Computational PDEs

(Theorie und Numerik partieller Differentialgleichungen)

D. Gallistl

Lecture in Summer 2021 at the Universität Jena
Last update: 30th June 2023

Our summer semester has 14 weeks (weeks 15–28). Each section in this lecture notes contains material for one week, corresponding to two lectures and a problem session.

Contents

Topic 1: Linear elliptic problems

§1 Dirichlet problem and Dirichlet principle *(week 15)*

In this lecture we study partial differential equations (PDEs) and their numerical approximation. We confine ourselves to linear equations of second order. Let us first define what we mean by this.

Definition 1.1. Let $n \in \mathbb{N}$ and $\Omega \subseteq \mathbb{R}^n$ be an open subset. Let furthermore a map $F : \mathbb{R}^{n\times n} \times \mathbb{R}^n \times \mathbb{R} \times \Omega \to \mathbb{R}$ be given. We call the equation

$$F(D^2u(x), \nabla u(x), u(x), x) = 0 \quad \text{for all } x \in \Omega$$

a *partial differential equation of 2nd order.* Any function $u : \Omega \to \mathbb{R}$ satisfying the above relation is called a *solution.*

The foregoing definition is rather abstract. At the same time, it implicitly requires further properties (differentiability) of the solution, which are not stated explicitly. We will work with this basic definition and will proceed with examples. The equation is called *partial* differential equation because it involves partial derivatives of the solution (in contrast to *ordinary differential equations (ODEs)*, which only depend on one scalar variable. The notion of *2nd order* describes that the highest involved derivative of u has order 2. At this point, the function F can be arbitrarily nonlinear.

Example 1.2. For a given function $f \in C(\Omega)$ (usually referred to as *right-hand side*) and F given by $F(A, b, c, x) = \det A - f(x)$, we obtain the equation

$$\det D^2u(x) = f(x).$$

It is called *Monge–Ampère equation.*

Recall the Laplacian $\Delta u(x) = \operatorname{div} \nabla u(x) = \sum_{j=1}^{n} \partial_{jj} u(x) = \operatorname{tr} D^2u(x)$, where $\operatorname{tr} A$ denotes the trace of a matrix A.

Example 1.3. For $f \in C(\Omega)$ and $F(A, b, c, x) = \operatorname{tr} A + f(x)$ we obtain *Poisson's equation*

$$-\Delta u(x) = f(x).$$

What is the difference between these two examples? Poisson's equation is *linear.* This means that, given solutions u to the right-hand side f and v to the right-hand side g, the equation

$$-\Delta w(x) = \alpha f(x) + \beta g(x)$$

will be satisfied by the linear combination $w := \alpha u + \beta v$, $(\alpha, \beta \in \mathbb{R})$. This is easy to verify. It is also elementary to verify that the Monge–Ampère equation does not have this property. We expect in general that

$$\det D^2(u(x) + v(x)) \neq f(x) + g(x),$$

for solutions u and v to right-hand sides f and g, respectively. Convince yourself of this fact by setting up suitable examples.

Definition 1.4. A 2nd order PDE is called *linear*, if it is of the form

$$\sum_{|\alpha|\leq 2} a_\alpha(x)\partial^\alpha u(x) = f(x).$$

Here, a_α and f are given functions over Ω. The above sum runs over all multi-indices α of length ≤ 2, and ∂^α is the partial derivative with respect to α.

We will start this lecture by considering Poisson's equation, the most basic instance of a PDE that is rich enough to highlight all relevant concepts. Generally, we pose the questions of *existence* of a solution to a PDE and its *uniqueness*. Even for Poisson's equation we will quickly reach certain limits that we will later overcome with tools of linear functional analysis.
Clearly, solutions to Poisson's equation are not unique without any further constraints being imposed. For instance, any solution can be shifted by an arbitrary affine function and will still remain a solution. We will thus consider the *Dirichlet problem*, which imposes a zero boundary condition on the solution. This PDE is posed on a domain $\Omega \subseteq \mathbb{R}^n$ which is open, bounded, and connected.

Definition 1.5. Let $\Omega \subseteq \mathbb{R}^n$ be open, bounded, and connected. A function $u \in C^2(\Omega)\cap C(\bar{\Omega})$ is said to solve the Dirichlet problem with right-hand side $f \in C(\Omega)$ if it satisfies

$$-\Delta u = f \text{ in } \Omega \quad \text{und} \quad u = 0 \text{ on } \partial\Omega.$$

It will often be important to impose more structure on the boundary $\partial\Omega$.

Definition 1.6. A domain Ω has a *Lipschitz boundary*, if there are finitely many open sets $U^1, \dots, U^N \subseteq \mathbb{R}^n$ that cover a neighbourhood of the boundary $\partial\Omega$ and have the property that, for any $j \in \{1, \dots, N\}$, the set $\partial\Omega \cup U^j$ can be represented as the graph of a Lipschitz function so that γ_j such that $\Omega \cap U^j$ lies on exactly one side of the graph.
Let us give a formal definition in two dimensions. There exists some $N \in \mathbb{N}$, finitely many open sets $U^1, \dots, U^N$, open intervals $I_j \subset \mathbb{R}$ and Lipschitz continuous functions $\gamma_j : \bar{I}_j \to \mathbb{R}^2$ on $\bar{I}_j$ with the following property: The U^j cover a neighbourhood U of $\partial\Omega$, i.e. $U \subseteq \cup_{j=1}^N U^j$, and any U^j satisfies (after some shift and rotation of the coordinate system)

- $U^j \cap \partial\Omega = \{(y, \gamma(y)) : y \in I\}$
- $U^j \cap \Omega \subseteq \{(y, z) : y \in I, z > \gamma(y)\}$

□

In our examples we will mostly deal with simple Lipschitz domains with boundaries consisting of piecewise smooth curve segments. It is known that Lipschitz domains posses, almost everywhere on the boundary, a well-defined outer unit normal vector ν. The divergence theorem teaches us the following: For a bounded Lipschitz domain Ω and a vector field $v \in C^1(\Omega; \mathbb{R}^n)$ we have

$$\int_{\partial\Omega} v \cdot \nu \, ds = \int_\Omega \operatorname{div} v \, dx.$$

Here (and throughout this text) we denote integration with respect to the n-dimensional Lebesgue measure by the symbol "dx" while integration with respect to the $(n-1)$-dimensional surface measure is indicated by "ds". The divergence theorem implies the formula of integration by parts: For two differentiable functions u and v we have

$$\int_\Omega (u\,\partial_j v + v\,\partial_j u)dx = \int_{\partial\Omega} uv\,\nu_j ds$$

for any $j \in \{1,\dots,n\}$, where ν_j is the jth component of the outer unit normal. A variant thereof is called *Green's formula*

$$\int_\Omega (u\Delta v + \nabla u \cdot \nabla v)dx = \int_{\partial\Omega} u\frac{\partial v}{\partial \nu} ds,$$

where $v \in C^2(\Omega) \cap C^1(\bar\Omega)$ is assumed.

Theorem 1.7 (uniqueness)**.** *The solution to the Dirichlet problem is unique.*

Proof. Let u and v be two solutions to the Dirichlet problem with right-hand side f. The linearity then implies that the difference $w := u - v$ satisfies $-\Delta w = 0$. Integration by parts results in

$$0 = -\int_\Omega w\Delta w\,dx = \int_\Omega |\nabla w|^2\,dx.$$

Here, no boundary term occurs because w vanishes on the boundary. Therefore, ∇w equals zero almost everywhere in Ω. Hence, w is constant; and since $w = 0$ on $\partial\Omega$, we have that w is the constant zero function whence $u = v$ in Ω. □

At this point we are not yet in the position to formulate a satisfactory existence theory. With the spaces of differentiable functions used above we may lose control over derivatives, which makes it often difficult to justify numerical methods. Let us discuss the following illustrative example.

Example 1.8. Let $\Omega = (-1,1)^2 \setminus ([0,1] \times [-1,0])$ be the Γ-shaped (or L-shaped) domain. Let u be given by

$$u(x,y) = (1-x^2)(1-y^2)r^{2/3}\sin\left(\frac{2\varphi}{3}\right).$$

Here, we use polar coordinates $0 < r < 1$ and $0 < \varphi < 3\pi/2$; note that $x = r\cos\varphi$ and $y = r\sin\varphi$. One can verify that u satisfies $-\Delta u = f$ for some $f \in C^0(\bar\Omega)$ and $u|_{\partial\Omega} = 0$. But u does not possess bounded derivatives and, thus, does not belong to $C^1(\bar\Omega)$.

We will now characterize the Dirichlet problem as an optimization problem. To this end, we will employ basic methods from the calculus of variations. The most important tool is the following.

Lemma 1.9 (fundamental lemma of calculus of variations)**.** *(a) Let the function $g \in C^0(\Omega)$ satisfy*

$$\int_\Omega g\psi\,dx = 0$$

for all $\psi \in C_c^\infty(\Omega)$ (smooth functions with compact support). Then $g = 0$ holds in the whole domain Ω.

(b) The assertion of (a) remains valid if $g \in L^1_{\text{loc}}(\Omega)$ (with the same conclusion almost everywhere in Ω).

Proof. Exercise. □

Theorem 1.10 (Dirichlet principle)**.** *Let $\Omega \subseteq \mathbb{R}^n$ be a bounded Lipschitz domain. Let $V = \{v \in C^1(\bar{\Omega}) : v = 0 \text{ on } \partial\Omega\}$ and $f \in C^0(\Omega)$. A function $u \in V$ satisfies $-\Delta u = f$ in Ω if and only if it minimizes the functional $J : V \to \mathbb{R}$ given by*

$$J(v) = \frac{1}{2}\int_\Omega |\nabla v|^2 dx - \int_\Omega f v \, dx \qquad (v \in V)$$

over V. Here, $|\cdot|$ denotes the Euclidean norm. Any solution to the Dirichlet problem in particular satisfies the necessary condition

$$\int_\Omega \nabla u \cdot \nabla v dx = \int_\Omega f v dx \quad \text{for all } v \in V.$$

Note that above we chose $V \subseteq C^1(\bar{\Omega})$ to ensure that the integral involving the gradient is finite. One can weaken this requirement.

Proof. If $u \in V$ minimizes the functional J, then a necessary criterion is that $J \leq J(u + tv)$ for small perturbations $t > 0$, $v \in V$. This means that $J(u + tv)$ has a minimum at $t = 0$, which implies for the directional derivative that

$$0 = \frac{d}{dt} J(u + tv).$$

The chain rule implies

$$\left(\frac{d}{dt}\frac{1}{2}\int_\Omega |\nabla(u + tv)|^2 dx\right)\bigg|_{t=0} = \int_\Omega \nabla u \cdot \nabla v dx.$$

It is furthermore elementary to verify

$$\left(\frac{d}{dt}\int_\Omega f\,(u + tv)\, dx\right)\bigg|_{t=0} = \int_\Omega f v dx.$$

This shows that necessary condition

$$\int_\Omega \nabla u \cdot \nabla v dx = \int_\Omega f v dx \quad \text{for all } v \in V. \tag{1}$$

Furthermore, integration by parts (Green's formula) implies

$$\int_\Omega \nabla u \cdot \nabla v dx = \int_\Omega (-\Delta u) v \, dx.$$

Due to the fact that $v \in V$, no boundary term occurs. Altogether, we have shown that

$$\int_\Omega (-\Delta u - f)v \, dx = 0 \quad \text{for all } v \in V.$$

We thus conclude with the fundamental lemma of calculus of variations that $-\Delta u = f$ holds pointwise in Ω.
Let us now assume that $u \in V$ solves Poisson's equation. Green's formula then implies (1) for every $v \in V$. A direct computation, with arbitrary $v \in V$, results in

$$\begin{aligned} J(u+v) - J(u) &= \frac{1}{2} \int_\Omega \left(|\nabla(u+v)|^2 - |\nabla u|^2 \right) \, dx - \int_\Omega f v \, dx \\ &= \int_\Omega \nabla u \cdot \nabla v \, dx - \int_\Omega f v \, dx + \frac{1}{2} \int_\Omega |\nabla v|^2 \, dx = \frac{1}{2} \int_\Omega |\nabla v|^2 \, dx \geq 0. \end{aligned}$$

For the first identity we have used the elementary formula $|b|^2 - |a|^2 = 2a \cdot (b-a) + |a-b|^2$; for the second identity we have used relation (1). Thus, J is minimal at u. □

Dirichlet's principle shows that solving Poisson's equation with boundary conditions is equivalent to solving the corresponding minimization problem. In terms of calculus of variations we say that Poisson's equation is the *Euler–Lagrange equation* corresponding to the minimization problem. Thus, we ask the question under which conditions and in which spaces minimizers of the functional J exist. This will lead us (later in this course) to the concept of *Sobolev spaces*. For the moment we just remark that the formulation as a minimization problem requires weaker conditions on u than the original Dirichlet problem: We only need *first derivatives* to exist. The Laplacian does not explicitly show up in the functional J. It is contained implicitly or *weakly* in that formulation. We will formalize this via the concept of *weak derivatives*.

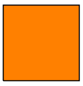
Problem 1. Let the following function be given

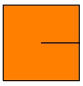
$$\Phi(x) = \begin{cases} -\frac{1}{2\pi} \log |x| & \text{if } n = 2 \\ \frac{1}{n(n-2)\alpha(n)} \frac{1}{|x|^{n-2}} & \text{if } n \geq 2. \end{cases}$$

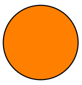
Here, $\alpha(n) \neq 0$ is some real number. Show that $\Delta\Phi(x) = 0$ holds for all $x \in \mathbb{R}^n \setminus \{0\}$.

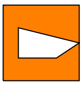
Problem 2. Prove, based on the divergence theorem, the formula of integration by parts as well as Green's formula.

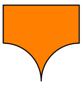
Problem 3. Which of the following domains possess a Lipschitz boundary?

(a) 

(b) 

(c) 

(d) 

(e) 

The lower part of the boundary of (e) is parametrized by $y = \sqrt{|x|}$.

Problem 4. Prove that the Laplacian is represented in polar coordinates (r, φ) as follows

$$\Delta = \frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \varphi^2}.$$

Problem 5. Verify the statements from Example 1.8.

Problem 6. Prove the fundamental lemma of calculus of variations.

§2 Weak derivatives and discrete functions *(week 16)*

The Dirichlet principle showed us that we can understand derivatives in some weaker sense.

Definition 1.11 (weak derivative)**.** Let $\Omega \subseteq \mathbb{R}^n$ be open. Let $v \in L^1_{\text{loc}}(\Omega)$ and $j \in \{1, \dots, n\}$. If there exists a function $g \in L^1_{\text{loc}}(\Omega)$ with the property

$$\int_\Omega v \partial_j \psi \, dx = - \int_\Omega g \psi \, dx \quad \text{for all } \psi \in C_c^\infty(\Omega),$$

then this function g is called the *weak partial derivative* of v with respect to the direction j, and it is denoted by $\partial_j v$. The vector of all partial derivatives is denoted (provided it exists) by ∇v.

Remark 1.12. The weak derivative is unique (see problems).

The idea behind this definition is to extend the common notion of differentiability. If v is differentiable, then the weak and the classical derivatives coincide. There are, however, functions that are not differentiable in the classical sense, but possess a weak derivative.

Example 1.13. The absolute value function $v(x) = |x|$ on $\Omega = (-1, 1)$ is not differentiable on $(-1, 1)$. Yet, its weak derivative is given by

$$v'(x) = \begin{cases} -1 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0. \end{cases} \tag{2}$$

Note that we can modify elements of $L^1_{\text{loc}}(\Omega)$ at $x = 0$ to any value.

From the example we see that functions with certain kinks can be weakly differentiable.

Example 1.14. We subdivide the interval $(-1, 1)$ into finitely many sub-intervals $[x_j, x_{j+1}]$, where

$$-1 = x_1 < \cdots < x_N = 1 \quad \text{and } \cup_{j=1}^{N-1} [x_j, x_{j+1}] = \bar{\Omega} = [-1, 1],$$

and consider the globally continuous functions that are affine when restricted to any of the sub-intervals $[x_j, x_{j+1}]$. Any such function is weakly differentiable.

The functions from the foregoing example allow for a very simple representation, and so they are generally suited for numerical computations. It is easy to verify that any such function can be characterized by the vector $(v(x_j))_{j=1}^N$ of its values at the points x_j. Between these nodal points, the values are interpolated by straight lines.
It is possible to generalize this construction to higher space dimensions. We only consider the case $n = 2$ in this lecture in order to minimize the technical efforts. Let the domain $\bar{\Omega}$ be subdivided in triangles. We consider the space of functions that are globally continuous and that are affine when restricted to any of the triangles. In order to define such spaces, we introduce a suitable class of triangular partitions.

Definition 1.15 (triangle)**.** A subset $T \subseteq \mathbb{R}^2$ is called *triangle* if there exists $(z_1, z_2, z_3) \in (\mathbb{R}^2)^3$ such that T is the convex hull of z_1, z_2, z_3 and these three points do not belong on one straight line. The points z_1, z_2, z_3 are called *vertices*. The line segments between z_j, z_k for $j \neq k$ are called *edges*.

Definition 1.16 (regular triangulation)**.** Let $\mathcal{T} \subset 2^{\bar{\Omega}}$ be a finite set of triangles in $\bar{\Omega}$ ($2^{\bar{\Omega}}$ denotes the power set). The set $\mathcal{T}$ is called a *regular triangulation* of Ω if die the triangles cover the domain $\bar{\Omega}$, i.e., $\bigcup_{T\in\mathcal{T}} = \bar{\Omega}$, and if any pair $(T, K) \in \mathcal{T}^2$ satisfies one of the following relations:

(i) $T \cap K = \emptyset$

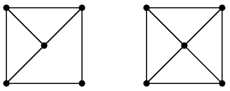
(ii) $T \cap K$ is a common vertex

(iii) $T \cap K$ is a common edge

(iv) $T = K$.

This means that the elements of a regular triangulation may only meet under certain rules.

Example 1.17. A non-regular and a regular triangulation of the square:



In what follows, $\mathcal{T}$ will always denote a regular triangulation of Ω. Let $T \in \mathcal{T}$ be a triangle. The affine functions over T are denoted by

$$P_1(T) := \{v \in L^\infty(T) : \exists (a, b, c) \in \mathbb{R}^3 \forall x \in T,\ v(x) = a + bx_1 + cx_2\}.$$

The functions that are piecewise affine with respect to $\mathcal{T}$ (but possibly globally discontinuous) are denoted by

$$P_1(\mathcal{T}) := \{v \in L^\infty(\Omega) : \forall T \in \mathcal{T}, v|_T \in P_1(T)\}.$$

Finally, the continuous and piecewise affine functions are denoted by

$$S^1(\mathcal{T}) := C^0(\Omega) \cap P_1(\mathcal{T})$$

and the subspace with zero boundary conditions reads

$$S^1_0(\mathcal{T}) := \{v \in S^1(\mathcal{T}) : v|_{\partial\Omega} = 0\}.$$

The letter S shall remind us of *splines*; a notion that is possibly known from one-dimensional interpolation.
The following property is very important, and its proof is discussed in the problems below.

Lemma 1.18. *The elements of $S^1(\mathcal{T})$ are weakly differentiable.*

Proof. Exercise. □

The set of vertices (or *nodes*) of a triangle is denoted by $\mathcal{N}(T)$ and the set of all vertices is

$$\mathcal{N} := \{z \in \bar{\Omega} : \text{there exists } T \in \mathcal{T} \text{ having } z \text{ as a vertex}\} = \bigcup_{T\in\mathcal{T}} \mathcal{N}(T).$$

The basis we choose for $S^1(\mathcal{T})$ or $S^1_0(\mathcal{T})$ is the *nodal basis*. First, we define the nodal basis of $S^1(\mathcal{T})$ (no boundary conditions) by $(\varphi_z)_{z\in\mathcal{N}}$, where for any $z \in \mathcal{N}$ the function $\varphi_z \in S^1(\mathcal{T})$ is defined by the property

$$\varphi_z(y) = \delta_{yz} = \begin{cases} 1 & \text{if } y = z \\ 0 & \text{if } y \in \mathcal{N} \setminus \{z\}. \end{cases} \tag{3}$$

These functions are usually referred to as “hat functions”. It will be shown in the exercises that these function indeed form a basis.

In order to define a nodal basis of $S^1_0(\mathcal{T})$, one omits the hat functions belonging to boundary vertices. To this end, we define the boundary vertices by $\mathcal{N}(\partial\Omega) := \partial\Omega \cap \mathcal{N}$ and the inner vertices by $\mathcal{N}(\Omega) := \mathcal{N} \setminus \mathcal{N}(\partial\Omega)$. The nodal basis of $S^1_0(\mathcal{T})$ then reads

$$(\varphi_z : z \in \mathcal{N}(\Omega)).$$

As in classical Lagrange interpolation, the coefficients with respect to the nodal basis are given by the nodal values. This means that any function $v_h \in S^1(\mathcal{T})$ can be expanded as follows

$$v_h = \sum_{z\in\mathcal{N}} v_h(z)\varphi_z.$$

The spaces $S^1(\mathcal{T})$ and $S^1_0(\mathcal{T})$ are called *finite element spaces*. Any continuous function $v \in C(\bar{\Omega})$ can be approximated by its interpolation $Iv \in S^1(\mathcal{T})$ as follows

$$Iv := \sum_{z\in\mathcal{N}} v(z)\varphi_z.$$

The map $I : C(\bar{\Omega}) \to S^1(\mathcal{T})$ is called *interpolation operator*. For the case of zero boundary conditions, the definition is analogous.

