$. ■

Usual function spaces on $T \subset \mathbb{R}^2$ are

$$\begin{aligned} P^p &:= \text{span}\{x^i y^j : 0 \le i,\, 0 \le j,\, i + j \le p\} \qquad \text{if T is a triangle,} \\ Q^p &:= \text{span}\{x^i y^j : 0 \le i \le p,\, 0 \le j \le p\} \qquad \text{if T is a quadrilateral.} \end{aligned}$$

Remark. The definition of the functionals in Ψ_T directly influences the nodal basis. Finite elements with point evaluation functionals (such as in the previous examples) are called **Lagrange finite elements**. We note that since $\psi_T^j(v) = v(x_j)$ for some node $x_j \in T$, the corresponding local nodal interpolation operator is the classical nodal interpolation (see lecture notes *Numerical Computation*) in the nodes x_j, i.e.,

$$I_T v := \sum_{j=1}^{N_T} v(x_j)\varphi_T^j.$$

Elements using also evaluations of derivatives are called **Hermite finite elements**. ■

In the 1D example, we actually worked with polynomials of the same degrees on sub-intervals. In practice, one actually wants to construct the local shape functions once and reuse them on each sub-interval. The following definition is motivated by this idea.

Definition 4.2. *Two finite elements (T, V_T, Ψ_T) and $(\widehat{T}, V_{\widehat{T}}, \Psi_{\widehat{T}})$ are called* **equivalent**, *if there exists an invertible function F such that*

- $T = F(\widehat{T})$
- $V_T = \{\hat{v} \circ F^{-1} : \hat{v} \in V_{\widehat{T}}\}$
- $\Psi_T = \{\psi_i^T : V_T \to \mathbb{R} : v \to \psi_i^{\hat{T}}(v \circ F)\}$

Two elements are called affine equivalent, if F is an affine-linear function, i.e., $F = A \cdot x + b$.

Lagrangian finite elements defined above are affine equivalent. The Hermite elements are not equivalent. However they are **interpolation equivalent**, meaning that there holds

$$I_T(v) \circ F = I_{\widehat{T}}(v \circ F).$$

Lagrangian finite elements are also interpolation equivalent due to the following lemma.

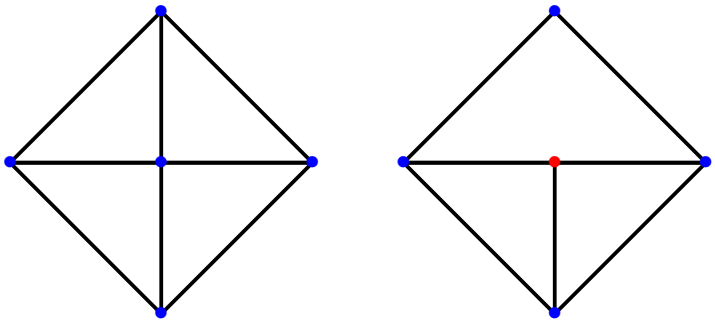
Lemma 4.3. *Equivalent elements are interpolation equivalent.*

Now, we can proceed with the description of the idea in 1D of subdivision of a domain Ω into small pieces.

Definition 4.4. *A regular mesh $\mathcal{T} = \{T_1, \dots, T_M\}$ of a domain $\Omega \subset \mathbb{R}^n$ (with $n \in \{1, 2, 3\}$) is the subdivision of a domain Ω into closed elements T_i such that $\overline{\Omega} = \cup T_i$ and $T_i \cap T_j$ is*

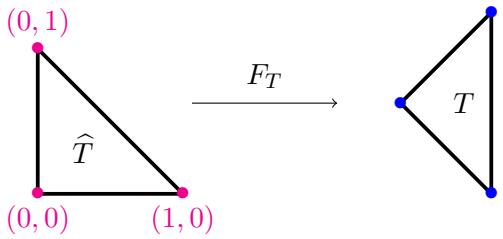
- *either empty*
- *or either exactly one common vertex/edge/face of T_i and T_j*
- *or $T_i = T_j$ in the case $i = j$.*

The condition on the intersection of two elements is called **conformity** and should ensure that there are no so called hanging nodes (see picture below, where a regular trianglulation is on the left an a non-conform triangulation on the right). This is not necessary, but leads to some very useful simplifications (as otherwise nodal bases can not be employed for all nodes).



Remark. We may also call a mesh a **triangulation** as this is the most common case of meshes used in finite element methods. A general mesh may consist of different element shapes such as line segments, triangles, quadrilaterals, tetrahedra, hexhedra, prisms, pyramids. ■

A finite element complex $\{(T, V_T, \Psi_T)\}$ is a set of finite elements defined on the geometric elements of the triangulation $\mathcal{T}$.
It is convenient to construct finite element complexes such that all its finite elements are affine equivalent to one **reference finite element** $(\widehat{T}, \hat{V}_T, \hat{\Psi}_T)$. Then, each element $T \in \mathcal{T}$ is given by using the transformation F_T as $T = F_T(\widehat{T})$.



Example. Using the mesh $\mathcal{T} := \{T_i, i = 1, \dots, N\}$, where $T_i = (x_{i-1}, x_i)$ and $x_i = \frac{i}{N}$ on (0,1), the reference element is given by $\widehat{T} = (0, 1)$ and the element transformations F_{T_i} read as

$$F_{T_i}(t) = x_{i-1} + t \cdot (x_i - x_{i-1}),$$

which is an affine function. ■

Now, one can combine all the local basis functions to a **global basis** and in turn to the space V_h needed for the Galerkin method by gluing the local basis functions together in an appropriate way (i.e., in order to get a subspace of $H^1(\Omega)$, one needs to ensure continuity).
More precisely, the local interpolation operators can be combined to a global interpolation operator $I_{\mathcal{T}}$ defined on $C^m(\overline{\Omega})$ as

$$I_{\mathcal{T}} v_{|T} = I_T v_T \qquad \forall\, T \in \mathcal{T},$$

which allows to define the finite element space.

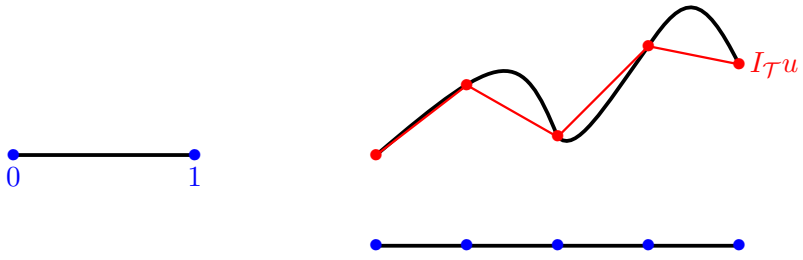
Definition 4.5. *The* **finite element space** *(FEM-space) is given by*

$$V_{\mathcal{T}} := \{v = I_{\mathcal{T}} w : w \in C^m(\overline{\Omega})\}$$

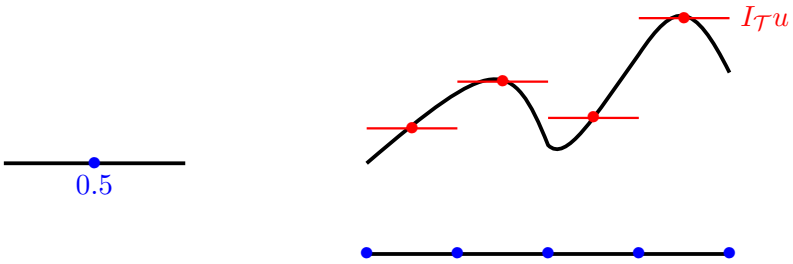
We say that $V_{\mathcal{T}}$ has **regularity** *r, if $V_{\mathcal{T}} \subset C^r(\overline{\Omega})$. If $V_{\mathcal{T}} \nsubseteq C(\Omega)$, the regularity is defined as -1.*

Example.

1. Lagrange finite elements with polynomials of degree 1 in 1D using point evaluations at the end points have regularity 0.



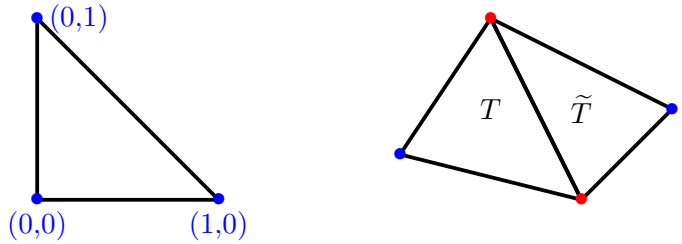
2. Finite elements in 1D using point evaluation in the midpoints, i.e., the reference element $\widehat{T} = (0, 1)$, $\widehat{V}_T = P^0(T)$, $\widehat{\psi}^T(v) = v(1/2)$ have regularity -1.



In the same way taking $\widehat{V}_T = P^1(T)$, $\widehat{\psi}_1^T(v) = v(1/3)$, $\widehat{\psi}_2^T(v) = v(2/3)$ produces discontinuous interpolation and consequently regularity -1.

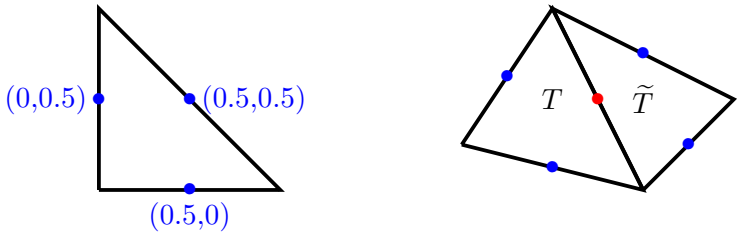
In order to obtain regularity of regularity > 0, one actually needs to use finite elements on the reference elements with point evaluations of the function and higher order derivatives (compare Hermite interpolation).

3. Taking $\widehat{T}$ as the triangle with vertices $(0,0)$, $(1,0)$ and $(0,1)$ and $\widehat{V}_T = P^1(T)$, $\widehat{\psi}_1^T(v) = v(0,0)$, $\widehat{\psi}_2^T(v) = v(1,0)$, $\widehat{\psi}_3^T(v) = v(0,1)$, this produces Lagrangian finite elements in 2D of regularity 0.



One actually obtains regularity 0 (i.e., continuity), since the functionals ψ_i^T and $\psi_j^{\widetilde{T}}$ corresponding to point evaluations at the same point are the same and an affine function (in 2D) is uniquely defined by 2 point values (in the picture the red points)!

4. Taking $\widehat{T}$ as the triangle with vertices $(0,0)$, $(1,0)$ and $(0,1)$ and $\widehat{V}_T = P^1(T)$, $\widehat{\psi}_1^T(v) = v(0.5,0.5)$, $\widehat{\psi}_2^T(v) = v(0.5,0)$, $\widehat{\psi}_3^T(v) = v(0,0.5)$ (i.e. mid point evaluation), this produces finite elements in 2D of regularity -1.



As in this setting neighboring elements only have one common evaluation point, in general, one can not have continuity of the linear interpolation. ■

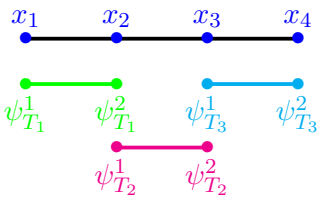
We now want to describe a **global basis** of the FEM-space:
Functionals ψ_T^i and $\psi_{\tilde{T}}^j$ of different elements sitting in the same location (e.g. the point evaluations in the red dots in the previous examples) are equivalent.
Collecting all linearly independent local functionals together gives a set of global functionals $\Psi = \{\psi_1, \dots, \psi_N\}$. The nodal basis for the global finite element space is the basis in $V_{\mathcal{T}}$ dual to the global functionals ψ_j, i.e.,

$$\psi_j(\varphi_i) = \delta_{ij}.$$

Global functionals can be described by local functionals by means of the connectivity matrix $C_T \in \mathbb{R}^{N \times N_T}$ by

$$\Psi_T(u) = C_T^t \Psi(u)$$

Example. We consider $\Omega = (0,1)$ and a grid of 3 subintervals and Lagrange finite elements of degree 1.



From the picture, one can actually see that $\psi_{T_1}^2$ and $\psi_{T_2}^1$ are equivalent and $\psi_{T_2}^2$ and $\psi_{T_3}^1$ are equivalent. A global basis can then be written as

$$\psi = \{\psi_{T_1}^1, \psi_{T_1}^2, \psi_{T_2}^2, \psi_{T_3}^2\} =: \{\psi_1, \psi_2, \psi_3, \psi_4\}.$$

The connectivity matrices for the elements actually read as

$$C_{T_1}^t = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \qquad C_{T_2}^t = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \qquad C_{T_3}^t = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

since e.g.

$$\begin{pmatrix} \psi_{T_1}^1(u) \\ \psi_{T_1}^2(u) \end{pmatrix} = C_{T_1}^t \begin{pmatrix} \psi_1(u) \\ \psi_2(u) \\ \psi_3(u) \\ \psi_4(u) \end{pmatrix}.$$

■

There holds

$$\begin{aligned} \varphi_i|_T &= I_T \varphi_i = \sum_{j=1}^{N_T} \psi_T^j(\varphi_i) \varphi_T^j \\ &= \sum_{j=1}^{N_T} (C_T^t \psi(\varphi_i))_j \varphi_T^j = \sum_{j=1}^{N_T} (C_T^t e_i)_j \varphi_T^j = \sum_{j=1}^{N_T} C_{T,ij} \varphi_T^j. \end{aligned}$$

which means that a global basis function restricted to an element is a combination of local basis function. In the previous example, e.g., one would get $\psi_2|_{T_2} = \psi_{T_2}^1$.

The preceding definition of finite elements and FEM spaces is quite general and can be employed for lots of different problems that require different types of finite elements (e.g., Maxwell's equations require something different than Lagrangian finite elements, so called Nedelec elements). However, the most common finite elements are Lagrangian finite of order 1. Let $\Omega \subset \mathbb{R}^n$ and $\mathcal{T}$ a regular mesh of Ω. Then, one can actually also write the FEM spaces as

$$V_{\mathcal{T}} = \mathcal{P}^1(\mathcal{T}) := \{u \in C(\Omega) \ : \ u|_T \text{ is a polynomial of degree } \leq 1 \forall T \in \mathcal{T}\}.$$

A basis of this space can again be constructed by using hat function characterized by

$$\phi_i(x_j) = \delta_{ij},$$

where x_j are the nodes of the mesh. One can also incorporate homogeneous Dirichlet boundary conditions by defining

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

A basis is again given by hat functions, but only corresponding to interior nodes of the triangulation (i.e. nodes that are not located on the boundary $\partial\Omega$).

4.2 Finite element system assembling

Now that we have a FEM space and a global basis at hand, we can actually rewrite the FEM into a linear system of equations.
The finite element problem reads as

$$\text{Find } u_h \in V_{\mathcal{T}} \text{ such that } : A(u_h, v_h) = f(v_h) \qquad \forall\, v_h \in V_{\mathcal{T}}. \tag{4.1}$$

The nodal basis and the dual functionals provides the one to one relation between $\mathbb{R}^N$ and $V_{\mathcal{T}}$:

$$\mathbb{R}^N \ni \mathbf{u} \leftrightarrow u_h \in V_{\mathcal{T}} \qquad \text{with} \qquad u_h = \sum_{i=1}^{N} \varphi_i \mathbf{u}_i \qquad \text{and} \qquad \mathbf{u}_i = \psi_i(u_h).$$

Using the nodal basis expansion of u_h in (4.1), and testing only with the set of basis functions, one has

$$A\left(\sum_{i=1}^{N} u_i \varphi_i, \varphi_j\right) = f(\varphi_j) \qquad \forall\, j = 1 \dots N$$

With the **stiffness matrix** $\mathbf{A} \in \mathbb{R}^{N \times N}$ and the **load vector** $\mathbf{f} \in \mathbb{R}^N$ defined by

$$\mathbf{A}_{ji} = A(\varphi_i, \varphi_j) \qquad \text{and} \qquad \mathbf{f}_j = f(\varphi_j),$$

the variational formulation of the FEM is equivalent to the linear system of equations

$$\mathbf{A}\mathbf{u} = \mathbf{f}.$$

The preferred way to compute the matrix A and vector f is a sum over element contributions. Consider now the Poisson equation with Robin boundary conditions. Then, the restrictions of the bilinear and linear form to the elements are

$$A_T(u, v) = \int_T \nabla u \cdot \nabla v \, dx + \int_{\partial\Omega \cap T} \alpha \operatorname{tr} u \operatorname{tr} v \, ds$$

and

$$f_T(v) = \int_T f v \, dx + \int_{\partial\Omega \cap T} u_R \operatorname{tr} v \, ds.$$

Then,

$$A(u,v) = \sum_{T \in \mathcal{T}} A_T(u,v) \qquad f(v) = \sum_{T \in \mathcal{T}} f_T(v)$$

On each element, one defines the $N_T \times N_T$ **element matrix** and **element vector** in terms of the local basis on T:

$$\mathbf{A}_{T,k\ell} = A_T(\varphi_T^k, \varphi_T^\ell) \qquad \mathbf{f}_{T,\ell} = \mathbf{f}_T(\varphi_T^\ell).$$

Then, the global matrix and the global vector are computed as

$$\mathbf{A} = \sum_{T \in \mathcal{T}} C_T \mathbf{A}_T C_T^t$$

and

$$\mathbf{f} = \sum_{T \in \mathcal{T}} C_T \mathbf{f}_T.$$

This is true due to the following calculation

$$\begin{aligned} \mathbf{f}_i &= f(\varphi_i) = \sum_{T \in \mathcal{T}} f_T(\varphi_i|_T) = \sum_{T \in \mathcal{T}} f_T(\sum_\ell C_{T,i\ell} \varphi_T^\ell) \\ &= \sum_{T \in \mathcal{T}} \sum_\ell C_{T,i\ell} f_T(\varphi_T^\ell) = \sum_{T \in \mathcal{T}} \sum_\ell C_{T,i\ell} \mathbf{f}_\ell \end{aligned}$$

and

$$\begin{aligned} \mathbf{A}_{ji} &= \sum_{T \in \mathcal{T}} A(\varphi_i|_T, \varphi_j|_T) = \sum_{T \in \mathcal{T}} A\left(\sum_\ell C_{T,i\ell} \varphi_T^\ell, \sum_k C_{T,jk} \varphi_T^k \right) \\ &= \sum_{T \in \mathcal{T}} \sum_\ell \sum_k C_{T,i\ell} \mathbf{A}_{T,\ell k} C_{T,jk}. \end{aligned}$$

On the elements T, the integrands are smooth functions. Thus, numerical integration rules can be applied.

In the case of Dirichlet boundary conditions, let $\gamma_D \subset \{1, \ldots, N\}$ correspond to the vertices x_i at the Dirichlet boundary, and $\gamma_f = \{1, \ldots N\} \setminus \gamma_D$.
We have the equations

$$\sum_{i \in \gamma_D} A_{ji} u_i + \sum_{i \in \gamma_f} A_{ji} u_i = f_j \qquad \forall\, j \in \gamma_f$$

Inserting $u_i = u_D(x_D)$ for $i \in \gamma_i$ results in the reduced system

$$\sum_{i \in \gamma_f} A_{ji} u_i = f_j - \sum_{i \in \gamma_D} A_{ji} u_D(x_i).$$

4.2.1 Assembly in 2D

We are now looking to assemble the stiffness matrix and load vector for the Poisson equation in 2D. By the previous discussion, we actually have to compute the element stiffness matrix, element load vector and the connectivity matrices.

Let $\Omega \subset \mathbb{R}^2$ and $\mathcal{T}$ a regular triangulation of Ω. We assume that our finite element complex is based on the reference element $\widehat{T}$ being the triangle with vertices $(0,0),(1,0),(0,1)$ and consider Lagrangian finite elements of degree 1. This leads to the FEM formulation: Find $u \in \mathcal{P}_0^1(\mathcal{T})$ such that

$$A(u_N, v_N) = \int_\Omega \nabla u_N \cdot \nabla v_N \, dx = \int_\Omega f v_N \, dx = l(v_N) \quad \forall v_N \in \mathcal{P}_0^1(\mathcal{T}).$$

Using the definition of the stiffness matrix and the load vector and the discussion of the preceding subsection, we start with the computation of the element stiffness matrix given by

$$A_T(\varphi_i, \varphi_j) = \int_T \nabla \varphi_i \cdot \nabla \varphi_j \, dx.$$

By definition of the finite element complex, for each $T \in \mathcal{T}$, there exists an affine linear mapping $F_T : \widehat{T} \to T$. Thus, we can use the transformation theorem to transform the element stiffness matrices to the reference element, which gives

$$\int_T \nabla \varphi_i \cdot \nabla \varphi_j \, dx = \int_{\widehat{T}} \nabla \widehat{\varphi_i} \cdot (F_T')^{-1} (F_T')^{-t} \nabla \widehat{\varphi_j} \left| \det F_T' \right| \, dx$$

where we defined $\widehat{\varphi_i} := \varphi_i|_T \circ F_T$ and used the chain rule. The crucial gain of this transformation is that – due to the affine linearity of the maps F_T, we now that $\varphi_i|_T \circ F_T$ has to be a linear function on the reference element (a local basis function). The hat functions on the reference element can be easily written as

$$N_1(x,y) = 1 - x - y, \qquad N_2(x,y) = x, \qquad N_3(x,y) = y,$$

and we obtain that – provided $\operatorname{supp}(\varphi) \cap K \neq \emptyset$ – there holds

$$\varphi_i|_T \circ F_T \in \{N_1, N_2, N_3\}.$$

Therefore, one may compute the integrals (collected into the element stiffness matrix $\mathbf{A}^T \in \mathbb{R}^{3\times 3}$)

$$\mathbf{A}^T_{k\ell} := \int_{\widehat{T}} \nabla N_\ell \cdot (F_T')^{-1} (F_T')^{-t} \nabla N_k \left| \det F_T' \right| \, dx, \qquad \ell, k = 1,2,3$$

and afterwards match each $\varphi_i|_K$ with the corresponding function N_ℓ on K
The same can be done for the load vector, i.e., it can be computed as

$$l(\varphi) = \sum_{T \in \mathcal{T}} \int_T \widehat{f} \widehat{\varphi_i} \left| \det F_T' \right| \, dx.$$

As in the discussion above, one therefore can compute the element load vector $\mathbf{f}^T \in \mathbb{R}^3$ given by

$$\mathbf{f}_i^T = \int_{\widehat{T}} \widehat{f} N_i \left| \det F_T' \right| \, dx.$$

The matching of local and global basis function can either be realized by the connectivity matrices or by a *local-to-global mapping* of indices, i.e. a mapping L_T from $\{1,2,3\}$ to $\{1,\dots,N\}$ such that

$$\varphi_{L_T(i)} \circ F_T = N_i.$$

It remains to specify the element mappings F_T. As F_T is affine linear (i.e. translation and stretching) mapping the reference triangle to the given triangle T, it can be written as

$$F_T(x) = a + Bx \qquad a \in \mathbb{R}^2,\ B \in \mathbb{R}^{2\times 2}.$$

Consequently, the quantities $(F_T')^{-1} = B^{-1}$ and $\det F_T' = \det B$ can directly be computed from knowing B (which in turn can be easily computed from knowing the vertices of T).

Example. Let T be a triangle with vertices $(0,2)$, $(1,1)$ and $(1,3)$. Then, the first step in the determination of F_T is a translation of one vertex to the origin, e.g., a translation by the vector $\begin{pmatrix}0\\-2\end{pmatrix}$. Then, the translated triangle has the vertices $(0,0)$, $(1,-1)$ and $(1,1)$. Now, putting the last to vertices into a matrix and translating back defines the sought transformation as

$$F_T(x) = \begin{pmatrix}1 & 1\\ -1 & 1\end{pmatrix} \cdot \begin{pmatrix}x_1\\x_2\end{pmatrix} + \begin{pmatrix}0\\2\end{pmatrix}.$$

As for the computations of the element stiffness matrix and the element load vector only the derivative is needed, one only needs to compute the matrix B, whose entries are given by differences between vertices of the triangle. ■

Summing up all this steps produces an algorithm with pseudo code as follows:

Algorithm 4.6:

```
input:  mesh 𝒯, data f
initialize:  A := 0, f := 0
for T ∈ 𝒯 do
   compute F_T'
   compute A^T and f^T
   compute the local-to-global mapping L_T (or compute matrices C_T)
   for i=1,2,3
      f_{L_T(i)} := f_{L_T(i)} + f_i^T
      i=1,2,3
         A_{L_T(i)L_T(j)} := A_{L_T(i)L_T(j)} + A_ij^T
      end
   end
end
```

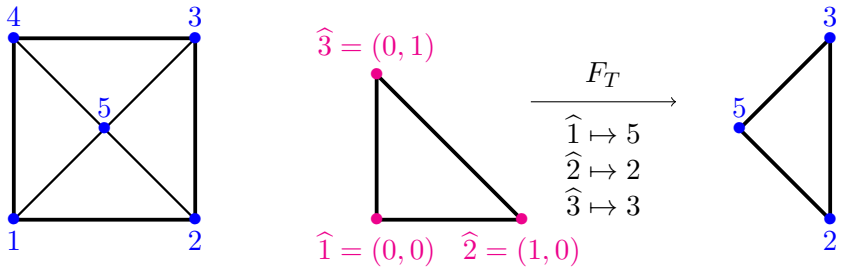
As usual, for Dirichlet boundary conditions lines and columns corresponding to nodal basis functions with nodes on the Dirichlet boundary of the matrix $\mathbf{A}$ can be omitted.

Similar to the 1D case one obtains a sparse matrix, however the resulting matrices $\mathbf{A}$ are not tridiagonal any more!

In practice, the grid information is usually provided by a list of nodes given by their coordinates (in 2D this can be an $N \times 2$-array `coordinates`) and a list of elements, where each element is described by the numbering of its nodes (in 2D this would be an $N_{\mathcal{T}} \times 3$-array `elements`). Now, one can just (in any way you want) map each vertex of each triangle to a vertex of the reference element. This naturally induces the element mapping, as well as the local-to-global mapping.

Example. Taking Ω the unit square and its vertices as well as its center as grid points. Then, this reads as

$$\texttt{coordinates} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 0.5 & 0.5 \end{pmatrix}, \qquad \texttt{elements} = \begin{pmatrix} 1 & 2 & 5 \\ 2 & 3 & 5 \\ 3 & 4 & 5 \\ 4 & 1 & 5 \end{pmatrix}$$



Abbreviating the matrix `coordinates` by c, the element map reads as

$$F_T(x) = \begin{pmatrix} c(2,1) - c(5,1) & c(3,1) - c(5,1) \\ c(2,2) - c(5,2) & c(3,2) - c(5,2) \end{pmatrix} x + \begin{pmatrix} c(5,1) \\ c(5,2) \end{pmatrix} = \begin{pmatrix} 0.5 & 0.5 \\ -0.5 & 0.5 \end{pmatrix} x + \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$$

and the local to global mapping is given by $\widehat{1} \mapsto 5$, $\widehat{2} \mapsto 2$, $\widehat{3} \mapsto 3$.

4.2.2 Higher order Lagrangian finite elements

We now consider finite element complexes based on reference elements $(\widehat{T}, \hat{V}_T, \hat{\Psi}_T)$, where $\widehat{T}$ is the unit triangle and $\hat{V}_T$ is now a space of polynomials with maximal degree p (where $p > 1$) and $\hat{\Psi}_T$ are point evaluation functionals.
Provided the point evaluation functionals are chose such that the regularity is 0, one can, for a mesh $\mathcal{T}$, define the FEM spaces as continuous higher order splines

$$\mathcal{P}^p(\mathcal{T}) := \{u \in C(\Omega) \ : \ u|_T \in P^p(T)\},$$
$$\mathcal{P}^p_0(\mathcal{T}) := \mathcal{P}^p(\mathcal{T}) \cap H^1_0(\Omega),$$

where $P^p(T)$ denotes the space of polynomials of maximal degree p on the triangle T.

We now want to provide a nodal basis for the FEM space.

The 1D case

Fix $p > 1$, then the space $P^p(T)$ has dimension $p+1$. As continuity is required for the global FEM-space, on may construct a basis on the reference element $\widehat{T} = (0,1)$ with the properties:

$$N_1(x) = x, \quad N_2(x) = 1 - x, \quad N_i(1) = N_i(0) = 0 \qquad i = 3, \ldots, p+1.$$

In fact, there are many choices possible for the functions N_i, e.g., $N_i(x) = x(1-x)x^{i-3}$. However, if p gets large, then this choice leads to ill-conditioned matrices. It is more advisable to use the choice

$$N_i(x) = \int_0^x L_{i-2}(t)\, dt,$$

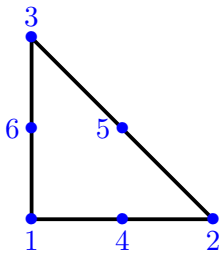
where L_i denotes the i-th Legendre polynomial (see lecture notes *Numerical Computation* for their definition).
One should note that, if higher order basis functions are employed, the element stiffness matrices have to be computed using quadrature rules. Choosing the quadrature order large enough (e.g. using Gaussian quadrature with p quadrature points), this can be done exactly.

The 2D case

We now study the case of a reference triangle $\widehat{T}$ with vertices $(0,0)$, $(1,0)$ and $(0,1)$.

The choice $p = 2$

Take $p = 2$, then the space $P^p(\widehat{T})$ has dimension 6. We also note that a 1D-polynomial of degree 2 is uniquely defined by prescribing 3 nodal values. Therefore, one may fix 3 values on each edge of the reference triangle. Choosing the point evaluation functionals as evaluations at the vertices and mid points of the edges defines a nodal basis of regularity 0 as follows:



$$\begin{aligned}
N_1(x,y) &= 1 - x - y \\
N_2(x,y) &= x \\
N_3(x,y) &= y \\
N_4(x,y) &= x(1-x-y) \\
N_5(x,y) &= xy \\
N_6(x,y) &= y(1-x-y)
\end{aligned}$$

The functions N_1, N_2, N_3 are classical hat functions (polynomial degree 1!) associated with the vertices of the triangle. The functions N_4, N_5, N_6 are chosen such that they vanish on exactly two edges and consequently, an all blue marked points except one, i.e., we have a nodal basis. Therefore, the basis is often written as

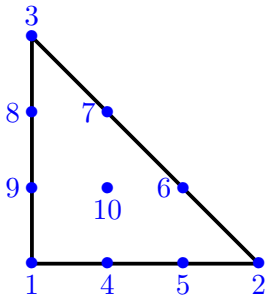
$$\mathcal{B} = \mathcal{B}^v \cup \mathcal{B}^e,$$

where $\mathcal{B}^v$ denote the hat functions and $\mathcal{B}^e$ the remaining basis functions associated with the edges (so called edge form functions). Note also that the basis functions for $p = 2$ consist of hat functions and all products of hat functions (this may be used in implementations).

Finally, we note the taking the mid point of the edge (i.e. choosing the point symmetrically in terms of the edge) for evaluation is not necessary, but makes the implementation significantly more easy!

The choice $p = 3$

Taking $p = 3$ leads to a polynomial space of dimension 10. Following the discussion from above, we take a nodal basis corresponding to the vertices, add now 2 basis functions per edge (in a symmetric) way. This however, only fixes 9 basis functions. The final basis function may now be associated with a note inside the triangle, e.g., the center of mass.



N_1, N_2, N_3 as in cases $p = 1, 2$
N_4, N_5 only non zero on bottom edge
N_6, N_7 only non zero on right edge
N_8, N_9 only non zero on left edge
$N_{10}(x, y) = xy(1 - x - y)$

Here, one may write the basis as

$$\mathcal{B} = \mathcal{B}^v \cup \mathcal{B}^e \cup \mathcal{B}^{\text{int}},$$

where $\mathcal{B}^v$ denote the hat functions, $\mathcal{B}^e$ the edge form functions and $\mathcal{B}^{\text{int}}$ the basis functions associated with interior nodes.

4.3 Finite element error analysis

Let u be the solution of the variational problem, and u_h its Galerkin approximation in the finite element sub-space V_h. Cea's Lemma bounds the finite element error $u - u_h$ by the best approximation error

$$\|u - u_h\|_V \leq C \inf_{v_h \in V_h} \|u - v_h\|_V.$$

The constant factor C is the ratio of continuity and coercivity constant the bilinear form $A(\cdot, \cdot)$. Provided that the solution u is sufficiently smooth, we can take the finite element interpolant to bound the best approximation error:

$$\inf_{v \in V_h} \|u - v_h\|_V \leq \|u - I_{\mathcal{T}} u\|_V$$

In the following, we will bound the interpolation error.
We start with estimating the norm of the element transformations F_T.

Lemma 4.7. *Let $\widehat{T}$ and T be d-dimensional domains related by the invertible affine linear transformation $F_T : \widehat{T} \to T$*

$$F_T(x) = a + B_T x,$$

where $a \in \mathbb{R}^n$ and B_T is a regular matrix in $\mathbb{R}^{n \times n}$. Then, there holds:

$$\|u \circ F_T\|_{L_2(\widehat{T})} = (\det B_T)^{-1/2} \|u\|_{L_2(T)}, \tag{4.2}$$

$$|u \circ F_T|_{H^m(\widehat{T})} \leq C (\det B_T)^{-1/2} \|B_T\|^m |u|_{H^m(T)}. \tag{4.3}$$

Proof: Transformation of integrals, chain rule (exercise). □

In order to estimate the matrices B_T additional assumption have to be made. We define the diameter of the element T, also called **local mesh width**,

$$h_T = \operatorname{diam} T = \sup\{|x - y| \ : \ x, y \in T\}.$$

Definition 4.8. *A triangulation is called* **shape regular***, if all its elements fulfill*

$$|T| \geq \gamma h_T^d$$

with a so called shape-regularity constant $\gamma > 0$ that should be moderate ~ 1.

A triangulation is called **quasi − uniform***, if all elements are essentially of the same size, i.e., there exists one global $h > 0$ such that*

$$h \simeq h_T \qquad \forall\, T \in \mathcal{T}.$$

Shape regularity can also be formulated as bounding the quotient of the local mesh width and the apothem of T (radius of the largest inscribable circle in T). This actually provides a bound for all angles of a triangle/tetrahedron from below, i.e., triangles may not get degenerated in one direction.

If one studies convergence, one considers families of triangulations with decreasing element sizes h_T. In that case, the family of triangulations is called shape regular, if there is a common constant $\gamma > 0$ such that all elements of all triangulations fulfill $|T| \geq \gamma h_T^d$.

Lemma 4.9. *Let $F_T = a + B_T x$ be the mapping from the reference triangle to the triangle T. Let $|T| \geq \gamma h_T^d$. Then, there holds*

$$\begin{aligned} \|B_T\| &\simeq h_T, \\ \|B_T^{-1}\| &\simeq h_T^{-1}. \end{aligned}$$

To bound the finite element interpolation error, we will transform functions from the elements T to the reference element $\widehat{T}$ and use the Bramble-Hilbert lemma.

Theorem 4.10. *Let $\mathcal{T}$ be a shape regular triangulation of Ω and a finite element complex of regularity 0 based on the reference element $(\widehat{T}, V_{\widehat{T}}, \Psi_{\widehat{T}})$, where $V_{\widehat{T}}$ contains at least the polynomials $P^1(\widehat{T})$. Then, there holds*

$$\|v - I_{\mathcal{T}} v\|^2_{L_2(\Omega)} \leq C \sum_{T \in \mathcal{T}} h_T^4 |v|^2_{H^2(T)} \qquad \forall\, v \in H^2(\Omega)$$

$$|v - I_{\mathcal{T}} v|^2_{H^1(\Omega)} \leq C \sum_{T \in \mathcal{T}} h_T^2 \, |v|^2_{H^2(T)} \qquad \forall\, v \in H^2(\Omega)$$

Proof. We prove the H^1 estimate, the L_2 one follows with the same arguments. The interpolation error on each element is transformed to the interpolation error on the reference element. Using the interpolation equivalence of the elements to the reference element gives

$$\begin{aligned} |v - I_{\mathcal{T}} v|^2_{H^1(\Omega)} &= \sum_{T\in\mathcal{T}} |v_T - I_T v_T|^2_{H^1(T)} \\ &\overset{\text{Lem. 4.7}}{\leq} C \sum_{T\in\mathcal{T}} (\det B_T)\, \|B_T^{-1}\|^2\, |(v_T - I_T v_T)\circ F_T|^2_{H^1(\widehat{T})} \\ &= \sum_{T\in\mathcal{T}} (\det B_T) \|B_T^{-1}\|^2 |v_T \circ F_T - I_{\widehat{T}}(v_T\circ F_T)|^2_{H^1(\widehat{T})}. \end{aligned}$$

On the reference element $\widehat{T}$ we apply the Bramble-Hilbert lemma, meaning that $|v_T \circ F_T - I_{\widehat{T}}(v_T \circ F_T)|_{H^1(\widehat{T})} \leq C|v_T\circ F_T|_{H^2(\widehat{T})}$. Then, we transform back to the individual elements, which gives

$$\begin{aligned} |v - I_{\mathcal{T}} v|^2_{H^1(\Omega)} &\leq C \sum_{T\in\mathcal{T}} (\det B_T) \|B_T^{-1}\|^2 |v_T\circ F_T|^2_{H^2(\widehat{T})} \\ &\leq C \sum_{T\in\mathcal{T}} (\det B_T)\, \|B_T^{-1}\|^2\, (\det B_T^{-1})\, \|B_T\|^4\, |v_T|^2_{H^2(T)} \\ &= C \sum_{T\in\mathcal{T}} h_T^2 |v|^2_{H^2(T)}. \end{aligned}$$

This gives the estimate in the H^1-seminorm. ■

If the mesh $\mathcal{T}$ is additionally quasi-uniform with mesh-width h, there hold the interpolation error estimates

$$\begin{aligned} \|u - I_{\mathcal{T}} u\|_{L_2(\Omega)} &\leq Ch^2\, |u|_{H^2(\Omega)} \\ |u - I_{\mathcal{T}} u|_{H^1(\Omega)} &\leq Ch\, |u|_{H^2(\Omega)} \end{aligned}$$

Now, together with the Cea-lemma, we actually have derived convergence rates for the FEM.

Theorem 4.11 (Finite element convergence). *Assume that*

- *the weak solution of the model problem is in $H^2(\Omega)$,*
- *the triangulation $\mathcal{T}$ is quasi-uniform with mesh-size h,*
- *the local finite element spaces contain P^1.*

Then, the finite element error is bounded by

$$\|u - u_h\|_{H^1(\Omega)} \leq Ch\, |u|_{H^2(\Omega)}.$$

The previous theorem shows that, if one decreases the mesh-size $h \to 0$, one obtains convergence of the FEM-solution to the true solution. Moreover, the best possible rate of convergence is order 1. Improving the finite element approximation by reducing the mesh-size is also called **h-FEM**.

Higher order elements

If one uses polynomials of degree > 1, one can obtain faster convergence rates. However, the exact weak solutions have to be smoother (i.e., in higher order Sobolev spaces).

Theorem 4.12. *Assume that*

- *the weak solution of the model problem is in $H^m(\Omega)$ for $m \geq 2$,*
- *the triangulation $\mathcal{T}$ is quasi-uniform with mesh-size h,*
- *the local finite element spaces contain polynomials P^p for $p \geq 1$.*

Then, there holds

$$\|u - u_h\|_{H^1(\Omega)} \leq C h^{\min\{m-1,p\}} p^{1-m} \|u\|_{H^m(\Omega)}.$$

The constant in the estimates depend on the Sobolev index m but not on the polynomial degree and mesh-size.

The previous theorem allows the following observations: Convergence of the FEM solution to the exact solution can be obtained by

- fixing p and taking $h \to 0$, the higher p is fixed, the faster the convergence will be (h-FEM);
- fixing h and sending $p \to \infty$. This is called **p-FEM**.

However, a fair comparison between methods is only given by comparing the rate of convergence with the growth of the dimension of the FEM-space (obviously, a larger FEM space induces a smaller error...)!
For the space $\mathcal{P}^p(\mathcal{T})$ of continuous piecewise polynomials of maximal degree p, one actually has

$$\dim \mathcal{P}^p(\mathcal{T}) \simeq p^d \# \mathcal{T} \simeq p^d h^{-d}.$$

Regularity of weak solutions

The theorems in the previous subsection made the assumption that the weak solution u is in $H^2(\Omega)$ (or $H^m(\Omega)$ for higher order elements). As we have seen, e.g, in the exercises for the 1D-example $u(x) = x^{3/4}(1-x)$ this may not always hold true. The following theorem provides criteria, for which H^2-regularity always hold.

Theorem 4.13 (Shift theorem). *Let u be the weak solution of $-\Delta u = f$ with homogeneous Dirichlet boundary conditions. Then,*

1. *if Ω is a convex set and $f \in L^2(\Omega)$, then $u \in H^2(\Omega)$;*
2. *if Ω is a smooth domain (i.e can be parametrized by a C^∞-function, e.g. a circle), then $f \in H^k(\Omega)$ implies $u \in H^{k+2}(\Omega)$ for all $k \in \mathbb{N}$.*

Proof. Proof for 2D square in exercise. ■

Note that the previous theorem does not cover the case of non-convex polygons (see Figure... below). Polygonal domains are important in applications as most mesh generating software actually produces a polygonal approximation to a given geometry Ω, which then will be used in the FEM-Code. For such domains in 2D, additional information is available.

Assume that $\Omega \subset \mathbb{R}^2$ is a non-convex polygon. This means that there exists at least one corner z_j with an interior angle $\omega_j > \pi$ at the corner (a so called re-entrant corner). Let $\beta > \max\{1 - \frac{\pi}{\omega_j}\}$, where the maximum is taken over all re-entrant corners. Then, the solution u is not in $H^2(\Omega)$, but one can bound a weighted second order Sobolev norm

$$\|r^\beta D^2 u\|_{L_2(\Omega)} < \infty,$$

where $r(x) = \min_j\{\text{dist}(x, z_j)\}$. One may choose a mesh such that

$$h_T \simeq H r_T^\beta, \qquad \forall\, T \in \mathcal{T}$$

where r_T is the distance of the center of the element to the singular corner, and $H \in \mathbb{R}^+$ is a global mesh size parameter. This is called a **graded mesh** with **mesh grading factor** $\widehat{\beta} = \frac{1}{1-\beta}$.

We bound the interpolation error:

$$\begin{aligned} \|u - I_{\mathcal{T}} u\|_{H^1}^2 &\leq C \sum_{T\in\mathcal{T}} h_T^2 |u|_{H^2(T)} \simeq \sum_{T\in\mathcal{T}} H^2 \, |r^\beta D^2 u|_{L_2(T)} \\ &\simeq H^2 \, \|r^\beta D^2 u\|_{L_2(\Omega)}^2 \leq C\, H^2. \end{aligned}$$

The number of elements in the domain can be roughly estimated by the integral over the density of elements. The density is number of elements per unit volume, i.e., the inverse of the area of the element:

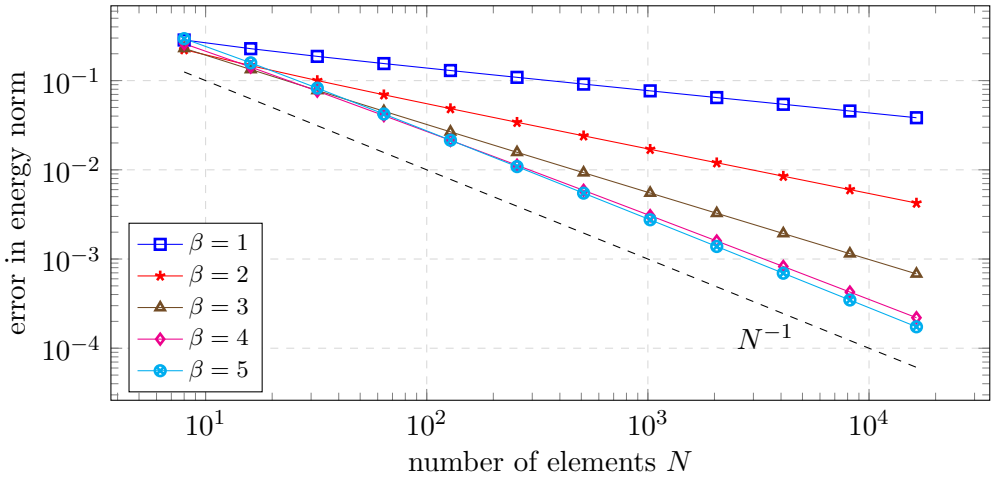
$$N_{el} \simeq \int_\Omega |T|^{-1}\, dx = \int_\Omega H^{-2} r^{-2\beta}\, dx = H^{-2} \int r^{-2\beta}\, dx \simeq C H^{-2}$$

In two dimensions, and $\beta \in (0, 1)$, the integral is finite.
Combining the two estimates, one obtains a relation between the error and the number of elements:

$$\|u - I_{\mathcal{T}} u\|_V^2 \leq C N_{el}^{-1}$$

This is the same order of convergence as in the $H^2(\Omega)$-regular case !

Example. On $\Omega = (0, 1)$, we consider the 1D Poisson equation $-u'' = f$ with homogeneous Dirichlet boundary conditions, where f is given such that the exact solution is given by $u(x) = x^{3/4}(1 - x)$.



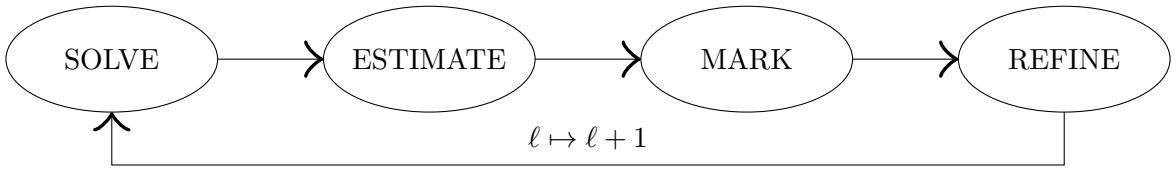
The plot shows the convergence on graded meshes in 1D. In fact, taking a grading factor of $\widehat{\beta} = 4$ gives rate N^{-1}. Grading factors $\widehat{\beta} < 4$ produce reduced convergence rates and grading factors $\widehat{\beta} > 4$ do not improve the rate of N^{-1}.

Chapter 5

Adaptivity

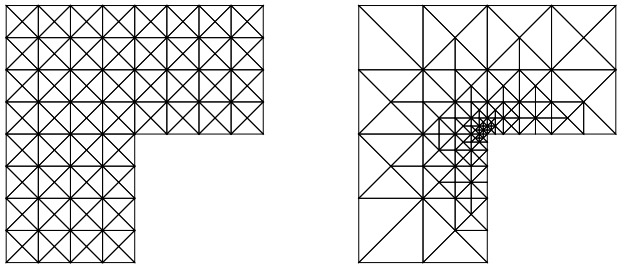
At the end of the last section, we observed that using suitably designed meshes (there so called graded meshes) might improve the speed of convergence, if the solutions to the PDE are not smooth. However, the discussed graded meshes are designed by previously knowledge where the singularities of the solution are. Here, we present an alternative approach that does this in an automatic ("black-box") fashion.