Let us now briefly discuss how to operate with triangulations and finite element functions on a computer (using Python).

We describe a triangulation by prescribing a list of nodes and a list of triangles. The nodes are put in a list `coord` $\in \mathbb{R}^{N\times 2}$. The x and y coordinate of the jth node are written to the jth row. In the example of Figure 1 this means

```
coord = np.asarray([[0,0],
                    [1,0],
                    [1,1],
                    [0,1],
                    [.5,.5]])
```

for the unit square $(0,1)^2$. Here, we use the library `numpy`:

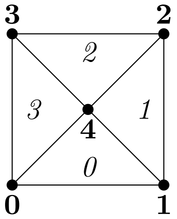


Figure 1: Triangulation of the square $(0,1)^2$ in four triangles. The bold numbers indicate the node numbers wile the numbers of the triangles are displayed in italic.

```
import numpy as np
```

Now we form triangles out of the node numbers. We use convention that the numbering is counterclockwise. The list `triangles` $\in \mathbb{R}^{N\times 3}$ contains in its jth row the three node numbers of triangle number j. In the example from Figure 1 this reads

```
triangles = np.asarray([[0,1,4],
                        [1,2,4],
                        [2,3,4],
                        [3,0,4]])
```

We finally save the node pairs of the boundary edges on the Dirichlet boundary

```
dirichlet= np.array([[0,1],
                     [1,2],
                     [2,3],
                     [3,0]])
```

We will comment on (and make use of) this later. In Python we can now plot our triangulation by:

```
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.tri as mtri

plt.triplot(mtri.Triangulation(coord[:,0], coord[:, 1], triangles))
plt.show()
```

If we want to generate a surface plot of a piecewise affine function from $S^1(\mathcal{T})$, we can use `trisurf`. Figure 2 shows a complete example.
The triangulation in the above example is very coarse. Finer triangulations can be obtained by mesh refinement. A very simple refinement rule is called *red refinement*. Here, every triangle is subdivided in four congruent sub-triangles by connecting the edge midpoints by straight lines. We provide a routine `red_refine.py` on the lecture webpage. We do not care about the actual code, but we just use it. It can be used as follows

```
import numpy as np
import matplotlib.pyplot as plt
import matplotlib.tri as mtri
from mpl_toolkits import mplot3d
from mpl_toolkits.mplot3d import Axes3D

coord = np.asarray([[0,0],[1,0],[1,1],[0,1],[.5,.5]])
triangles = np.asarray([[0,1,4],[1,2,4],[2,3,4],[3,0,4]])
dirichlet= np.array([[0,1],[1,2],[2,3],[3,0]])
# show triangulation
plt.triplot(mtri.Triangulation(coord[:,0], coord[:, 1], triangles))
plt.show()
# plot the interpolation of the function x+y
func = lambda x, y:  x + y
func2=np.vectorize(func)
z=func2(coord[:,0],coord[:,1])
fig = plt.figure(figsize =(14, 9))
ax = plt.axes(projection ='3d')
trisurf = ax.plot_trisurf(coord[:,0],coord[:,1],z,
                          triangles = triangles,
                          cmap =plt.get_cmap('summer'),
                          edgecolor='Gray');
plt.show()
```

Figure 2: Sample use of `triplot` and `trisurf`

```
neumann=np.zeros([0, 2])
coord, triangles, dirichlet,_,_,_ = \
            red_refine(coord, triangles, dirichlet, neumann)
```

Here, neumann is just an empty list that, at this stage, has no importance. Later in the lecture we will also consider problems with a second type of boundary condition (so-called Neumann boundary condition), but for the moment we can ignore it; we also do not care about the three ignored output arguments of the function.

Problem 7. Show that the weak derivative is unique. (Hint: fundamental lemma of calculus of variations)

Problem 8. Show that the function $v(x) = \log(|\log(|x|)|)$ on the unit disk $\Omega = \{x \in \mathbb{R}^2 : |x| < 1\}$ is weakly differentiable but neither bounded nor continuous on Ω.

Problem 9. Show that the notions of classical and weak derivative coincide for continuously differentiable functions.

Problem 10. Draw a regular triangulation of the square $(0, 1)^2$ with 7 triangles.

Problem 11. Let K, T be triangles that intersect in one point $z = T \cap K$. The point z is vertex to T but not to K. Such point is called a *hanging node*. Draw a picture of this situation and convince yourself that regular triangulations cannot contain any hanging node.

Problem 12. Prove the assertions from Example 1.14. Draw plots of such piecewise affine function for some examples.

Problem 13. Prove the claims from Example 1.13.

Problem 14. Is the sign function from (2) weakly differentiable?

Problem 15. Show that the nodal basis $(\varphi_z)_{z\in\mathcal{N}}$ forms a partition of unity.

Problem 16. Show that the functions φ_z are uniquely defined by (3) and that they form as basis of $S^1(\mathcal{T})$. Draw the graph of one of the basis functions φ_z on an example triangulation.

Problem 17 (barycentric coordinates)**.** Let $T \subseteq \mathbb{R}^2$ be a triangle with vertices z_1, z_2, z_3. Show that to any point $x \in T$ there exist unique real numbers $\lambda_1(x)$, $\lambda_2(x)$, $\lambda_3(x)$ with the properties

$$x = \lambda_1(x)z_1 + \lambda_2(x)z_2 + \lambda_3(x)z_3 \qquad \text{and} \qquad \lambda_1(x) + \lambda_2(x) + \lambda_3(x) = 1.$$

The λ_j are called *barycentric coordinates*. Show furthermore that the barycentric coordinates (as functions of x) coincide with the three nodal basis functions for the vertices of T.

Problem 18. Start from the example triangulation from Figure 1 and plot the interpolation of the function $u(x, y) = \sin(12\pi x)y^2$ on a sequence of 6 red-refined triangulations.

§3 The finite element method *(week 17)*

So far we did not study any existence theory to the Dirichlet problem of Poisson's equation. Instead we first introduce the central numerical method of this lecture: the finite element method (FEM). We want to use it to approximately solve Poisson's and other equations. In the course of this lecture we will then justify the method by convergence theory. But for the moment we motivate the scheme in a purely heuristic manner in order to be able to quickly proceed with practical results.

The point of departure for the FEM is that the energy functional J from the Dirichlet principle (Theorem 1.10) is well defined for elements from $S^1(\mathcal{T})$ or $S^1_0(\mathcal{T})$. Indeed, when considering the functional

$$J(v) = \frac{1}{2}\int_\Omega |\nabla v|^2 dx - \int_\Omega f v\, dx$$

we see that, for $v \in S^1(\mathcal{T})$, the gradient ∇v is defined in the sense of weak derivatives. It belongs to $L^2(\Omega)$ (it is even piecewise constant). Thus, the first part of the sum is finite. The second term is finite as well if we impose the (fairly weak) condition $f \in L^2(\Omega)$: the Hölder (or Cauchy-Schwarz) inequality then reveals

$$\left|\int_\Omega f v\, dx\right| \leq \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)}.$$

Since we are considering the homogeneous Dirichlet problem (i.e., a zero boundary condition), we restrict the attention to approximations from the subspace $S^1_0(\mathcal{T})$. In our notation, we indicate that we are dealing with "discrete functions" by adding the index h to the variables. As an approximation to the solution u to the Dirichlet problem with right-hand side $f \in L^2(\Omega)$ we seek $u_h \in S^1_0(\mathcal{T})$ minimizing the functional J over the finite-dimensional space $S^1_0(\mathcal{T})$, written

$$u_h \in \operatorname*{arg\,min}_{v_h \in S^1_0(\mathcal{T})} J(v_h). \tag{4}$$

Theorem 1.19. *Let $\Omega \subseteq \mathbb{R}^2$ be an open, bounded, polygonal Lipschitz domain with a regular triangulation $\mathcal{T}$. Given any $f \in L^2(\Omega)$, there exists a unique $u_h \in S^1_0(\mathcal{T})$ solving the discrete optimization problem* (4)*. Furthermore,* (4) *is equivalent to the condition*

$$\int_\Omega \nabla u_h \cdot \nabla v_h dx = \int_\Omega f v_h dx \quad \textit{for all } v_h \in S^1_0(\mathcal{T}). \tag{5}$$

Proof. Like in the proof of Theorem 1.10 one can show that the minimization problem (4) leads to the (in this case discrete) Euler–Lagrange equation (5) as a necessary condition. Details will be worked out in the problem sessions. Let us now have a closer look at (5). Note carefully that the left-hand side of (5) is a positive definite bilinear form, while the right-hand side is a linear form (that is, an element of the dual space). This is shown as an exercise. It is actually immediate to see that the left-hand side is bilinear and positive semidefinite. The definiteness follows from the Dirichlet boundary condition. For a better overview, we now formulate (5) in terms of vectors and matrices.

Let the dimension of the finite-dimensional space $S^1_0(\mathcal{T})$ be denoted by $N \in \mathbb{N}$. Then $S^1_0(\mathcal{T})$ has the nodal basis $(\varphi_1, \dots \varphi_N) \in (S^1_0(\mathcal{T}))^N$. With respect to this basis, we can represent the parts

$$a(u_h, v_h) := \int_\Omega \nabla u_h \cdot \nabla v_h dx \quad \text{und} \quad F(v_h) := \int_\Omega f v_h dx$$

from (5) in the usual way as matrices as follows: Let $v_h = \sum_{j=1}^N x_j \varphi_j$ and $w_h = \sum_{k=1}^N y_k \varphi_k$ be elements of $S^1_0(\mathcal{T})$. We then have

$$a(v_h, w_h) = x^\top A y \quad \text{for } A \in \mathbb{R}^{N \times N} \text{ where } A_{jk} = a(\varphi_j, \varphi_k) \quad (j, k = 1, \dots, N).$$

Similarly, we have for $F \in V_h^*$ that

$$F(w_h) = b^\top y \quad \text{for } b \in \mathbb{R}^N \text{ where } b_k = F(\varphi_k) \quad (k = 1, \dots, N).$$

Here, $x, y \in \mathbb{R}^n$ are the vectors with entries x_j, y_k. If we expand $u_h = \sum_j x_j \varphi_j$ with respect to the given basis, then the coefficients x of the solution u_h (if existent) satisfy the system

$$A^\top x = b. \tag{6}$$

This means that we have transformed (5) in the equivalent matrix-vector system (6). For this system it is immediate that it is uniquely solvable because A is positive definite (because $a(\cdot, \cdot)$ is). Thus, there exists a unique solution $u_h \in S^1_0(\mathcal{T})$ to (5). What is left to be shown is that u_h minimizes the functional J. Since A is symmetric, we derive with (6) the relation

$$\begin{aligned} J(u_h + w_h^y) - J(u_h) &= \frac{1}{2}(x+y)^\top A^\top (x+y) - (x+y)^\top b \\ &= \frac{1}{2} x^\top A^\top x - x^\top b + y^\top A^\top y \\ &= J(u_h) + y^\top A^\top y \geq J(u_h) \quad \text{for each } y \in \mathbb{R}^N \text{ and } w_h^y = \sum_{k=1}^N y_k \varphi_k, \end{aligned}$$

where the last estimate follows with the positive definiteness of A. Hence, J attains a minimum at u_h. The minimum is unique as the necessary condition (5) is satisfied by exactly one minimizer (namely u_h). □

In summary, we have formulated the finite-dimensional problem (4) as the linear system (6). Let us remark that the form (6) is more general in that it does not require symmetry but only definiteness. Let us now describe how to implement the FEM on the computer. For this, we need to assemble the matrix A and the vector b from the foregoing proof within a computer program.

In order to discretize Poisson's equation (subject to homogeneous Dirichlet boundary conditions) with the FEM, we need

- a triangulation $\mathcal{T}$, described through the data structures `coord`, `triangles`, `dirichlet`,
- the right-hand side f, e.g. given through values at certain points or as function,

```
import numpy as np
from mpl_toolkits import mplot3d
import matplotlib.tri as mtri
import matplotlib.pyplot as plt
from mpl_toolkits import mplot3d
from mpl_toolkits.mplot3d import Axes3D
from mpl_toolkits.mplot3d import proj3d
import math
import pylab
from red_refine import red_refine #our refinement routine
import scipy.sparse
import scipy.sparse.linalg
from scipy.sparse import csr_matrix
```

Figure 3: The required packages in Python

```
def FEM(coord,triangles,dirichlet,f):
    nnodes=np.size(coord,0)
    A=stiffness_matrix(coord,triangles)
    b=RHS_vector(coord,triangles,f)
    dbnodes=np.unique(dirichlet)
    dof=np.setdiff1d(range(0,nnodes),dbnodes)
    A_inner=A[np.ix_(dof,dof)]
    b_inner=b[dof]
    x=np.zeros(nnodes)
    x[dof]=scipy.sparse.linalg.spsolve(A_inner,b_inner)
    return x
```

Figure 4: The basic FEM routine.

- a vector b representing the linear functional $\int_\Omega f \bullet dx$ with respect to the nodal basis of $S^1(\mathcal{T})$,
- the so-called *stiffness matrix* A, i.e., the matrix representing the bilinear form from Poisson's equation with respect to the nodal basis of $S^1(\mathcal{T})$.

With these objects at hand, we can solve (6). It is important to restrict the matrices to the *degrees of freedom.* In our case, these correspond to the inner nodes (as the values for the boundary nodes are already fixed by the value 0). The list of degrees of freedom is usually given the variable name `dof`.

We start by specifying all required packages, see Figure 3. The structure of the program is displayed in Figure 4.

It remains to describe the routines for assembling the stiffness matrix A and the right-hand side vector b. We start with A. First, we build up *local* stiffness matrices for each triangle T

$$A_T^{loc} := (\int_T \nabla\varphi_j \cdot \nabla\varphi_k \, dx)_{j,k=1,2,3}.$$

Here, the vertices of T are locally numbered by $1, 2, 3$. Since the φ_j are affine functions, their gradients are constant so that we arrive at the formula

$$A_T^{loc} = \operatorname{area}(T) \begin{bmatrix} \nabla\varphi_1^\top \\ \nabla\varphi_2^\top \\ \nabla\varphi_3^\top \end{bmatrix} \begin{bmatrix} \nabla\varphi_1 \ \nabla\varphi_2 \ \nabla\varphi_3 \end{bmatrix}.$$

The area is easily computed as follows. With the three vertices $z_1, z_2, z_3 \in \mathbb{R}^2$ of T, we have that

$$\operatorname{area}(T) = \frac{1}{2}\det[z_2 - z_1, z_3 - z_1].$$

For the computation of $\nabla\varphi_j$ we observe that the basis functions (or barycentric coordinates) satisfy the system

$$\underbrace{\begin{bmatrix} 1 & 1 & 1 \\ z_1 & z_2 & z_3 \end{bmatrix}}_{\in\mathbb{R}^{3\times 3}} \begin{bmatrix} \varphi_1(x) \\ \varphi_2(x) \\ \varphi_3(x) \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \\ x \end{bmatrix}}_{\in\mathbb{R}^{3\times 1}}$$

for any T. If we take derivatives (w.r.t. x) on both sides, we arrive at

$$\underbrace{\begin{bmatrix} 1 & 1 & 1 \\ z_1 & z_2 & z_3 \end{bmatrix}}_{\in\mathbb{R}^{3\times 3}} \underbrace{\begin{bmatrix} \nabla\varphi_1^\top \\ \nabla\varphi_2^\top \\ \nabla\varphi_3^\top \end{bmatrix}}_{\in\mathbb{R}^{3\times 2}} = \underbrace{\begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}}_{\in\mathbb{R}^{3\times 2}}.$$

Therefore

$$\begin{bmatrix} \nabla\varphi_1^\top \\ \nabla\varphi_2^\top \\ \nabla\varphi_3^\top \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ z_1 & z_2 & z_3 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

We compute all local stiffness matrices in a loop

```
nelems=np.size(triangles,0)
Alocal=np.zeros((nelems,3,3))

for j in range(0,nelems):
    nodes_loc=triangles[j,:]
    coord_loc=coord[nodes_loc,:]
    T=np.array([coord_loc[1,:]-coord_loc[0,:] ,
           coord_loc[2,:]-coord_loc[0,:] ])
    area = 0.5 * ( T[0,0]*T[1,1] - T[0,1]*T[1,0] )
    T= np.concatenate((np.array([[1,1,1]]), coord_loc.T),axis=0)
    T1= np.array([[0,0],[1,0],[0,1]])
    grads = np.linalg.solve(T,T1)
    Alocal[j,:,:]=area* np.matmul(grads,grads.T)
```

Now we need to assemble the local stiffness matrices into the global stiffness matrix. The entry A_{jk} of the global stiffness matrix is given by

$$A_{jk} = \int_\Omega \nabla\varphi_j \cdot \nabla\varphi_k \, dx = \sum_{T\in\mathcal{T}} \int_T \nabla\varphi_j \cdot \nabla\varphi_k \, dx = \sum_{\substack{T\in\mathcal{T} \\ \text{nodes } j,k \\ \text{belong to } K}} \int_T \nabla\varphi_j \cdot \nabla\varphi_k \, dx.$$

This means that, for any triangle, we save the index pairs (j, k in the above sum) assigning the global node numbers to the entries of the local stiffness matrix. We write these indices into the arrays I1, I2. We then build up a sparse matrix based on these indices (note that repeated indices imply summation).

```
nelems=np.size(triangles,0)
nnodes=np.size(coord,0)
I1=np.zeros((nelems,3,3))
I2=np.zeros((nelems,3,3))

for j in range(0,nelems):
    nodes_loc=triangles[j,:]
    I1[j,:,:] = np.concatenate((np.array([nodes_loc]),\
          np.array([nodes_loc]),np.array([nodes_loc])),axis=0)
    I2[j,:,:] = np.concatenate((np.array([nodes_loc]).T,\
            np.array([nodes_loc]).T,np.array([nodes_loc]).T),axis=1)

Alocal=np.reshape(Alocal,(9*nelems,1)).T
I1=np.reshape(I1,(9*nelems,1)).T
I2=np.reshape(I2,(9*nelems,1)).T
A=csr_matrix((Alocal[0,:],(I1[0,:],I2[0,:])),shape = (nnodes,nnodes))
```

The full routine for the stiffness matrix can be found in Figure 5. We now proceed with the assembling of the right-hand side. We again run a loop over all elements. Since b is not sparse, we can just update the vector in each loop iteration. For approximating the integral, we use the midpoint rule

$$\int_T f\varphi_j \, dx \approx \text{area}(T) f(m)\varphi_j(m),$$

where $m = \frac{1}{3}(z_1 + z_2 + z_3)$ is the midpoint (barycentre) of T. Since φ_j is affine, we can easily compute $\varphi_j(m) = 1/3$. This results in the routine of Figure 6.
For testing the FEM code, we use the above data for the unit square. In the code they will be loaded by a function geom_square. We use the following right-hand side for validation

$$f(x) = 2(x_1(1 - x_1) + x_2(1 - x_2)).$$

The exact solution reads

$$u(x) = x_1(x_1 - 1)x_2(x_2 - 1).$$

For a very basic convergence test for the L^∞ norm we now execute the following lines of code

```
def stiffness_matrix(coord,triangles):
    nelems=np.size(triangles,0)
    nnodes=np.size(coord,0)
    Alocal=np.zeros((nelems,3,3))
    I1=np.zeros((nelems,3,3))
    I2=np.zeros((nelems,3,3))

    for j in range(0,nelems):
        nodes_loc=triangles[j,:]
        coord_loc=coord[nodes_loc,:]
        T=np.array([coord_loc[1,:]-coord_loc[0,:] ,
               coord_loc[2,:]-coord_loc[0,:] ])
        area = 0.5 * ( T[0,0]*T[1,1] - T[0,1]*T[1,0] )
        tmp1= np.concatenate((np.array([[1,1,1]]), coord_loc.T),axis
           =0)
        tmp2= np.array([[0,0],[1,0],[0,1]])
        grads = np.linalg.solve(tmp1,tmp2)
        Alocal[j,:,:]=area* np.matmul(grads,grads.T)
        I1[j,:,:] = np.concatenate((np.array([nodes_loc]),np.array([
           nodes_loc]),np.array([nodes_loc])),axis=0)
        I2[j,:,:] = np.concatenate((np.array([nodes_loc]).T,np.array
           ([nodes_loc]).T,np.array([nodes_loc]).T),axis=1)

    Alocal=np.reshape(Alocal,(9*nelems,1)).T
    I1=np.reshape(I1,(9*nelems,1)).T
    I2=np.reshape(I2,(9*nelems,1)).T
    A=csr_matrix((Alocal[0,:],(I1[0,:],I2[0,:])),shape = (nnodes,
       nnodes))
    return A
```

Figure 5: Routine for the stiffness matrix.

```
def RHS_vector(coord,triangles,f):
    nelems=np.size(triangles,0)
    nnodes=np.size(coord,0)
    b=np.zeros(nnodes)
    for j in range(0,nelems):
        nodes_loc=triangles[j,:]
        coord_loc=coord[nodes_loc,:]
        tmp=np.array([coord_loc[1,:]-coord_loc[0,:] ,
               coord_loc[2,:]-coord_loc[0,:] ])
        area = 0.5 * ( tmp[0,0]*tmp[1,1] - tmp[0,1]*tmp[1,0] )
        mid=1/3*(coord_loc[0,:]+coord_loc[1,:]+coord_loc[2,:])
        b[nodes_loc]=b[nodes_loc]+area/3*f(mid[0],mid[1])
    return b
```

Figure 6: Routine for the right-hand side vector.

```
fun = lambda x, y:  (x-x**2)*(y- y**2)
u_exact=np.vectorize(fun)
f = lambda x, y:  2* ((x-x**2)+(y- y**2) )
coord, triangles, dirichlet, neumann = get_geom()
max_err=np.zeros(5)
for j in range(0,5):
    coord, triangles, dirichlet,_,_,_ = \
           red_refine(coord, triangles, dirichlet, neumann)
    x=FEM(coord, triangles, dirichlet,f)
    u_at_nodes=u_exact(coord[:,0],coord[:,1])
    max_err[j]=np.max(np.abs(u_at_nodes-x))
print(max_err)
```

Problem 19. Prove the unproven assertions from Theorem 1.19.