The idea is to start with a mesh $\mathcal{T}_0$ and wrap a loop (with running index ℓ) around the FEM code that does essentially four steps



- SOLVE: Compute your FEM-solution on the current mesh $\mathcal{T}_\ell$.
- ESTIMATE: Compute for each $T \in \mathcal{T}_\ell$ a quantity $\eta_\ell(T)$ that somewhat represents the error on T (note that the true solution in general is unknown, so the true error can not be computed!).
- MARK: Mark the elements in $\mathcal{T}_\ell$ with largest indicators for refinement.
- REFINE: Bisect the (at least) the marked elements.

The module SOLVE is already discussed in the previous section, so we will focus on the remaining three steps in the loop and then write down an implementable adaptive FEM algorithm. A key observation is that by doing the marking step, one may actually use finite element meshes that use different element sizes on different parts of Ω. Thus, you can "zoom in" to regions, where more computational effort is needed (see the following picture).



5.1 A posteriori error estimator

We now discuss the module ESTIMATE in the loop mentioned above. As mentioned, we want to have an estimate for the FEM error that is computable without knowing the true solution.
Such ***a posteriori* error indicators** may use the finite element solution u_h, and input data such as the source term f, i.e., they are functions

$$\eta(u_h, f).$$

Definition 5.1. *An error estimator is called* **reliable**, *if it is an upper bound for the error, i.e., there exists a constant C_1 such that*

$$\|u - u_h\|_V \leq C_1\, \eta(u_h, f). \tag{5.1}$$

An error estimator is **efficient**, *if it is a lower bound for the error, i.e., there exists a constant C_2 such that*

$$\|u - u_h\|_V \geq C_2\, \eta(u_h, f). \tag{5.2}$$

The constants may depend on the domain, and the shape of the triangles, but may not depend on the source term f, or the (unknown) solution u.

The usual error estimators are defined as sum over element contributions:

$$\eta^2(u_h, f) = \sum_{T \in \mathcal{T}} \eta_T^2(u_h, f),$$

where $\eta_T^2(u_h, f)$ are quantities that are only defined on T (and maybe neighboring elements) and somehow should correspond to the local error on T.

In the following, we consider the Poisson equation $-\Delta u = f$ with homogenous Dirichlet boundary conditions $u = 0$ on $\partial\Omega$. We choose piecewise linear Lagrangian finite elements on triangles.

The Zienkiewicz Zhu error estimator

The simplest a posteriori error estimator is the one by Zienkiewicz and Zhu, the so called ZZ error estimator.
We look at the error in the H^1-seminorm

$$\|\nabla(u - u_h)\|_{L^2(\Omega)},$$

where u is the exact solution to the Poisson equation, and u_h is its piecewise linear FEM approximation.
Define the gradient $p = \nabla u$ and the discrete gradient $p_h = \nabla u_h$. The discrete gradient p_h is by definition of the FEM space constant on each element. Let $\tilde{p}_h$ be the piecewise linear and continuous finite element function obtained by averaging the element values of p_h in the vertices:

$$\tilde{p}_h(x_i) = \frac{1}{|\{T : x_i \in T\}|} \sum_{T: x_i \in T} p_{h|T} \qquad \text{for all vertices } x_i.$$

Note that we only defined the nodal values of $\tilde{p}_h$, but a piecewise linear function in the FEM spaces is uniquely defined by the nodal values.
The hope is that the averaged gradient is a much better approximation to the true gradient, i.e.,

$$\|p - \tilde{p}_h\|_{L^2(\Omega)} \leq \alpha \, \|p - p_h\|_{L^2(\Omega)} \tag{5.3}$$

holds with a small constant $\alpha \ll 1$. This property is known as *super-convergence*. It is indeed true on (locally) uniform meshes and smooth right-hand sides f.
The ZZ error estimator replaces the true gradient (which is unknown) in the error $p - p_h$ by the good approximation $\tilde{p}_h$:

$$\eta(u_h) = \|\tilde{p}_h - p_h\|_{L^2(\Omega)}$$

If the super-convergence property (5.3) is fulfilled, than the ZZ error estimator is reliable:

$$\begin{aligned} \|\nabla(u - u_h)\|_{L^2(\Omega)} &= \|p - p_h\|_{L^2(\Omega)} \leq \|p_h - \widetilde{p}_h\|_{L^2(\Omega)} + \|p - \widetilde{p}_h\|_{L^2(\Omega)} \\ &\leq \|p_h - \widetilde{p}_h\|_{L^2(\Omega)} + \alpha \|p - p_h\|_{L^2(\Omega)}, \end{aligned}$$

and

$$\|\nabla(u - u_h)\|_{L^2(\Omega)} \leq \frac{1}{1-\alpha} \|p_h - \widetilde{p}_h\|_{L^2(\Omega)}.$$

It is also efficient, due to a similar short application of the triangle inequality. By definition of the L^2-norm, one can also easily deduce the local element contributions

$$\eta(u_h)^2 = \|\tilde{p}_h - p_h\|^2_{L^2(\Omega)} = \sum_{T \in \mathcal{T}} \|\tilde{p}_h - p_h\|^2_{L^2(T)} =: \sum_{T \in \mathcal{T}} \eta_T^2(u_h)^2.$$

There is a rigorous analysis of the ZZ error estimator, e.g., by showing equivalence to the following residual error estimator.

The residual error estimator

The idea is to compute the residual of the Poisson equation

$$f + \Delta \, u_h,$$

in the natural norm $H^{-1}(\Omega)$ (the norm of the dual space of $H^1_0(\Omega)$). The classical Δ-operator cannot be applied to u_h, since the first derivatives, ∇u_h, are non-continuous across element boundaries. One can compute the residuals on the elements

$$f_{|T} + \Delta \, u_{h|T} \qquad \forall \, T \in \mathcal{T},$$

and one can also compute the violation of the continuity of the gradients on the edge $E = T_1 \cap T_2$. We define the normal-jump term

$$\left[\frac{\partial u_h}{\partial n}\right] := \frac{\partial u_h}{\partial n_1}|_{T_1} + \frac{\partial u_h}{\partial n_2}|_{T_2}.$$

The residual error estimator is

$$\eta^{res}(u_h, f)^2 := \sum_T \eta_T^{res}(u_h, f)^2$$

with the element contributions

$$\eta_T^{res}(u_h, f)^2 := h_T^2 \|f + \Delta u_h\|_{L^2(T)}^2 + \sum_{\substack{E: E \subset T \\ E \subset \Omega}} h_E \left\| \left[\frac{\partial u_h}{\partial n}\right] \right\|_{L^2(E)}^2.$$

The scaling with h_T comes from replacing the dual norm in H^{-1} (which is not computable) by the "equivalent" weighted L^2-norm.

Theorem 5.2. *The residual error estimator is reliable:*

$$\|u - u_h\|_V \leq C_1 \eta^{res}(u_h, f).$$

If the source term f is a piecewise polynomial on the mesh $\mathcal{T}$, then the error estimator η^{res} is also efficient:

$$\|u - u_h\|_V \geq C_2 \eta^{res}(u_h, f).$$

We note that there are many more possible choices for error indicators (e.g. equilibrated flux or multilevel indicators).

5.1.1 Marking strategies

In the marking step, a set $\mathcal{M}_\ell \subset \mathcal{T}_\ell$ of elements of the mesh $\mathcal{T}_\ell$ is chosen based on the local estimator contributions η_T. We present two different strategies for that.

The first idea is to select elements such that the corresponding local error indicators make up (at least) a fixed percentage of the total error indicator, which is called **Dörfler marking**. More precisely, for $\theta \in (0, 1)$, one seeks a minimal set $\mathcal{M}_\ell$ such that

$$\theta \eta^2 \leq \sum_{T \in \mathcal{M}_\ell} \eta_T^2(u_h, f).$$

A way to obtain this minimal set is to sort all element indicators $\eta_T^2(u_h, f)$ from largest to smallest and add elements to the sum on the right-hand side until the criterium is fulfilled.

The second strategy is the so called **bulk criterion**, which, for $\theta \in (0, 1)$, marks all elements $T \in \mathcal{T}_\ell$ that satisfy

$$\theta \max\{\eta_{T'}(u_h, f) \ : \ T' \in \mathcal{T}\} \leq \eta_T(u_h, f),$$

i.e., all elements that are bigger than θ times the largest element are refined.

Remark. The choice of marking parameter θ is a balancing act: Choosing θ very small means that only few elements are refined and therefore more steps in the loop may be needed to obtain a sought tolerance. However, choosing θ close to 1 means that you get closer to uniform refinement, which is inefficient for non-smooth solutions. In practice, a typical choice is $\theta = 0.25$.

5.1.2 Refinement algorithms

The mesh refinement algorithm has to take care of

- generating a sequence of regular meshes;
- generating a sequence of shape regular meshes.

Especially the requirement of keeping the shape-regularity constant γ (roughly) the same independent of the done refinements is challenging.

In the following, we present refinement strategies based on so called edge-refinement. The idea hereby is to follow the rules:

- If an element T is marked for refinement, all edges of the element are marked.
- If a non-marked element has an edge that is marked (e.g. it is sharing an edge to a marked element), we also mark its *longest edge* for refinement.

Then, each element $T \in \mathcal{T}_\ell$ will be refined according to the following rules: Edges will always be halved and:

- If no edge of T is marked, the element will not be refined, i.e., $T \in \mathcal{T}_{\ell+1}$;
- If all edges of T are marked, we use so called **red-refinement**, where the triangle is decomposed into 4 triangles;
- If one edge of T is marked, we use so called **green-refinement**, where the triangle is dexomposed into 2 triangles;
- If two edges of T are marked, we use so called **blue-refinement**, where the triangle is split into 3 triangles.

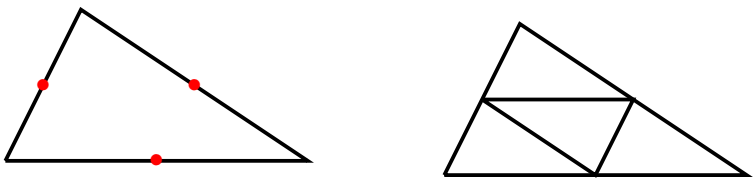


Figure 5.1: Red-refinement.

The following lemma shows that RGB refinement produces regular and shape regular triangulations.

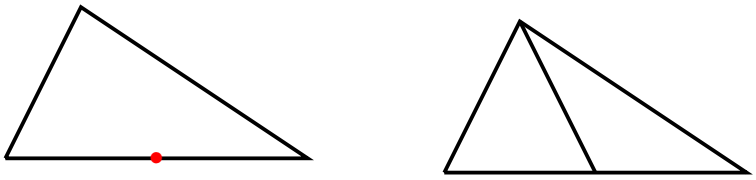


Figure 5.2: Green-refinement.

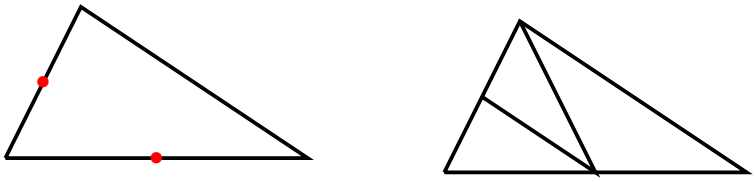


Figure 5.3: Blue-refinement.

Lemma 5.3. *Let $\mathcal{T}_0$ be a regular triangulation and $\widehat{\gamma}$ be the smallest angle of all triangles in $\mathcal{T}_0$. Let $\mathcal{T}_\ell$ be a mesh obtained by ℓ-steps of RGB refinement of some (arbitrary) marked elements. Then, $\mathcal{T}_\ell$ is regular and the smallest angle of all triangles is at least $\widehat{\gamma}/2$.*

Newest vertex bisection:
We start with an initial triangulation and, for each element, we label an edge of the triangle for refinement, e.g., the longest edge, to be the *reference edge*.
As previously done for the RGB refinement, we mark all edges of elements that are marked by the marking step.
Then, refinement is done as follows:

- The midpoint of the reference edge will become a new node in $\mathcal{T}_{\ell+1}$ and the triangle will be split into two sons T_1, T_2;
- The reference edges of the new triangles are opposite to the newly generated vertex;
- If more than one edge of an element is marked, we do this process until all marked edges are refined;

In the end some non-marked elements may also have to be refined to keep a mesh without hanging nodes.

Both of these refinement algorithms are also possible in 3D.

There are other possible refinement algorithms (e.g. red-green refinement or longest edge bisection). Red-green-blue refinement is rather simple to implement, but newest vertex bisection has some big theoretical advantages (like that refinement increases the amout of similar triangles - e.g. triangles with the same angles - by a fixed factor of 4).

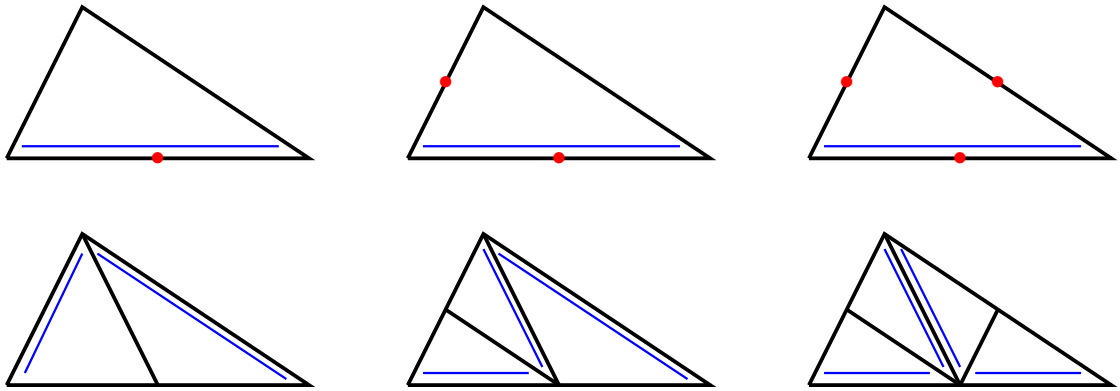


Figure 5.4: Newest vertex bisection,reference edges in blue (top initial triangle, bottom refined triangles).

5.1.3 The adaptive algorithm

A posteriori error estimates are used to control recursive mesh refinement:

Algorithm 5.4:

Input: initial mesh $\mathcal{T}_0$, tolerance TOL, marking parameter $\theta \in (0, 1]$
$\ell := 0$
Loop over ℓ
    compute FEM solution u_ℓ on $\mathcal{T}_\ell$
    compute local error contributions $\eta_T(u_h, f)$
    sum the local contributions $\eta^2 = \sum_T \eta_T^2(u_h, f)$
    if $\eta \leq$ TOL then stop
    choose a marking set $\mathcal{M}_\ell \subset \mathcal{T}_\ell$
    generate the mesh $\mathcal{T}_{\ell+1}$ by refining all elements in $\mathcal{M}_\ell$ (and maybe some more
      to remove hanging nodes)
    $\ell := \ell + 1$
end

For the classical FEM, convergence of the method followed directly from the error estimate for $h \to 0$, which in fact used, that all elements will get sufficiently small as long as h gets small. For adaptive FEM this is not obvious.

Theorem 5.5. *Let η_ℓ be either the ZZ-estimator or the residual estimator and assume that marking is either done by Dörfler marking or the bulk criterion. Then, the approximation u_ℓ computed with the adaptive algorithm converges to the exact solution, i.e.,*

$$\|u - u_\ell\|_{H^1(\Omega)} \to 0 \quad \text{for } \ell \to \infty.$$

For adaptive FEM one can even show that (under some more assumptions) that the convergence happens with the *optimal algebraic rate*. This means that, even if the solution $u \notin H^2(\Omega)$, the AFEM approximation with first order Lagrangian finite elements converges with rate 1. Under the same assumptions as for the rate optimality one can also show that the overal computational complexity of the method is optimal (provided you have a good solver for the linear system of equations).

Chapter 6

Mixed Formulations

6.1 Inf-sup stable variational problems

In the previous discussions, we studied continuous and coercive bilinear forms, for which the Lax-Milgram Lemma provides a unique solution. However, The coercivity condition is by no means a necessary condition for a stable solvable system.

Example. Let $V = \mathbb{R}^2$ and define

$$B(u, v) = u_1 v_1 - u_2 v_2.$$

Then, $B(\cdot, \cdot)$ is not coercive, but the solution to the variational formulation $B(u, v) = f^T v$ for all $v \in \mathbb{R}^2$ exists and is $u_1 = f_1$ and $u_2 = -f_2$. ■

In the following, we will follow the convention to call coercive bilinear forms $A(\cdot, \cdot)$, and the more general ones $B(\cdot, \cdot)$. Moreover, we also allow ansatz- and test-space to be different!

Definition 6.1. *Let V and W be Hilbert spaces. We call a bilinear form $B(\cdot, \cdot) : V \times W \to \mathbb{R}$* ***continuous****, if*

$$B(u, v) \leq \beta_2 \|u\|_V \|v\|_W \qquad \forall\, u \in V,\ \forall\, v \in W. \tag{6.1}$$

We say that $B(\cdot, \cdot)$ satisfies the **inf-sup condition***, if there is a constant $\beta_1 > 0$ such that*

$$\inf_{\substack{u \in V \\ u \neq 0}} \sup_{\substack{v \in W \\ v \neq 0}} \frac{B(u, v)}{\|u\|_V \, \|v\|_W} \geq \beta_1. \tag{6.2}$$

By definition, we have that coercivity implies the inf-sup condition.

Reformulating the problem as an operator equation by defining the linear operator $B : V \to W^*$ by $\langle Bu, v \rangle_{W^* \times W} = B(u, v)$, the inf-sup condition reads as

$$\sup_{v \in W} \frac{\langle Bu, v \rangle}{\|v\|_W} \geq \beta_1 \|u\|_V, \qquad \forall\, u \in V.$$

We immediately obtain that B is one to one (injective), since

$$Bu = 0 \Rightarrow u = 0$$

However, the inf-sup condition (6.2) does not imply that B is onto W^*. To insure that, we can pose an inf-sup condition the other way around:

$$\inf_{\substack{v\in W\\ v\neq 0}} \sup_{\substack{u\in V\\ u\neq 0}} \frac{B(u,v)}{\|u\|_V \, \|v\|_W} \geq \beta_2. \tag{6.3}$$

It will be sufficient to impose the following weaker condition:

Definition 6.2. *Let V and W be Hilbert spaces. We say that a bilinear form $B(\cdot,\cdot)$ satisfies the* **non-degeneracy condition**, *if*

$$\sup_{\substack{u\in V\\ u\neq 0}} \frac{B(u,v)}{\|u\|_V \, \|v\|_W} > 0 \qquad \forall\, v \in W. \tag{6.4}$$

We note that coercivity also implies the non-degeneracy condition.

We have motivated the following theorem for existence and uniqueness of weak solutions.

Theorem 6.3. *Assume that the continuous bilinear form $B(\cdot,\cdot)$ fulfills the inf-sup condition (6.2) and condition (6.4). Then, the variational problem: find $u \in V$ such that*

$$B(u,v) = f(v) \qquad \forall\, v \in W \tag{6.5}$$

has a unique solution. The solution depends continuously on the right hand side:

$$\|u\|_V \leq \beta_1^{-1} \|f\|_{W^*}$$

Example. We check the inf-sup and non-degeneracy condition to the continuous bilinear form $B(u,v) = u_1v_1 - u_2v_2$ with $V = W = \mathbb{R}^2$. In this case, checking (6.2) and (6.3) is the same and consequently the non-degeneracy condition follows from the inf-sup condition. We compute

$$\begin{aligned}
\inf_{\substack{u\in V\\ u\neq 0}} \sup_{\substack{v\in W\\ v\neq 0}} \frac{B(u,v)}{\|u\|_V \, \|v\|_W} &= \inf_{\substack{u\in V\\ u\neq 0}} \sup_{\substack{v\in W\\ v\neq 0}} \frac{u_1v_1 - u_2v_2}{\|u\|_V \, \|v\|_W} \\
&\geq \inf_{\substack{u\in V\\ u\neq 0}} \frac{u_1^2 + u_2^2}{\|(u_1,u_2)\|_V \, \|(u_1,-u_2)\|_W} = 1,
\end{aligned}$$

where we chose $v = (u_1, -u_2)$ in the penultimate step. ■

6.1.1 Approximation of inf-sup stable variational problems

Again, to approximate (6.5), we pick finite dimensional subspaces $V_h \subset V$ and $W_h \subset W$, and pose the finite dimensional variational problem: find $u_h \in V_h$ such that

$$B(u_h, v_h) = f(v_h) \qquad \forall\, v_h \in W_h.$$

But now, in contrast to the coercive case, the solvability of the finite dimensional equation does not follow from the solvability conditions of the original problem on $V \times W$.

Example. We continue with the previous example in $\mathbb{R}^2$. Taking the subspace $V_h = W_h = \text{span}\{(1,1)\}$, i.e., all points in $\mathbb{R}^2$ lying on the line $y = x$. Then, we have

$$B(u_h, v_h) = u_1 v_1 - u_2 v_2 = u_1 v_1 - u_1 v_1 = 0$$

for all $u_h, v_h \in V_h$. Therefore, if f is not zero, the problem is not solvable. ■

In order to obtain a solvable formulation, we have to pose an extra inf-sup condition for the discrete problem:

$$\inf_{\substack{u_h \in V_h \\ u_h \neq 0}} \sup_{\substack{v_h \in W_h \\ v_h \neq 0}} \frac{B(u_h, v_h)}{\|u_h\|_V \, \|v_h\|_W} \geq \beta_{1h}. \tag{6.6}$$

On a finite dimensional space, one to one is equivalent to onto, and therefore the discrete inf-sup condition already implies the non-degeneracy condition, which we therefore can omit.

As in the previous section, we directly obtain unique solvability provided the discrete inf-sup condition holds. In fact, one also obtains a Cea-Lemma (quasi-optimality) for such problems.

Theorem 6.4. *Assume that $B(\cdot,\cdot)$ is continuous with bound β_2, and $B(\cdot,\cdot)$ fulfills the discrete inf-sup condition with bound β_{1h}. Then, there holds the quasi-optimal error estimate*

$$\|u - u_h\|_V \leq (1 + \beta_2/\beta_{1h}) \inf_{v_h \in V_h} \|u - v_h\|_V \tag{6.7}$$

Proof. Again, there holds the Galerkin orthogonality $B(u, w_h) = B(u_h, w_h)$ for all $w_h \in V_h$. Choose an arbitrary $v_h \in V_h$ and estimate

$$\begin{aligned}
\|u - u_h\|_V &\leq \|u - v_h\|_V + \|v_h - u_h\|_V \\
&\leq \|u - v_h\|_V + \beta_{1h}^{-1} \sup_{w_h \in W_h} \frac{B(v_h - u_h, w_h)}{\|w_h\|_V} \\
&= \|u - v_h\|_V + \beta_{1h}^{-1} \sup_{w_h \in W_h} \frac{B(v_h - u, w_h)}{\|w_h\|_V} \\
&\leq \|u - v_h\|_V + \beta_{1h}^{-1} \sup_{w_h \in W_h} \frac{\beta_2 \|v_h - u\|_V \, \|w_h\|_W}{\|w_h\|_W} \\
&= (1 + \beta_2/\beta_{1h}) \|u - v_h\|_V.
\end{aligned}$$

■

6.2 Mixed Methods

A mixed method is a variational formulation involving two function spaces and a bilinear-form of a special saddle point structure. Usually, it is obtained from variational problems *with constraints*. The motivation behind the following formulations is again given by energy minimization under constraints. Seek the minimum of

$$J(v) := \frac{1}{2}a(v,v) - f(v)$$

in a Hilbert space V under the side constraint

$$b(v,q) = g(q) \qquad \text{for all } q \in Q$$

where Q is also a Hilbert space, $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ are bilinear forms and $f(\cdot), g(\cdot)$ are linear forms. The solution of this minimization problem actually satisfies the system of variational equations, which can be obtained by employing the method of Lagrangian multipliers, i.e., one searches for stationary points (which turn out to be saddle points, exercise!) of the functional

$$L(u,\lambda) := J(u) + b(u,\lambda) - g(\lambda).$$

This leads to the problem of finding $u \in V$ and $p \in Q$ such that

$$\begin{array}{lcccll} a(u,v) & + & b(v,p) & = & f(v) & \forall\, v \in V, \\ b(u,q) & & & = & g(q) & \forall\, q \in Q. \end{array} \tag{6.8}$$

6.2.1 Weak formulation of Neumann problem

We start with the Poisson problem

$$-\Delta u = f \qquad \text{in } \Omega, \tag{6.9}$$

and boundary conditions

$$\frac{\partial u}{\partial n} = 0 \quad \text{on } \partial\Omega.$$

As already observed in the exercises, the weak formulation

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v dx \qquad \forall v \in H^1(\Omega)$$

is not uniquely solvable. However, it is uniquely solvable, if one imposes the side constraint $\int_\Omega u \, dx = 0$. Instead of adding this constraint to the function space $H^1(\Omega)$, one could also define the bilinear form

$$b(\cdot,\cdot) : H^1(\Omega) \times \mathbb{R},\ b(u,q) = q \int_\Omega u \, dx$$

and use the saddle point formulation motivated above, i.e., seek $(u,p) \in H^1(\Omega) \times \mathbb{R}$ such that

$$\begin{array}{lcccll} \int_\Omega \nabla u \cdot \nabla v \, dx & + & p \int_\Omega v \, dx & = & \int_\Omega f v dx & \forall\, v \in H^1(\Omega), \\ q \int_\Omega u dx & & & = & 0 & \forall\, q \in \mathbb{R}. \end{array} \tag{6.10}$$

If $(u,0)$ solves the previous system, we actually have found the unique solution of the Neumann problem that satisfies the side-constraint.

We note that a similar formulation can also be made to impose non-homogeneous Dirichlet boundary conditions (exercise!).

6.2.2 A mixed method for the flux

Again, we consider the Poisson problem now with mixed Dirichlet-Neumann boundary conditions

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega, \\ u &= u_D && \text{on } \Gamma_D \\ \frac{\partial u}{\partial n} &= g && \text{on } \Gamma_N. \end{aligned}$$

Recalling that $\Delta u = \operatorname{div}(\nabla u)$, we introduce the flux variable $\sigma := \nabla u$ to rewrite the equations as: Find u and σ such that

$$\sigma - \nabla u = 0, \tag{6.11}$$
$$\operatorname{div} \sigma = -f, \tag{6.12}$$

and boundary conditions

$$\begin{aligned} u &= u_D && \text{on } \Gamma_D \\ \sigma \cdot n &= g && \text{on } \Gamma_N. \end{aligned}$$

We want to derive a variational formulation for the system of equations. For that, we multiply the first equations by vector-valued test functions τ, the second equation by test functions v, and integrate:

$$\begin{aligned} \textstyle\int_\Omega \sigma \cdot \tau \, dx \quad - \quad \int_\Omega \tau \cdot \nabla u \, dx &= 0 && \forall \tau \\ \textstyle\int_\Omega \operatorname{div} \sigma \, v \, dx \qquad\qquad\quad &= -\int_\Omega f v \, dx && \forall v. \end{aligned}$$

We would like to have the second term of the first equation of the same structure as the first term in the second equation (as it is required by the saddle-point formulation in the motivation). This can be obtained by integration by parts applied to either one of them. The interesting case is to integrate by parts in the first line to obtain:

$$\int_\Omega \sigma \cdot \tau \, dx + \int_\Omega \operatorname{div} \tau \, u \, dx - \int_{\Gamma_D} \tau \cdot n \, u \, ds - \int_{\Gamma_N} \tau \cdot n \, u \, ds = 0.$$

Here, we make use of the boundary conditions. On the Dirichlet boundary, we know $u = u_D$, and use that in the equation. The Neumann boundary condition $\sigma \cdot n = g$ must be put into the space V to be fixed later. Thus, it is enough to choose test functions of the sub-space fulfilling $\tau \cdot n = 0$. The problem is now the following: Find $\sigma \in V, \sigma \cdot n = g$ on Γ_N, and $u \in Q$ such that

$$\begin{aligned} \textstyle\int_\Omega \sigma \cdot \tau \, dx \quad + \quad \int_\Omega \operatorname{div} \tau \, u \, dx &= \int_{\Gamma_D} u_D \tau \cdot n \, ds && \forall \tau : \ \tau \cdot n = 0 \text{ on } \Gamma_N \\ \textstyle\int_\Omega \operatorname{div} \sigma \, v \, dx \qquad\qquad &= -\int_\Omega f v \, dx && \forall v \end{aligned}$$

The derivatives are put onto the flux unknown σ (and its test function τ). We don't have to derive the primal unknown u. This will give us better approximation for the fluxes than for the scalar u. That is one of the reasons to use this mixed method.

6.3 Abstract theory

In the following we briefly present results regarding unique solvability of mixed formulations. A mixed variational formulation involves two Hilbert spaces V and Q, bilinear-forms

$$a(u,v) : V \times V \to \mathbb{R}, \qquad b(u,q) : V \times Q \to \mathbb{R},$$

and continuous linear-forms

$$f(v) : V \to \mathbb{R}, \qquad g(q) : Q \to \mathbb{R}.$$

The problem is to find $u \in V$ and $p \in Q$ such that

$$\begin{array}{rcccll} a(u,v) & + & b(v,p) & = & f(v) & \forall\, v \in V, \\ b(u,q) & & & = & g(q) & \forall\, q \in Q. \end{array} \tag{6.13}$$

The two examples from above are of this form.

Instead of considering this as a system of equations, one can look at the mixed method as one variational problem on the product spaces $V \times Q$. For this, simply add both lines, and search for $(u,p) \in V \times Q$ such that

$$a(u,v) + b(u,q) + b(v,p) = f(v) + g(q) \qquad \forall\, (v,q) \in V \times Q.$$

Define the big bilinear-form $B(\cdot,\cdot) : (V \times Q) \times (V \times Q) \to \mathbb{R}$ as

$$B((u,p),(v,q)) = a(u,v) + b(u,q) + b(v,p),$$

to write the whole system as single variational problem

$$\text{Find } (u,p) \in V \times Q : \qquad B((u,p),(v,q)) = f(v) + g(q) \qquad \forall\, (v,q) \in V \times Q.$$

In the interesting examples, the bilinear form $b(\cdot,\cdot)$ has a kernel, defined as

$$V_0 := \{v \in V \ : \ b(v,q) = 0 \ \ \forall q \in Q\}.$$

Example. For the case of the Neumann problem for the Poisson equation, the bilinear form

$$b(v,q) = q \int_\Omega v \, dx$$

has the kernel $V_0 = \overline{H}^1(\Omega) := \{v \in H^1(\Omega) \ : \ \int_\Omega v \, dx = 0\}$. ■

Now, we will give conditions to ensure a unique solution of a mixed problem. In fact, one could apply Theorem 6.3 to the "big" bilinear form $B(\cdot,\cdot)$. However, this would not take advantage of the special structure of the given system, which is used in the following theorem to devise weaker assumptions.

Theorem 6.5 (Brezzi's theorem). *Assume that $a : V \times V \to \mathbb{R}$ and $b : V \times Q \to \mathbb{R}$ are continuous bilinear-forms. Assume there holds coercivity of $a(\cdot,\cdot)$ on the kernel of $b(\cdot,\cdot)$, i.e.,*

$$a(u,u) \geq \alpha_1 \|u\|_V^2 \qquad \forall\, u \in V_0, \tag{6.14}$$

and there holds the LBB (Ladyshenskaja-Babuška-Brezzi) condition

$$\sup_{u \in V} \frac{b(u,q)}{\|u\|_V} \geq \beta_1 \|q\|_Q \quad \forall\, q \in Q. \tag{6.15}$$

Then, the mixed problem is uniquely solvable. The solution fulfills the stability estimate

$$\|u\|_V + \|p\|_Q \leq c\{\|f\|_{V^*} + \|g\|_{Q^*}\},$$

with the constant c depending on α_1, β_1 and the continuity constants of $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$.

6.4 Analysis of the model problems

Now, we apply the abstract framework to the two model problems.

6.4.1 Weak formulation of the Neumann problem

The problem is well posed for the spaces

$$V = H^1(\Omega) \qquad \text{and} \qquad Q = \mathbb{R}$$

The bilinear-forms are

$$\begin{aligned} a(u,v) &= \int_\Omega \nabla u \cdot \nabla v \, dx \\ b(u,q) &= (q,u)_{L^2(\Omega)} = q\,(1,u)_{L^2(\Omega)}\,. \end{aligned}$$

Theorem 6.6. *The mixed problem* (6.10) *has a unique solution $u \in H^1(\Omega)$ and $\lambda \in \mathbb{R}$.*

Proof. The spaces V and Q, and the bilinear-forms $a(.,.)$ and $b(\cdot,\cdot)$ fulfill the assumptions of Theorem 6.5. The kernel space V_0 is given by the previous example as H^1-functions with vanishing integral.
The continuity of $a(\cdot,\cdot)$ on V is clear. Coercivity on V does not hold, but, due to Friedrichs inequality, coercivity on V_0 holds true.
The bilinear-form $b(\cdot,\cdot)$ is continuous on $V \times Q$:

$$|b(u,q)| = |q||\,(1,u)_{L^2(\Omega)}\,| \leq C|q|\|u\|_{L^2(\Omega)} \leq C\|q\|_Q\|u\|_V.$$

The LBB - condition of $b(\cdot,\cdot)$ also follows easily by taking $u \equiv q$. Then, $\|u\|_{H^1(\Omega)} = |q||\Omega|^{1/2}$, and

$$\|q\|_Q = |q| = \frac{q \cdot q}{|q|} = \frac{1}{|\Omega|^{1/2}} \frac{b(u,q)}{\|u\|_V}.$$

■

6.4.2 Mixed method for the fluxes

Looking at the weak formulation for the mixed method for the flux, one observes that actually the function σ should be in $L^2(\Omega)^d$ and additionally only its divergence has to be in $L^2(\Omega)$.
This motivates the function space $H(\mathrm{div}, \Omega)$:

Definition 6.7. *A measurable function g is called the* **weak divergence** *of σ on $\Omega \subset \mathbb{R}^d$, if there holds*

$$\int_\Omega g\,\varphi\,dx = -\int_\Omega \sigma \cdot \nabla\varphi\,dx \qquad \forall\,\varphi \in C_0^\infty(\Omega).$$

The **function space** $H(\mathrm{div}, \Omega)$ *is defined as*

$$H(\mathrm{div}, \Omega) := \{\sigma \in [L_2(\Omega)]^d : \mathrm{div}\ \sigma \in L^2(\Omega)\},$$

its norms is

$$\|\sigma\|_{H(\mathrm{div},\Omega)} = \left(\|\sigma\|^2_{L^2(\Omega)} + \|\,\mathrm{div}\ \sigma\|^2_{L^2(\Omega)}\right)^{1/2}.$$

The mixed method is now formulated on the spaces

$$V = H(\mathrm{div}, \Omega) \qquad Q = L^2(\Omega)$$

and the bilinear-forms are

$$\begin{aligned} a(\sigma,\tau) &= \int_\Omega \sigma\tau\,dx \qquad \forall\,\sigma,\tau \in V \\ b(\sigma, v) &= \int_\Omega \mathrm{div}(\sigma)\,v\,dx \qquad \forall\,\sigma \in V,\ \forall\,v \in Q. \end{aligned}$$

Theorem 6.8. *The mixed problem for the fluxes is well posed.*

Proof. We check the conditions of the theorem of Brezzi: The bilinear-forms are bounded:

$$a(\sigma,\tau) = \int_\Omega \sigma\tau\,dx \le \|\sigma\|_{L^2(\Omega)}\,\|\tau\|_{L^2(\Omega)} \le C\|\sigma\|_V\,\|\tau\|_V$$

and

$$b(\sigma, v) = \int_\Omega \mathrm{div}\ \sigma v\,dx \le \|\,\mathrm{div}\,\sigma\|_{L^2(\Omega)}\,\|v\|_{L^2(\Omega)} \le \|\sigma\|_V\,\|v\|_Q.$$

The kernel space $V_0 = \{\tau : b(\tau, v) = 0\ \forall\, v \in Q\}$ is

$$V_0 = \{\tau \in H(\mathrm{div}, \Omega) : \mathrm{div}\ \tau = 0\}.$$

There holds ellipticity of $a(\cdot,\cdot)$ on V_0: Let $\tau \in V_0$, then

$$a(\tau,\tau) = \int_\Omega |\tau|^2\,dx = \|\tau\|^2_{H(\mathrm{div},\Omega)}.$$

We are left to verify the LBB condition

$$\sup_{\sigma\in H(\mathrm{div},\Omega)} \frac{\int_\Omega \mathrm{div}\ \sigma\, v\,dx}{\|\sigma\|_{H(\mathrm{div},\Omega)}} \geq C\|v\|_{L^2(\Omega)} \qquad \forall\, v \in L^2(\Omega). \tag{6.16}$$

For given $v \in L^2(\Omega)$, we will construct a flux σ satisfying the inequality. For this, we solve the artificial Poisson problem $-\Delta\varphi = v$ with Dirichlet boundary conditions $\varphi = 0$ on $\partial\Omega$. The solution satisfies $\|\nabla\varphi\|_{L^2(\Omega)} \leq C\|v\|_{L^2(\Omega)}$. Set $\sigma = -\nabla\varphi$. There holds $\mathrm{div}\ \sigma = v$. Its norm is

$$\|\sigma\|^2_{H(\mathrm{div},\Omega)} = \|\sigma\|^2_{L^2(\Omega)} + \|\,\mathrm{div}\ \sigma\|^2_{L^2(\Omega)} = \|\nabla\varphi\|^2_{L^2(\Omega)} + \|v\|^2_{L^2(\Omega)} \leq C\|v\|^2_{L^2(\Omega)}.$$

Using it in (6.16), we get the result

$$\frac{\int_\Omega \mathrm{div}\ \sigma\, v\,dx}{\|\sigma\|_{H(\mathrm{div},\Omega)}} = \frac{\int_\Omega v^2\,dx}{\|\sigma\|_{H(\mathrm{div},\Omega)}} \geq C\|v\|_{L^2(\Omega)}.$$

Now Brezzi's theorem gives the statement. ■

6.4.3 The function space $H(\mathrm{div},\Omega)$

The mixed formulation has motivated the definition of the function space $H(\mathrm{div},\Omega)$. In the following, we will study some properties of this space and afterwards construct finite elements for the approximation of functions in $H(\mathrm{div},\Omega)$.

For a function in $H^1(\Omega)$, the boundary values are well defined by the trace operator. For a vector-valued function in $H(\mathrm{div},\Omega)$, only the normal-component is well defined on the boundary:

Theorem 6.9. *There exists a normal-trace operator*

$$\mathrm{tr}_n : H(\mathrm{div}) \to (H^{1/2}(\partial\Omega))^*$$

such that for $\sigma \in H(\mathrm{div},\Omega) \cap [C(\overline{\Omega})]^d$ it coincides with its normal component

$$\mathrm{tr}_n\,\sigma = \sigma\cdot n \qquad \text{on } \partial\Omega.$$

Lemma 6.10. *There holds integration by parts*

$$\int_\Omega \sigma\cdot\nabla\varphi\,dx + \int_\Omega (\mathrm{div}\ \sigma)\varphi\,dx = \langle \mathrm{tr}_n\,\sigma, \mathrm{tr}\,\varphi\rangle$$

for all $\sigma \in H(\mathrm{div})$ and $\varphi \in H^1(\Omega)$.

We have seen that for $\Omega = \overline{\Omega_1} \cup \overline{\Omega_2}$ being decomposed in at least two subsets, $H^1(\Omega)$-functions are obtained by being in $H^1(\Omega_i)$ and continuous across the common interface $\overline{\Omega_1} \cap \overline{\Omega_2}$. A similar property holds for functions in $H(\mathrm{div}, \Omega)$.

Lemma 6.11. *Let $\sigma \in [L_2(\Omega)]^d$ such that $\sigma|_{\Omega_i} \in H(\mathrm{div}, \Omega_i)$ and*

$$\mathrm{tr}_{n,i}\, \sigma|_{\Omega_i} = -\,\mathrm{tr}_{n,j}\, \sigma|_{\Omega_j} \qquad \text{on } \overline{\Omega_1} \cap \overline{\Omega_2}.$$

Then $\sigma \in H(\mathrm{div}, \Omega)$, and

$$(\mathrm{div}\ \sigma)|_{\Omega_i} = \mathrm{div}\,(\sigma|_{\Omega_i}).$$

Now, we want to define proper finite elements for this space. The previous characterization by sub-domains allows the definition of finite element sub-spaces of $H(\mathrm{div}, \Omega)$ as one only has to take care of continuity in the normal direction of each edge.

The cheapest element for $H(\mathrm{div}, \Omega)$ is the lowest order **Raviart-Thomas element RT0**. The finite element $(T, V_T, \{\psi_i\})$ with $T \subset \mathbb{R}^2$ being a triangle is defined by

$$V_T = \left\{ \begin{pmatrix} a \\ b \end{pmatrix} + c \begin{pmatrix} x \\ y \end{pmatrix} : a, b, c \in \mathbb{R} \right\},$$

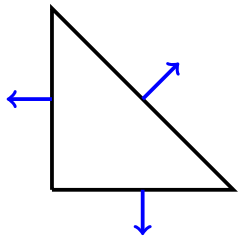
with the linear functionals are the integrals of the normal components on the three edges e_i of the triangle

$$\psi_i(\sigma) = \int_{e_i} \sigma \cdot n \, ds \qquad i = 1, 2, 3.$$

The three functionals are linearly independent on V_T. This means, for each choice of $\sigma_1, \sigma_2, \sigma_3$, there exists three unique numbers $a, b, c \in \mathbb{R}$ such that

$$\sigma = \begin{pmatrix} a \\ b \end{pmatrix} + c \begin{pmatrix} x \\ y \end{pmatrix}$$

satisfies $\psi_i(\sigma) = \sigma_i$.



Exercise 4. Compute the corresponding "nodal" basis functions for the RT0 - element on the reference triangle. □

The global finite element functions are defined as follows. Given one value σ_i for each edge e_i of the triangulation. The corresponding RT0 finite element function σ is defined by

$$\sigma|_T \in V_T \qquad \text{and} \qquad \int_{e_i} \sigma|_T \cdot n_{e_i} \, ds = \sigma_i$$

for all edges $e_i \subset T$ and all triangles $T \in \mathcal{T}$.

We have to verify that this construction gives a function in $H(\mathrm{div}, \Omega)$. For each element, $\sigma|_T$ is a linear polynomial, and thus in $H(\mathrm{div}, T)$. The normal components must be continuous. By construction, there holds

$$\int_e \sigma|_{T_i} \cdot n \, ds = \int_e \sigma|_{T_j} \cdot n \, ds$$

for the edge $e = T_i \cap T_j$. The normal component is continuous since $\sigma \cdot n_e$ is constant on an edge: Points (x, y) on the edge e fulfill $xn_x + yn_y$ is constant. There holds

$$\sigma \cdot n_e = \left[\begin{pmatrix} a \\ b \end{pmatrix} + c \begin{pmatrix} x \\ y \end{pmatrix} \right] \cdot \begin{pmatrix} n_x \\ n_y \end{pmatrix} = an_x + bn_y + c(xn_x + yn_y) = \text{constant}.$$

The global RT0-basis functions φ_i^{RT} are associated to the edges e_i, and satisfy

$$\int_{e_i} \varphi_j^{RT} \cdot n_e \, ds = \delta_{ij} \qquad \forall \, i, j = 1, \ldots N_{edges}.$$

By this basis, we can define the Raviart-Thomas - interpolation operator

$$I_h^{RT} \sigma = \sum_{edges \; e_i} \left(\int_{e_i} \sigma \cdot n_e \, ds \right) \varphi_i^{RT}.$$

It is a projection on V_h. The interpolation operator preserves the divergence *in mean*:

Lemma 6.12. *The RT0 - interpolation operator satisfies*

$$\int_T \mathrm{div} \, I_h^{RT} \sigma \, dx = \int_T \mathrm{div} \, \sigma dx$$

for all triangles $T \in \mathcal{T}$.

Let P_h be the $L^2(\Omega)$-orthogonal projection onto piecewise constant finite element functions. This is: Let $Q_h = \{q \in L^2(\Omega) : q|_T = \text{const} \; \forall T \in \mathcal{T}\}$. Then, $P_h p$ is defined by $P_h p \in Q_h$ and

$$\int_\Omega P_h p \, q_h \, dx = \int_\Omega p \, q_h \, dx \qquad \forall \, q_h \in Q_h.$$

In fact, this is equivalent to $P_h p$ satisfies $P_h p \in Q_h$ and

$$\int_T P_h p \, dx = \int_T p \, dx \qquad \forall \, T \in \mathcal{T}.$$

The Raviart-Thomas finite elements are piecewise linear. Thus, the divergence is piecewise constant. From $\mathrm{div} \, I_h \sigma \in Q_h$ and Lemma 6.12 there follows

$$\mathrm{div} \; I_h \sigma = P_h \, \mathrm{div} \; \sigma.$$

This relation is known as commuting diagram property:

$$\begin{array}{ccc} H(\mathrm{div}) & \xrightarrow{\mathrm{div}} & L^2 \\ \big\downarrow I_h & & \big\downarrow P_h \\ V_h^{RT} & \xrightarrow{\mathrm{div}} & Q_h \end{array} \tag{6.17}$$

The analysis of the approximation error is based on the transformation to the reference element. For $H^1(\Omega)$-finite elements, interpolation on the element T is equivalent to interpolation on the reference element $\widehat{T}$, i.e., $(I_h v) \circ F_T = \hat{I}_h(v \circ F_T)$. This is not true for the $H(\mathrm{div}, \Omega)$ elements: The transformation F changes the direction of the normal vector. Thus $\int_e \sigma \cdot n \, ds \neq \int_{\hat{e}} \hat{\sigma} \cdot \hat{n} \, ds$.
The Piola transformation is the remedy:

Definition 6.13 (Piola Transformation). *Let $F : \widehat{T} \to T$ be the mapping from the reference element $\widehat{T}$ to the element T. Let $\hat{\sigma} \in L^2(\widehat{T})$. Then, the Piola transformation*

$$\sigma = \mathcal{P}(\hat{\sigma})$$

is defined by

$$\sigma(F(\hat{x})) = (\det DF)^{-1} DF \hat{\sigma}(\hat{x}).$$

The functionals $\psi_i(\sigma) = \int_e \sigma \cdot n \, ds$ are preserved by the Piola transformation:

Lemma 6.14. *Let $\hat{\sigma} \in H(\mathrm{div}, \widehat{T})$, P be the Piola transformation and $\sigma = \mathcal{P}(\hat{\sigma})$. Then, there holds*

$$(\mathrm{div}\ \sigma)(F(\hat{x})) = (\det F')^{-1} \,\mathrm{div}\ \hat{\sigma}.$$

Let $\hat{e}$ be an edge of the reference element, and $e = F(\hat{e})$. Then,

$$\int_e \sigma \cdot n \, ds = \int_{\hat{e}} \hat{\sigma} \cdot \hat{n} \, ds.$$

Proof. Essentially transformation theorem, exercise! ■

Lemma 6.15. *The Raviart-Thomas triangle T and the Raviart-Thomas reference triangle are interpolation equivalent:*

$$I_h^{RT} \mathcal{P}(\hat{\sigma}) = \mathcal{P}(\widehat{I}_h^{RT} \hat{\sigma})$$

Proof. The element spaces are equivalent, i.e., $V_T = \mathcal{P}\{V_{\widehat{T}}\}$, and the functionals are preserved by the Piola transformation. ■

For the error, one actually obtains convergence of order 1.