Problem 20. Compute the kernel of the local stiffness matrix.

Problem 21. Study all the routines of this section line by line and convince yourself that they are doing what they are expected to do.

Problem 22. Do a convergence study for the unit square and the right-hand side $f(x) = 2(x_1(1-x_1) + x_2(1-x_2))$ (exact solution see above) with respect to the following error (semi-)norm

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

similar to that from the above convergence test. For computing the gradient of u_h on a given element T, use the local representation in terms of the nodal basis. The gradients of the basis vectors were already computed in the loop for the stiffness matrix. Perform an analogous convergence study for the error in the L^2 norm and compare the convergence rates (with respect to the maximal diameter of the triangles in the triangulations, the so-called mesh size). Visualize the results in a loglog-diagram (horizontal axis: mesh size, vertical axis: error in the different norms).

Problem 23. Given a triangle T with barycentric coordinates (nodal basis functions) φ_1, φ_2, φ_3, prove the formula

$$\int_T \varphi_1^a \varphi_2^b \varphi_3^c \, dx = 2|T| \frac{a!b!c!}{(a+b+c+2)!}$$

for any $a, b, c \in \mathbb{N} \cup \{0\}$. (It is enough to show it on the reference triangle with vertices $(0,0)$, $(1,0)$, $(0,1)$ and to then argue by transformation.)

§4 Elementary properties of Sobolev spaces *(week 18)*

We introduce spaces of functions that posses appropriate weak derivatives. It will turn out that these are suited for a sound theory of Poisson's equation (and similar problems). We shall prove many, but not all of the stated results.

Definition 1.20 (Sobolev spaces)**.** Let $\Omega \subseteq \mathbb{R}^2$ be bounded and open. Define

$$H^1(\Omega) := \{v \in L^2(\Omega) : \forall j \in \{1,2\}\ \partial_j v \in L^2(\Omega)\}.$$

That is, the functions from $H^1(\Omega)$ belong to $L^2(\Omega)$; their first weak derivatives exist and belong to $L^2(\Omega)$ as well.

Remark 1.21. For L^p spaces instead of L^2 one can analogously define Sobolev spaces, which are commonly denoted by $W^{1,p}(\Omega)$. We will not consider such spaces in this lecture.

Remark 1.22. In many cases it will be enough for our purposes to confine ourselves to polygonal Lipschitz domains. Most of the results will, however, hold under weaker conditions.

Sobolev functions have far more structure than generic L^2 functions. Recall that elements from $L^2(\Omega)$ are equivalence classes (up to equality almost everywhere) and that point evaluations are not well defined. This is generally the case for Sobolev function, too. Yet, we will see that such functions possess boundary values in some generalized sense. We first study an important property, namely that $H^1(\Omega)$ can equivalently be defined by a completion process. Let us define the following norm on $H^1(\Omega)$,

$$\|v\|_{H^1(\Omega)} := \sqrt{\|v\|^2_{L^2(\Omega)} + \|\nabla v\|^2_{L^2(\Omega)}}.$$

Remark 1.23. We use the convention that $\|\nabla v\|^2_{L^2(\Omega)} = \int_\Omega |\nabla v|^2\,dx$ for the Euclidean norm $|\cdot|$.

Theorem 1.24. *Let $\Omega \subseteq \mathbb{R}^2$ be open and bounded. The space $H^1(\Omega)$ is complete with respect to the norm $\|\cdot\|_{H^1(\Omega)}$, i.e. a Banach space.*

Proof. The proof is left as an exercise (Problem 28). □

In the following, we will use arguments involving the open covering

$$U_j := \{x \in \Omega : \frac{\operatorname{diam}(\Omega)}{2} 2^{-j} \leq \operatorname{dist}(x, \partial\Omega) \leq 2\operatorname{diam}(\Omega)2^{-j}\}$$

for the open and bounded set Ω. This covering is locally finite in the sense that any of the sets U_j has a nonempty intersection with only finitely many sets U_k (exercise). It is known from multivariate analysis that there exists a corresponding smooth partition of unity, that is a family $(\eta_j)_j$ of nonnegative functions $\eta_j \in C^\infty_c(U_j)$ with the property

$$\sum_j \eta_j(x) = 1 \quad \text{for all } x \in \Omega.$$

Theorem 1.25 (approximation by smooth functions I). *Let $\Omega \subseteq \mathbb{R}^2$ be open and bounded. Then the space $H^1(\Omega)$ is the completion of*

$$C^\infty(\Omega) \cap H^1(\Omega)$$

with respect to the norm $\|\cdot\|_{H^1(\Omega)}$. In other words: Given any $v \in H^1(\Omega)$ there exists a sequence $(v_j)_j$ in $C^\infty(\Omega) \cap H^1(\Omega)$ with the property that $\|v - v_j\|_{H^1(\Omega)} \to 0$ for $j \to \infty$.

Proof. Let $v \in H^1(\Omega)$ and $\varepsilon > 0$. The proof uses approximation by convolution, which is known from integration theory. Let $(\psi_\varepsilon)_{\varepsilon>0}$ be a standard Dirac sequence. For $v \in H^1(\Omega)$ we define the approximation

$$v_\varepsilon(x) := (\psi_\varepsilon * 1_\Omega v)(x) = \int_\Omega \psi_\varepsilon(x-y)v(y)dy,$$

where "dy" means integration w.r.t. the Lebesgue measure and the variable y. Given any open subset $D \subseteq \Omega$ with $\delta := \operatorname{dist}(D, \partial\Omega) > 0$ we then have

$$v_\varepsilon \in H^1(D) \cap C^\infty(D) \quad \text{provided } \varepsilon < \delta.$$

Let us prove this claim. It is known that $v_\varepsilon \in C^\infty(D)$ as well as $v_\varepsilon \in L^2(D)$. Moreover, the partial derivatives satisfy due to rotational symmetry of ψ_ε and Lebesgue's theorem

$$\partial_j v_\varepsilon(x) = \int_\Omega \frac{\partial}{\partial x_j}\psi_\varepsilon(x-y)v(y)dy = -\int_\Omega \frac{\partial}{\partial y_j}\psi_\varepsilon(x-y)v(y)dy.$$

We observe that, for any $x \in \Omega$ with $\operatorname{dist}(x, \partial\Omega) > \varepsilon$, the function $y \mapsto \psi_\varepsilon(x-y)$ belongs to $C_c^\infty(\Omega)$. By definition of the weak derivative ∂_j we thus have for such x, after integration by parts, that,

$$\partial_j v_\varepsilon(x) = \int_\Omega \psi_\varepsilon(x-y)\partial_j v(y)dy = (\psi_\varepsilon * 1_\Omega \partial_j v)(x).$$

In other words: The derivative of the regularization is the regularized derivative. By known results from measure and integration theory related to approximation by convolution we thus infer convergence $v_\varepsilon \to v$ in $L^2(D)$ and $\partial_j v_\varepsilon \to \partial_j v$ in $L^2(D)$, and therefore $v_\varepsilon \to v$ in $H^1(D)$ for $\varepsilon \to 0$.

In order to show the result on Ω (and not just on the subsets D) another technical step is required. Let $(U_k)_{k\in\mathbb{N}}$ be a locally finite open covering of Ω and $(\eta_k)_{k\in\mathbb{N}}$ be a corresponding smooth partition of unity as constructed above. Owing to the above results we can find, for any k and any $\varepsilon > 0$, an approximation $v_{k,\varepsilon} \in C^\infty(U_k)$ by convolution such that

$$\|v - v_{k,\varepsilon}\|_{H^1(U_k)} \leq \frac{1}{10}\frac{\varepsilon}{2^k(1 + \|\eta_k\|_{C^1(\bar\Omega)})}.$$

We combine these local approximations and define

$$v_{\Omega,\varepsilon} := \sum_{k\in\mathbb{N}_0} \eta_k v_{k,\varepsilon}.$$

We compute with the product rule (ses also Problem 29)

$$\partial_j(v - v_{\Omega,\varepsilon}) = \sum_{k\in\mathbb{N}_0} \partial_j(\eta_k(v - v_{k,\varepsilon})) = \sum_{k\in\mathbb{N}_0} (\partial_j\eta_k(v - v_{k,\varepsilon}) + \eta_k\partial_j(v - v_{k,\varepsilon}))$$

and obtain with the triangle inequality

$$\begin{aligned}\|\partial_j(v - v_{\Omega,\varepsilon})\|_{L^2(\Omega)} &\leq \sum_{k\in\mathbb{N}_0} (\|\partial_j\eta_k(v - v_{k,\varepsilon})\|_{L^2(U_k)} + \|\eta_k\partial_j(v - v_{k,\varepsilon})\|_{L^2(U_k)}) \\ &\leq 2\sum_{k\in\mathbb{N}_0} \|\eta_k\|_{C^1(\bar{\Omega})}\|v - v_{k,\varepsilon}\|_{H^1(U_k)}\end{aligned}$$

Here we have estimated the terms containing η_k by their maxima; we furthermore used the elementary estimate $a + b \leq 2\sqrt{a^2 + b^2}$ for real a, b. With the above choice of $v_{k,\varepsilon}$ and the geometric series we arrive at

$$\|\partial_j(v - v_{\Omega,\varepsilon})\|_{L^2(\Omega)} \leq \frac{2}{5}\varepsilon.$$

For the L^2 norm we obtain in a similar fashion the direct estimate

$$\|v - v_{\Omega,\varepsilon}\|_{L^2(\Omega)} \leq \sum_{k\in\mathbb{N}_0} \|\eta_k\|_{C^1(\bar{\Omega})}\|v - v_{k,\varepsilon}\|_{L^2(U_k)} \leq \varepsilon/5.$$

Altogether

$$\|v - v_{\Omega,\varepsilon}\|_{H^1(\Omega)} \leq \|v - v_{\Omega,\varepsilon}\|_{L^2(\Omega)} + \sum_{j=1}^{2} \|\partial_j(v - v_{\Omega,\varepsilon})\|_{L^2(\Omega)} \leq \frac{\varepsilon}{5} + \frac{2\varepsilon}{5} + \frac{2\varepsilon}{5} \leq \varepsilon.$$

We have shown that, given any $v \in H^1(\Omega)$, there exists an approximation $v_{\Omega,\varepsilon}$ that converges in the $H^1(\Omega)$-Norm towards v as $\varepsilon \to 0$. □

Theorem 1.26 (approximation by smooth functions II)**.** *Let $\Omega \subseteq \mathbb{R}^2$ be an open and bounded Lipschitz domain. Then, $H^1(\Omega)$ is the completion of*

$$C^\infty(\bar{\Omega})$$

with respect to the norm $\|\cdot\|_{H^1(\Omega)}$. In other words: Given any $v \in H^1(\Omega)$ there exists a sequence $(v_j)_j$ in $C^\infty(\bar{\Omega})$ with the property that $\|v - v_j\|_{H^1(\Omega)} \to 0$ for $j \to \infty$.

Proof. In contrast to Theorem 1.25 we need some regularity of the boundary. Since the domain has a Lipschitz boundary, there are open sets $U^1, \ldots, U^N$ covering a neighbourhood U of $\partial\Omega$ and having the property (after some shift and rotation of the coordinate system) that

$$\Omega \cap U^j \subset \{x \in U^j : x_2 > \gamma(x_1)\} \qquad (7)$$

as well as $\partial\Omega \cap U^j = \mathrm{Graph}(\gamma)$ for some Lipschitz function γ. We choose smooth functions $\phi_j \in C_c^\infty(U^j)$, that form a partition of unity on U (i.e., $\sum_{j=1}^N \phi_j = 1$ in U). In order to cover the inner part of Ω, we choose an open set $U^0 \subset\subset \Omega$ such that $\Omega \subseteq \cup_{j=0}^N U_j$ and set $\phi_0 := 1 - \sum_{j=1}^N \phi_j$. It follows that there is an open domain $\hat{\Omega} \supset\supset \Omega$ for which $(\phi_j : j = 0, \dots, N)$ is a partition of unity. Now we claim that for any $v \in H^1(\Omega)$, any $\varepsilon > 0$, and any $j = 0, \dots, N$ there exists some $w_j \in C_c^\infty(\mathbb{R}^2)$ such that $\|\phi_j v - w_j\|_{H^1(U_j)} \leq \varepsilon/(N+1)$. Assuming for the moment this property, we define $w := \sum_{j=0}^N w_j$. We then immediately infer $w \in C_c^\infty(\mathbb{R}^2)$ and, in particular, $w|_\Omega \in C^\infty(\bar{\Omega})$. The triangle inequality furthermore implies

$$\|v - w\|_{H^1(\Omega)} \leq \sum_{j=0}^N \|\phi_j v - w_j\|_{H^1(\Omega)} \leq \varepsilon$$

which proves the assertion of the theorem. Let us now prove the above claim. For $j = 0$ the claim follows as in Theorem 1.25. Given any $j \in \{1, \dots, N\}$, we choose a local coordinate system according to (7). We extend $v_j := \phi_j v$ by zero to the whole $\mathbb{R}^2$. For small $t > 0$ we then consider the shifted function $v_{j,t}(x) = v_j(x + t \begin{pmatrix} 0 \\ 1 \end{pmatrix})$, whose support locally overlaps beyond $\partial\Omega$. As in the proof of Theorem 1.25 we can approximate $v_{j,t}$ via convolution by functions $\psi_\eta * v_{j,t}$ (these satisfy $(\psi_\eta * v_{j,t})|_{U^j} \in C^\infty(\bar{U}_j)$). In particular, we have $\|v_{j,t} - \psi_\eta * u_{j,t}\|_{H^1(\Omega)} \to 0$ for $\eta \to 0$. On the other hand, we have $\|v_{j,t} - v_j\|_{H^1(\Omega)} \to 0$ for $t \to 0$ (see Problem 30). For any prescribed $\delta > 0$ we thus conclude

$$\|v_{j,t} - v_j\|_{H^1(\Omega)} < \delta/2 \quad \text{for sufficiently small } t > 0.$$

Next, we choose $\eta > 0$ small enough such that

$$\|v_{j,t} - \psi_\eta * v_{j,t}\|_{H^1(\Omega)} < \delta/2$$

and obtain with the triangle inequality that

$$\|v_j - \psi_\eta * v_{j,t}\|_{H^1(\Omega)} < \delta.$$

□

Problem 24. Show that the finite element space satisfies $S^1(\mathcal{T}) \subseteq H^1(\Omega)$.

Problem 25. Let $\mathcal{T}$ be a regular triangulation of $\Omega \subseteq \mathbb{R}^2$ and let $v \in P_1(\mathcal{T})$ be a piecewise affine function. For each interior edge F with adjacent triangles T_+ and T_- (i.e., $F = T_+ \cap T_-$), the jump across F is defined by $[v]_F := v|_{T_+} - v|_{T_-}$.

(a) Prove that

$$v \in H^1(\Omega) \iff [v]_F = 0 \quad \text{for all interior edges } F.$$

(b) The space $H(\mathrm{div}, \Omega)$ is defined by

$$H(\mathrm{div}, \Omega) := \left\{ v \in L^2(\Omega; \mathbb{R}^2) \;\middle|\; \begin{array}{l} \exists g \in L^2(\Omega) \text{ such that for all } \varphi \in C_c^\infty(\Omega) \\ \int_\Omega v \cdot \nabla\varphi \, dx = -\int_\Omega g\varphi \, dx \end{array} \right\}.$$

Prove that

$$v \in H(\mathrm{div}, \Omega) \iff [v \cdot \nu_F]_F = 0 \quad \text{for all interior edges } F$$

where ν_F is some normal vector of F.

Problem 26. Show that $\|\cdot\|_{H^1(\Omega)}$ is a norm on $H^1(\Omega)$. Does $\|\nabla \cdot\|_{L^2(\Omega)}$ define a norm on $H^1(\Omega)$ as well?

Problem 27. Let $v(x) = \log(|\log(|x|)|)$ be given on the disc $\Omega = \{x \in \mathbb{R}^2 : |x| < 1/\exp(1)\}$. Prove $v \in H^1(\Omega)$ (cf. Problem 8).

Problem 28. Prove that $H^1(\Omega)$ is complete with respect to $\|\cdot\|_{H^1(\Omega)}$. (*Hint:* You may use the result that $L^2(\Omega)$ is complete (Fischer-Riesz Theorem)).

Problem 29. Let $v, w \in H^1(\Omega)$. Prove that the product vw is weakly differentiable and satisfies $\partial_j(vw) = (\partial_j v)w + v(\partial_j w)$. Does vw belong to $H^1(\Omega)$?

Problem 30. Let $\Omega \subset \hat{\Omega}$ be bounded open sets and let $u \in H^1(\hat{\Omega})$ be a function with compact support within $\hat{\Omega}$. Prove that the shifted function $u_t(x) = u(x + tb)$ for some fixed vector $b \in \mathbb{R}^2$ satisfies $\|u_t - u\|_{H^1(\Omega)} \to 0$ for $t \to 0$. *(Hint: Approximate u in the L^2 norm by a smooth function ϕ and use uniform continuity of ϕ.)*

Problem 31. Let T be a triangle with with set of vertices $\mathcal{N}(T)$. Given $y \in \mathcal{N}(T)$, denote by $\varphi_y \in P_1(T)$ local hat function with

$$\varphi_y(z) = \delta_{yz} \quad \text{for all } z \in \mathcal{N}(T).$$

Compute the following 3×3 matrices

$$\begin{aligned}
M_T &:= \left(\int_T \varphi_y \varphi_z \, dx \right)_{(y,z) \in (\mathcal{N}(T))^2} && \text{(local mass matrix)} \\
C_T &:= \left(\int_T \varphi_y (\beta \cdot \nabla \varphi_z) \, dx \right)_{(y,z) \in (\mathcal{N}(T))^2} && \text{(local convection matrix with given } \beta \in \mathbb{R}^2\text{)} \\
S_T &:= \left(\int_T \nabla \varphi_y \cdot \nabla \varphi_z \, dx \right)_{(y,z) \in (\mathcal{N}(T))^2} && \text{(local stiffness matrix).}
\end{aligned}$$

You may use the formula from Problem 23. The gradients can be assumed to be given as a matrix $[\nabla \varphi_j^\top]_{j=1}^3$ as in prior sections.

§5 Traces; the Dirichlet problem in Sobolev spaces *(week 19)*

In the previous lecture we have seen hat suitable smooth functions are dense in $H^1(\Omega)$. As a first application of this result we will show that we can assign boundary values to functions from $H^1(\Omega)$ in a consistent fashion. Such property is, obviously, impossible to achieve for mere $L^2(\Omega)$ functions.

Theorem 1.27 (trace identity and trace inequality for triangles)**.** *Let $T \subseteq \mathbb{R}^2$ be a triangle with some edge $F \subseteq T$ and opposite vertex $P \in T$. Any function $v \in C^1(T)$ then satisfies*

$$\frac{|T|}{|F|}\int_F v\,ds = \int_T v\,dx + \frac{1}{2}\int_T (\bullet - P)\cdot\nabla v\,dx$$

and

$$\|v\|^2_{L^2(F)} \leq \frac{3|F|}{2|T|}\|v\|^2_{L^2(T)} + \frac{|F|}{2|T|}\operatorname{diam}(T)^2\|\nabla v\|^2_{L^2(T)}.$$

Here, $|T|$ denotes the area of T and $|F|$ denotes the length of F.