Theorem 6.16. *The Raviart-Thomas interpolation operator satisfies the approximation properties*

$$\|\sigma - I_h^{RT}\sigma\|_{L^2(\Omega)} \le Ch\,\|\nabla\sigma\|_{L^2(\Omega)},$$
$$\|\operatorname{div}\sigma - \operatorname{div} I_h^{RT}\sigma\|_{L^2(\Omega)} \le Ch\,\|\nabla\operatorname{div}\sigma\|_{L^2(\Omega)}.$$

Proof. Transformation to the reference element, using that the interpolation preserves constant polynomials, and the Bramble Hilbert lemma. The estimate for the divergence uses the commuting diagram property

$$\|\operatorname{div}(\sigma - I_h^{RT}\sigma)\|_{L^2(\Omega)} = \|(I-P_h)\operatorname{div}\sigma\|_{L^2(\Omega)} \le Ch\,\|\nabla\operatorname{div}\sigma\|_{L^2(\Omega)}.$$

■

6.5 Approximation of mixed systems

We apply a Galerkin-approximation for the mixed system. For this, we choose (finite element) sub-spaces $V_h \subset V$ and $Q_h \subset Q$, and define the Galerkin approximation $(u_h, p_h) \in V_h \times Q_h$ by

$$B((u_h,p_h),(v_h,q_h)) = f(v_h) + g(q_h) \qquad \forall\, v_h \in V_h\ \forall\, q_h \in Q_h.$$

Now, Theorem 6.4 together with the observation that on finite dimensional spaces the inf-sup condition implies the non-degeneracy condition provides a best-approximation under the assumption of the discrete inf-sup condition

$$\inf_{v\in V_h, q\in Q_h}\sup_{u\in V_h, p\in Q_h}\frac{B((u,p),(v,q))}{(\|v\|_V+\|q\|_Q)(\|u\|_V+\|p\|_Q)} \ge \beta. \tag{6.18}$$

for $B(\cdot,\cdot)$, i.e.,

$$\|u-u_h\|_V + \|p-p_h\|_Q \le C \inf_{v_h\in V_h, q_h\in Q_h}\{\|u-v_h\|_V + \|p-q_h\|_Q\}.$$

However, the inf-sup stability on the continuous level $V\times Q$ does not imply the discrete stability! Usually, one checks the conditions of Brezzi on the discrete level to prove stability of $B(\cdot,\cdot)$ on the discrete levels. The continuity of $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ are inherited from the continuous levels. The stability conditions have to be checked separately: The discrete ellipticity of $a(\cdot,\cdot)$ on the kernel of $b(\cdot,\cdot)$ reads as

$$a(v_h,v_h) \ge C\|v_h\|_V^2 \qquad \forall\, v_h \in V_{0h} := \{v_h \in V_h : b(v_h,q_h) = 0\ \forall\, q_h \in Q_h\}, \tag{6.19}$$

and the discrete LBB condition reads as

$$\sup_{u_h\in V_h}\frac{b(u_h,q_h)}{\|u_h\|_V} \ge C\|q_h\|_Q \qquad \forall\, q_h \in Q_h. \tag{6.20}$$

The discrete LBB condition is posed for less dual variables q_h in $Q_h \subset Q$, but the space in the supremum is also smaller. It does not follow from the LBB condition on the continuous levels.

There is a canonical technique to derive the discrete LBB condition from the continuous one:

Lemma 6.17. *Assume there exists an (interpolation) operator, the so called* **Fortin operator**, $\Pi_h : V \to V_h$ *that is continuous*

$$\|\Pi_h v\|_V \leq C\|v\|_V \qquad \forall\, v \in V,$$

and satisfies

$$b(\Pi_h v, q_h) = b(v, q_h) \qquad \forall\, q_h \in Q_h.$$

Then, the continuous LBB condition implies the discrete one.

Proof. For all $p_h \in Q_h$ there holds

$$\sup_{v_h \in V_h} \frac{b(v_h, p_h)}{\|v_h\|_V} \geq \sup_{v \in V} \frac{b(\Pi_h v, p_h)}{\|\Pi_h v\|_V} \geq C \sup_{v \in V} \frac{b(v, p_h)}{\|v\|_V} \geq C\|p_h\|_Q.$$

■

6.5.1 Approximation of the mixed method for the flux

We choose the pair of finite element spaces, the Raviart Thomas spaces for the flux

$$V_h = \{v \in H(\mathrm{div}, \Omega) : v|_T \in V_T^{\mathrm{RT}}\} \quad \subset \quad V = H(\mathrm{div}, \Omega)$$

and the space of piecewise constants for the scalar

$$Q_h = \{q \in L^2(\Omega) : q|_T \in P^0(T)\} \subset Q = L^2(\Omega).$$

Pose the discrete mixed problem: Find $(\sigma_h, u_h) \in V_h \times Q_h$ such that

$$\begin{array}{lcll} \int_\Omega \sigma_h \cdot \tau_h \, dx \;+\; \int_\Omega \mathrm{div}\, \tau_h \, u_h \, dx & = & \int_{\Gamma_D} u_D \tau_n \, ds & \forall\, \tau_h \in V_h \\ \int_\Omega \mathrm{div}\, \sigma_h \, v_h \, dx & = & -\int_\Omega f v_h \, dx & \forall\, v_h \in Q_h. \end{array} \tag{6.21}$$

Lemma 6.18 (Discrete Stability). *The discrete mixed variational problem (6.21) is well posed.*

Proof. By Brezzi's theorem. Continuity of the bilinear-form and the linear-form follow from the continuous level. We prove the kernel ellipticity: Since

$$\mathrm{div}\; V_h \subset Q_h,$$

there holds

$$\int_\Omega \mathrm{div}\; \sigma_h \, q_h \, dx = 0 \quad \forall\, q_h \in Q_h \quad \Rightarrow \quad \mathrm{div}\; \sigma_h = 0,$$

and thus $V_{0h} \subset V_0$. In this special case, the discrete kernel ellipticity is simple the restriction of the continuous one to V_{0h}. We are left with the discrete LBB condition. We would like to apply Lemma 6.17. The Fortin operator is the Raviart-Thomas interpolation operator I_h^{RT}. The abstract condition

$$b(I_h^{RT}\sigma, v_h) = b(\sigma, v_h) \qquad v_h \in Q_h$$

reads as

$$\int_T \operatorname{div} I_h^{RT}\sigma \, dx = \int_T \operatorname{div} \sigma dx,$$

which was proven in Lemma 6.12. (Technically, the operator is not continuous in $H(\operatorname{div}, \Omega)$, but on its subspace $H^1(\Omega)^d$, which suffices.) ■

Finally, we present an error estimate.

Theorem 6.19 (A priori estimate). *The mixed finite element method for the flux satisfies the error estimates*

$$\|\sigma - \sigma_h\|_{L^2(\Omega)} + \|\operatorname{div}(\sigma - \sigma_h)\|_{L^2(\Omega)} + \|u - u_h\|_{L^2(\Omega)} \leq Ch \left(\|\sigma\|_{H^1(\Omega)} + \|u\|_{H^1(\Omega)} + \|f\|_{H^1(\Omega)}\right). \tag{6.22}$$

Proof. By discrete stability, one can bound the discretization error by the best approximation error

$$\|\sigma - \sigma_h\|_{H(\operatorname{div},\Omega)} + \|u - u_h\|_{L^2(\Omega)} \leq C \inf_{\substack{\tau_h \in V_h \\ v_h \in Q_h}} \left(\|\sigma - \tau_h\|_{H(\operatorname{div},\Omega)} + \|u - v_h\|_{L^2(\Omega)}\right).$$

The best approximation error is bounded by the interpolation error. The first term is (using the commuting diagram property and $\operatorname{div}\sigma = f$)

$$\begin{aligned}\inf_{\tau_h \in V_h} \left(\|\sigma - \tau_h\|_{L^2(\Omega)} + \|\operatorname{div}(\sigma - \tau_h)\|_{L^2(\Omega)}\right) &\leq \|\sigma - I_h^{RT}\sigma\|_{L^2(\Omega)} + \|(I - P^0)\operatorname{div}\sigma\|_{L^2(\Omega)} \\ &\leq Ch \left(\|\sigma\|_{H^1(\Omega)} + \|f\|_{H^1(\Omega)}\right).\end{aligned}$$

The second term is

$$\inf_{v_h \in Q_h} \|u - v_h\|_{L^2(\Omega)} \leq \|u - P^0 u\|_{L^2(\Omega)} \leq Ch \, \|u\|_{H^1(\Omega)},$$

which shows the first order estimate. ■

The smoothness requirements onto the solution of (6.22) are fulfilled for problems on convex domains. There holds $\|u\|_{H^2} \leq C\|f\|_{L_2}$. Since $\sigma = \nabla u$, there follows $\|\sigma\|_{H^1(\Omega)} \leq C\|f\|_{L^2(\Omega)}$. The mixed method requires more smoothness onto the right hand side data, $f \in H^1(\Omega)$. It can be reduced to $H^1(\Omega)$ on sub-domains, what is a realistic assumption.

Chapter 7

Applications of finite elements

We investigate finite element methods for some other PDEs.

7.1 The Stokes Equation

The Stokes equation simulating a stationary incompressible Newtonian fluid (with high viscosities and low Reynolds numbers) is given by

$$\begin{aligned} -\mu\Delta u + \nabla p &= f, \\ \operatorname{div} u &= 0, \end{aligned}$$

where $u : \Omega \to \mathbb{R}^d$ is the velocity field, $p : \Omega \to \mathbb{R}$ is the pressure and $\mu \in \mathbb{R}$ is the viscosity constant. A common type of boundary conditions are no-slip conditions, i.e., homogeneous Dirichlet boundary conditions for the velocity field.

For simplicity, we set $\mu = 1$ in the following. The weak formulation reads as: Find $u \in [H_0^1(\Omega)]^d$ and $p \in L^2(\Omega)$ such that

$$\begin{array}{rcccll} \int_\Omega \nabla u : \nabla v \, dx & + & \int_\Omega \operatorname{div} v \, p \, dx & = & \int_\Omega f v \, dx & \forall\, v \in [H_0^1(\Omega)]^d \\ \int \operatorname{div} u \, q & & & = & 0 & \forall\, q \in L^2(\Omega). \end{array} \tag{7.1}$$

Note that here ∇u is a matrix and the product $\nabla u : \nabla v$ is understood as inner product for matrices, i.e., $\nabla u : \nabla v := \sum_{i,j} (\nabla u)_{ij} (\nabla v)_{ij}$.
The pressure p is only unique up to a constant, which we fix by changing the function space for p into

$$L_0^2(\Omega) := \{ q \in L^2(\Omega) \ : \ \int_\Omega q \, dx = 0 \}.$$

7.1.1 Stability of the continuous equation

Solvability follows from Brezzi's theorem. The only non-trivial part is the LBB condition:

$$\sup_{v \in [H_0^1(\Omega)]^d} \frac{\int_\Omega \operatorname{div} v \, p \, dx}{\|v\|_{H^1(\Omega)}} \geq \beta \|p\|_{L^2(\Omega)} \qquad \forall\, p \in L_0^2(\Omega).$$

Proof. We sketch a proof: The LBB condition becomes simple if we skip the Dirichlet conditions:

$$\sup_{v\in[H^1(\Omega)]^d} \frac{\int_\Omega \operatorname{div} v\, p\, dx}{\|v\|_{H^1(\Omega)}} \geq \beta\|p\|_{L^2(\Omega)} \qquad \forall\, p \in L^2_0(\Omega).$$

Take $p \in L^2(\Omega)$ and extend it by 0 to $L^2(\mathbb{R}^d)$. Now compute a right-inverse of the div-operator via Fourier transform: Since $\operatorname{div} v = p$ translates to $i\xi \cdot \hat{v} = \hat{p}$, we have

$$\begin{aligned} \hat{v}(\xi) &= \frac{-i\xi}{|\xi|^2}\hat{p}(\xi) \\ v(x) &= \mathcal{F}^{-1}(\hat{v}) = \mathcal{F}^{-1}(\frac{-i\xi}{|\xi|^2}\hat{p}(\xi)) \end{aligned}$$

Furthermore, $|v|_{H^1(\Omega)} = \|i\xi\hat{v}\|_{L^2(\Omega)} \leq C\|\hat{p}\|_{L^2(\Omega)} = \|p\|_{L^2(\Omega)}$. We restrict this v to Ω and use it in the LBB-condition. The $L^2(\Omega)$-part of $\|v\|_{H^1(\Omega)}$ follows from the Poincare inequality after subtracting the mean value.
The technical part is to ensure Dirichlet - boundary conditions. One can build an extension operator $\mathcal{E}$ from $L^2(\mathbb{R}^d \setminus \Omega)$ onto $\mathbb{R}^d$, which commutes with the div-operator and set

$$v_{final} := v - \mathcal{E}v$$

This v satisfies $v = 0$ on $\partial\Omega$. Since $\operatorname{div} v = p = 0$ outside of Ω, the correction did not change the divergence inside Ω. ■

7.1.2 Finite Elements for the Stokes equation

Now, we turn to the discrete system posed on $V_h \subset V$ and $Q_h \subset Q$. In order to formulate a well-posed FEM for the Stokes equations, we need to ensure that the discrete LBB condition holds, i.e., the spaces V_h and Q_h can not be arbitrary.

Elements with discontinuous pressure

The simplest pair one could think of is taking piecewise linear finite elements for the velocity field and piecewise constant functions for the pressure, i.e.,

$$V_h = [P^1_0(\mathcal{T})]^d, \qquad Q_h = P^0(\mathcal{T}) \cap L^2_0(\Omega).$$

However, this pair is not inf-sup stable:

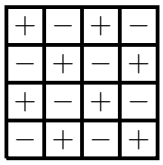
Example. Let $\Omega = (0,1)^2$ and take a mesh consisting of 4 quadrilaterals of side length $1/2$. Let $p \in P^0(\mathcal{T})$ be given as $p|_T = \pm 1$ according to the following picture. Then, $p \in L^2_0(\Omega)$ as well.

1	−1
−1	1

Then, there holds (exercise!)

$$\int_\Omega p \operatorname{div} u \, dx = 0 \qquad \forall u \in [P_0^1(\mathcal{T})]^2.$$

Consequently, inf-sup stability can not hold. This example can easily be generalized to arbitrary fine meshes by using the checkerboard structure (+ denotes 1 and − denotes −1) for p:



■

A remedy for this problem is to make the space V_h bigger by increasing the polynomial degree, i.e., we take the so called **Taylor-Hood type element**

$$V_h = [P_0^2(\mathcal{T})]^d, \qquad Q_h = P^0(\mathcal{T}) \cap L_0^2(\Omega).$$

Then, the following Lemma gives discrete inf-sup stability (all inf-sup stability results in the following can be proved by constructing suitable Fortin operators).

Lemma 7.1. *Let $\mathcal{T}$ be a regular, shape-regular triangulation of Ω and $V_h = [P_0^2(\mathcal{T})]^d$, $Q_h = P^0(\mathcal{T}) \cap L_0^2(\Omega)$. Then, there exists a constant $\beta > 0$ such that*

$$\inf_{p \in Q_h} \sup_{u \in V_h} \frac{b(u,p)}{\|p\|_{L^2(\Omega)} \|u\|_{H^1(\Omega)}} \geq \beta > 0.$$

Error estimates are

$$\|u - u_h\|_{H^1(\Omega)} + \|p - p_h\|_{L^2(\Omega)} \leq C \inf_{v_h, q_h} \|u - v_h\|_{H^1(\Omega)} + \|p - q_h\|_{L^2(\Omega)} = O(h).$$

Although we approximate u_h with P^2-elements, the bad approximation of p leads to first order convergence, only. This element is considered to be sub-optimal.

Elements with continuous pressure

Although the pressure p is only in $L^2(\Omega)$, we may approximate it with continuous elements. The so called **MINI-element** is

$$V_h = [P_0^1(\mathcal{T}) + B_3]^d \qquad Q_h = P^1(\mathcal{T}) \cap L_0^2(\Omega),$$

where $B_3 := \{v \in C(\overline{\Omega}) \ : \ v|_T = c b_T, \ c \in \mathbb{R}\}$ and b_T is the cubic bubble function given as $b_T = \varphi_1^T \varphi_2^T \varphi_3^T$ with the hat functions φ_i^T associated with the vertices of T.

The MINI-element is inf-sup stable by the following lemma.

Lemma 7.2. *Let $\mathcal{T}$ be a regular, shape-regular triangulation of Ω and $V_h = [P_0^1(\mathcal{T}) + B_3]^d$, $Q_h = P^1(\mathcal{T}) \cap L_0^2(\Omega)$. Then, there exists a constant $\beta > 0$ such that*

$$\inf_{p \in Q_h} \sup_{u \in V_h} \frac{b(u,p)}{\|p\|_{L^2(\Omega)} \|u\|_{H^1(\Omega)}} \geq \beta > 0.$$

This method is $O(h)$ convergent.

Another (essentially) inf-sup stable pair is the so called **Taylor-Hood element**

$$V_h = [P_0^2(\mathcal{T})]^d \qquad Q_h = P^1(\mathcal{T}) \cap L_0^2(\Omega).$$

Its analysis is more involved and requires the additional assumption that no two edges of one element are on the domain boundary. Its convergence rate is $O(h^2)$.

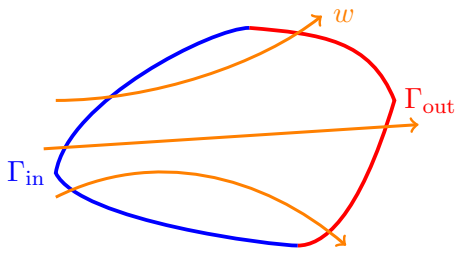
7.2 Convection dominated problems

In the following, we consider second order PDEs with lower order terms that actually dominate. More precisely, with a smooth vector field w with $\operatorname{div} w = 0$, we consider the problem

$$\begin{aligned} -\varepsilon \Delta u + w \cdot \nabla u &= f && \text{in } \Omega \\ u &= u_D && \text{on } \partial\Gamma_D \\ \nabla u \cdot n &= u_N && \text{on } \partial\Gamma_N, \end{aligned}$$

where the boundary $\partial\Omega = \Gamma_D \cup \Gamma_N$ may consist of a Dirichlet and a Neumann part.
Here, $\varepsilon \in \mathbb{R}^+$ will be small, i.e, $\varepsilon << 1$. The solution of the problem will therefore mainly be characterized by the vector field w. For this, we define the **inflow boundary** and the **outflow boundary** as

$$\begin{aligned} \Gamma_{\text{in}} &:= \{x \in \partial\Omega \ : \ w \cdot n < 0\} \\ \Gamma_{\text{out}} &:= \{x \in \partial\Omega \ : \ w \cdot n \geq 0\}. \end{aligned}$$

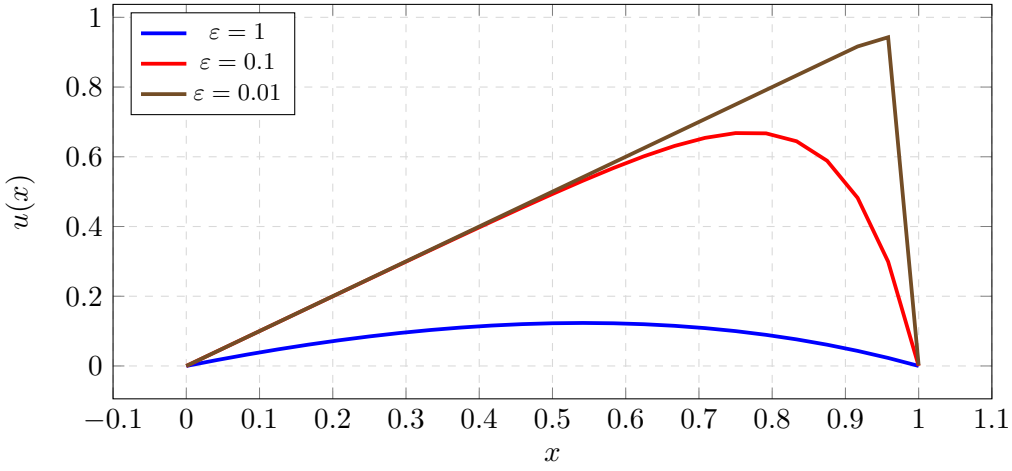


A critical quantity in the analysis will be the ratio $\frac{\varepsilon}{\|w\|_\infty}$ as can also be seen in the following example:

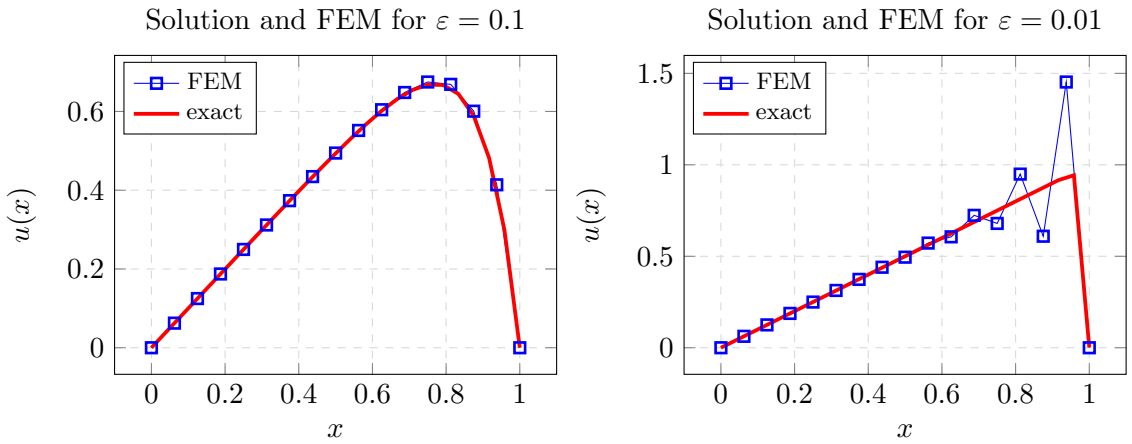
Example. Consider $\Omega = (0,1)$, take $w = 1$ and homogeneous Dirichlet boundary conditions, i.e., $u = 0$ on $\partial\Omega$. In 1D the equation then reads as

$$-\varepsilon u'' + u' = 1,$$

which has the exact solution $u(x) = x + \frac{1}{e^{1/\varepsilon} - 1}(1 - e^{\frac{x}{\varepsilon}})$.



As can be seen from the plot, the solution tends to get high gradients close to $x = 1$ and exhibits a so called **boundary layer behavior**. This means that taking the limit $\varepsilon \to 0$, the limiting equation is $u' = 1$. However, the problem now is that one can not impose TWO boundary conditions. Therefore, for small ε, one expects a sharp gradient close to the boundary (which can be quantified as $\sim \varepsilon$ away from $x = 1$ here).



Doing a FEM here can get very badly, if this boundary layer is not resolved by the mesh, i.e., if $h >> \varepsilon$. ■

We start our analysis of these phenomenon with the weak formulation of the model problem (with homogeneous boundary conditions). Multiplying with a test-function $v \in H^1_D(\Omega)$ and integration by parts gives: Find $u \in H^1_D(\Omega)$ such that

$$\int_\Omega \varepsilon \nabla u \cdot \nabla v + (w \cdot \nabla u) v \, dx = \int_\Omega f v \, dx \quad \forall v \in H^1_D(\Omega).$$

It will be convenient to write this as

$$A(u, v) := a(u, v) + c(u, v) = l(v)$$

with

$$a(u,v) := \int_\Omega \varepsilon \nabla u \cdot \nabla v dx, \quad c(u,v) := \int_\Omega (w \cdot \nabla u) v \, dx, \quad l(v) := \int_\Omega f v \, dx.$$

The following theorem based on the Lax-Milgram lemma gives existence and uniqueness of solution as well as explicit coercivity and continuity constants.

Theorem 7.3. *Assume $|\Gamma_D| > 0$ and let $w \cdot n \geq 0$ on Γ_N (this means that the Neumann boundary is an outflow boundary). Then, we have*

$$A(u,v) \leq (\varepsilon + \|w\|_\infty C_F) \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)},$$
$$A(u,u) \geq \varepsilon \|\nabla u\|^2_{L^2(\Omega)},$$

where C_F denotes the constant in the Poincaré-Friedrichs inequality. Consequently, the weak formulation has a unique solution in $u \in H^1_D(\Omega)$.

Proof. Continuity of the bilinear forms follow from the Cauchy-Schwarz and Poincaré-Friedrichs inequality:

$$\begin{aligned} |a(u,v)| &\leq \varepsilon \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)}, \\ |c(u,v)| &\leq \int_\Omega |(w \cdot \nabla u) v| dx \leq \|w\|_{L^\infty(\Omega)} \|\nabla u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \\ &\leq C_F \|w\|_{L^\infty(\Omega)} \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)}. \end{aligned}$$

For the coercivity, we integrate by parts and use $\operatorname{div}(w) = 0$ to obtain

$$c(u,u) = \int_\Omega (w \cdot \nabla u) u \, dx = -\int_\Omega uw \cdot \nabla u - u^2 \operatorname{div}(w) \, dx + \int_{\Gamma_N} u^2 w \cdot n \, ds,$$

which implies $c(u,u) = \frac{1}{2} \int_{\Gamma_N} u^2 w \cdot n \, ds$. As the Neumann boundary is an outflow boundary, we conclude

$$a(u,u) + c(u,u) = a(u,u) + \frac{1}{2} \int_{\Gamma_N} u^2 w \cdot n \, ds \geq a(u,u) = \varepsilon \|\nabla u\|^2_{L^2(\Omega)}.$$

Unique solvability follows from the Lax-Milgram lemma. ■

Now we take a standard FEM formulation based on some discrete FEM-space $V_h \subset H^1_D(\Omega)$. Find $u_h \in V_h$ such that

$$A(u_h, v_h) = l(v_h) \qquad \forall v_h \in V_h.$$

As usual for coercive problems, existence and uniqueness is inherited from the continuous setting. We now apply the Cea-Lemma from the abstract FEM-chapter, which gives the quasi-bestapproximation result

$$\begin{aligned} \|\nabla u - \nabla u_h\|_{L^2(\Omega)} &\leq \frac{\varepsilon + \|w\|_\infty C_F}{\varepsilon} \inf_{v_h \in V_h} \|\nabla u - \nabla v_h\|_{L^2(\Omega)} \\ &= (1 + \mathcal{P}) \inf_{v_h \in V_h} \|\nabla u - \nabla v_h\|_{L^2(\Omega)}, \end{aligned}$$

where $\mathcal{P} := \frac{\|w\|_\infty C_F}{\varepsilon}$ is the so called **Péclet-number**. For $\varepsilon \to 0$, we observe $\mathcal{P} \to \infty$, which makes this estimate useless for small $\varepsilon << 1$.

We now aim to improve on that bound. Using Galerkin-orthogonality, we obtain

$$\begin{aligned}\varepsilon\|\nabla(u-u_h)\|^2_{L^2(\Omega)} &\leq a(u-u_h, u-u_h) + c(u-u_h, u-u_h)\\ &= a(u-u_h, u-Iu) + c(u-u_h, u-Iu)\\ &\leq \varepsilon\|\nabla(u-u_h)\|_{L^2(\Omega)}\|\nabla(u-Iu)\|_{L^2(\Omega)} + \|w\|_\infty\|\nabla(u-u_h)\|_{L^2(\Omega)}\|u-u_h\|_{L^2(\Omega)}.\end{aligned}$$

Division by $\varepsilon\|\nabla(u-u_h)\|_{L^2(\Omega)}$ and approximation properties of the interpolation operator Iu gives

$$\begin{aligned}\|\nabla(u-u_h)\|_{L^2(\Omega)} &\leq \|\nabla(u-Iu)\|_{L^2(\Omega)} + \frac{\|w\|_\infty}{\varepsilon}\|u-u_h\|_{L^2(\Omega)}\\ &\leq h\|u\|_{H^2(\Omega)} + \frac{\|w\|_\infty}{\varepsilon}h^2\|u\|_{H^2(\Omega)}\\ &\leq \left(1 + \frac{\|w\|_\infty h}{\varepsilon}\right)h\|u\|_{H^2(\Omega)}.\end{aligned}$$

Here $\mathcal{P}_h := \frac{\|w\|_\infty h}{\varepsilon}$ is called the **mesh-Péclet-number** (for a uniform mesh of mesh-width h). For non-uniform meshes, one can also define the element Péclet-numbers $\mathcal{P}_{h_T} := \frac{\|w\|_{\infty,T} h_T}{\varepsilon}$. The crucial observation here is that, provided $\mathcal{P}_h$ is moderate, we still obtain a properly convergent method. This also suggests that one should take mesh-sizes $h \leq \varepsilon$ (which also corresponds to resolving the boundary layer).

7.2.1 The streamline-upwind Petrov-Galerkin formulation

We now aim to present a numerical method that also works for large mesh-Péclet-number. The idea of the streamline-upwind Petrov-Galerkin formulation (SUPG), which is also called streamline-diffusion (SD) method in literature, is to add stabilizing terms in the direction of the streamlines and then still use standard H^1-based finite element spaces.

Let $s_h(\cdot,\cdot) : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ be a bilinear form (the stabilization) and $f_s(\cdot) : H^1(\Omega) \to \mathbb{R}$ be a linear form (modified right-hand side) . The construction of s_h and f_s will be elementwise (and afterwards everything will be summed up) and consists of

- a stabilization parameter γ_T,
- the strong form of the residual $-\varepsilon\Delta u_h + w \cdot \nabla u_h - f$ on T,
- the "streamline derivative" of the testfunction $w \cdot \nabla v$.

More precisely, we define

$$\begin{aligned}s_h(u_h, v) &= \sum_T \gamma_T \int_T (-\varepsilon\Delta u_h + w \cdot \nabla u_h) w \cdot \nabla v \, dx\\ f_s(v) &= \sum_T \gamma_T \int_T f w \cdot \nabla v \, dx.\end{aligned}$$

We now consider the SUPG formulation: Find $u_h \in V_h \subset H^1_D(\Omega)$ such that

$$A_h(u_h, v_h) := A(u_h, v_h) + s_h(u_h, v_h) = l(v_h) + f_s(v_h) \qquad \forall v_h \in V_h.$$

For simplification, we only apply the stabilization, where the convection dominates, i.e., if $\mathcal{P}_{h_T} >> 1$, i.e., we assume that there is a constant $C_{P_\ell} > 1$ such that

$$\mathcal{P}_{h_T} < C_{P_\ell} \qquad \Longrightarrow \gamma_T = 0 \qquad \text{for } T \in \mathcal{T}.$$

The following lemma shows a Galerkin orthogonality for the SUPG method.

Lemma 7.4. *Let $u \in H^1_D(\Omega)$ be the exact weak solution to the model problem and assume additionally $u|_T \in H^2(T)$ for all $T \in \mathcal{T}_h$. Then,*

$$A_h(u - u_h, v_h) = 0 \qquad \forall v_h \in V_h.$$

Proof. Integration by parts together with $s_h(u, v_h) - f_s(v_h) = 0$ (exercise!). ■

In order to obtain coercivity of the discrete formulation with a good constant, we introduce the norm on V_h

$$|\!|\!| v_h |\!|\!|^2 := \varepsilon \|\nabla v_h\|^2_{L^2(\Omega)} + \sum_T \gamma_T \|w \cdot \nabla v_h\|^2_{L^2(T)} \qquad \forall v_h \in V_h.$$

as well as a so-called inverse inequality

$$\|\Delta u_h\|_{L^2(T)} \leq C_{\text{ie}} h_T^{-1} \|\nabla u_h\|_{L^2(T)} \qquad \forall u_h \in V_h, \tag{7.2}$$

where the constant $C > 0$ does not depend on the size of T.

Lemma 7.5. *Let $V_h = \mathcal{P}^p(\mathcal{T})$ be a Lagrangian finite element space of order $p \geq 1$. With the constant C_{ie} from (7.2) assume for the stability parameter that*

$$\gamma_T \leq \frac{h_T^2}{C_{\text{ie}}^2 \varepsilon},$$

Then, for all $u_h \in V_h$, there holds

$$A_h(u_h, u_h) \geq \frac{1}{2} |\!|\!| u_h |\!|\!|^2 .$$

Proof. We compute

$$A_h(u_h, u_h) = A(u_h, u_h) + s_h(u_h, u_h) = \varepsilon \|\nabla u_h\|^2_{L^2(\Omega)} + \sum_T \gamma_T \left(\|w \cdot \nabla u_h\|^2_{L^2(T)} - (\varepsilon \Delta u_h, w \cdot \nabla u_h)_{L^2(T)} \right).$$

We estimate the last term from above (thus this gives an estimate from below for (-1)-times this term)

$$\begin{aligned}\left|(\varepsilon\Delta u_h, w\cdot\nabla u_h)_{L^2(T)}\right| &\le \varepsilon\|\Delta u_h\|_{L^2(T)}\|w\cdot\nabla u_h\|_{L^2(T)}\\ &\overset{(7.2)}{\le} \varepsilon\frac{C_{\text{ie}}}{h_T}\|\nabla u_h\|_{L^2(T)}\|w\cdot\nabla u_h\|_{L^2(T)}\\ &\le \frac{1}{2}\|w\cdot\nabla u_h\|^2_{L^2(T)} + \frac{1}{2}\varepsilon^2\frac{C_{\text{ie}}^2}{h_T^2}\|\nabla u_h\|^2_{L^2(T)}.\end{aligned}$$

Inserting this into the above equality gives

$$A_h(u_h,u_h) \ge \sum_T\left(\varepsilon - \frac{1}{2}\varepsilon^2\frac{C_{\text{ie}}^2}{h_T^2}\gamma_T\right)\|\nabla u_h\|^2_{L^2(T)} + \sum_T\gamma_T\frac{1}{2}\|w\cdot\nabla u_h\|^2_{L^2(T)} \ge \frac{1}{2}\left|\!\left|\!\left|u_h\right|\!\right|\!\right|^2,$$

where the last estimate follows from the assumption on the stabilization parameter. ■

In order to show continuity, one introduces yet another norm

$$\left|\!\left|\!\left|v\right|\!\right|\!\right|_\star^2 := \left|\!\left|\!\left|v\right|\!\right|\!\right|^2 + \sum_T\left(\gamma_T\|\varepsilon\Delta v\|^2_{L^2(T)} + \frac{1}{\gamma_T}\|v\|^2_{L^2(T)}\right).$$

Lemma 7.6. *For all $u,v\in H^1_D(\Omega)$ satisfying $u|_T, v|_T\in H^2(T)$ there holds*

$$A_h(u,v) \le 2\left|\!\left|\!\left|u\right|\!\right|\!\right|_\star\left|\!\left|\!\left|v\right|\!\right|\!\right|.$$

Proof. Exercise. ■

Now, combining coercivity and continuity gives a type of a Cea-Lemma.

Lemma 7.7. *Let u be the weak solution to the model problem and u_h its approximation by the SUPG-method. Under the assumptions of the previous lemmas there holds*

$$\left|\!\left|\!\left|u-u_h\right|\!\right|\!\right| \le 5\inf_{v_h\in V_h}\left|\!\left|\!\left|u-v_h\right|\!\right|\!\right|_\star.$$

Proof. The triangle inequality gives

$$\left|\!\left|\!\left|u-u_h\right|\!\right|\!\right| \le \left|\!\left|\!\left|u-v_h\right|\!\right|\!\right| + \left|\!\left|\!\left|v_h-u_h\right|\!\right|\!\right|.$$

Coercivity and Galerkin-orthogonality implies

$$\left|\!\left|\!\left|v_h-u_h\right|\!\right|\!\right|^2 \le 2A_h(u_h-v_h,u_h-v_h) = 2A_h(u-v_h,u_h-v_h) \le 4\left|\!\left|\!\left|u-v_h\right|\!\right|\!\right|_\star\left|\!\left|\!\left|u_h-v_h\right|\!\right|\!\right|.$$

Inserting that in the above estimate gives the result using $\|\!|\cdot|\!\| \leq \|\!|\cdot|\!\|_\star$. ■

As usual, in order to derive convergence rates, we employ interpolation error estimates.

Lemma 7.8. *Let u be the exact weak solution to the model problem and assume $u \in H^{k+1}(\Omega)$ for $k \geq 1$. Let $I_\mathcal{T}$ be the Lagrange interpolation operator. Under the assumptions of this section there holds*

$$\|\!|u - I_\mathcal{T} u|\!\|_\star^2 \leq C \sum_{T \in \mathcal{T}} \left(\varepsilon + \gamma_T \|w\|_{\infty,T}^2 + \gamma_T^{-1} h_T^2\right) h_T^{2k} \|u\|_{H^{k+1}(T)}^2.$$

Proof. Standard interpolation estimates (as in Section 4.3) applied to each term in the norm give $\|\!|\cdot|\!\|_\star$ gives

$$\|\!|u - I_\mathcal{T} u|\!\|_\star^2 \leq C \sum_{T \in \mathcal{T}} \left(\varepsilon + \gamma_T(\|w\|_{\infty,T}^2 + \varepsilon^2 h_T^{-2}) + \gamma_T^{-1} h_T^2\right) h_T^{2k} \|u\|_{H^{k+1}(T)}^2,$$

where only the third term is left to treat. By choice of γ_T, we have that $\gamma_T \neq 0$ only if $\mathcal{P}_{h_T} \geq C_{P_\ell}$, which gives

$$C_{P_\ell} \leq \frac{\|w\|_{\infty,T} h_T}{\varepsilon} \qquad \Longrightarrow \varepsilon^2 h_T^{-2} < \|w\|_{\infty,T}^2 C_{P_\ell}^{-2}.$$

Consequently,

$$(\|w\|_{\infty,T}^2 + \varepsilon^2 h_T^{-2}) \leq C \|w\|_{\infty,T}^2,$$

which shows the estimate. ■

Choice of stabilization parameter

It remains to discuss how to choose the stabilization parameter.

In the convection dominated case we have $\varepsilon \leq C\|w\|_{\infty,T} h_T$. In order for all terms in the previous estimate to scale in the same way (as $\|w\|_{\infty,T} h_T$), we have to choose

$$\gamma_T = \gamma_0 \frac{h_T}{\|w\|_{\infty,T}}.$$

In order to fix the multiplicative constant γ_0, we note that we also have to fulfill $\gamma_T \leq \frac{h_T^2}{C_{\text{ie}}^2 \varepsilon}$, which gives

$$\gamma_0 \leq \frac{h_T \|w\|_{\infty,T}}{C_{\text{ie}}^2 \varepsilon} = \frac{\mathcal{P}_{h_T}}{C_{\text{ie}}^2},$$

which can e.g. be ensured if the quantity

$$\frac{C_{P_\ell}}{\gamma_0}$$

is sufficiently large.

A priori error estimate

Now, collecting everything together, we obtain an estimate for the rate of convergence of the SUPG method.
Choosing the stabilization γ_T as described above, there holds by the Cea-type estimate together with the interpolation estimate

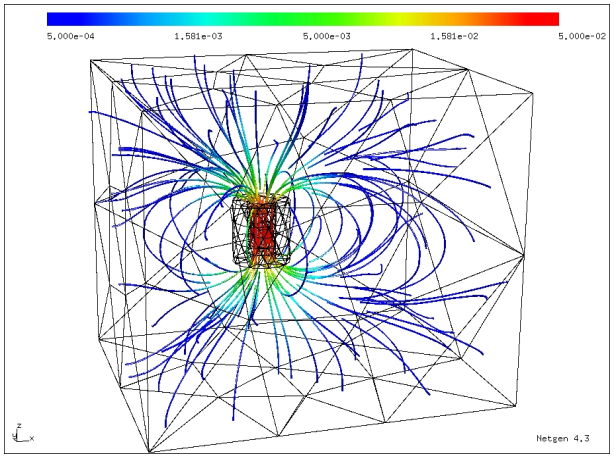
$$|||u - u_h|||^2 \leq C \sum_T (\varepsilon + h_T \|w\|_{\infty,T})\, h_T^{2k} \|u\|^2_{H^{k+1}(T)}.$$

If the mesh is quasi-uniform, i.e., $h_T \simeq h$ for all $T \in \mathcal{T}_h$, there holds the simplified estimate

$$|||u - u_h||| \leq C \|w\|_\infty^{1/2} h^{k+1/2} \|u\|_{H^{k+1}(\Omega)}.$$

7.3 Maxwell equations

Maxwell equations describe electro-magnetic fields (a magnetic field is caused by an electric current). In the following, we consider the special case of stationary magnetic fields (i.e. no time dependency in the equation). Maxwell equations are three-dimensional.
The picture below shows the magnetic field caused by a tangential current density in a coil:



We suppose that a current density (right-hand side)

$$j \in [L^2(\Omega)]^3$$

is given. (Stationary) currents do not have sources, i.e., div $j = 0$. Then, we seek an electric field intensity E such that

$$\operatorname{curl} \mu^{-1} \operatorname{curl} E = j. \tag{7.3}$$

As $\operatorname{curl} \nabla\varphi = 0$ for any function φ, the equation is not uniquely solvable. To obtain a unique solution, the so called Coulomb-Gauging can be applied, i.e., we seek a vector field E that is orthogonal to gradient fields.

In principle, Maxwell equations are valid in the whole $\mathbb{R}^3$. For simulation, we have to truncate the domain and have to introduce artificial boundary conditions. Here we take so called perfectly conducting boundary conditions

$$E \times n = 0 \qquad \text{on } \partial\Omega.$$

We also assume that μ is a constant. The coefficient μ is 10^3 to 10^4 times larger in iron (and other ferro-magnetic metals) as in most other media (air).
As usual, we go over to the weak form (we will specify the proper functions spaces later on). Equations (7.3) and the orthogonality side-constraint together become: Find E such that

$$\int_\Omega \mu^{-1} \operatorname{curl} E \operatorname{curl} v \, dx = \int_\Omega j \cdot v \, dx \qquad \forall\, v$$
$$\int_\Omega E \cdot \nabla \psi \, dx = 0 \qquad \forall\, \psi.$$

To obtain a symmetric system, we add a new scalar variable φ and make a saddle-point formulation. The problem is now: Find $E \in V$ and $\varphi \in Q$ such that

$$\begin{array}{rcccll} \int_\Omega \mu^{-1} \operatorname{curl} E \cdot \operatorname{curl} v \, dx & + & \int_\Omega \nabla \varphi \cdot v \, dx & = & \int_\Omega j \cdot v \, dx & \forall\, v \in V, \\ \int_\Omega E \cdot \nabla \psi \, dx & & & = & 0 & \forall\, \psi \in Q. \end{array} \tag{7.4}$$

For the function space Q, we would like to choose $H^1(\Omega)$, but then uniqueness can not hold as $\varphi + c$ will also be solution, if φ is a solution. Therefore, we take

$$Q = \overline{H}^1(\Omega) := \{\varphi \in H^1(\Omega) \ : \ \int_\Omega \varphi \, dx = 0\}.$$

The proper function space V is the $H(\operatorname{curl}, \Omega)$:

$$H(\operatorname{curl}, \Omega) = \left\{ v \in [L^2(\Omega)]^3 : \operatorname{curl} v \in [L^2(\Omega)]^3 \right\}.$$

Again, the differential operator curl is understood in the weak sense. The norm on the space is

$$\|v\|^2_{H(\operatorname{curl},\Omega)} := \|v\|^2_{L^2(\Omega)} + \|\operatorname{curl} v\|^2_{L^2(\Omega)}.$$

Finally, we take V as functions that also fulfill the boundary conditions, i.e.,

$$V = H_0(\operatorname{curl}, \Omega) = \{v \in H(\operatorname{curl}, \Omega) \ : \ v \times n = 0 \text{ on } \partial\Omega\}.$$

The function space $H(\operatorname{curl}, \Omega)$

Similar to $H^1(\Omega)$ and $H(\operatorname{div}, \Omega)$, there exists a trace operator for $H(\operatorname{curl}, \Omega)$. Now, only the tangential components of the boundary values are well defined:

Theorem 7.9 (Trace theorem). *There exists a tangential trace operator* $\operatorname{tr}_\tau v : H(\operatorname{curl}, \Omega) \to L^2(\partial\Omega)$ *such that*

$$\operatorname{tr}_\tau v = n \times v|_{\partial\Omega}$$

for smooth functions $v \in [C(\overline{\Omega})]^3$.

As for the spaces $H^1(\Omega)$ and $H(\operatorname{div}, \Omega)$, one can characterize $H(\operatorname{curl}, \Omega)$ functions on subsets by requiring continuity of the tangential trace.

Theorem 7.10. *Let* $\Omega = \Omega_1 \cup \Omega_2$. *Assume that* $u|_{\Omega_i} \in H(\operatorname{curl}, \Omega_i)$, *and the tangential traces are continuous across the interfaces* $\gamma_{ij} = \overline{\Omega_1} \cap \overline{\Omega_2}$. *Then,* $u \in H(\operatorname{curl}, \Omega)$.

The theorems are similar to the ones we have proven for $H(\mathrm{div}, \Omega)$. However, the proofs (in $\mathbb{R}^3$) are more involved.

The gradient operator ∇ relates the space $H^1(\Omega)$ and $H(\mathrm{curl}, \Omega)$:

$$\nabla : H^1(\Omega) \to H(\mathrm{curl}, \Omega).$$

Furthermore, the kernel space

$$H_c(\mathrm{curl}, \Omega) = \{v \in H(\mathrm{curl}, \Omega) : \mathrm{curl}\ v = 0\}$$

is exactly the range of the gradient:

$$H_c(\mathrm{curl}, \Omega) = \nabla\, H^1(\Omega).$$

Theorem 7.11. *The mixed system (7.4) is a well posed problem on* $H(\mathrm{curl}, \Omega) \times \overline{H}^1(\Omega)$.