Proof. We have $\operatorname{div}(\bullet - P) = 2$ (in two space dimensions). Integration by parts therefore reveals

$$\int_T v\,dx + \frac{1}{2}\int_T (\bullet - P)\cdot\nabla v\,dx = \int_{\partial T} v\,(\bullet - P)\cdot\nu\,ds,$$

where ν is the outer unit normal of T. We observe that, on the two edges of T different from F, the vector $(\bullet - P)$ is tangential to ∂T and, thus, its product with ν equals zero. Hence,

$$\int_{\partial T} v\,(\bullet - P)\cdot\nu\,ds = \int_F v\,(\bullet - P)\cdot\nu\,ds.$$

Since furthermore ν is constant along F, the quantity $(\bullet - P)\cdot\nu$ is constant on F as well, and its value corresponds to the orthogonal projection of $(\bullet - P)$ in direction of ν. This is precisely the length of the height on F, which by elementary geometry takes the value $2|T|/|F|$. This proves the first assertion.
In order to show the second claimed property, we apply the trace identity to v^2. Note that $\nabla(v^2) = 2v\nabla v$. We thus infer

$$\frac{|T|}{|F|}\int_F v^2\,ds = \int_T v^2\,dx + \int_T (\bullet - P)\cdot v\nabla v\,dx \leq \int_T v^2\,dx + \operatorname{diam}(T)\int_T |v|\,|\nabla v|\,dx,$$

where in the second step we have estimated the length of $(\bullet - P)$ by the diameter of T. After rearranging the identity we obtain

$$\|v\|^2_{L^2(F)} \leq \frac{|F|}{|T|}\|v\|^2_{L^2(T)} + \frac{|F|}{|T|}\operatorname{diam}(T)\int_T |v|\,|\nabla v|\,dx.$$

We use the Cauchy-Schwarz inequality and the Young inequality $2ab \leq a^2 + b^2$ to estimate the second integral as follows

$$\begin{aligned}\operatorname{diam}(T)\frac{|F|}{|T|}\int_T |v|\,|\nabla v|\,dx &= \frac{|F|}{|T|}\int_T |v|\,\big(\operatorname{diam}(T)|\nabla v|\big)\,dx \\ &\leq \frac{|F|}{2|T|}(\|v\|^2_{L^2(T)} + \operatorname{diam}(T)^2\|\nabla v\|^2_{L^2(T)}).\end{aligned}$$

This implies the second assertion. □

Theorem 1.28. *Let $\Omega \subseteq \mathbb{R}^2$ be an open, bounded domain with polygonal Lipschitz boundary. Then, there exists a unique continuous and linear map $S : H^1(\Omega) \to L^2(\partial\Omega)$ with the property*

$$Sv = v|_{\partial\Omega} \quad \textit{for all } v \in H^1(\Omega) \cap C^0(\bar{\Omega}).$$

Remark 1.29. A linear map $T : H^1(\Omega) \to L^2(\partial\Omega)$ is said to be continuous if there exists a constant $C_T < \infty$ such that

$$\|Tv\|_{L^2(\partial\Omega)} \leq C_T\|v\|_{H^1(\Omega)} \quad \text{for all } v \in H^1(\Omega).$$

The theorem states the following. The operation of taking boundary values, which is well defined for functions from $C^0(\bar{\Omega})$, has a unique continuation to functions from $H^1(\Omega)$. Taking such generalized boundary values still leads to functions in $L^2(\partial\Omega)$, and we interpret these as boundary values of functions from $H^1(\Omega)$. This concept turns out important if we wish to pose the Dirichlet problem in Sobolev spaces. The operator S is called *trace operator*, and Sv is called the trace of v on $\partial\Omega$.

Proof of Theorem 1.28. We tesselate $\bar{\Omega}$ with a regular triangulation $\mathcal{T}$. Given any $v \in C^1(\bar{\Omega})$, we can apply Theorem 1.27 to any boundary edge $F \subseteq \partial\Omega$ of the triangulation and obtain

$$\|v\|^2_{L^2(F)} \leq \frac{3|F|}{2|T_F|}\|v\|^2_{L^2(T_F)} + \frac{|F|}{2|T_F|}\operatorname{diam}(T_F)^2\|\nabla v\|^2_{L^2(T_F)}.$$

Here, T_F is the uniquely defined triangle containing F as an edge. Since $\mathcal{T}$ has finitely many elements, the constant

$$C_{\mathcal{T}} := \max\left\{\max\left\{\frac{3|F|}{2|T_F|}, \frac{|F|}{2|T_F|}\operatorname{diam}(T_F)^2\right\} : F \subseteq \partial\Omega \text{ edge with triangle } T_F\right\}$$

is finite and we have the estimate

$$\|v\|^2_{L^2(F)} \leq C_{\mathcal{T}}(\|v\|^2_{L^2(T_F)} + \|\nabla v\|^2_{L^2(T_F)}) \quad \text{for all } v \in C^1(\bar{\Omega}).$$

For the whole boundary $\partial\Omega$ we then obtain

$$\|v\|^2_{L^2(\partial\Omega)} = \sum_{\substack{F \text{ boundary edge} \\ \text{of } \mathcal{T}}} \|v\|^2_{L^2(F)} \leq C_{\mathcal{T}} \sum_{\substack{F \text{ boundary edge} \\ \text{of } \mathcal{T}}} (\|v\|^2_{L^2(T_F)} + \|\nabla v\|^2_{L^2(T_F)}).$$

Obviously, every triangle can contain (at most) three boundary edges. Thus, any T_F occurs at most three times in the sum on the right hand side, and we can estimate

$$\|v\|^2_{L^2(\partial\Omega)} \leq 3C_{\mathcal{T}} \sum_{T\in\mathcal{T}} (\|v\|^2_{L^2(T)} + \|\nabla v\|^2_{L^2(T)}) = 3C_{\mathcal{T}} \|v\|^2_{H^1(\Omega)}.$$

Altogether, we have shown that there is a constant C such that

$$\|v\|_{L^2(\partial\Omega)} \leq C\|v\|_{H^1(\Omega)} \quad \text{for all } v \in C^1(\bar{\Omega}).$$

Thus, we have shown the desired estimate for the map $S : C^1(\bar{\Omega}) \to L^2(\Omega)$ assigning boundary values on the space $C^1(\bar{\Omega})$, which is dense in $H^1(\Omega)$ by Theorem 1.26. Thus, by an elementary result of linear functional analysis, there is a unique continuation of S to $H^1(\Omega)$. The continuity constant remains the same. □

As a consequence from the trace theorem, it makes sense to impose boundary values on functions from $H^1(\Omega)$. We will usually write $u|_{\partial\Omega}$ instead of Su etc., but we need to be aware that this function is only of class L^2 on $\partial\Omega$. For the Dirichlet problem, it is reasonable to consider the following subspace

$$H^1_0(\Omega) := \{v \in H^1(\Omega) : v|_{\partial\Omega} = 0\},$$

i.e., the space of Sobolev functions with zero boundary values. Sometimes, an alternative characterization turns out useful.

Theorem 1.30. *Let $\Omega \subseteq \mathbb{R}^2$ be an open and bounded Lipschitz domain. Then $H^1_0(\Omega)$ is the closure of $C^\infty_c(\Omega)$ with respect to the norm $\|\cdot\|_{H^1(\Omega)}$, i.e.,*

$$H^1_0(\Omega) := \overline{C^\infty_c(\Omega)}^{\|\cdot\|_{H^1(\Omega)}}.$$

Proof. We do not prove this technical result here. Its proof relies on similar techniques as Theorem 1.25 or Theorem 1.26 and can be found in the literature [Dob10, Eva10]. □

For functions from $H^1_0(\Omega)$, the L^2 norm can be controlled by the L^2 norm of the gradient. This result is called Friedrichs' inequality (sometimes Poincaré–Friedrichs inequality).

Theorem 1.31 (Friedrichs' inequality)**.** *Let Ω be an open, bounded, and connected Lipschitz domain. Then there exists a constant $C > 0$ such that*

$$\|v\|_{L^2(\Omega)} \leq C\|\nabla v\|_{L^2(\Omega)} \quad \textit{for all } v \in H^1_0(\Omega).$$

The constant is C proportional to the diameter of Ω.

Proof. The proof is left as an exercise. We sketch the basic idea. In view of Theorem 1.30, it is enough to consider $v \in C^\infty_c(\Omega)$ and then argue by density. We extend v by zero to some larger rectangular box containing Ω. After shifting coordinates, we may assume that

$\Omega \subseteq (0,L)^2$, $L > 0$. Then, v is of class $C_c^\infty((0,L)^2)$ with respect to this box. For any $x \in \Omega$, we can integrate

$$v(x) = v(x_1, x_2) = v(0, x_2) + \int_0^{x_1} \partial_1 v(t, x_2)\, dt.$$

We observe that the boundary term is zero. For the remaining term, we use the Cauchy-Schwarz/Hölder inequality and obtain

$$|v(x)|^2 \leq L \int_0^L |\partial_1 v(t, x_2)|^2\, dt.$$

We now intergrate with respect to x_1

$$\int_0^L |v(x)|^2 dx_1 \leq L^2 \int_0^L |\partial_1 v(t, x_2)|^2\, dt.$$

and thereafter integrate with respect to x_2

$$\int_0^L \int_0^L |v(x)|^2\, dx_1 dx_2 \leq L^2 \int_0^L \int_0^L |\partial_1 v(t, x_2)|^2\, dt dx_2.$$

Since the support of v lies inside Ω, this implies the asserted estimate for v. By a density argument, it is true for all functions from $H_0^1(\Omega)$. □

The most important implication of Friedrichs' inequality is that $\|\nabla \cdot\|_{L^2(\Omega)}$ defines a norm on $H_0^1(\Omega)$. (Convince yourself that this cannot be a norm on the larger space $H^1(\Omega)$ by considering constant functions.) Denoting the constant from Friedrichs' inequality by C_{F}, we indeed have the equivalence of norms

$$\|v\|_{H^1(\Omega)}^2 = \|v\|_{L^2(\Omega)}^2 + \|\nabla v\|_{L^2(\Omega)}^2 \leq (1 + C_{\mathrm{F}}^2)\|\nabla v\|_{L^2(\Omega)}^2 \leq (1 + C_{\mathrm{F}}^2)\|v\|_{H^1(\Omega)}^2. \tag{8}$$

We use the notation $|v|_1 = \|\nabla v\|_{L^2(\Omega)}$.
We are now in the position to formulate the Dirichlet problem in Sobolev spaces. It is based on the necessary condition from Theorem 1.10 (Dirichlet principle).

Definition 1.32. Let $\Omega \subseteq \mathbb{R}^2$ be an open and bounded Lipschitz domain. Given $f \in L^2(\Omega)$, the variational (or weak) formulation of the Dirichlet problem for Poisson's equation seeks $u \in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla u \cdot \nabla v\, dx = \int_\Omega f v\, dx \quad \text{for all } v \in H_0^1(\Omega).$$

This generalizes Poisson's equation in the sense that every classical solution will also be a solution to the variational formulation (see exercises).
We will now establish that the variational formulation possesses unique solutions. This will be an immediate consequence of the Riesz representation theorem, a fundamental result from the theory of Hilbert spaces. This topic is taught in any class on linear functional analysis, and we briefly recall basic details.

Let X be a (real) linear space. Given a symmetric and positive definite bilinear form $(\cdot,\cdot)_X$, we define $\|x\|_X = \sqrt{(x,x)_X}$ for any $x \in X$. It is elementary to establish the Cauchy–Schwarz inequality

$$(x,y)_X \leq \|x\|_X \, \|y\|_X \quad \text{for any } x, y \in X.$$

It can be shown that $\|\cdot\|_X$ defines a norm on X (thereby justifying the notation).

Definition 1.33 (Hilbert space)**.** A linear space X (over $\mathbb{R}$) equipped with a symmertic and positive definite bilinear form $(\cdot,\cdot)_X$ is called *Hilbert space* if it is complete with respect to the norm $\|\cdot\|_X := \sqrt{(\cdot,\cdot)_X}$.

Basically, Hilbert spaces are Banach spaces with an euclidean structure. We recall the dual space X^*, the space of continuous linear functionals over X. The Reisz representation theorem states that there exists an isometric isomorphism between X and X^*.

Theorem 1.34 (Riesz representation theorem)**.** *Let X be a Hilbert space with inner product $(\cdot,\cdot)_X$ and let $F \in X^*$ be a continuous linear functional. Then there exists a unique element $x \in X$ with the property*

$$(x,y)_X = F(y) \quad \textit{for all } y \in X.$$

The element x satisfies $\|x\|_X = \|F\|_{X^*}$.

Proof. The proof is taught in every course on linear functional analysis. □

We now use Hilbert space methods to show well-posedness of our variational formulation.

Lemma 1.35. *Let $\Omega \subseteq \mathbb{R}^2$ be an open and bounded Lipschitz domain. The space $H^1_0(\Omega)$ equipped with the bilinear form*

$$\int_\Omega \nabla v \cdot \nabla w \, dx$$

is a Hilbert space.

Proof. Friedrichs' inequality shows that the symmetric bilinear form is positive definite. The completeness with respect to $|\cdot|_1$ is a consequence of the equivalence of norms (8) and the fact that $H^1_0(\Omega)$ is a closed subspace of $H^1(\Omega)$. □

Theorem 1.36. *Let $\Omega \subseteq \mathbb{R}^2$ be an open and bounded Lipschitz domain and let $f \in L^2(\Omega)$. The variational formulation of the Dirichlet problem of Poisson's equation has a unique solution $u \in H^1_0(\Omega)$.*

Proof. We check that

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

is a continuous linear functional on the Hilbert space $H^1_0(\Omega)$. This follows from the Cauchy and the Friedrichs inequality

$$\int_\Omega f v \, dx \leq \|f\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)} C_{\mathrm{F}} |v|_1.$$

Hence, we are in the setting of the Riesz representation theorem, which states that there is a unique element $u \in H_0^1(\Omega)$ satisfying

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx \quad \text{for all } v \in H_0^1(\Omega).$$

□

By using elementary Hilbert space theory we could establish existence and uniqueness to the Dirichlet problem for any right-hand side $f \in L^2(\Omega)$. Note that this setting only needs the weak form of the Laplacian; furthermore even situations like Example 1.8 are covered by the theory. What is furthermore attractive about this approach is a direct characterization of the finite element error. We note that $S_0^1(\mathcal{T})$ is a finite-dimensional subspace of the Hilbert space $H_0^1(\Omega)$. It turns out that the finite element solution $u_h \in S_0^1(\mathcal{T})$ is the orthogonal projection of u to $S_0^1(\mathcal{T})$ and, thus, the best approximation in this space.

Theorem 1.37. *Let $\Omega \subset \mathbb{R}^2$ be an open, bounded, connected Lipschitz polygon with a triangulation $\mathcal{T}$. Given $f \in L^2(\Omega)$, the error between the solution $u \in H_0^1(\Omega)$ to the variational form of Poisson's equation and the finite element solution $u_h \in S_0^1(\mathcal{T})$ satisfies*

$$|u - u_h|_1 = \inf_{v_h \in S_0^1(\mathcal{T})} |u - v_h|_1.$$

Proof. We observe that, for any $v_h \in S_0^1(\mathcal{T}) \subseteq H_0^1(\Omega)$, we have

$$\int_\Omega \nabla(u - u_h) \cdot \nabla v_h \, dx = \int_\Omega \nabla u \cdot \nabla v_h \, dx - \int_\Omega \nabla u_h \cdot \nabla v_h \, dx = \int_\Omega f v_h \, dx - \int_\Omega f v_h \, dx = 0.$$

This property is also called *Galerkin orthogonality* because it describes that the error is orthogonal on discrete space. We compute

$$|u - u_h|_1^2 = \int_\Omega \nabla(u - u_h) \cdot \nabla u \, dx - \int_\Omega \nabla(u - u_h) \cdot \nabla u_h \, dx.$$

and see from the Galerkin orthogonality that the second term equals zero and remains zero if u_h on the right is replaced by any $v_h \in S_0^1(\mathcal{T})$. Thus,

$$|u - u_h|_1^2 = \int_\Omega \nabla(u - u_h) \cdot \nabla(u - v_h) \, dx \leq |u - u_h|_1 |u - v_h|_1$$

by Cauchy's inequality. The assertion then follows from dividing by $|u - u_h|_1$ and taking the infimum over v_h. □

We have seen that the finite element method is, in some sense, optimal. The result should illustrate the basic idea of the error analysis. In the next sections, we will generalize the theory to more general operators (not just the Laplacian) and see that the finite element method satisfies similar error bounds. It will turn out as a special case of *Galerkin approximations*.

Problem 32. Show that any function $u \in C^1(\bar{\Omega}) \cap C^2(\Omega)$, satisfying $-\Delta u = f$ for $f \in C^0(\bar{\Omega})$ and $u|_{\partial\Omega} = 0$, also satisfies the variational formulation.

Problem 33. Let $T \subseteq \mathbb{R}^2$ be a triangle and $v \in H^2(T) := \{w \in H^1(T) : \partial_j w \in H^1(T) \text{ for } j = 1, 2\}$ with norm

$$\|v\|_{H^2(T)} = \sqrt{\sum_{|\alpha|\leq 2} \|\partial^\alpha v\|^2_{L^2(T)}}.$$

(a) Consider a sub-triangle $t := \operatorname{conv}\{A, B, C\}$ with $E := \operatorname{conv}\{A, B\}$ and with tangent vector τ. Apply the trace inequality to $f|_E := \nabla v \cdot \tau$ and prove that

$$|v(B) - v(A)| \leq |E|^{1/2} \varrho^{-1/2} 2\big(1 + \operatorname{diam}(t)^2\big)^{1/2} \|v\|_{H^2(t)}$$

for $\varrho := 2|t|/|E|$.

(b) For any two points A and B in T there exists $C \in T$ such that (with $E := \operatorname{conv}\{A, B\}$ and $t := \operatorname{conv}\{A, B, C\}$), ϱ^{-1} is uniformly bounded by some constant $C(T)$ that depends only on T, but not on A, B, or t.

(c) Conclude that v is Hölder continuous with exponent $1/2$.

Remark: This shows the embedding $H^2(T) \hookrightarrow C^{0,1/2}(T)$ on a triangle.

Problem 34. (a) Prove that in one space dimension the approximation of the equation $u''(x) = 1$ (on the interval $(0, 1)$ with homogeneous Dirichlet boundary conditions) with the P_1 finite element method results in the nodal interpolation, that is $u_h = I_h u$.

(b) Convince yourself that this property cannot be valid in higher space dimensions (e.g., by a computational test case).

Problem 35. Let $f \in L^2(\Omega)$ and recall the energy functional

$$J(v) := \frac{1}{2}\|\nabla v\|^2_{L^2(\Omega)} - \int_\Omega f v \, dx \quad \text{for } v \in H^1_0(\Omega).$$

Prove that the error of the finite element method for the Poisson problem with right-hand side f satisfies

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

Problem 36. (a) Write the data structures for a triangulation of the L-shaped domain $\Omega := (-1, 1)^2 \setminus ([0, 1] \times [-1, 0])$ with Dirichlet boundary $\partial\Omega$.

(b) Plot the convergence history for $-\Delta u = 1$ on the L-shaped domain (cf. Problem 35; the exact solution satisfies $\|\nabla u\|^2 = 0.2140750232$). Compare the convergence rate with the results on the square domain.

§6 Finite element theory for linear coercive operators *(week 20)*

We can use Hilbert space methods to consider more complicated second-order operators than the Laplacian. In many applications, we encounter PDEs of the form

$$-\operatorname{div}(A\nabla u) + b \cdot \nabla u + cu = f$$

for a matrix field A, a vector field b, and a function c. These three terms are referred to as *diffusion*, *advection*, and *reaction*, respectively. As for the Laplacian, we can interpret the divergence operator weakly and derive the following variational formulation for $u \in H_0^1(\Omega)$:

$$\int_\Omega \Big((A\nabla u) \cdot \nabla v + (b \cdot \nabla u)v + c\,uv \Big)\,dx = \int_\Omega fv\,dx \quad \text{for all } v \in H_0^1(\Omega). \tag{9}$$

In this section we will study under which (sufficient) conditions this system has a unique solution. Note that the left-hand side need not be symmetric, and an immediate use of scalar products like in the case of Poisson's equation is not possible. Note furthermore that there need not be any related energy functional or Dirichlet principle.
The following important result extends, in some sense, the Riesz representation theorem to a class of nonsymmetric bilinear forms.

Theorem 1.38 (Lax–Milgram lemma)**.** *Let V be a real Hilbert space with inner product $(\cdot,\cdot)_V$ and let $a : V \times V \to \mathbb{R}$ be a bilinear form satisfying the following two properties*

- $\exists \beta > 0\, \forall (v,w) \in V^2 \quad |a(v,w)| \le \beta \|v\|_V \|w\|_V \qquad$ *(continuity)*
- $\exists \alpha > 0\, \forall v \in V \quad \alpha \|v\|_V^2 \le a(v,v) \qquad$ *(coercivity) .*

Then, there exists a unique map $T : V \to V$ with the property

$$a(w,v) = (Tw,v)_V \quad \text{for all } (v,w) \in V^2.$$

The map T is linear, continuous, and invertible with

$$\|T\|_{L(V,V)} \le \beta \quad \text{and} \quad \|T^{-1}\|_{L(V,V)} \le \frac{1}{\alpha}.$$

Proof. We will prove a more general result later in this class. It will imply the Lax–Milgram lemma. □

Corollary 1.39. *Let a be a continuous and coervice bilinear form on a Hilbert space V with inner product $(\cdot,\cdot)_V$. Given any $F \in V^*$, there is a unique $u \in V$ such that*

$$a(u,v) = F(v) \quad \text{for all } v \in V.$$

It satisfies $\|u\|_V \le \alpha^{-1} \|F\|_{L(V,V)}$.