Proof. The bilinear-forms

$$a(E, v) = \int_\Omega \mu^{-1}\, \mathrm{curl}\, E \cdot \mathrm{curl}\, v\, dx$$

and

$$b(v, \varphi) = \int_\Omega v \cdot \nabla\varphi\, dx$$

are continuous w.r.t. the norms of $V = H(\mathrm{curl}, \Omega)$ and $Q = \overline{H}^1(\Omega)$.
The LBB-condition in this case is trivial. Fix $\varphi \in Q$ and choose $v = \nabla\varphi$, then

$$\sup_{v \in H(\mathrm{curl},\Omega)} \frac{\int_\Omega v \cdot \nabla\varphi\, dx}{\|v\|_{H(\mathrm{curl},\Omega)}} \geq \frac{\int_\Omega \nabla\varphi \cdot \nabla\varphi\, dx}{\|\nabla\varphi\|_{H(\mathrm{curl},\Omega)}} = \frac{\|\nabla\varphi\|^2_{L^2(\Omega)}}{\|\nabla\varphi\|_{L^2(\Omega)}} = \|\nabla\varphi\|_{L^2(\Omega)} \simeq \|\varphi\|_Q,$$

where the last step follows from the Poincaré inequality.
Here, the difficult part is the kernel coercivity of $a(\cdot, \cdot)$. The norm involves also the $L^2(\Omega)$-norm, while the bilinear-form only involves the semi-norm $\|\,\mathrm{curl}\ v\|_{L^2(\Omega)}$. Coercivity cannot hold on the whole V: Take a gradient function $\nabla\psi$, then $\mathrm{curl}\,\nabla\psi = 0$. On the kernel

$$V_0 = \{v \in H(\mathrm{curl}, \Omega) : \int_\Omega v \cdot \nabla\varphi\, dx = 0 \ \ \forall\, \varphi \in H^1(\Omega)\},$$

one can actually bound the $L^2(\Omega)$-norm by the semi-norm:

$$\|v\|_{L^2(\Omega)} \leq C \|\,\mathrm{curl}\ v\|_{L^2(\Omega)} \qquad \forall\, v \in V_0,$$

which is a Poincaré-Friedrichs-like inequality. ■

7.3.1 Finite elements in $H(\mathrm{curl}, \Omega)$

We now construct finite elements in three dimensions. The trace theorem implies that functions in $H(\mathrm{curl})$ have continuous tangential components across element boundaries (=faces).

We design tetrahedral finite elements.

Nédélec elements

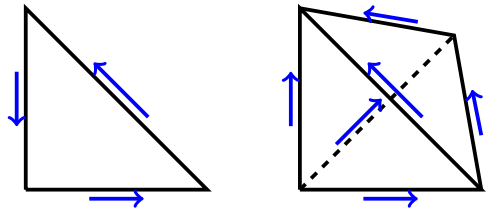
There is a cheaper element, called **Nédélec, or edge-element**, which is similar to the Raviart-Thomas element. It contains all constants, and some linear polynomials. The local finite element space is

$$V_T = \{a + b \times x : a, b \in \mathbb{R}^3\}.$$

These are 6 coefficients, i.e., we have dimension 6. For each of the 6 edges of a tetrahedron, one chooses the integral of the tangential component along the edge

$$\psi_{E_i}(u) = \int_{E_i} u \cdot \tau_{E_i} \, ds,$$

where τ_{E_i} denotes the unit vector in the direction of the edge E_i.



Lemma 7.12. *The basis function φ_{E_i} associated with the edge E_i is*

$$\varphi_{E_i} = \lambda_{E_i^1} \nabla \lambda_{E_i^2} - \nabla \lambda_{E_i^1} \lambda_{E_i^2},$$

where E_i^1 and E_i^2 are the two vertex numbers of the edge, and $\lambda_1, \ldots \lambda_4$ are the vertex shape functions (hat functions).

Proof.

- These functions are in V_T.
- If $i \neq j$, then $\psi_{E_j}(\varphi_{E_i}) = 0$.
- $\psi_{E_i}(\varphi_{E_i}) = 1$

Thus, edge elements belong to $H(\mathrm{curl}, \Omega)$. Next, we will see that they have also very interesting properties. ■

The de'Rham complex

The spaces $H^1(\Omega)$, $H(\mathrm{curl},\Omega)$, $H(\mathrm{div},\Omega)$, and $L^2(\Omega)$ form a sequence:

$$H^1(\Omega) \quad \xrightarrow{\nabla} \quad H(\mathrm{curl},\Omega) \quad \xrightarrow{\mathrm{curl}} \quad H(\mathrm{div},\Omega) \quad \xrightarrow{\mathrm{div}} \quad L^2(\Omega).$$

Since $\nabla H^1(\Omega) \subset [L^2(\Omega)]^3$, and $\mathrm{curl}\,\nabla = 0$, the gradients of $H^1(\Omega)$ functions belong to $H(\mathrm{curl},\Omega)$. Similar, since $\mathrm{curl}\, H(\mathrm{curl},\Omega) \subset [L^2(\Omega)]^3$, and $\mathrm{div}\,\mathrm{curl} = 0$, the curls of $H(\mathrm{curl},\Omega)$ functions belong to $H(\mathrm{div},\Omega)$.
The sequence is a **complete sequence**, which means that the kernel of the right differential operator is exactly the range of the left one (on simply connected domains). We have used this property already in the analysis of the mixed system.

The same property holds on the discrete level: Let

W_h	be the nodal finite element sub-space of $H^1(\Omega)$,
V_h	be the Nédélec (edge) finite element sub-space of $H(\mathrm{curl},\Omega)$,
Q_h	be the Raviart-Thomas (face) finite element sub-space of $H(\mathrm{div},\Omega)$,
S_h	be the piece-wise constant finite element sub-space of $L^2(\Omega)$.

Theorem 7.13. *The finite element spaces form a complete sequence*

$$W_h \quad \xrightarrow{\nabla} \quad V_h \quad \xrightarrow{\mathrm{curl}} \quad Q_h \quad \xrightarrow{\mathrm{div}} \quad S_h$$

Now, we discretize the mixed formulation (7.4) by choosing edge-finite elements for $H(\mathrm{curl},\Omega)$, and nodal finite elements for $H^1(\Omega)$: Find $E_h \in V_h$ and $\varphi_h \in W_h$ such that

$$\begin{array}{rcll} \int_\Omega \mu^{-1}\,\mathrm{curl}\,E_h \cdot \mathrm{curl}\,v_h\,dx \;+\; \int_\Omega \nabla\varphi_h \cdot v_h\,dx &=& \int_\Omega j \cdot v_h\,dx & \forall\, v_h \in V_h, \\ \int_\Omega E_h \cdot \nabla\psi_h\,dx &=& 0 & \forall\, \psi_h \in W_h. \end{array} \tag{7.5}$$

The stability follows (roughly) from the discrete sequence property. The verification of the LBB condition is the same as on the continuous level. The kernel of the $a(\cdot,\cdot)$- form are the discrete gradients, the kernel of the $b(\cdot,\cdot)$-form is orthogonal to the gradients. This implies solvability. The discrete kernel-coercivity (with h-independent constants) is true (nontrivial).

The complete sequences on the continuous level and on the discrete level are connected in the de'Rham complex: Choose the canonical interpolation operators (vertex-interpolation I^W, edge-interpolation I^V, face-interpolation I^Q, L_2-projection I^S). This relates the continuous level to the discrete level:

$$\begin{array}{ccccccc} H^1(\Omega) & \xrightarrow{\nabla} & H(\mathrm{curl},\Omega) & \xrightarrow{\mathrm{curl}} & H(\mathrm{div},\Omega) & \xrightarrow{\mathrm{div}} & L^2(\Omega) \\ \downarrow I^W & & \downarrow I^V & & \downarrow I^Q & & \downarrow I^S \\ W_h & \xrightarrow{\nabla} & V_h & \xrightarrow{\mathrm{curl}} & Q_h & \xrightarrow{\mathrm{div}} & S_h \end{array} \tag{7.6}$$

Theorem 7.14. *The diagram (7.6)* commutes*:*

$$I^V \nabla = \nabla I^W \qquad I^Q \operatorname{curl} = \operatorname{curl} I^V \qquad I^S \operatorname{div} = \operatorname{div} I^Q$$

Proof. We prove the first part. Note that the ranges of both, ∇I^W and $I^V \nabla$, are in V_h. Two functions in V_h coincide if and only if all functionals coincide. It remains to prove that

$$\int_E (\nabla I^W w) \cdot \tau \, ds = \int_E (I^V \nabla w) \cdot \tau \, ds.$$

Per definition of the interpolation operator I^V there holds

$$\int_E (I^V \nabla w) \cdot \tau \, ds = \int_E \nabla w \cdot \tau \, ds.$$

Integrating the tangential derivative gives the difference

$$\int_E \nabla w \cdot \tau \, ds = \int_E \frac{\partial w}{\partial \tau} \, ds = w(E^2) - w(E^1).$$

Starting with the left term, and using the property of the nodal interpolation operator, we obtain

$$\int_E (\nabla I^W w) \cdot \tau \, ds = (I^W w)(E^2) - (I^W w)(E^1) = w(E^2) - w(E^1).$$

We have already proven the commutativity of the $H(\operatorname{div}, \Omega) - L^2(\Omega)$ part of the diagram. The middle one involves Stokes´ theorem. ■

This is the key for interpolation error estimates. E.g., in $H(\operatorname{curl}, \Omega)$ there holds

$$\begin{aligned} \|u - I^V u\|^2_{H(\operatorname{curl},\Omega)} &= \|u - I^V u\|^2_{L^2(\Omega)} + \|\operatorname{curl}(I - I^V)u\|^2_{L^2(\Omega)} \\ &= \|u - I^V u\|^2_{L^2(\Omega)} + \|(I - I^Q)\operatorname{curl} u\|^2_{L^2(\Omega)} \\ &\le C \quad h^2 \|u\|^2_{H^1(\Omega)} + h^2 \|\operatorname{curl} u\|^2_{H^1(\Omega)}. \end{aligned}$$

Since the estimates for the $L^2(\Omega)$-term and the curl-term are separate, one can also scale each of them by an arbitrary coefficient.

The sequence is also compatible with transformations. Let $F : \widehat{T} \to T$ be an (element) transformation. Choose

$$\begin{aligned} w(F(x)) &= \hat{w}(x) \\ v(F(x)) &= (F')^{-T} \hat{v}(x) && \text{(covariant transformation)} \\ q(F(x)) &= (\det F')^{-1} (F') q(x) && \text{(Piola-transformation)} \\ s(F(x)) &= (\det F')^{-1} \hat{s}(x) \end{aligned}$$

Then

$$\begin{aligned} \hat{v} = \nabla \hat{w} &\Rightarrow v = \nabla w \\ \hat{q} = \operatorname{curl} \hat{v} &\Rightarrow q = \operatorname{curl}\ v \\ \hat{s} = \operatorname{div} \hat{q} &\Rightarrow s = \operatorname{div}\ q \end{aligned}$$

Using these transformation rules, the implementation of matrix assembling for $H(\text{curl})$-equations is very similar to the assembling for H^1 problems (mapping to reference element).

Chapter 8

Time dependent problems

We now consider time dependent problems such as the advection equation, the heat equation and the wave equation. For the advection equation, we introduce finite difference methods, for the heat and the wave equation we consider finite elements as well. In principle, there are two ways how to approach those: treat time as just a new variable and do FEM for all time and space variables together (so called space-time methods) or treat time and space variables separately (so called time stepping methods).

8.1 The advection equation – finite differences

In the following, we consider the 1D-advection equation given by

$$\frac{\partial u(x,t)}{\partial t} + \frac{\partial u(x,t)}{\partial x} = f \qquad \text{in } \mathbb{R} \times (0,\infty)$$

with the initial condition $u(x,0) = u_0(x)$. If $f = 0$, the solution is simply a translation of the initial data, i.e.,

$$u(x,t) = u_0(x-t).$$

In the following, we consider a **finite difference** (FD) method, similar to the one we introduced in Section 1. The idea hereby is to approximate the derivatives by some difference quotients.

Take an (infinite) mesh in space of $\mathbb{R}$ with points $x_j = jh$, $j \in \mathbb{Z}$ of mesh-width h and an (infinite) mesh in time of $(0,\infty)$ with $t_n = nk$, $n \in \mathbb{N}$ of time-step size k. The goal is to compute the solution to the PDE in the gridpoints, i.e., compute values

$$u_j^n \simeq u(x_j, t_n).$$

In order to approximate the derivatives by difference quotients, we have multiple options.

1. At first, we take the right-difference quotient in time and space, i.e., approximate a 1D-derivative by $(u(x+h) - u(x))/h$, which leads to

$$\frac{u_j^{n+1} - u_j^n}{k} + \frac{u_{j+1}^n - u_j^n}{h} = f(x_j, t_n).$$

   This is also called forward time/forward space method.

2. We can also take the right-difference quotient in time and the left difference quotient in space, which leads to
$$\frac{u_j^{n+1} - u_j^n}{k} + \frac{u_j^n - u_{j-1}^n}{h} = f(x_j, t_n).$$
This is also called forward time/backward space method.

3. Another choice would be to take the central difference quotient in space, which leads to
$$\frac{u_j^{n+1} - u_j^n}{k} + \frac{u_{j+1}^n - u_{j-1}^n}{2h} = f(x_j, t_n).$$
This is also called forward time/central space method.

4. Finally, we take the central difference quotient in space and replace in the time difference quotient the u_j^n by the mean of its neighbors, which gives
$$\frac{u_j^{n+1} - (u_{j+1}^n + u_{j-1}^n)/2}{k} + \frac{u_{j+1}^n - u_{j-1}^n}{2h} = f(x_j, t_n).$$
This is called Lax-Friedrichs method.

We now aim to analyze the finite difference methods. For simplicity, we take $f = 0$ in the following. Rearranging the first method gives
$$u_j^{n+1} = (1 + \frac{k}{h})u_j^n - \frac{k}{h}u_{j+1}^n.$$
By the formula above, the approximation to the exact solution u_h at the point x_0, t_0 depends only on the values $u_h(x_0, t_0 - k)$ and $u_h(x_0 + h, t_0 - k)$. Thus only values left of x_0 are used. This can be now iteratively worked back until one reaches $t = 0$ and one still only uses values left of x_0 (see picture below).

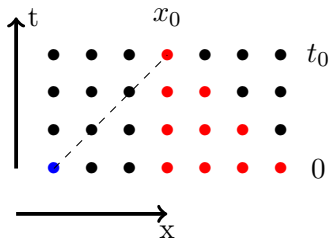


Figure 8.1: The forward time/forward space method.

However, the exact solution is given by $u(x_0, t_0) = u_0(x_0 - t_0)$ (indicated by the blue dot). If, e.g., the initial condition u_0 is 1 at the blue dot, but zero for $x \geq x_0$, the exact solution satisfies
$$u(x_0, t_0) = 1 \qquad \text{but} \qquad u_h(x_0, t_0) = 0.$$
This can not be avoided by taking small time-steps or a finer mesh in space! Therefore, the forward time/forward space method does not converge!

The previous discussion shows that the **numerical domain of dependence**, i.e., the set of all grid points, for which the value of the initial data influence the value of the solution, does not coincide with the domain of dependence for the true solution.
In fact, as $h, k \to 0$ this is a necessary condition for convergence of the finite difference method and called the **CFL-condition** (Courant, Friedrichs, Lax).

Exercise 5. Determine the numerical domain of dependence for the forward time/backward space method. Argue that the CFL-condition has to have the form $\frac{k}{h} \leq 1$. □

8.1.1 Stability analysis

We now make a stability analysis of the FD methods. Stability analysis is usually done using the (discrete) Fourier transform (this is also called von Neumann-analysis).

We now write $u_h^n(x)$ for a discrete function in x, where n denotes the timestep, meaning that $u_h^n(jh) = u_j^n$. All of the mentioned finite difference methods above can then be written in the form

$$u_h^{n+1} = Eu_h^n$$

with a so called propagation operator E. By iteration, one can now write the finite difference method as

$$u_h^n = E^n u_h^0,$$

and we call the method **stable**, if $\|E\| \leq 1$, where one takes the norm

$$\|E\| := \sup \frac{\|Ev\|_{\ell_h^1}}{\|v\|_{\ell_h^1}} \qquad \|v\|_{\ell_h^1} := \sum_{j\in\mathbb{Z}} h\,|v(jh)|\,.$$

The forward time/backward space method

Here, the propagation operator reads as

$$(Eu^n)_j = \left(1 - \frac{k}{h}\right) u_j^n + \frac{k}{h} u_{j-1}^n$$

For simplicity, we assume to have periodic solutions of period length 2π. Then, one can expand the solution into a Fourier series with basis functions

$$\phi_m(x) = e^{imx}$$

Inserting this functions into the propagation operator for the forward time/backward space method, we see that

$$E\phi_m = \left(1 - \frac{k}{h}\right)\phi_m + \frac{k}{h} e^{-imh}\phi_m = \left(1 - \frac{k}{h} + \frac{k}{h} e^{-imh}\right)\phi_m.$$

Thus, the eigenvalues of the propagation operator are $1 - \frac{k}{h} + \frac{k}{h}e^{-imh}$ and provided the CFL condition holds, one can see that

$$|1 - \frac{k}{h} + \frac{k}{h} e^{-imh}| \leq 1$$

and the method is stable.

The forward time/central space method

Here, the propagation operator reads as

$$(Eu^n)_j = u_j^n - \frac{k}{2h}\left(u_{j+1}^n - u_{j-1}^n\right).$$

Inserting the Fourier basis function e^{imx} gives the eigenvalues of the propagation operator as

$$1 - \frac{k}{2h}\left(e^{imh} - e^{-imh}\right) = 1 - i\frac{k}{h}\sin(mh).$$

All of those values have modulus bigger than 1, irrespectively of k, h, so the forward time/central space method is unstable!

8.1.2 Convergence analysis

Consistency is the error the numerical method makes when doing *one step*. More precisely, let $u(x,t)$ be the exact solution and $U_j^n := u(x_j, t_n)$, then the consistency error is defined as

$$\tau_j^{n+1} := \frac{1}{k}\left(U_j^{n+1} - (EU^n)_j\right) - \left(u_t(x_j,t_n) + u_x(x_j,t_n)\right).$$

We call the method **consistent**, if

$$\tau_j^{n+1} \to 0 \qquad \text{for } h,k \to 0.$$

For the forward time/backward space method this can easily be checked using Taylor expansion:

$$\begin{aligned} U_j^{n+1} &= u(x_j, t_n + k) = u(x_j,t_n) + ku_t(x_j,t_n) + \mathcal{O}(k^2) \\ U_{j-1}^n &= u(x_j - h, t_n) = u(x_j,t_n) - hu_x(x_j,t_n) + \mathcal{O}(h^2). \end{aligned}$$

Substituting that into the finite difference method gives

$$0 = \frac{1}{k}\left(U_j^{n+1} - U_j^n\right) + \frac{1}{h}\left(U_j^n - U_{j-1}^n\right) = u_t(x_j,t_n) + u_x(x_j,t_n) + \mathcal{O}(k) + \mathcal{O}(h),$$

which shows consistency of the method.
Convergence of FD methods is usually shown using the following equivalence (for linear problems).

Theorem 8.1. *A finite difference method is convergent, if and only if it is consistent and stable.*

In fact, for the forward time/backward space method we therefore obtain first order convergence ($\mathcal{O}(k+h)$) of the error between exact solution and FD-approximation (in the ℓ_h^1-norm).

8.2 Parabolic partial differential equations

We start with parabolic PDEs for which the equation involves first order derivatives in time, and an elliptic differential operator in space. The prototypical example of this form is the heat equation.

Let $\Omega \subset \mathbb{R}^d$, and $Q = \Omega \times (0,T)$ (also called space-time cylinder). Consider the initial-boundary value problem (here only Dirichlet conditions, but all other types can be done as well) for the heat equation

$$\begin{aligned} \frac{\partial u(x,t)}{\partial t} - \Delta_x u(x,t) &= f(x,t) && (x,t) \in Q \\ u(x,t) &= u_D(x,t) && (x,t) \in \partial\Omega \times (0,T), \\ u(x,0) &= u_0(x) && x \in \Omega. \end{aligned}$$

In order to derive a weak formulation, we first multiply the equation with a test-function $v = v(x)$ vanishing on the boundary $\partial\Omega$ and integrate by parts in the Laplacian-term. This gives

$$\int_\Omega \partial_t u(x,t) v(x)\,dx + \int_\Omega \nabla u(x,t) \cdot \nabla v(x)\,dx = \int_\Omega f(x,t) v(x)\,dx \qquad \forall\, v,\ t \in (0,T].$$

This second term is well-defined, if we require that for each t, $u(x,t)$ should be in $H_0^1(\Omega)$, i.e., $u : [0,T] \to H_0^1(\Omega)$. Moreover, if the time-derivative $\partial_t u$ is in $L^2(\Omega)$ for each t, i.e., $\partial_t u : [0,T] \to L^2(\Omega)$, the first term is well-defined.

One can also write this in abstract form: Find $u : [0,T] \to V$ s.t.

$$(u'(t), v)_{L^2(\Omega)} + a(u(t), v) = (f(t), v)_{L^2(\Omega)} \qquad \forall\, v \in V,\ t \in (0,T] \tag{8.1}$$

or in operator form (with $\langle Au, v\rangle = a(u,v)$):

$$u'(t) + Au(t) = f(t) \qquad \in V^*.$$

This can be understood as an ordinary differential equation in time!

For simplicity, we assume that – as a function in t – the solution is continuously differentiable. This is in accordance of the solution theory, we discussed on ODEs in the lecture *Applied Mathematics Foundations*.
Consequently, we at first define the normed space

$$C([0,T], V) \qquad \|u\|_{C([0,T],V)} := \sup_{t\in[0,T]} \|u(t)\|_V,$$

which means that for all fixed t, $u(t)$ is a function in V (in the spatial variables) and as a mapping in t, we have continuity. As this should also hold for the time derivative, we seek solutions in the space

$$X = C^1((0,T), V) := \{u \in C([0,T], V)\ :\ \partial_t u \in C((0,T), V)\}.$$

For solution to parabolic equations there holds a decay estimate.

Lemma 8.2. *Let $a(\cdot,\cdot)$ be a continuous, coercive bilinear form with coercivity constant μ. Then, there there exists a unique solution $u \in X$ to (8.1). Additionally, u satisfies the estimate*

$$\|u(t)\|_{L^2(\Omega)} \leq Ce^{-\mu t}\|u_0\|_{L^2(\Omega)} + \int_0^t e^{-\mu(t-s)}\|f(s)\|_{L^2(\Omega)}ds.$$

Proof. For simplicity, we only show the case $f = 0$. Choose the test functions $v = u(t)$, then

$$(u'(t), u(t))_{L^2(\Omega)} + a(u(t), u(t)) = (f(t), u(t))_{L^2(\Omega)} = 0.$$

Using

$$\frac{d}{dt}\|u(t)\|^2_{L^2(\Omega)} = 2(u'(t), u(t))_{L^2(\Omega)},$$

as well as the coercivity of $a(\cdot,\cdot)$ gives

$$\frac{1}{2}\frac{d}{dt}\|u(t)\|^2_{L^2(\Omega)} = -a(u(t), u(t)) \leq -\mu\|u(t)\|^2_{L^2(\Omega)}.$$

We aim to write this as an inequality for a derivative. Therefore, using the "integrating factor" $e^{2\mu t}$, we can write with the product rule

$$\frac{d}{dt}\left(e^{2\mu t}\|u(t)\|^2_{L^2(\Omega)}\right) = e^{2\mu t}\left(2\mu\|u(t)\|^2_{L^2(\Omega)} + \frac{d}{dt}\|u(t)\|^2_{L^2(\Omega)}\right) \leq 0$$

Now, integration of the equation from 0 to t gives

$$e^{2\mu t}\|u(t)\|^2_{L^2(\Omega)} - e^{2\mu 0}\|u(0)\|^2_{L^2(\Omega)} \leq 0$$

and moving the last term to the right shows the statement. ■

In physical terms this means, that solutions to parabolic PDEs with $f = 0$ are dissipative. Moreover, parabolic equations have a smoothing effect in time, i.e., if $u_0 \in L^2(\Omega)$, one has that $u(t)$ is infinitely times differentiable in time!