Proof. Let $f \in V$ denote the den Riesz representative of F in V, and let T denote the mapping from the Lax–Milgram lemma. Then, $u := T^{-1}f$ satisfies

$$F(v) = (f, v)_V = (TT^{-1}f, v)_V = (Tu, v)_V = a(u, v)$$

for any $v \in V$. The norm bound for u folllows from the bound on T^{-1} from the Lax–Milgram lemma. □

Example 1.40 (general elliptic operator)**.** Let

$$A \in [L^\infty(\Omega)]^{2\times 2}, \quad b \in [L^\infty(\Omega)]^2, \quad c \in L^\infty(\Omega)$$

be the coefficients of the above PDE with $f \in L^2(\Omega)$ und homogeneous Dirichlet boundary condition. After multipying with test functions an integrating (by parts) we obtain the following weak formulation: Seek $u \in H^1_0(\Omega)$ such that

$$a(u, v) = F(v) \quad \text{for all } v \in H^1_0(\Omega),$$

where

$$a(u, v) := \int_\Omega \Big((A\nabla u) \cdot \nabla v + (b \cdot \nabla u)v + c\,uv \Big)\,dx \quad \text{and} \quad F(v) := \int_\Omega fv\,dx.$$

We now apply, under further structural assumptions, the Lax–Milgram lemma to the above setting.

Theorem 1.41. *Let $\Omega \subseteq \mathbb{R}^2$ be an open, bounded, connected Lipschitz polygon. Let the coefficients A, b, c from Example 1.40 satisfy the following assumptions.*

- *The field A is pointwise symmetric and there exist real numbers $0 < a_0, a_1$ such that*

  $$a_0|\xi|^2 \leq (A(x)\xi) \cdot \xi \leq a_1|\xi|^2 \qquad \text{a.e. in } \Omega \text{ for all } \xi \in \mathbb{R}^2$$

  i.e., A is uniformly positive definite.

- *The vector field b is divergence-free,* $\operatorname{div} b = 0$ *(in the sense of the weak divergence, see Problem 25).*

- *The function $c \geq 0$ is nonnegative.*

Then there exists a unique solution $u \in H^1_0(\Omega)$ to the weak form from Example 1.40. It satisfies the bound

$$|u|_1 \leq C\|f\|_{L^2(\Omega)}$$

for some constant $C > 0$ that is independent of f and u.

Proof. The proof is left as an exercise. It is enough to verify that a satisfies the assumptions from the Lax–Milgram lemma. □

A finite element discretization of this problem is straight-forward: We restrict the bilinear form a and the right-hand side to finite element functions. The form a, however, need not be a scalar product, and the best-approximation property (as in the Laplacian case) is not valid in its original form. We will now study approximations in a more general setting.

Definition 1.42. Let a be a coercive and continuous bilinear form on a Hilbert space V, and let $V_h \subseteq V$ be a closed subspace. Given $F \in V^*$, let $u \in V$ solve

$$a(u, v) = F(v) \quad \text{for all } v \in V.$$

The unique solution $u_h \in V_h$ to

$$a(u_h, v_h) = F(v_h) \quad \text{for all } v_h \in V_h$$

is called the *Galerkin approximation* to u.

Remark 1.43. We remark that, in the foregoing definition, the Galerkin approximation indeed exists and is unique. This follows from the fact that closed subspaces of Hilbert spaces are again Hilbert spaces. It is immediate to see that the Lax–Milgram lemma applies on such subspaces as well.

Example 1.44. The finite element approximation to Poisson's equation is a Galerkin method based on the finite-dimensional subspace $V_h := S_0^1(\mathcal{T})$ of $V := H_0^1(\Omega)$. The finite element approximation to the operator from Example 1.40 is a Galerkin method as well.

We now formulate the basic error estimate for Galerkin approximations.

Theorem 1.45 (Céa's lemma)**.** *Let V be a Hilbert space and let $V_h \subseteq V$ be a closed subspace. Let $a : V \times V \to \mathbb{R}$ be a continuous and coervice bilinear form (with α, β as in the Lax–Milgram lemma) and let $F \in V^*$. Let $u \in V$ solve*

$$a(u, v) = F(v) \quad \textit{for all } v \in V.$$

The Galerkin approximation $u_h \in V_h$ solving

$$a(u_h, v_h) = F(v_h) \quad \textit{for all } v_h \in V_h$$

satisfies the following error bound

$$\|u - u_h\|_V \leq \frac{\beta}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|_V.$$

Proof. The coercivity reveals the following relation of the norm and the form a,

$$\alpha \|u - u_h\|_V^2 = a(u - u_h, u - u_h). \tag{10}$$

We observe that, due to the property $V_h \subseteq V$, the variational problems in V and V_h satisfy

$$a(u - u_h, v_h) = a(u, v_h) - a(u_h, v_h) = F(v_h) - F(v_h) = 0. \tag{11}$$

As a consequence, we can take an arbitrary $w_h \in V_h$ and compute

$$\begin{aligned} a(u-u_h, u-u_h) &= a(u-u_h, u) - \underbrace{a(u-u_h, u_h)}_{=0} = a(u-u_h, u) \\ &= a(u-u_h, u) - \underbrace{a(u-u_h, w_h)}_{=0} = a(u-u_h, u-w_h). \end{aligned}$$

We use this relation in the above formula (10) and deduce from the continuity that

$$\alpha \|u-u_h\|_V^2 = a(u-u_h, u-w_h) \leq \beta \|u-u_h\|_V \|u-w_h\|_V.$$

In case that $u - u_h = 0$, the assertion of the theorem is trivially satisfied. Otherwise, we can divide by the norm of $u - u_h$ and take the infimum over w_h. □

Due to the factor β/α in the error estimate, the Galerkin method is said to be quasi-optimal. We can apply the abstract setting to the finite element method for the second-order system from Example 1.40 and Theorem 1.41 and obtain the quasi-optimal bound

$$|u - u_h|_1 \leq C \inf_{v_h \in S_0^1(\mathcal{T})} |u - v_h|_1.$$

In this sense, the finite element method computes a near-best approximation in the space $S_0^1(\mathcal{T})$. We will quantify this approximation in the forthcoming sections.
Let us now discuss in which regards the theory and methods to more general situations.

Inhomogeneous Dirichlet values

It is often required to prescribe nonzero boundary values to the Dirichlet problem. For some given function $u_D : \partial\Omega \to \mathbb{R}$ one is interested in finding a function u satisfying

$$-\Delta u = f \text{ in } \Omega \quad \text{and} \quad u = u_D \text{ on } \partial\Omega.$$

In the variational form one seeks $u \in H^1(\Omega)$ with

$$a(u, v) = \int_\Omega f v \, dx \text{ for all } v \in H_0^1(\Omega) \quad \text{and} \quad u = u_D \text{ on } \partial\Omega \text{ in the sense of traces.}$$

Here, a is the Laplacian inner product, but the generalization to other operators is immediate. Of course, for this formulation to make sense, the data u_D must belong to the range of the trace operator, i.e., it must possess a continuation to a function from $H^1(\Omega)$. We denote the range of the trace operator by $H^{1/2}(\partial\Omega)$, without further characterizing it here. Let $\hat{u}_D \in H^1(\Omega)$ denote an extension of u_D to the domain Ω. The idea is to shift the solution by u_D and to seek for $w = u - u_D$, which then has zero boundary data. One solves for $w \in H_0^1(\Omega)$ such that

$$a(w, v) = \int_\Omega f v \, dx - a(u_D, v) \quad \text{for all } v \in H_0^1(\Omega).$$

It follows from the continuity of a that the right-hand side defines a continuous linear functional on $H_0^1(\Omega)$ and so there is a unique solution w. Then, $u := w + u_D$ solves the inhomogeneous boundary value problem. In a practical finite element implementation, we proceed analogously. It might be required to interpolate u_D with piecewise affine functions along the boundary so that it is the trace of a finite element function. We can extend this finite element function to the domain by setting it to zero at all interior vertices and denote the coefficient vector by x_D. The modified right-hand side then reads $\tilde{b} := b - A^\top x_D$, where b is the load vector related to f and A is the stiffness matrix. We then solve for x_0 (which is zero at the boundary vertices) by restricting the system $A^\top x_0 = \tilde{b}$ to the interior vertices as the degrees of freedom. Then, $x := x_0 + x_D$ is the coefficient vector of the finite element solution.

Neumann boundary values

In many applications, for example when u describes the heat distribution in some domain Ω, one wants to prescribe $(A\nabla u) \cdot \nu$ on the boundary rather than actual values for u. In the context of a heat distribution, this corresponds to the heat flux. The boundary is then subdivided in two disjoint parts

$$\partial\Omega = \Gamma_D \cup \Gamma_N$$

where Γ_D is relatively closed. The part Γ_D is called the Dirichlet boundary and Γ_N is called the Neumann boundary. Either of the parts is allowed to be empty. The boundary value problem in its strong form then reads

$$-\operatorname{div} A\nabla u + b \cdot \nabla u + cu = f \text{ in } \Omega, \quad u = u_D \text{ on } \Gamma_D, \quad (A\nabla u) \cdot \nu = g \text{ on } \Gamma_N$$

where u_D is the prescribed Dirichlet data and $g \in L^2(\Gamma_N)$ is a given function, the so-called Neumann data. Assume for simplicity that $u_D = 0$. As we have no homogeneous boundary condition on the whole boundary, we need to work with the space

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

Be aware that in our integration-by-parts arguments the boundary terms do not vanish any more because test functions may be nonzero along Γ_N. Indeed, we find

$$\int_\Omega (-\operatorname{div} A\nabla u) v \, dx = \int_\Omega (A\nabla u) \cdot \nabla v \, dx - \int_{\Gamma_N} (A\nabla u) \cdot \nu \, v \, ds$$

for any test function $v \in H_D^1(\Omega)$. The term $(A\nabla u) \cdot \nu$ is prescribed by the Neumann data g. The weak formulation with Neumann data thus reads: Find $u \in H_D^1(\Omega)$ such that

$$a(u, v) = \int_\Omega f v \, dx + \int_{\Gamma_N} g v \, ds.$$

In our Python code, we can prescribe the Neumann boundary by the structure `neumann`, which so far had been left empty. In order to show well-posedness with the Lax–Milgram lemma, we need coercivity of a on $H^1_D(\Omega)$. This means that we need a generalization of Friedrichs' inequality to the case where our functions only vanish on some part (of positive surface measure) of the boundary. In the pure Neumann case $\Gamma_N = \partial\Omega$ and $\Gamma_D = \emptyset$, it is easy to see that there will be no unique solution because solutions may be shifted by arbitrary constants. In this case we therefore need to normalize the solution

$$\int_\Omega u\,dx = 0 \quad \text{in case of pure Neumann bounary conditions.}$$

This condition guarantees coercivity, as will be shown in the next theorem. We use the notation

$$H^1(\Omega)/\mathbb{R} = \{v \in H^1(\Omega) : \int_\Omega v\,dx = 0\}.$$

Theorem 1.46 (generalized Poincaré and Friedrichs inequalities)**.** *Let $\Omega \subseteq \mathbb{R}^2$ be an open, bounded, connected Lipschitz domain. There is a constant $C_P > 0$ such that*

$$\|v\|_{L^2(\Omega)} \leq C_P \|\nabla v\|_{L^2(\Omega)} \quad \textit{for all } v \in H^1(\Omega)/\mathbb{R} \quad \textit{(Poincaré inequality).}$$

Let $\Gamma_D \subseteq \partial\Omega$ have positive surface measure. Then there is a constant $C_F > 0$ such that

$$\|v\|_{L^2(\Omega)} \leq C_F \|\nabla v\|_{L^2(\Omega)} \quad \textit{for all } v \in H^1_D(\Omega) \quad \textit{(Friedrichs inequality).}$$

The constants C_F and C_P are proportional to the diameter of the domain Ω.

Proof. The proof is based on a so-called compactness argument that will be presented in the next section. We postpone the proof to the exercises of that section. □

Problem 37. Prove the assertions of Theorem 1.41 and track the dependence of the constant C on the spectral bounds a_0, a_1 and the L^∞ norms of the coefficients.

Problem 38. Prove existence and uniqueness of solutions to the second-order elliptic problem in case of nontrivial Neumann boundary $\Gamma_N \neq 0$.

Problem 39. *(convection-diffusion equation)*

(a) Implement the finite element method for the convection-diffusion equation $-\varepsilon\Delta u + \beta \cdot \nabla u = f$.

(b) Consider the unit square $\Omega = (0,1)^2$ with homogeneous Dirichlet boundary conditions and the right-hand side f according to the exact solution

$$u(x) = \left(\frac{e^{r_1(x_1-1)} - e^{r_2(x_1-1)}}{e^{-r_1} - e^{-r_2}} + x_1 - 1 \right) \sin(\pi x_2)$$

with

$$r_1 = \frac{-1 + \sqrt{1 + 4\varepsilon^2\pi^2}}{-2\varepsilon} \quad \text{and} \quad r_2 = \frac{-1 - \sqrt{1 + 4\varepsilon^2\pi^2}}{-2\varepsilon}.$$

Run numerical computations for the following parameters

(i) $\varepsilon = 0.1$ and $\beta = (1, 0)^T$.

(ii) $\varepsilon = 0.001$ and $\beta = (1, 0)^T$.

Problem 40. Extend your finite element code to the case of inhomogeneous Dirichlet and zero Neumann boundary data. Use the follwing test case to validate your code (via comparison of the solution graphs and convergence tests): The domain $\Omega = (0, 1)^2$ is the unit square. The Neumann boundary is the line $\{(1, t) : 0 < t < 1\}$, and $\Gamma_D = \partial\Omega \setminus \Gamma_N$. The exact solution is given by

$$u(x, y) = 5 + \sin(\frac{\pi}{2}x)\sin(\pi y)$$

with $u_D = 5$ and $g = 0$. The right-hand side reads $f = \frac{5}{4}\pi^2 \sin(\frac{\pi}{2}x)\sin(\pi y)$.

Problem 41. *(nodal interpolation not L^2 or H^1 stable)* For a triangle $T \subseteq \mathbb{R}^2$, prove that there is no constant C such that the nodal P_1 interpolation I satisfies

$$\begin{aligned} &\|Iu\|_{L^2(T)} \le C\|u\|_{L^2(T)} \text{ for all } u \in C^\infty(T) \\ \text{or} \quad &\|\nabla Iu\|_{L^2(T)} \le C\|\nabla u\|_{L^2(T)} \text{ for all } u \in C^\infty(T). \end{aligned}$$

§7 Finite element error estimates *(week 21)*

We would like to quantify the right-hand side of Céa's lemma in terms of the mesh-size (maximum diameter of the triangles in $\mathcal{T}$). The idea is to plug in a suitable approximation in the infimum for which we then derive quantified bounds. To achieve this, we will use the finite element interpolation. It is, however, not a well defined on $H^1(\Omega)$ because it takes point evaluations, which need not exist without further assumptions (see Problem 8). This means that the interpolation operator, denoted by I_h, assigning the finite element interpolation $I_h v$ to any suitable (say continuous) function v, is not well defined on $H^1(\Omega)$, see Problem 41. We have seen in Problem 33 that point evaluations are well-defined in the space

$$H^2(\Omega) = \{v \in L^2(\Omega) : \text{all weak derivatives of } v \text{ up to order 2 exist as functions of } L^2(\Omega)\}$$

with norm

$$\|v\|_{H^2(\Omega)} = \sqrt{\sum_{|\alpha|\leq 2} \|\partial^\alpha v\|^2_{L^2(\Omega)}}.$$

The proof, which was shown for triangles, can be extended to polygonal domains.

Theorem 1.47. *Let $\Omega \subseteq \mathbb{R}^2$ be an open and bounded Lipschitz polygon. Then, we have the continuous embedding $H^2(\Omega) \hookrightarrow C(\bar{\Omega})$ and there exists a constant $C > 0$ such that*

$$\|v\|_{L^\infty(\Omega)} \leq C\|v\|_{H^2(\Omega)} \quad \textit{for any } v \in H^2(\Omega).$$

Proof. The proof is left as an exercise. The idea is to triangulate the domain and to use the bound that was proven for triangles. □

We have seen that we can apply the finite element interpolation I_h under the assumption that our solution u satisfies the stronger property $u \in H^1_0(\Omega) \cap H^2(\Omega)$. For a derivation of a quantitative bound on the interpolation under this assumption, we need an additional property from the theory of Sobolev spaces.

Theorem 1.48. *Let $\Omega \subset \mathbb{R}^2$ be an open and bounded Lipschitz domain. Then, the embedding $H^1(\Omega) \hookrightarrow L^2(\Omega)$ is compact. That is, any weakly convergent sequence $v_n \rightharpoonup v$ $(n \to \infty)$ in $H^1(\Omega)$ converges strongly in $L^2(\Omega)$.*

Proof. The proof is shown in advanced courses on linear functional analysis and is beyond the scope of this lecture. □

Remark 1.49. An iterative application of Theorem 1.48 shows that $H^2(\Omega)$ is compactly embedded in $H^1(\Omega)$.

With these tools we can now prove an interpolation error estimate on triangles. The constant appearing in the estimate will depend on the aspect ratio of the triangles.

Definition 1.50. Let $T \subseteq \mathbb{R}^2$ be a triangle. Let h_T denote its diameter and let ρ_T denote the diameter of the largest ball inscribed to T. The quantity h_T/ρ_T is called the *aspect ratio* of T.

The quantities ρ_T, h_T arise from the transformation of the domain. Observe that any pair of triangles T, $\hat{T}$ allows for some affine diffeomorphism $\Phi : \hat{T} \to T$, which can be written $\Phi(\hat{x}) = B\hat{x} + c$ with a 2×2 matrix $B = D\Phi$ and a vector c.

Lemma 1.51. *Let $\Phi(\hat{x}) = B\hat{x} + c$ denote the affine map from a triangle $\hat{T}$ to the triangle T. Then, the spectral norm $\|\cdot\|$ of B and B^{-1} satisfies*

$$\|B\| \leq \frac{h_T}{\rho_{\hat{T}}} \quad \textit{and} \quad \|B^{-1}\| \leq \frac{h_{\hat{T}}}{\rho_T}.$$

Proof. Given any vector $\xi \in \mathbb{R}^2$ of length $|\xi| = \rho_{\hat{T}}$, there exists pair of points $\hat{x}$, $\hat{y}$ inside $\hat{T}$ with $\hat{x} - \hat{y} = \xi$ because the full ball of diameter $\rho_{\hat{T}}$ is contained in $\hat{T}$. Since $\Phi(\hat{x})$ and $\Phi(\hat{y})$ belong to T, the image under B satisfies $B\xi = B(\hat{x} - \hat{y}) = \Phi(\hat{x}) - \Phi(\hat{y})$ and its length is bounded by the diameter h_T. We thus compute

$$\|B\| = \sup_{\xi \in \mathbb{R}^2, |\xi|=1} |B\xi| = \sup_{\xi \in \mathbb{R}^2, |\xi|=\rho_{\hat{T}}} \frac{1}{\rho_{\hat{T}}} |B\xi| \leq \frac{h_T}{\rho_{\hat{T}}}.$$

The second asserted estimate follows from interchanging the roles of T and $\hat{T}$. □

We now prove the interpolation error estimate.

Theorem 1.52 (interpolation error estimate)**.** *There exists a constant $C > 0$ such that for any triangle $T \subseteq \mathbb{R}^2$ the interpolation error satisfies*

$$\begin{aligned} \|\nabla(v - I_h v)\|_{L^2(T)} &\leq C \frac{h_T}{\rho_T} h_T \|D^2 v\|_{L^2(T)} \\ \textit{and} \qquad \|v - I_h v\|_{L^2(T)} &\leq C h_T^2 \|D^2 v\|_{L^2(T)} \end{aligned}$$

for any $v \in H^2(T)$. Here, $\|D^2 v\|_{L^2(T)} = \sqrt{\int_T \sum_{j,k=1}^2 |\partial_{jk} v|^2 \, dx}$.

Proof. We first prove an auxiliary estimate on some fixed reference triangle $\hat{T}$. We claim that there is a constant $\hat{C}$ such that any $w \in H^2(\hat{T})$ satisfies

$$\|w\|_{H^1(\hat{T})} \leq \hat{C}(\|D^2 w\|_{L^2(\hat{T})} + \sum_{z \in \mathcal{N}(\hat{T})} |w(z)|).$$

Assume for contradiction that the statement is wrong. Then there is a sequence $w_n \in H^2(\hat{T})$ with

$$\|w_n\|_{H^1(\hat{T})} \geq n(\|D^2 w_n\|_{L^2(\hat{T})} + \sum_{z \in \mathcal{N}(\hat{T})} |w_n(z)|) \quad \text{for all } n \in \mathbb{N}.$$

After normalizing the sequence to $\|w_n\|_{H^1(\hat{T})} = 1$ we obtain

$$\|D^2 w_n\|_{L^2(\hat{T})} + \sum_{z \in \mathcal{N}(\hat{T})} |w_n(z)| \leq 1/n \quad \text{for all } n \in \mathbb{N}.$$

The space $H^2(T)$ is reflexive, whence there exists a weakly convergent subsequence of this bounded sequence with some weak limit $w \in H^2(\hat{T})$. We do not relabel the subsequence and still denote it by w_n. The compact embedding of Theorem 1.48 shows that we have $w_n \to w$ in $H^1(\hat{T})$. It is even a Cauchy sequence in $H^2(\hat{T})$ because

$$\begin{aligned} \|w_j - w_k\|^2_{H^2(\hat{T})} &= \|w_j - w_k\|^2_{H^1(\hat{T})} + \|D^2(w_j - w_k)\|^2_{L^2(\hat{T})} \\ &\le \|w_j - w_k\|^2_{H^1(\hat{T})} + \|D^2 w_j\|^2_{L^2(\hat{T})} + \|D^2 w_k\|^2_{L^2(\hat{T})} \end{aligned}$$

and the norms of the Hessian converge to 0. Therefore we have strong convergence $w_n \to w$ in $H^2(\hat{T})$, and $D^2 w = 0$. Thus, w is an affine function. By continuity, we furthermore see that $w(z) = 0$ at the vertices of $\hat{T}$. Thus, w is the zero function. But this contradicts $\|w_n\|_{H^1(\hat{T})} = 1$. This proves the claimed auxiliary estimate.

Now, let T be an arbitrary triangle. Then, there is an affine transformation

$$\Phi : \hat{T} \to T$$

from the reference triangle to T. We denote by $e := v - I_h v$ the interpolation error and observe from the change-of-variables formula that

$$\|\nabla e\|^2_{L^2(T)} = \int_T |\nabla e|^2 \, dx = \int_{\hat{T}} |(\nabla e) \cdot \Phi|^2 |\det D\Phi| \, dx$$

We use notation $\hat{e} := e \circ \Phi$. The chain rule reveals for any $\hat{x} \in \hat{T}$ that

$$\nabla \hat{e}(\hat{x}) = D\Phi(\hat{x})^\top \nabla e|_{\Phi(\hat{x})}.$$

Multiplying with the inverse of $D\Phi^\top$ and taking squares thus leads to

$$|(\nabla e) \circ \Phi|^2 = |D\Phi^{-\top} \nabla \hat{e}|^2 \le \|D\Phi^{-1}\|^2 |\nabla \hat{e}|^2$$

where $\|\cdot\|$ denotes the (pointwise) spectral matrix norm.

We observe that $D\Phi$ is constant on $\hat{T}$ (because Φ is affine). We thus obtain

$$\|\nabla e\|^2_{L^2(T)} \le \|D\Phi^{-1}\|^2 |\det D\Phi| \|\nabla \hat{e}\|^2_{L^2(\hat{T})}.$$

By the auxiliary result, there exists a constant $\hat{C}$, depending on $\hat{T}$, such that

$$\|\nabla \hat{e}\|^2_{L^2(\hat{T})} \le \hat{C}^2 \|D^2 \hat{e}\|^2_{L^2(\hat{T})}.$$

Here, we have used that e, the interpolation error vanishes at the vertices of T, and so does the transformed function on the vertices of $\hat{T}$. So far we have shown

$$\|\nabla e\|^2_{L^2(T)} \le \hat{C}^2 \|D\Phi^{-1}\|^2 \int_{\hat{T}} |D^2 \hat{v}|^2 |\det D\Phi| \, dx.$$

The chain rule shows

$$D^2 \hat{v}(\hat{x}) = D\Phi(\hat{x})^\top D^2 v|_{\Phi(\hat{x})} D\Phi(\hat{x}).$$

We thus find

$$|D^2\hat{v}|^2 \leq \|D\Phi(\hat{x})\|^4 |(D^2 v)\circ\Phi|^2.$$

After transforming back to T we thus obtain

$$\|\nabla e\|^2_{L^2(T)} \leq \hat{C}^2 \|D\Phi^{-1}\|^2 \|D\Phi\|^4 \|D^2\hat{v}\|^2_{L^2(T)}.$$

The norms of $D\Phi$ and its inverse can be estimated with Lemma 1.51 as follows

$$\|D\Phi^{-1}\|^2\|D\Phi\|^4 \leq \frac{h^2_{\hat{T}}}{\rho^2_T}\frac{h^4_T}{\rho^4_{\hat{T}}} = \frac{h^2_{\hat{T}}}{\rho^4_{\hat{T}}}\frac{h^4_T}{\rho^2_T}.$$

The terms related to $\hat{T}$ are independent of T and can be estimated by some universal constant. We thus obtain (after taking squareroots) the asserted bound on the norm of the gradient. The bound on the L^2 norm is left as an exercise. □

We see from the interpolation error estimate of Theorem 1.52 that the interpolation error is proportional to h_T provided the aspect ratio of the triangle is bounded. If we take, for instance, any fixed triangle and refine it uniformly with the red refinement rule, the aspect ratio is bounded by a universal constant. We say that a family of triangulations with bounded aspect ratio is *shape-regular.* The approximation of an H^2 function is then determined by the mesh-size h_T and thus improved under mesh-refinement.

Corollary 1.53 (global interpolation error estimate)**.** . *Let $\Omega \subseteq \mathbb{R}^2$ be an open and bounded polygonal Lipschitz domain. Let $\{\mathcal{T}_h\}_h$ be a shape-regular family of triangulations. Then, there is a constant $C > 0$ such that for any $v \in H^2(\Omega)$ the finite element interpolation I_h with respect to a mesh $\mathcal{T}_h$ satisfies*

$$\|\nabla(v - I_h v)\|_{L^2(\Omega)} \leq Ch\|D^2 v\|_{L^2(\Omega)} \quad \text{and} \quad \|v - I_h v\|_{L^2(\Omega)} \leq Ch^2\|D^2 v\|_{L^2(\Omega)}$$

for the maximal mesh-size $h = \max T \in \mathcal{T}_h h_T$.

Proof. This follows from writing the L^2 norm as

$$\|\cdot\|_{L^2(\Omega)} = \sqrt{\sum_{T\in\mathcal{T}_h} \|\cdot\|^2_{L^2(T)}}$$

and using the local bounds of Theorem 1.52. □

We have seen that any $v \in H^2(\Omega)$ is approximated with order h by the finite element interpolation the H^1 norm and with order h^2 in the L^2 norm.
The combination with Céa's lemma now yields a quantified bound for the finite element approximation on our PDE.

Corollary 1.54. *Let $\Omega \subseteq \mathbb{R}^2$ be an open, bounded, connected Lipschitz polygon. Let the coefficients A, b, c satisfy the assumptions from Theorem 1.41. Assume that the solution $u \in H_0^1(\Omega)$ to the weak form from Example 1.40 additionally satisfies*

$$u \in H_0^1(\Omega) \cap H^2(\Omega).$$

Then, the error between u and the finite element approximation u_h with respect to a triangulation $\mathcal{T}_h$ from a shape-regular family satisfies

$$|u - u_h|_1 \leq Ch\|D^2u\|_{L^2(\Omega)}.$$

Proof. Céa's lemma (Theorem 1.45) yields

$$|u - u_h|_1 \leq \frac{\beta}{\alpha} \inf_{v_h \in \mathcal{S}_0^1(\mathcal{T}_h)} |u - v_h|_1$$

where α, β denote the coercivity and continuity constant, respectively. We now plug the choice $v_h := I_h u$ in the infimum. Note that the interpolation exists because $u \in H^2(\Omega)$ was assumed. The assertion then follows from the interpolation error estimate of Corollary 1.53. □

As an immediate question we ask under what condition the assumption $u \in H^2(\Omega)$ is satisfied. This is the topic of *regularity theory* and is beyond our discussion within this lecture. We only mention one important result here for the case of the Laplacian.

Theorem 1.55 (regularity on convex domains)**.** *Let $\Omega \subset \mathbb{R}^2$ be an open convex domain. Given any $f \in L^2(\Omega)$, the solution to the Dirichlet problem of the Laplacian (Poisson's equation) satisfies $u \in H_0^1(\Omega) \cap H^2(\Omega)$ with the bound*

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

Proof. See the literature, e.g., [Dob10, Eva10]. □

When the domain is nonconvex, the solution may fail to belong to $H^2(\Omega)$. This is for instance the case in Example 1.8.

We can prove an improved bound for the error in the L^2 norm.

Theorem 1.56 (L^2 error bound)**.** *Let $\Omega \subset \mathbb{R}^2$ be an open convex domain. Given any $f \in L^2(\Omega)$, the solution to the Dirichlet problem of the Laplacian (Poisson's equation) and its finite element approximation satisfy*

$$\|u - u_h\|_{L^2(\Omega)} \leq Ch^2\|D^2u\|_{L^2(\Omega)} \leq Ch^2\|f\|_{L^2(\Omega)}$$

Proof. The technique employed in the proof is known as the *Aubin-Nitsche duality trick.* The idea is to solve for a solution $z \in H_0^1(\Omega)$ an auxiliary problem whose right-hand side is given by the error $e := u - u_h$. Let z solve

$$\int_\Omega \nabla z \cdot \nabla v \, dx = \int_\Omega ev \, dx \quad \text{for all } v \in H_0^1(\Omega).$$

We test the equation with $v := e$ and obtain

$$\|e\|_{L^2(\Omega)}^2 = \int_\Omega e\,e\,dx \quad \text{for all } v \in H_0^1(\Omega) = \int_\Omega \nabla e \cdot \nabla z\,dx.$$

We now use the Galerkin orthogonality and plug in the finite element approximation z_h to z,

$$\int_\Omega \nabla e \cdot \nabla z\,dx = \int_\Omega \nabla(u-u_h) \cdot \nabla z\,dx = \int_\Omega \nabla(u-u_h) \cdot \nabla(z-z_h)\,dx.$$

Corollary 1.54 implies for the finite element errors that

$$\begin{aligned} &\|\nabla(u-u_h)\|_{L^2(\Omega)} \leq C\|D^2 u_h\|_{L^2(\Omega)} \\ \text{and} \qquad &\|\nabla(z-z_h)\|_{L^2(\Omega)} \leq C\|D^2 z_h\|_{L^2(\Omega)} \leq C\|e\|_{L^2(\Omega)}. \end{aligned}$$

We now combine the above formulas and divide by the norm of e to arrive at the first asserted estimate. The second one follows from Theorem 1.55. □

For simplicity, we have considered right-hand sides $f \in L^2(\Omega)$. The argument in Theorem 1.41 however even applies to any right-hand side F in the dual space of $H_0^1(\Omega)$. This dual space is denoted by

$$H^{-1}(\Omega) := [H_0^1(\Omega)]^*$$

where we use the norm $|\cdot|_1$. Accordingly, the norm in $H^{-1}(\Omega)$ reads

$$\|F\|_{H^{-1}(\Omega)} = \sup_{v \in H_0^1(\Omega)\setminus\{0\}} \frac{\langle F, v\rangle}{\|\nabla v\|_{L^2(\Omega)}}.$$

The space $L^2(\Omega)$ is embedded in $H^{-1}(\Omega)$ by the identification of $f \in L^2(\Omega)$ with the functional $T_f \in H^{-1}(\Omega)$ defined by

$$\langle T_f, v\rangle = \int_\Omega f v\,dx \quad \text{for all } v \in H_0^1(\Omega).$$

The map $f \mapsto T_f$ is injective and continuous, thus an embedding. The injectivity follows from the fact that

$$\int_\Omega g v\,dx \quad \text{for all } v \in H_0^1(\Omega)$$

implies $g = 0$ (by density of $H_0^1(\Omega)$ in $L^2(\Omega)$) whence T_g is zero in $H^{-1}(\Omega)$. Continuity follows from the Friedrichs inequality as follows:

$$\|T_f\|_{H^{-1}(\Omega)} = \sup_{v \in H_0^1(\Omega)\setminus\{0\}} \frac{\int_\Omega f v\,dx}{\|\nabla v\|_{L^2(\Omega)}} \leq \sup_{v \in H_0^1(\Omega)\setminus\{0\}} \frac{\|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)}}{\|\nabla v\|_{L^2(\Omega)}} \leq C_F\|f\|_{L^2(\Omega)}.$$

This map thus leads to

$$H_0^1(\Omega) \subseteq L^2(\Omega) \hookrightarrow H^{-1}(\Omega)$$

and the embedding is often interpreted as the inclusion $L^2(\Omega) \subseteq H^{-1}(\Omega)$.

Problem 42. Prove Theorem 1.47.

Problem 43. Let $\mathcal{T}$ be a triangulation. Prove that the aspect ratio of the triangles stays bounded under iterative red refinement.

Problem 44. Prove that there exists a constant C only dependent on the shape regularity such that any finite element function $v_h \in S^1(\mathcal{T})$ satisfies

$$\|\nabla v_h\|_{L^2(T)} \leq C h_T^{-1} \|v_h\|_{L^2(T)} \quad \text{for all } T \in \mathcal{T}.$$

This estimate is called *inverse inequality. (Hint: Use transformation to a reference element $\hat{T}$. Use equivalence-of-norms argument in the finite dimensional space $P_1(\hat{T})$ with a constant $C(\hat{T})$ only depending on $\hat{T}$. Afterwards, transform back.)*

Problem 45. Prove for the solution u from Example 1.8 that $u \notin H^2(\Omega)$.

Problem 46. A family of triangulations satisfies the *minimal angle condition* if there is a lower bound $0 < \alpha_0$ to all interior angles of the triangles from that family. Prove that the *minimal angle condition* implies shape regularity.

Topic 2: Saddle-point problems

§8 Abstract saddle-point problems and the Stokes equation *(week 22)*

We now draw our attention to problems under linear constraints. For example, the velocity field $u : \Omega \to \mathbb{R}^2$ of a very viscous fluid under some volume force $f : \Omega \to \mathbb{R}^2$ is modelled by the following constrained minimization problem:

$$J(v) := \frac{1}{2}\int_\Omega |Dv|^2\,dx - \int_\Omega f\cdot v\,dx \to \min \quad \text{subject to } \operatorname{div} u = 0 \quad \text{and } u|_{\partial\Omega} = 0.$$

The energy is measured with the Dirichlet functional for vector-valued variables. The constraint $\operatorname{div} u = 0$ models that the fluid under consideration is incompressible. In order to —at least formally— compute a corresponding PDE as an Euler–Lagrange equation, we need to reformulate it as a problem on the whole nonconstrained space. The enegry J is well-defined on $[H^1_0(\Omega)]^2$, the space of vector fields whose components belong to $H^1_0(\Omega)$. This can be done via Lagrange multipliers. The idea is to add a term involving the constraint to the functional and to minimize

$$\frac{1}{2}\int_\Omega |Dv|^2\,dx - \int_\Omega p\operatorname{div} v - \int_\Omega f\cdot v\,dx$$

for all $v \in [H^1_0(\Omega)]^2$ and some p in the range of the divergence operator. The latter is then called the space of Lagrange multipliers. It will turn out that the range is precisely

$$L^2_0(\Omega) = \{v \in L^2(\Omega) : \int_\Omega v\,dx = 0\}.$$

As in the proof of the Dirichlet principle (Theorem 1.10), we can compute the derivatives in the directions of perturbations to v and to q and arrive at the following necessary condition

$$\begin{aligned} \int_\Omega Du : Dv\,dx - \int_\Omega p\operatorname{div} v\,dx &= \int_\Omega f\cdot v\,dx && \text{for all } v \in [H^1_0(\Omega)]^2,\\ -\int_\Omega q\operatorname{div} u\,dx &= 0 && \text{for all } q \in L^2_0(\Omega) \end{aligned}$$

for the solution pair $(u,p) \in [H^1_0(\Omega)]^2 \times L^2_0(\Omega)$. (The notation $A : B = \sum_{j=1}^2 A_{jk}B_{jk}$ is used the inner product of matrices.) The physical interpretation of the Lagrange multiplier p is the role of a pressure variable. This system is referred to as the *Stokes equations.* It is easy to see that this system is symmetric but not coercive with respect to the product space (chose for instance $p = \lambda\operatorname{div} u$ for sufficiently large λ). Systems of this structure are called *saddle-point* problems. We will lay the functional analytic basis for studying such saddle-point problems.

We now want to characterize isomorphisms between certain Banach spaces. For a closed subspace $U \subseteq V^*$ of the dual of a Banach space V we define the polar set

$${}^\circ U := \{v \in V : \langle u, v\rangle = 0 \text{ for all } u \in U\} \subseteq V.$$

Here, we use the notation $\langle F, u\rangle = F(u)$. It is the space of all elements of V on which all functionals from U vanish. We recall that for Banach spaces V and W and a continuous linear map $L : V \to W$ the dual $L^* : W^* \to V^*$ is defined by

$$L^*(F) = \langle F, L\,\cdot\rangle \in V^*.$$

We recall the closed range theorem.

Theorem 2.57 (closed range theorem)**.** *Let $L : V \to W$ be a continuous linear map between Banach spaces V and W. The range $L(V)$ is closed in W if and only if $L(V) = {}^\circ(\ker L^*)$.*

Proof. See linear functional analysis. □

Recall that a Banach space Y is called reflexive if the map

$$J : Y \to Y^{**}, \quad Y \ni y \mapsto \langle\cdot, y\rangle$$

from Y to its bidual Y^{**} is an isomorphism.

Lemma 2.58 (Banach–Babuška–Nečas lemma)**.** *Let X be a Banach space and let Y be a reflexive Banach space. A linear map $L : X \to Y^*$ is an isomorphism if and only if the following three conditions are satisfied:*

(1) Continuity: $\|Lx\|_{Y^*} \le C\|x\|_X$ for a constant $C > 0$ and all $x \in X$.

(2) There exists $\gamma > 0$ such that for all $x \in X$

$$\gamma\|x\|_X \le \|Lx\|_{Y^*}$$

(3) For every nonzero $y \in Y \setminus \{0\}$ there exists some $x \in X$ such that $\langle Lx, y\rangle \neq 0$.

Proof. Let conditions (1)–(3) be satisfied. Then, by (1), L is continuous and, by (2), it is injective because $Lx = 0$ implies $x = 0$. Hence, L is bijective as a map from X to its range $L(X)$. The inverse $L^{-1} : L(X) \to X$ is continuous because, by (2),

$$\|L^{-1}z\|_X \le \gamma^{-1}\|LL^{-1}z\|_{Y^*} = \gamma^{-1}\|z\|_{Y^*}.$$

The continuity of L and L^{-1} implies that $L(X)$ is closed. The closed range theorem then teaches

$$L(X) = {}^\circ(\ker L^*) \subseteq Y^*. \tag{12}$$

Let us write down the polar set of $\ker L^* \subseteq Y^{**}$ explicitly:

$${}^\circ(\ker L^*) := \{v \in Y^* : \langle u, v\rangle = 0 \text{ for all } u \in \ker L^*\}.$$

We furthermore observe from the definition of L^* that

$$u \in \ker L^* \iff \langle L^*u, x\rangle = 0 \text{ for all } x \in X \iff \langle u, Lx\rangle = 0 \text{ for all } x \in X.$$

Since Y is reflexive, we see that

$$u \in \ker L^* \iff \langle Lx, J^{-1}u\rangle = 0 \text{ for all } x \in X$$

for $J^{-1}u \in Y$ and the isomorphism J. Property (3) therefore implies that implies that $Ju = 0$ and so $\ker L^* = \{0\}$. By (12), we then have that $L(X) = Y^*$. Thus, L is an isomorphism.
The proof of the converse direction is immediate and left as an exercise to the readers. □

Remark 2.59. An immediate consequence of Lemma 2.58 is the Lax–Milgram Lemma. For the proof, let $X = Y$ be a Hilbert space and let L be defined by the bilinear form a as follows

$$X \in x \mapsto a(x, \cdot) := L(x) \in X^*.$$

It is then an exercise to show that continuity and coercivity of a imply the assumptions of the Banach–Babuška–Nečas lemma, see Problem 48.

We can formulate an analogue of Céa's lemma for Galerkin methods.

Lemma 2.60 (quasi-optimality of Galerkin methods)**.** *Let X be a Banach space and let Y be a reflexive Banach space and let $L : X \to Y^*$ with $\|Lx\|_{Y^*} \leq C$ be a continuous linear map. Let $z \in X$ satisfy $Lz = F$ for some $F \in Y^*$. Let $X_h \subseteq X$ and $Y_h \subseteq Y$ be closed subspaces satisfying*

(i) There exists $\gamma_h > 0$ such that for all $x_h \in X_h$

$$\gamma_h \|x_h\|_X \leq \|Lx_h\|_{Y_h^*}$$

(ii) For every nonzero $y_h \in Y_h \setminus \{0\}$ there exists some $x_h \in X_h$ such that $\langle Lx_h, y_h\rangle \neq 0$.