Remark. One could also weak the notion of solution in accordance to the theory we discussed in the previous chapters, i.e., one could also impose the differentiablity in time in a weak sense and seek solutions in the space

$$H^1((0,T), V) = \{v \in L^2((0,T), V) : v' \in L^2((0,T), V^*)\}$$

with norm

$$\|v\|^2_{H^1} := \|v\|^2_X + \|v'\|^2_{X^*}.$$

Provided the bilinear form $a(\cdot,\cdot)$ satisfies the assumptions of the Lax-Milgram lemma, this formulation is uniquely solvable.

8.2.1 Semi-discretization

We now aim to derive a computable formulation in some finite dimensional space.
We start with a discretization in the space variables. Choose a (finite element) sub-space $V_h \subset V$. The Galerkin discretization is: Find $u : [0,T] \to V_h$ such that

$$(u_h'(t), v_h)_{L^2(\Omega)} + a(u_h(t), v_h) = (f(t), v_h)_{L^2(\Omega)} \qquad \forall\, v_h \in V_h,\ \forall\, t \in (0,T],$$

and initial conditions

$$(u_h(0), v_h)_{L^2(\Omega)} = (u_0, v_h)_{L^2(\Omega)} \qquad \forall\, v_h \in V_h.$$

We now choose a basis $\{\varphi_1, \dots \varphi_N\}$ of V_h and expand the solution with respect to this basis:

$$u_h(x,t) = \sum_{i=1}^{N} u_i(t)\varphi_i(x).$$

Using the basis functions as test-functions in the weak formulation, i.e., $v = \varphi_j$, this leads to the mass matrix M and the stiffness matrix A (in space) defined as

$$M = \big((\varphi_j, \varphi_i)_{L^2(\Omega)}\big)_{i,j=1,\dots,N} \qquad A = \left(a(\varphi_j, \varphi_i)\right)_{i,j=1,\dots,N},$$

and the t-dependent load vector

$$f(t) = \big((f(t), \varphi_j)_{L^2(\Omega)}\big)_{i=1,\dots,N}.$$

With this definitions the PDE in time and space reduces (exercise!) to the system of ODEs

$$Mu'(t) + Au(t) = f(t), \qquad u(0) = u_0.$$

Thus, we now only have to numerically solve the ODE, which – by the Picard-Lindelöf theorem – has a unique solution.

8.2.2 Time integration methods

Next, we discuss methods for solving the system of ODEs:

$$\begin{aligned} Mu'(t) + Au(t) &= f(t) \\ u(0) &= u_0. \end{aligned} \tag{8.2}$$

We focus on simple time integration rules and the specific properties arising from the space-discretization of parabolic PDEs. Let

$$0 = t_0 < t_1 < t_m = T,$$

be a partitioning of the interval $[0,T]$. Define $k_j = t_{j+1} - t_j$. Integrating (8.2) over the sub-interval $[t_j, t_{j+1}]$ leads to

$$M\left(u(t_{j+1}) - u(t_j)\right) + \int_{t_j}^{t_{j+1}} Au(s)\, ds = \int_{t_j}^{t_{j+1}} f(s)\, ds.$$

Next, we replace the integrals by numerical integration rules, which gives a computable approximation. The left-sided rectangle rule (i.e. evaluating the integrand at t_j) leads to

$$M(u(t_{j+1}) - u(t_j)) + k_j A u(t_j) = k_j f(t_j).$$

With the notation $u_j = u(t_j)$, $f_j = f(t_j)$, this leads to the sequence of linear equations

$$M u_{j+1} = M u_j + k_j (f_j - A u_j)$$

or for invertible M (usually the inversion/solution of the linear system with M can be cheaply done, in some cases the mass matrix is even diagonal!)

$$u_{j+1} = u_j + k_j (M^{-1} f_j - M^{-1} A u_j).$$

This is called the **explicit Euler method**.

Using the right-sided rectangle rule for the integration (i.e., evaluation of the integrand at t_{j+1}) leads to

$$M(u_{j+1} - u_j) + k_j A u_{j+1} = k_j f_{j+1},$$

or

$$(M + k_j A) u_{j+1} = M u_j + k_j f_{j+1}.$$

Here, a linear system must be solve in any case. Thus, this method is a so called implicit time integration method, the **implicit Euler method**.

A third simple choice is the trapezoidal quadrature rule leading to

$$(M + \frac{k_j}{2} A) u_{j+1} = M u_j + \frac{k_j}{2} (f_j + f_{j+1} - A u_j),$$

which is the so called **Crank-Nicholson method**.
It is also an implicit method. Since the trapezoidal integration rule is more accurate, we expect a more accurate method for approximating the ODE.

8.2.3 Stability and Error Analysis

In order to do an error analysis, we analyze the semi-discretization error first. The idea hereby is to split the error into two parts using the so called **Ritz projector** (also called elliptic projector)

$$R_h : V \to V_h : \qquad a(R_h u, v_h) = a(u, v_h) \qquad \forall\, u \in V, \forall\, v_h \in V_h.$$

Theorem 8.3. *There holds the estimate for the error*

$$\max_{t\in[0,T]} \|u(t) - u_h(t)\|_{L^2(\Omega)} \leq C \left(\max_{t\in[0,T]} \|u(t) - R_h u(t)\|_{L^2(\Omega)} + \max_{t\in[0,T]} \|(u - R_h u)'(t)\|_{L^2(\Omega)} \right).$$

Proof. The error is split into two parts:

$$u(t) - u_h(t) = \underbrace{u(t) - R_h u(t)}_{\rho(t)} + \underbrace{R_h u(t) - u_h(t)}_{\Theta_h}$$

The first part, $u(t) - R_h u(t)$ is the elliptic discretization error, which can be bounded by Cea's lemma. To bound the second term, we use the properties for the continuous and the discrete formulation:

$$\begin{aligned}(f, v_h)_{L^2(\Omega)} &= (u', v_h)_{L^2(\Omega)} + a(u, v_h) = (u', v_h)_{L^2(\Omega)} + a(R_h u, v_h) \\ &= (u_h', v_h)_{L^2(\Omega)} + a(u_h, v_h),\end{aligned}$$

i.e.,

$$(u' - u_h', v_h)_{L^2(\Omega)} + a(R_h u - u_h, v_h) = 0,$$

or

$$(R_h u' - u_h', v_h)_{L^2(\Omega)} + a(R_h u - u_h, v_h) = (R_h u' - u', v_h)_{L^2(\Omega)}.$$

With the abbreviations from above, we obtain the discrete parabolic equation for Θ_h:

$$\begin{aligned}(\Theta_h', v_h)_{L^2(\Omega)} + a(\Theta_h, v_h) &= (\rho', v_h)_{L^2(\Omega)} \\ \Theta_h(0) &= R_h u(0) - u_h(0).\end{aligned}$$

Using Θ_h as a test function together with

$$(\Theta_h', v_h)_{L^2(\Omega)} = \frac{1}{2}\frac{d}{dt}\|\Theta_h\|^2_{L^2(\Omega)} = \|\Theta_h\|_{L^2(\Omega)}\frac{d}{dt}\|\Theta_h\|_{L^2(\Omega)}$$

gives

$$\|\Theta_h\|_{L^2(\Omega)}\frac{d}{dt}\|\Theta_h\|_{L^2(\Omega)} + a(\Theta_h, \Theta_h) = (\rho', \Theta_h)_{L^2(\Omega)} \leq \|\rho'\|_{L^2(\Omega)}\|\Theta_h\|_{L^2(\Omega)}.$$

Since $a(\Theta_h, \Theta_h) \geq 0$, this implies

$$\frac{d}{dt}\|\Theta_h\|_{L^2(\Omega)} \leq \|\rho'\|_{L^2(\Omega)}.$$

Integration in time from 0 to T gives

$$\begin{aligned}\|\Theta_h(t)\|_{L^2(\Omega)} &\leq \|\Theta_h(0)\|_{L^2(\Omega)} + \int_0^T \|\rho'\|_{L^2(\Omega)} dt = \|\rho(0)\|_{L^2(\Omega)} + \int_0^T \|\rho'\|_{L^2(\Omega)} dt \\ &\leq \max_{t\in[0,T]} \|\rho(t)\|_{L^2(\Omega)} + T \max_{t\in[0,T]} \|\rho'(t)\|_{L^2(\Omega)},\end{aligned}$$

which proves the estimate. ■

As the error $u - R_h u$ is just a FEM error for an elliptic PDE, we can use error estimates deduced in previous sections, e.g., $\|u - R_h u\|_{L^2(\Omega)} \leq h^2 \|u\|_{H^2(\Omega)}$. Together with the observation that $\partial_t R_h u$ is the Ritz projection of ∂u (which follows directly from differentiation of the definition of R_h in time), we deduce that

$$\max_{t\in[0,T]} \|u(t) - u_h(t)\|_{L^2(\Omega)} \leq Ch^2 \left(\|u\|_{C([0,T],H^2(\Omega))} + \|\partial_t u\|_{C([0,T],H^2(\Omega))} \right).$$

We note that the regularity requirements on u made here can be significantly weakened by showing a more refined estimate in the previous theorem.

Stability of time integration methods

We now proceed to include the time discretization. Before we analyze the fully discrete method, we briefly only look at the time integration methods. Consider the ODE

$$y' = \lambda y.$$

Then all mentioned time integration methods can be written as $y^1 = R(k\lambda)y^0$ or

$$y^n = R(k\lambda)^n y^0.$$

The function $R(\cdot)$ is called the **stability function** of the time integration method. For the presented methods we have

- $R(z) = 1 + z$ for the explicit Euler-method;
- $R(z) = \frac{1}{1-z}$ for the explicit Euler-method;
- $R(z) = \frac{1+z/2}{1-z/2}$ for the Crank-Nicoloson-method.

We call methods satisfying

$$|R(z)| \leq 1 \qquad \forall z \in (-\infty, 0]$$

A-stable. We note that A-stability actually implies

$$|y^n| = |R(k\lambda)|^n \left|y^0\right| \leq \left|y^0\right| \qquad \forall k > 0, \lambda \leq 0,$$

which means that discrete solutions remain bounded. From our stability estimate for the continuous problem, we actually observe that the exact solution to the heat equation converges to zero for $t \to \infty$. Therefore, the time-integration method should also capture this behaviour.
This is true for so called **L-stable** methods that satisfy

$$|R(z)| \to 0 \qquad z \to -\infty$$

since

$$|y^n| = |R(k\lambda)|^n \left|y^0\right| \to 0 \qquad \text{for } \lambda < 0 \text{ and } k \to \infty.$$

The implicit Euler method is A-stable and L-stable, while the explicit Euler method is not A-stable. In fact, the convergence of the time integration methods fit into the framework of the previous subsection, i.e., the time integration method is convergent if it is consistent and stable. Consistency holds for all three methods. The Euler-methods converge with order 1, while the Crank-Nicholson method is of order 2.

We now turn back to the ODE, we obtained from the semi-discretization (and take $f = 0$ for simplicity)

$$Mu'(t) + Au(t) = 0.$$

Denoting the eigenvalues of the matrix $M^{-1}A$ by

$$\sigma_h := \{\lambda_j \ : \ j = 1, \ldots, N\},$$

and the corresponding eigenvectors by φ_j, $j = 1, \dots, N$, one can see that, if u would be $u = v_j(t)\varphi_j$ one obtains the equation

$$u'(t) + M^{-1}Au(t) = u'(t) + \lambda_j u = 0.$$

Thus, we obtain an equation of the form $y' = -\lambda_j y$ and the eigenvalues λ_j are always non-negative (since the bilinear form $a(\cdot, \cdot)$ that induces the matrix A is coercive). Expressing the function u by a linear combination over eigenvalues and eigenfunctions, this now shows that we are indeed in the setting of the previous discussion and we actually should employ an – at least – A-stable method. In order to estimate the error, we use the following lemma.

Lemma 8.4. *Let $u_h(t)$ be the semi-discrete solution. Then, there holds*

$$\|u_h(t_n) - u_h^n\|_{L^2(\Omega)} \leq \sup_{\lambda \in \sigma_h} |F_n(k\lambda)| \, \|u_h(0)\|_{L^2(\Omega)},$$

where

$$F_n(z) := e^{-nz} - (R(-z))^n.$$

The use of this lemma is that one only has to estimate $F_n(z)$. For the implicit Euler method, one can, e.g., show

$$|F_n(\lambda k)| := \left| e^{-n\lambda k} - \frac{1}{(1+z)^n} \right| \leq C k t_n^{-1}.$$

Combining the error estimate for the semi-discretization and the time-stepping method, one obtains

$$\|u(t_n) - u_h^n\|_{L^2(\Omega)} \leq \|u(t) - u_h(t_n)\|_{L^2(\Omega)} + \|u_h(t_n) - u_h^n\|_{L^2(\Omega)} \leq C(h^2 + k).$$

In fact, for a consistent, A-stable and L-stable method that additionally satisfies $R(z) \leq q < 1$ for all $z < -z_0 < 0$ one can always estimate

$$|F_n(\lambda k)| \leq C k^p t_n^{-p},$$

where p is the order of the time-stepping method.

The explicit Euler method is not A-stable and does not fit into the previous discussion. We write it as

$$u^{n+1} = (I - kM^{-1}A)u^n.$$

Here the "amplification factor" $(I - kM^{-1}A)$ might lead to an instable method. In terms of the stability function, there has to hold

$$|R(-\lambda k)| = |1 - \lambda k| \leq 1.$$

Thus, since $\lambda > 0$, there actually has to hold

$$\max_{\lambda \in \sigma_h} \lambda k \leq 2.$$

For the heat equation, the eigenvalues λ correspond to discrete eigenvalues for the Laplacian. Using Lagrangian finite elements of order 1 on a uniform mesh, one can actually bound

$$\max_{\lambda \in \sigma_h} \lambda \leq C h^{-2}$$

with a constant $C > 0$ depending only on the domain Ω. Therefore, one obtains conditional stability under the CFL-type condition $C \frac{k}{2h^2} \leq 1$.

Remark. The implicit and explicit Euler method combined with piecewise linear finite elements in space give first order convergence in time and space. Higher order methods can also be employed. As usual, in space one takes Lagrangian finite elements of degree $p > 1$. For the time integration, the Crank-Nicholoson method is of second order, higher order methods can be obtained by so-called **Runge-Kutta methods**.

8.3 Hyperbolic partial differential equations

In the following, we consider second order hyperbolic PDE. The prototype of such equations is the wave equation

$$\frac{\partial^2 u(x,t)}{\partial t^2} - \Delta u = f.$$

Hyperblox equations require two initial conditions

$$\begin{aligned} u(x,0) &= u_0(x), \\ \frac{\partial u(x,t)}{\partial t} &= v_0(x). \end{aligned}$$

Again, we consider the time-stepping semi-dscretization method.
Space discretization is accoring to parabolic problems, and lead to the second order ODE

$$M u''(t) + A u(t) = f,$$

and initial conditions

$$u(0) = u_0 \qquad u'(0) = v_0.$$

By introducing a new function $v = u'$, the second order ODE can be reduced to the first order system

$$\begin{aligned} u' &= v \\ M v' &= f - Au, \end{aligned}$$

and initial conditions

$$u(0) = u_0 \qquad v(0) = v_0.$$

Time integration methods for first order systems can be applied.

8.3.1 Time-stepping methods for wave equations

We consider the method of lines, where we first discretize in space, and then apply some time-stepping method for the ODE. In principal, one can reduce the second order ODE to a first order system, and apply some standard time-stepping method for it. This will in general require the solution of linear systems of twice the size. In addition, the structure (symmetric and positive definite) may be lost, which makes it difficult to solve.
We consider two approaches specially taylored for wave equations

(a) for the second order equation,

(b) for first order systems.

The Newmark time-stepping method

We consider the ordinary differential equation

$$Mu'' + Au = f$$

as well as single-step methods: From given state $u_n \approx u(t_n)$ and velocity $u'_n \approx u'(t_n)$, we compute u_{n+1} and u'_{n+1}. The acceleration $u''_n = M^{-1}(f_n - Au_n)$ with $f_n = f(t_n)$ follows from the equation. The Newmark method is based on a Taylor expansion for u and u', where second order derivatives are approximated from old and new accelerations. The real parameters β and γ will be fixed later, k is the time-step:

$$\begin{aligned} u_{n+1} &= u_n + ku'_n + k^2\left[(\tfrac{1}{2}-\beta)u''_n + \beta u''_{n+1}\right] && (8.3)\\ u'_{n+1} &= u'_n + k\left[(1-\gamma)u''_n + \gamma u''_{n+1}\right]. && (8.4) \end{aligned}$$

Inserting the formula for u_{n+1} into $Mu'' + Au = f$ at time t_{n+1} we obtain

$$Mu''_{n+1} + A\left(u_n + ku'_n + k^2\left[(\tfrac{1}{2}-\beta)u''_n + \beta u''_{n+1}\right]\right) = f_{n+1}.$$

Now, we keep unknows left and put known variables to the right:

$$\left[M + \beta k^2 A\right] u''_{n+1} = f_{n+1} - A\left(u_n + ku'_n + k^2(\tfrac{1}{2}-\beta)u''_n\right).$$

The Newmark method requires to solve one linear system with the (symmetric, positive definite) matrix $M + k^2\beta A$, for which efficient direct or iterative methods are available. After computing the new acceleration, the new state u_{n+1} and velocity u'_{n+1} are computed from the explicit formulas (8.3) and (8.4).

The Newmark method satisfies a discrete energy conservation (take $f = 0$):

$$\left[\tfrac{1}{2}u'Mu' + \tfrac{1}{2}u^T A_{eq} u\right]_n^{n+1} = -(\gamma - \tfrac{1}{2})(u_{n+1} - u_n)A_{eq}(u_{n+1} - u_n),$$

where

$$A_{eq} = A + (\beta - \tfrac{1}{2}\gamma)k^2 A M^{-1} A,$$

using the notation $[E]_a^b := E(b) - E(a)$. From this, we get the conservation of a modified energy with the so called *equivalent* stiffness matrix A_{eq}. Depending on the parameter γ we get

- $\gamma = \frac{1}{2}$: conservation
- $\gamma > \frac{1}{2}$: damping
- $\gamma < \frac{1}{2}$: growth of energy (unstable)

If A_{eq} is positive definite, then this conservation proves stability. This is unconditionally true if $\beta \geq \frac{1}{2}\gamma$ (the method is called unconditionally stable). If $\beta < \frac{1}{2}\gamma$, the allowed time step is limited by

$$k^2 \leq \lambda_{max}(M^{-1}A)^{-1} \frac{1}{\frac{1}{2}\gamma - \beta}.$$

For second order problems we have $\lambda_{max}(M^{-1}A) \simeq h^{-2}$, and thus $k \leq Ch$ which is a CFL type condition.
Choices for β and γ of particular interests are:

- $\gamma = \frac{1}{2}, \beta = \frac{1}{4}$: unconditionally stable, conservation of original energy ($A_{eq} = A$)
- $\gamma = \frac{1}{2}, \beta = 0$: conditionally stable. We have to solve

  $$Mu''_{n+1} = f_{n+1} - A(u_n + ku'_n + \tfrac{\tau^2}{2}u''_n),$$

  which is explicit, if M is cheaply invertible.

For $\gamma > 1/2$, the Newark method is of first order, for $\gamma = 1/2$, one even has quadratic convergence in time.

Methods for the first order system

We reduce the wave equation

$$u'' - \Delta u = f$$

to a first order system of PDEs. We introduce $\sigma = \int_0^t \nabla u$. Then,

$$\begin{aligned} \sigma' &= \nabla u \\ u' - \operatorname{div}\sigma &= \tilde{f} \end{aligned}$$

with the integrated source $\tilde{f} = \int_0^t f$. In the following, we set $f = 0$.
A mixed variational formulation in $H(\operatorname{div}, \Omega) \times L^2(\Omega)$, for given initial conditions $u(0)$ and $\sigma(0)$, is:

$$\begin{aligned} (\sigma', \tau)_{L^2(\Omega)} &= -(u, \operatorname{div}\tau)_{L^2(\Omega)} \qquad \forall\, \tau \in H(\operatorname{div}, \Omega) \\ (u', v)_{L^2(\Omega)} &= (v, \operatorname{div}\sigma)_{L^2(\Omega)} \qquad \forall\, v \in L^2(\Omega). \end{aligned}$$

With mass matrices M_σ (corresponding to evaluation of $(\sigma, \tau)_{L^2(\Omega)}$ with basis functions of $H(\operatorname{div}, \Omega)$-finite elements, e.g. Raviart-Thomas elements) and M_u (corresponding to evaluation of $(u, v)_{L^2(\Omega)}$ with basis functions in $L^2(\Omega)$, e.g., piecewise constant functions) as well as the stiffness matrix B (corresponding to evaluation of the bilinear form $b(u, \operatorname{div}\tau)$ at the (already) chosen FE-bases in $L^2(\Omega) \times H(\operatorname{div}, \Omega)$), after space discretization we obtain the system of ODEs

$$\begin{pmatrix} M_\sigma & 0 \\ 0 & M_u \end{pmatrix} \begin{pmatrix} \sigma' \\ u' \end{pmatrix} = \begin{pmatrix} 0 & -B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \sigma \\ u \end{pmatrix}.$$

Conservation of energy is now seen from

$$\frac{d}{dt}\left[\tfrac{1}{2}\sigma^T M_\sigma \sigma + \tfrac{1}{2}u^T M_u u\right] = \sigma^T M_\sigma \sigma' + u^T M_u u' = -\sigma^T B^T u + u^T B\sigma = 0.$$

Methods taylored for the skew-symmetric (Hamiltonian) structure are so called **symplectic methods**: The symplectic Euler method is

$$\begin{aligned} M_\sigma \frac{\sigma_{n+1} - \sigma_n}{k} &= -B^T u_n, \\ M_u \frac{u_{n+1} - u_n}{k} &= B\sigma_{n+1}. \end{aligned}$$

For updating the second variable, the new value of the first variable is used. For the analysis, we can reduce the large system to 2×2 systems, where β are singular values of $\widetilde{B} := M_\sigma^{-1/2} B M_u^{-1/2}$ (square-roots of eigenvalues of $\widetilde{B}^T\widetilde{B}$):

$$\sigma' = -\beta u \qquad u' = \beta\sigma.$$

The symplectic Euler method can be written as

$$\begin{pmatrix} \sigma_{n+1} \\ u_{n+1} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 \\ k\beta & 1 \end{pmatrix}\begin{pmatrix} 1 & -k\beta \\ 0 & 1 \end{pmatrix}}_{T=\begin{pmatrix} 1 & -k\beta \\ k\beta & 1-(k\beta)^2 \end{pmatrix}} \begin{pmatrix} \sigma_n \\ u_n \end{pmatrix}$$

The eigenvalues of T satisfy $\lambda_1\lambda_2 = \det(T) = 1$, and, if $k\beta \le \sqrt{2}$, they are conjugate complex, and thus $|\lambda_1| = |\lambda_2| = 1$. Thus, the discrete solution is oscillating without damping or growth.
Again, diagonal mass matrices M_u and M_σ would render explicit methods efficient.

The symplectic Euler method is of first order.