Then there exists a unique $z_h \in X_h$ solving

$$\langle Lz_h, y_h\rangle = \langle F, y_h\rangle \quad \textit{for all } y_h \in Y_h.$$

It satisfies the quasi-optimal error estimate

$$\|z - z_h\|_X \leq (1 + \frac{C}{\gamma_h}) \inf_{x_h \in X} \|z - x_h\|_X.$$

Proof. The unique solvability follows from an application of Lemma 2.58 to the discrete spaces. For an arbitrary $w_h \in X_h$, the triangle inequality leads to

$$\|z - z_h\|_X \leq \|z - w_h\|_X + \|w_h - z_h\|_X.$$

Note that $(w_h - z_h) \in X_h$ and thus we have due to assumption (i)

$$\gamma_h \|w_h - z_h\|_X \leq \|L(w_h - z_h)\|_{Y_h^*} = \sup_{Y_h \in Y_h \setminus \{0\}} \frac{\langle L(w_h - z_h), y_h\rangle}{\|y_h\|_Y}.$$

From the solution property of z_h and z we obtain

$$\langle Lz_h, y_h\rangle = \langle F, y_h\rangle = \langle Lz, y_h\rangle \quad \text{for any } y_h \in Y_h \subseteq Y.$$

This is a generalization of Galerkin orthogonality. With the continuity of L we therefore obtain

$$\begin{aligned}\gamma_h\|w_h - z_h\|_X = \sup_{Y_h\in Y_h\setminus\{0\}} \frac{\langle L(w_h - z), y_h\rangle}{\|y_h\|_Y} &= \|L(w_h - z)\|_{Y_h^*} \\ &\leq \|L(w_h - z)\|_{Y^*} \leq C\|w_h - z\|_X.\end{aligned}$$

Since w_h was arbitrary, we combine the above estimates and take the infimum to conclude the proof. □

Saddle-point problems

So far, we have seen an abstract characterization of isomorphisms. The conditions are, in general, difficult to verify. We want to focus on problems with some additional structure. We consider two Hilbert spaces V, M and continuous bilinear forms

$$a : V \times V \to \mathbb{R}, \qquad b : V \times M \to \mathbb{R}.$$

Given $F \in V^*$ and $G \in M^*$, we seek $(u, p) \in V \times M$ such that

$$\begin{aligned} a(u, v) + b(v, p) &= F(v) \quad \text{for all } v \in V \\ b(u, q) &= G(q) \quad \text{for all } q \in M. \end{aligned} \tag{13}$$

This is the abstract *saddle-point problem.* Obviously, the above Stokes equations belong to this class. For such problems, we will formulate criteria that guarantee solvability of the system.

The bilinear form b induces a linear map

$$B : V \to M^*$$

from V to the dual of M, which acts as follows

$$V \ni v \mapsto Bv = b(v, \cdot) \in M^*.$$

The adjoint operator

$$B' : M \to V^*$$

is given by

$$M \ni \mu \mapsto B'\mu = b(\cdot, \mu) \in V^*.$$

The kernel of B reads

$$Z := \ker(B) = \{v \in V : \forall \mu \in M \; b(v, \mu) = 0\}.$$

In the following, we will write (with the isomorphism J from V to its bidual)

$$Z^\circ := {}^\circ J(Z) = \{F \in V^* : \langle F, z\rangle = 0 \text{ for all } z \in Z\}$$

for the space of linear functionals vanishing on the kernel Z. We now study under which circumstances B is an isomorphism. The central assumption on b is the following condition

$$0 < \beta = \inf_{\mu\in M\setminus\{0\}} \frac{\|B'\mu\|_{V^*}}{\|\mu\|_M} = \inf_{\mu\in M\setminus\{0\}} \|b(\cdot,\mu)\|_{V^*}\|\mu\|_M = \inf_{\mu\in M\setminus\{0\}} \sup_{v\in V\setminus\{0\}} \frac{b(v,\mu)}{\|\mu\|_M\|v\|_V} \tag{14}$$

called the *inf-sup condition.*

Lemma 2.61. *Let b satisfy the inf-sup condition. Then:*

(i) $B' : M \to Z^\circ$ is an isomorphism with $\|(B')^{-1}\|_{\mathcal{L}(Z^\circ,M)} \le \beta^{-1}$.

(ii) $B : Z^\perp \to M^*$ is an isomorphism with $\|B^{-1}\|_{\mathcal{L}(M,Z^\perp)} \le \beta^{-1}$.

Proof. Let us prove (i). We observe that the range of B' is indeed a subset of Z° because $\langle B'\mu, z\rangle = b(z,\mu) = 0$ for all $z \in Z$. By the above assumptions, B' is continuous (property (1) of Lemma 2.58) and the inf-sup condition implies that property (2) of Lemma 2.58 is satisfied. As in the proof of Lemma 2.58 we therefore see that B' and its inverse are continuous. The closed range theorem then shows

$$B'(M) = {}^\circ(\ker((B')^*)).$$

It is direct to verify

$$u \in \ker((B')^*) \iff J^{-1}u \in Z.$$

Thus the range equals Z° and B' is an isomorphism between M and Z°. The bound on the norm is left as an exercise.
For the proof of (ii), let $v \in Z^\perp$ be given and observe that

$$g : V \to \mathbb{R} \text{ defined by } w \mapsto (v,w)_V$$

is a continuous linear functional in Z°. Since, by (i), B' is an isomorphism, there exists $\mu \in M$ with $\|g\|_{V^*} = \|v\|_V$ such that

$$b(\cdot,\mu) = g.$$

The norm bound from (i) shows

$$\beta\|\mu\|_M \le \|B'\mu\|_{V^*} = \|g\|_{V^*} = \|v\|_V.$$

We divide this relation by the norm of μ and multiply with the norm of v and obtain

$$\beta\|v\|_V \le \|v\|_V \frac{\|\overbrace{B'\mu}^{=g}\|_{V^*}}{\|\mu\|_M} = \frac{\overbrace{(v,v)_V}^{\langle Bv,\mu\rangle}}{\|\mu\|_M} \le \|B\|_{M^*}.$$

Thus (1) and (2) from Lemma 2.58 are satisfied. Furthermore, for any $v \in V$ and its orthogonal projection to Πv to $Z^\perp$ we have $\|\Pi v\|_V \leq \|v\|_V$ by standard results on orthogonal projections. Thus, the inf-sup condition shows

$$0 < \beta = \inf_{\mu \in M\setminus\{0\}} \sup_{v \in V\setminus\{0\}} \frac{b(v,\mu)}{\|\mu\|_M \|v\|_V} \leq \inf_{\mu \in M\setminus\{0\}} \sup_{w \in Z^\perp\setminus\{0\}} \frac{b(w,\mu)}{\|\mu\|_M \|w\|_V}$$

because the Z-part of $v \in V$ is not seen by b. This shows that (3) from Lemma 2.58 is satisfied. The same lemma then states that B is an isomorphism from $Z^\perp$ to M^*. Again, the norm bound is left as an exercise. □

We now formulate the main criterion for saddle-point problems.

Theorem 2.62 (Brezzi splitting)**.** *Let V, M be Hilbert spaces with continuous bilinear forms a on $V \times V$ and b on $V \times M$. If, in the above notation, a is coercive on the kernel Z and b satisfies the inf-sup condition* (14)*, then for any $F \in V^*$ and $G \in M^*$ the saddle-point problem* (13) *has a unique solution $(u,p) \in V \times M$. It satisfies*

$$\begin{aligned} \|u\|_V &\leq \alpha^{-1}\|F\|_{V^*} + \beta^{-1}(1 + \frac{C_a}{\alpha})\|G\|_{M^*}, \\ \|p\|_M &\leq \beta^{-1}(1 + \frac{C_a}{\alpha})\|F\|_{V^*} + \beta^{-1}(1 + \frac{C_a}{\alpha})\frac{C_a}{\beta}\|G\|_{M^*}. \end{aligned}$$

Here, α is the coercivity constant and C_a the continuity constant of a.

Proof. Property (ii) of Lemma 2.61 shows that there exists some $u_0 \in Z^\perp$ with $Bu_0 = G$ and $\|u_0\|_V \leq \beta^{-1}\|G\|_{M^*}$. Upon defining $w := u - u_0$, we reformulate (13) into

$$\begin{aligned} a(w,v) + b(v,p) &= F(v) - a(u_0,v) \qquad &&\text{for all } v \in V \\ b(w,q) &= 0 &&\text{for all } q \in M. \end{aligned}$$

Since a is coercive on Z, there exists (by the Lax–Milgram lemma) a unique $w \in Z$ satisfying

$$a(w,v) = F(v) - a(u_0,v) \quad \text{for all } v \in Z$$

with $\|w\|_V \leq \alpha^{-1}(\|F\|_{V^*} + C_a\|u_0\|_V)$. Since $F - a(u_0 + w, \cdot) \in Z^\circ$, property (i) from Lemma 2.61 yields the existence of $p \in M$ with

$$b(v,p) = F(v) - a(u_0 + w, v) \quad \text{for all } v \in V$$

and

$$\|p\|_M \leq \beta^{-1}(\|F\|_{V^*} + C_a\|u_0 + w\|_V).$$

Hence, $u := u_0 + w$ and p satisfy (13). The asserted norm bounds follow from directly tracing the constants in the above estimates. □

We will use the Brezzi-splitting theorem to prove well-posedness of the Stokes equations. Obviously, the Stokes equations are equivalent to (13) with the choices $V = [H^1_0(\Omega)]^2$, $M := L^2_0(\Omega)$ and

$$a(v,w) = \int_\Omega Dv : Dw \, dx, \quad b(v,q) = -\int_\Omega q \operatorname{div} v \, dx, \quad F(v) = \int_\Omega f \cdot v \, dx, \quad G = 0.$$

Theorem 2.63. *Let $\Omega \subseteq \mathbb{R}^2$ be an open, bounded, and connected domain with polygonal Lipschitz boundary. Given any $f \in L^2(\Omega)$, there exists a unique solution $(u,p) \in V \times M$ to the Stokes equations.*

Proof. Since a is coercive (by Friedrichs' inequality), it remains to prove the inf-sup condition

$$0 < \beta = \inf_{q \in L^2_0(\Omega)\setminus\{0\}} \sup_{v \in [H^1_0(\Omega)]^2 \setminus\{0\}} \frac{\int_\Omega q \operatorname{div} v \, dx}{\|Dv\|_{L^2(\Omega)} \|q\|_{L^2(\Omega)}}$$

for some β. This result is not shown here because its proof is far beyond the scope of this lecture. It can be found in the literature. □

Problem 47. Prove that the Stokes equations are a necessary condition for any minimizer of the constrained energy minimization problem. Show that, for sufficiently regular solutions, the Stokes equations can be written as $-\Delta u + \nabla p = f$ and $\operatorname{div} u = 0$. Here, Δ is the component-wise action of the Laplacian.

Problem 48. Use the Banach–Babuška–Nečas lemma to prove the Lax–Milgram lemma.

Problem 49. Prove that div maps $[H^1_0(\Omega)]^2$ to a subspace of $L^2_0(\Omega)$.

Problem 50. *(Euler's formulae)*
Let $\mathcal{T}$ be a regular triangulation of the simply-connected bounded domain $\Omega \subseteq \mathbb{R}^2$ with vertices $\mathcal{N}$, edges $\mathcal{E}$ and interior edges $\mathcal{E}(\Omega)$. Prove that

$$\#\mathcal{N} + \#\mathcal{T} = 1 + \#\mathcal{E}$$

and

$$2\,\#\mathcal{T} + 1 \ = \#\mathcal{N} + \#\mathcal{E}(\Omega)$$

($\#A$ denotes the cardinality of a set A).

§9 A finite element method for the Stokes system *(week 23)*

For the approximation of saddle-point problems, we aim at choosing finite element subspaces $V_h \subseteq V$ and $M_h \subseteq M$. Since these are closed subspaces, they are again Hilbert spaces and the Brezzi splitting can be used to study the solvability of the resulting discrete problem. In contrast to the coercivity in the Lax–Milgram setting, however, the inf-sup condition is usually not inherited by the discrete spaces. It needs to be imposed as an additional condition. We call

$$0 < \beta_h = \inf_{q_h \in M_h \setminus \{0\}} \sup_{v_h \in V_h \setminus \{0\}} \frac{b(v_h, q_h)}{\|v_h\|_V \|q_h\|_M}$$

the *discrete inf-sup condition.* The construction of discrete spaces satisfying this property turns out nontrivial. The standard finite element space $[S_0^1(\mathcal{T})]^2$ is not suited for a discretization involving the constraint on the divergence, see Problem 51 and Problem 52. We postpone the abstract analysis of Galerkin discretizations of saddle-point problems to later sections and focus for the moment on a practical discretization of the Stokes equations, the so-called Mini finite element. We will define the corresponding spaces, comment on the implementation, and show that it satisfies a discrete inf-sup condition.
Given a triangulation $\mathcal{T}$ of the domain Ω and any $T \in \mathcal{T}$ with vertices z_1, z_2, z_3, we define the so-called element bubble b_T by

$$b_T = \varphi_{z_1} \varphi_{z_2} \varphi_{z_3}$$

where the φ_{z_j} are the hat functions associated to the vertices of T. It is immediate to verify that

- $b_T|_T$ is a cubic polynomial on T,
- b_T is positive in the interior of T,
- b_T vanishes on $\Omega \setminus T$,
- $b_T \in H_0^1(\Omega)$.

We denote the space of bubble functions by

$$\mathcal{B}(\mathcal{T}) := \operatorname{span}\{b_T : T \in \mathcal{T}\}.$$

We then have $\dim(\mathcal{T}) = \#\mathcal{T}$ where $\#$ denotes the cardinality of a set. The Mini finite element discretization is based on the discrete spaces

$$V_h := [S_0^1(\mathcal{T})]^2 \oplus [\mathcal{B}(\mathcal{T})]^2 \quad \text{and} \quad M_h := S^1(\mathcal{T}) \cap L_0^2(\Omega).$$

We clearly have the inclusions $V_h \subseteq V$ and $M_h \subseteq M$. For a practical implementation, we need (local) matrix representations of the bilinear forms a and b.

Local matrices of the MINI finite element

Denote by $\varphi_1, \varphi_2, \varphi_3$ the three nodal basis functions on a triangle T and recall the cubic bubble function $b_T := \varphi_1\varphi_2\varphi_3$. Define the local basis functions of the velocity part of the MINI finite element by

$$\psi_1 = \begin{pmatrix}\varphi_1\\0\end{pmatrix}, \psi_2 = \begin{pmatrix}\varphi_2\\0\end{pmatrix}, \psi_3 = \begin{pmatrix}\varphi_3\\0\end{pmatrix}, \psi_4 = \begin{pmatrix}0\\\varphi_1\end{pmatrix}, \psi_5 = \begin{pmatrix}0\\\varphi_2\end{pmatrix}, \psi_6 = \begin{pmatrix}0\\\varphi_3\end{pmatrix},$$
$$\psi_7 = \begin{pmatrix}b_T\\0\end{pmatrix}, \psi_8 = \begin{pmatrix}0\\b_T\end{pmatrix}.$$

The local basis functions for the pressure component are $\varphi_1, \varphi_2, \varphi_3$. The local matrices then read as

$$A_T = \left[\int_T D\psi_j : D\psi_k \, dx\right]_{j,k=1,\dots,8} \quad \text{and} \quad B_T = \left[-\int_T \varphi_j \operatorname{div} \psi_k \, dx\right]_{\substack{j=1,\dots,3\\k=1,\dots,8}}.$$

We can then compute the entries as follows.

Lemma 2.64. *(a) A_T has the following block structure*

$$A_T = \begin{bmatrix} S & 0 & 0\\ 0 & S & 0\\ 0 & 0 & R\end{bmatrix}$$

for

$$S = \left[\int_T \nabla\varphi_j \cdot \nabla\varphi_k \, dx\right]_{j,k=1,2,3} \quad \textit{and} \quad R = \frac{|T|}{180}\sum_{j=1}^{3} |\nabla\varphi_j|^2 \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}.$$

(b) B_T has the following block structure

$$B_T = |T| \left[\begin{array}{c|c} L & \\ L & G \\ L & \end{array}\right]$$

for

$$L = -\frac{1}{3}\begin{pmatrix}\partial_x\varphi_1 & \partial_x\varphi_2 & \partial_x\varphi_3 & \partial_y\varphi_1 & \partial_y\varphi_2 & \partial_y\varphi_3\end{pmatrix} \quad \textit{and} \quad G = \frac{1}{60}\begin{bmatrix}\partial_x\varphi_1 & \partial_y\varphi_1\\ \partial_x\varphi_2 & \partial_y\varphi_2\\ \partial_x\varphi_3 & \partial_y\varphi_3\end{bmatrix}.$$

Proof. Exercise (recall the integration formula from Problem 23). □

Assembling the global system matrix

These local matrices need to be assembled to the global system matrix of the Mini FEM. We need to fix some numbering of degrees of freedom. Let $N_N = \operatorname{card}(\mathcal{N}(\Omega))$ denote the number of interior vertices and $N_T = \operatorname{card}(\mathcal{T})$ denote the number of triangles in the triangulation $\mathcal{T}$ of the 2D domain Ω. With the nodal basis functions $(\lambda_z)_{z\in\mathcal{N}}$ define the following basis functions $\psi_1, \dots, \psi_{2N_N+2N_T}$ for the velocity by

$$(\psi_1, \dots, \psi_{N_N}) = \left[\begin{pmatrix}\varphi_z\\ 0\end{pmatrix}\right]_{z\in\mathcal{N}(\Omega)}, \quad (\psi_{N_N+1}, \dots, \psi_{2N_N}) = \left[\begin{pmatrix}0\\ \varphi_z\end{pmatrix}\right]_{z\in\mathcal{N}(\Omega)},$$
$$(\psi_{2N_N+1}, \dots, \psi_{2N_N+N_T}) = \left[\begin{pmatrix}b_T\\ 0\end{pmatrix}\right]_{T\in\mathcal{T}}, \quad (\psi_{2N_N+N_T+1}, \dots, \psi_{2N_N+2N_T}) = \left[\begin{pmatrix}0\\ b_T\end{pmatrix}\right]_{T\in\mathcal{T}}.$$

and for the pressure component define

$$(q_1, \dots, q_{N_N}) = [\varphi_z]_{z\in\mathcal{N}(\Omega)}.$$

The global matrices then read as

$$A = \left[\int_\Omega D\psi_j : D\psi_k \,dx\right]_{j,k=1,\dots,2(N_N+N_T)} \quad \text{and} \quad B = \left[-\int_\Omega q_j \operatorname{div}\psi_k \,dx\right]_{\substack{j=1,\dots,N_N\\ k=1,\dots,2(N_N+N_T)}}.$$

We then observe:

Lemma 2.65. *(a) The global system matrix M of the Mini FEM discretization of the saddle-point formulation has the block structure*

$$M = \begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}$$

so that the discrete equation reads as

$$\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}\begin{bmatrix} u_h \\ p_h \end{bmatrix} = \begin{bmatrix} F \\ 0 \end{bmatrix}.$$

(b) The matrix M has a nontrivial kernel, namely globally constant pressure modes.

Proof. Exercise. □

So far we did not enforce the discrete pressure variable to belong to $L^2_0(\Omega)$. This constraint is best included via a Lagrange multiplier. Details can be found in the code provided on the course webpage, which follows the implementation outlined above.

The discrete inf-sup condition for the Mini element

We now verify the discrete inf-sup condition for the Mini FEM. The main technical tool is a so-called quasi-interpolation operator. In contrast to the nodal interpolation from prior sections, the quasi-interpolation is well defined for any function from $H^1(\Omega)$. The idea is to replace point evaluations by certain volume averages around the vertices.
Given a triangulation $\mathcal{T}$ of the domain Ω, we define the nodal patch of any vertex $z \in \mathcal{N}$ by

$$\omega_z := \cup\{T \in T : z \in T\}$$

as the domain of all elements containing z, and the element patch

$$\omega_T := \cup_{z\in\mathcal{N}(T)}\omega_z$$

as the domain of all elements surrounding T. Given any $v \in H^1_0(\Omega)$, we then define its quasi-interpolation $R_h v \in S^1_0(\mathcal{T})$ by

$$R_h v := \sum_{z\in\mathcal{N}(\Omega)} \fint_{\omega_z} v\,dx\;\varphi_z$$

where we use the notation $\fint_{\omega_z} \cdot\,dx = |\omega_z|^{-1}\int_{\omega_z} \cdot\,dx$ for the integral mean. The operator $R_h : H^1_0(\Omega) \to S^1_0(\mathcal{T})$ is called *quasi-interpolation operator.* Its difference to the nodal interpolation is that the point evaluation $v(z)$ is replaced by the computation of some average around z.

Theorem 2.66. *Let $\mathcal{T}$ be a regular triangulation of some open and bounded Lipschitz polygon Ω. There exists a constant $C > 0$ such that the quasi-interpolation R_h satisfies for any $v \in H^1_0(\Omega)$ the local error estimates*

$$h_T^{-1}\|v - R_h v\|_{L^2(T)} + \|\nabla(v - R_h v)\|_{L^2(T)} \le C\|\nabla v\|_{L^2(\omega_T)}.$$

The constant C depends on the shape-regularity and on the shapes (but not the size) of the nodal patches.

Proof. We postpone the proof to later sections. □

Theorem 2.67. *The Mini finite element satisfies the discrete inf-sup condition. On a shape-regular sequence of triangulations $(\mathcal{T}_h)_h$, the discrete inf-sup constant β_h is uniformly bounded from below by some $\beta_0 > 0$.*

Proof. Given any $q_h \in M_h \subseteq M$, the inf-sup condition for the spaces V and M shows that there exists some $u \in V$ such that $\operatorname{div} u = q_h$ and $\|Du\|_{L^2(\Omega)} \le \beta^{-1}\|q_h\|_{L^2(\Omega)}$. we set

$$u_h = u_h(q_h) := R_h u + \sum_{T\in\mathcal{T}} \frac{b_T}{\int_T b_T\,dx}\int_T (u - R_h u)\,dx \in V_h.$$

It is shown as an exercise that

$$\|Du_h\|_{L^2(\Omega)} \le C\|Du\|_{L^2(\Omega)}$$

for some constant $C > 0$. It is furthermore immediate to see that u_h satisfies $\int_T (u_h - u) = 0$ for all $T \in \mathcal{T}$. Therefore, integration by parts reveals

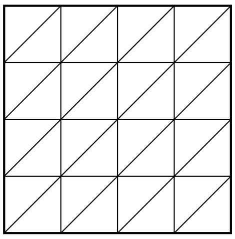
$$b(u_h - u, q_h) = \int_\Omega \operatorname{div}(u_h - u) q_h \, dx = -\int_\Omega (u_h - u) \cdot \nabla q_h \, dx = 0$$

because ∇q_h is piecewise constant. We compute, by plugging in the particular candidate $u_h(q_h)$ in the supremum, that

$$\begin{aligned} \inf_{q_h \in M_h \setminus \{0\}} \sup_{v_h \in V_h \setminus \{0\}} \frac{b(v_h, q_h)}{\|v_h\|_V \|q_h\|_M} &\geq \inf_{q_h \in M_h \setminus \{0\}} \frac{b(u_h(q_h), q_h)}{\|u_h(q_h)\|_V \|q_h\|_M} \\ &= \inf_{q_h \in M_h \setminus \{0\}} \frac{b(u(q_h), q_h)}{\|u_h(q_h)\|_V \|q_h\|_M} \\ &= \frac{1}{C} \inf_{q_h \in M_h \setminus \{0\}} \frac{b(u(q_h), q_h)}{\|u(q_h)\|_V \|q_h\|_M} \geq \frac{\beta}{C}. \end{aligned}$$

The choice $\beta_h := \beta / C$ completes the proof. □

Problem 51. *(conforming divergence-free functions are trivial)*
Let $\mathcal{T}$ be the criss triangulation of the unit square and let $u_h \in [S_0^1(\mathcal{T})]^2$ with $\operatorname{div} u_h = 0$. Prove that $u_h = 0$.
Hint: The criss triangulation is


.

Problem 52. *(standard FEMs are unstable for Stokes)*
Let $\Omega = (0,1)^2$. Prove that the following discretizations of the Stokes equations lead to unstable saddle-point problems (i.e., the discrete inf-sup condition is violated): $V_h := [S_0^1(\mathcal{T})]^2$ and $M_h := P_0(\mathcal{T}) \cap L_0^2(\Omega)$ on the criss triangulation $\mathcal{T}_h$.
Hint: Use a dimension argument with the formulae from Problem 50.

Problem 53. Prove Lemma 2.64.

Problem 54. Prove Lemma 2.65.

Problem 55. Prove the bound on $\|Du_h\|_{L^2(\Omega)}$ from the proof of Theorem 2.67.

Problem 56. Implement the Mini finite element. As a test example, use the following data on the square $\Omega = (-1,1)^2$ (not the unit square): The right-hand side $f = 0$ is zero and the exact solution is

$$u(x_1, x_2) = \begin{pmatrix} 20x_1x_2^4 - 4x_1^5 \\ 20x_1^4x_2 - 4x_2^5. \end{pmatrix}$$

Choose the inhomogeneous Dirichlet data u_D according to u. Create convergence history plots for the error in the u variable.
Hint: An example of an implementation can be found on the course webpage.

§10 Error estimates *(week 24)*

We start by reformulating Lemma 2.60 in the context of saddle-point problems. If we add the two equations of the saddle-point problem (13) we arrive at the equivalent formulation: Find $(u, p) \in V \times M$ such that

$$\mathcal{A}(u, p; v, q) = F(v) + G(q) \quad \text{for all } (v, p) \in V \times M \tag{15}$$

where

$$\mathcal{A}(u, p; v, q) = a(u, p) + b(v, p) + b(u, q)$$

is a continuous bilinear form on $V \times M$. The Brezzi splitting theorem (Theorem 2.62) states that

$$(u, p) \mapsto \mathcal{A}(u, p; \cdot, \cdot)$$

is an isomorphism from $V \times M$ to its dual $V^* \times M^*$ provided a is coercive on the kernel Z of b, and b satisfies the inf-sup condition.
Let $V_h \subseteq V$ and $M_h \subseteq M$ be closed subspaces. The restriction of $\mathcal{A}$ to $V_h \times M_h$ defines a map

$$(u_h, p_h) \mapsto \mathcal{A}(u_h, p_h; \cdot, \cdot)$$

from $V_h \times M_h$ to its dual.

Theorem 2.68. *Assume $\mathcal{A}$ is an isomorphism. Let $V_h \subseteq V$ and $M_h \subseteq M$ be closed subspaces, let a be coercive on the discrete kernel*

$$Z_h := \{v_h \in V_h : b(v_h, q_h) = 0 \textit{ for all } q_h \in M_h\}$$

and let b satisfy the discrete inf-sup condition. Then, given any $F \times G \in V^* \times M^*$, there exists a unique $(u_h, p_h) \in V \times M$ such that

$$\mathcal{A}(u_h, p_h; v_h, q_h) = F(v_h) + G(q_h) \quad \textit{for all } (v_h, q_h) \in V \times M.$$

We have the quasi-optimal error estimate

$$\|(u - u_h, p - p_h)\|_{V \times M} \leq (1 + \frac{C_{\mathcal{A}}}{\gamma_h}) \inf_{v_h \in V_h} \inf_{q_h \in M_h} \|(u - v_h, p - q_h)\|_{V \times M}$$

where $C_{\mathcal{A}}$ is the continuity constant of $\mathcal{A}$ and γ_h^{-1} is the continuity constant of the discrete inverse to $\mathcal{A}_h$.

Proof. From Theorem 2.62 applied to the discrete spaces, we deduce that $\mathcal{A}_h$ is an isomorphism from $V_h \times M_h$ to its dual. With the continuity constant $\gamma_h > 0$ of the inverse operator we have in particular that

$$\gamma_h \|(v_h, q_h)\|_{V_h \times M_h} \leq \|\mathcal{A}_h(v_h, q_h)\|_{V_h^* \times M_h^*}.$$

The existence and uniqueness of the approximate solution (u_h, p_h) as well as the error estimate follow from the abstract bound of Lemma 2.60. □

Warning 2.69. In general we expect $Z_h \not\subseteq Z$, i.e., the kernel spaces are not nested.

Checking the discrete inf-sup condition can be a difficult task. In the proof of Theorem 2.67 we have constructed a bounded operator from V to V_h, $u \mapsto u_h$, with the property $b(u_h - u, q_h) = 0$ for all $q_h \in M_h$. Such an operator is called *Fortin operator.* Constructing a Fortin operator often is a suitable method for verifying the inf-sup condition.

Lemma 2.70 (Fortin criterion)**.** *Let the bilinear form $b : V \times M \to \mathbb{R}$ satisfy the inf-sup condition. Assume that for closed subspaces $V_h \subseteq V$ and $M_h \subseteq M$ there exists a bounded linear map $\Pi_h : V \to V_h$ with the property that*

$$b(v - \Pi_h v, q_h) = 0 \quad \textit{for all } q_h \in M_h.$$

Then, the discrete inf-sup constant is satisfied with a constant proportional to the inverse of the continuity constant of Π_h.

Proof. We repeat the argument from Theorem 2.67 in this abstract framework. The inf-sup condition for V and M and the properties of Π_h show for any $q_h \in M_h \subseteq M$ that

$$\begin{aligned}\beta \|q_h\|_M \leq \sup_{v \in V \setminus \{0\}} \frac{b(v, q_h)}{\|v\|_V} &= \sup_{v \in V \setminus \{0\}} \frac{b(\Pi_h v, q_h)}{\|v\|_V} \\ &\leq C \sup_{\substack{v \in V \setminus \{0\} \\ \Pi_h v \neq 0}} \frac{b(\Pi_h v, q_h)}{\|\Pi_h v\|_V} \leq C \sup_{v_h \in V_h \setminus \{0\}} \frac{b(v_h, q_h)}{\|v_h\|_V}.\end{aligned}$$

Here, we denoted the continuity constant of Π_h by C. □

Error bounds for the quasi-interpolation

Proof of Theorem 2.66. We fix $T \in \mathcal{T}$ and $v \in H^1_0(\Omega)$. Then the error $R_h v$ restricted to T has the representation

$$R_h v|_T = \sum_{z \in \mathcal{N}(T)} (R_h v)(z) \varphi_z.$$

We use that the φ_z sum up to 1 on T and the Young inequality $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ that

$$\|v - R_h v\|^2_{L^2(T)} = \| \sum_{z \in \mathcal{N}(T)} (v - (R_h v)(z)) \varphi_z \|^2_{L^2(T)} \leq 3 \sum_{z \in \mathcal{N}(T)} \|v - (R_h v)(z)\|^2_{L^2(T)}$$

where we used $\|\varphi_z\|_{L^\infty(T)} = 1$. If $z \in \Omega$ is an interior vertex, we obtain from the Poincaré inequality that

$$\|v - (R_h v)(z)\|_{L^2(T)} \leq \|v - \fint_{\omega_z} v \, dx\|_{L^2(T)} \leq \|v - \fint_{\omega_z} v \, dx\|_{L^2(\omega_z)} \leq C C_P(\omega_z) h_T \|\nabla v\|_{L^2(\omega_z)}.$$

If $z \in \Omega$ is a boundary vertex, then $(R_h v)(z) = 0$ and z belongs to a boundary edge, so that v has zero boundary conditions on a part of $\partial\omega_z$ of positive surface measure. Thus, Friedrichs' inequality implies

$$\|v - (R_h v)(z)\|_{L^2(T)} = \|v\|_{L^2(T)} \leq \|v\|_{L^2(\omega_z)} \leq h_T C C_F \|\nabla v\|_{L^2(\omega_z)}.$$

Thus, we have the local L^2 bound

$$\|v - R_h v\|_{L^2(T)} \leq C h_T \|\nabla v\|_{L^2(\omega_T)}.$$

In order to show the local bound on the gradient, we again use the above representation of R_h and the fact that the $\nabla\varphi_z$ sum up to zero on T. From this we see that

$$\begin{aligned}\|\nabla R_h v\|^2_{L^2(T)} &= \| \sum_{z\in\mathcal{N}(T)} (v - (R_h v)(z))\nabla\varphi_z\|^2_{L^2(T)} \\ &\leq 3 \sum_{z\in\mathcal{N}(T)} \|v - (R_h v)(z)\|^2_{L^2(T)} \|\nabla\varphi_z\|_{L^\infty(T)} \\ &\leq C h_T^{-1} \sum_{z\in\mathcal{N}(T)} \|v - (R_h v)(z)\|^2_{L^2(T)}.\end{aligned}$$

Here we used $\|\nabla\varphi_z\|_{L^\infty(T)} \leq C h_T^{-1}$, see Problem 59. The above bounds on $v - (R_h v)(z)$ in the L^2 norms then imply the assertion. Note that the constants $C_P(\omega_z)$ and $C_F(\omega_z)$ are independent of the diameter of $C_P(\omega_z)$. They only depend on the shape of the patch. □

Let us conclude the study of the Stokes equations by summarizing our findings for the Mini finite element as an error estimate.

Theorem 2.71. *The mini finite element discretization (u_h, p_h) to the Stokes equations satisfies*

$$\begin{aligned}&\|D(u - u_h)\|_{L^2(\Omega)} + \|p - p_h\|_{L^2(\Omega)} \\ &\qquad \leq C\Big(\inf_{v_h \in [S^1_0(\mathcal{T}) + \mathcal{B}(\mathcal{T})]^2} \|D(u - v_h)\|_{L^2(\Omega)} + \inf_{q_h \in P_0(\mathcal{T}) \cap L^2_0(\Omega)} \|p - q_h\|_{L^2(\Omega)}\Big)\end{aligned}$$

for some constant C that is independent of the mesh size.

Remark 2.72. In case that Ω is convex, it is known that additionally we have $u \in [H^2(\Omega)]^2$ and $p \in H^1(\Omega)$ with

$$\|D^2 u\|_{L^2(\Omega)} + \|\nabla p\|_{L^2(\Omega)} \leq C\|f\|_{L^2(\Omega)}.$$

Together with suitable (quasi-)interpolation estimates we therefore conclude that the error of the Mini FEM is bounded by some $\tilde{C}h\|f\|_{L^2(\Omega)}$ on convex domains and thus converges at order h.

Problem 57 (H^1 stability of the L^2 projection)**.** Let $\Pi_h : H^1_0(\Omega) \to S^1_0(\mathcal{T})$ denote the L^2 projection, i.e., for given $v \in H^1_0(\Omega)$, the function $\Pi_h v \in S^1_0(\mathcal{T})$ satisfies

$$\int_\Omega (\Pi_h v) w_h \, dx = \int_\Omega v w_h \, dx \quad \text{for all } w_h \in S^1_0(\mathcal{T}).$$

Then, clearly, $\|\Pi_h v\|_{L^2(\Omega)} \leq \|v\|_{L^2(\Omega)}$. Prove that for a family of red-refined triangulations there exists a constant C such that

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

i.e., the L^2 projection is H^1 stable.
Hint: Given v, prove $\Pi_h v = \Pi_h(R_h v - v) + R_h v$. For the first term, use an inverse estimate (Problem 44), the L^2 stability of the projection $\Pi_h v$, and the approximation and stability properties of the quasi-interpolation R_h.

Problem 58 (density of finite element spaces)**.** Prove that the finite element spaces $S^1_0(\mathcal{T}_j)$ with respect to a sequence $\mathcal{T}_j$ of red refined triangulations are dense in $H^1_0(\Omega)$.
Hint: Given v, approximate it by a smooth function v_ε and interpolate v_ε by a finite element function on a sufficiently fine mesh.

Problem 59. Prove that the nodal basis functions φ_z on a triangle T satisfy

$$\|\nabla \varphi_z\|_{L^2(T)} \leq C_1 \quad \text{and} \quad \|\nabla \varphi_z\|_{L^\infty(T)} \leq C_2 h_T^{-1}$$

with constants C_1, C_2 only depending on the shape regularity.

Problem 60. *(backward facing step)* Use the Mini element to simulate the flow over a backward facing step. Print the computed velocity and pressure and present the plots in the tutorial session. The parameters are:

- Domain: $\Omega = ((-2, 8) \times (-1, 1)) \setminus ([-2, 0] \times [-1, 0])$ (see Figure 7)
- Forcing term: $f = 0$,
- Dirichlet data: $u_D(x, y) = \begin{cases} (0,0) & \text{for } -2 < x < 8 \\ (-y(y-1)/10, 0) & \text{for } x = -2 \\ (-(y+1)(y-1)/80, 0) & \text{for } x = 8. \end{cases}$

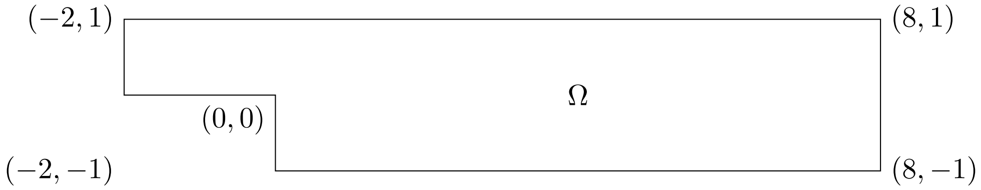


Figure 7: The backward facing step.

§11 Mixed finite elements for Poisson's equation *(week 25)*

Poisson's equation can be written as a saddle-point problem as follows. For simplicity we assume a homogeneous pure Dirichlet boundary condition on $\Gamma_D := \partial\Omega.$¸ We introduce an additional vector variable σ for the gradient and write

$$\sigma = \nabla u, \qquad -\operatorname{div}\sigma = f,$$

which is obviously equivalent to Poisson's equation for u because $\Delta = \operatorname{div}\nabla$. Given $f \in L^2$, we see that σ is required to satisfy certain differentiability properties, namely $\sigma \in [L^2(\Omega)]^2$ and $\operatorname{div}\sigma \in L^2$. We say that $\sigma \in [L^2(\Omega)]^2$ has a weak divergence if there is some $g \in L^1_{\mathrm{loc}}(\Omega)$ such that

$$\int_\Omega \sigma \cdot \nabla\psi \, dx = -\int_\Omega g\psi \, dx \quad \text{for all } \psi \in C_c^\infty(\Omega).$$

We write $\operatorname{div}\sigma = g$ for the weak divergence. The reasoning behind this definition is the same as for the weak partial derivative.

Example 2.73. Let $u \in H^1_0(\Omega)$ be the solution to the weak form of the Dirichlet problem for Poisson's equation with some right-hand side $f \in L^2(\Omega)$. The gradient ∇u satisfies

$$\int_\Omega \nabla u \cdot \nabla\psi \, dx = \int_\Omega f\psi \, dx \quad \text{for all } \psi \in C_c^\infty(\Omega)$$

in the variational formulation. Thus, ∇u possesses the weak divergence $-f$ and we have $-\operatorname{div}\nabla u = f$ as an equality of L^2 functions.

Definition 2.74. The function space of L^2 vector fields with weak divergence in $L^2(\Omega)$ is denoted by

$$H(\operatorname{div},\Omega) := \left\{ v \in L^2(\Omega;\mathbb{R}^2) \ : v \text{ has a weak divergence in } L^2(\Omega) \right\}.$$

It is endowed with the norm

$$\|v\|_{H(\operatorname{div},\Omega)} := \sqrt{\|v\|^2_{L^2(\Omega)} + \|\operatorname{div} v\|^2_{L^2(\Omega)}}.$$

It was already considered in Problem 25. For the variable σ above we require $\sigma \in H(\operatorname{div},\Omega)$ and

$$\int_\Omega \operatorname{div}\sigma \, v \, dx = -\int_\Omega f v \, dx \quad \text{for all } v \in L^2(\Omega).$$

The relation $\sigma = \nabla u$ and integration by parts reveal for any $\tau \in H(\operatorname{div},\Omega)$ that

$$\int_\Omega \sigma \cdot \tau \, dx = -\int_\Omega \operatorname{div}\tau u \, dx + \int_\Omega u\tau \cdot \nu \, ds.$$

Assuming a homogeneous Dirichlet boundary condition for u we thus find

$$\int_\Omega \sigma \cdot \tau \, dx + \int_\Omega \operatorname{div}\tau u \, dx = 0 \quad \text{for all } \tau \in H(\operatorname{div},\Omega).$$

For this equation it is enough that $u \in L^2(\Omega)$. We thus seek $\sigma \in H(\operatorname{div},\Omega)$ and $u \in L^2(\Omega)$ such that

$$\begin{aligned} \int_\Omega \sigma \cdot \tau \, dx + \int_\Omega \operatorname{div} \tau u \, dx &= 0 && \text{for all } \tau \in H(\operatorname{div},\Omega), \\ \int_\Omega \operatorname{div} \sigma \, v \, dx &= -\int_\Omega f v \, dx && \text{for all } v \in L^2(\Omega). \end{aligned}$$

In this way we have formulated Poisson's equation as a saddle-point problem. This formulation is referred to as *mixed formulation* because variables from different function spaces are involved. As an exercise it is shown that the system satisfies the properties from Brezzi's splitting theorem and is therefore well-posed. We remark that we have explicitly imposed the $H(\operatorname{div},\Omega)$ regularity for the vector variable but now merely ask u to belong to $L^2(\Omega)$. The property that σ is the weak gradient of u is implicitly contained in the first row of the system.

We want to identify appropriate finite element spaces leading to a stable discretization of the mixed Laplacian. Since $L^2(\Omega)$ function do not require any continuity, a reasonable choice is to discertize it by the subspace $P_0(\mathcal{T})$ of piecewise constant (possibly discontinuous) functions with respect to a regular triangulation. For piecewise polynomial discretizations of $H(\operatorname{div},\Omega)$ we have seen in Problem 25 that for each edge of the triangulation the component of the piecewise polynomial vector field must be continuous in the normal direction with respect to the edge. We thus will use the normal directions at the edges as the degrees of freedom. We begin with the construction on a single triangle. We set

$$RT_0(T) := \{v \in [L^2(T)]^2 : v(x) = \begin{pmatrix} a \\ b \end{pmatrix} + cx \text{ for } a,b,c \in \mathbb{R}\}.$$

The vector fields of $RT_0(T)$ belong to a subset of the vector fields that are affine in each component. Obviously $\dim RT_0(T) = 3$. For the standard P_1 finite element, the degrees of freedom were the point evaluations at the vertices and we worked with the nodal basis of hat functions. Since here we want to enforce continuity of the normal component across edges we seek a basis $(\psi_E)_{E \in \mathcal{E}(T)}$, where $\mathcal{E}(T)$ is the set of edges of T, such that

$$⨍_F \psi_E \cdot \nu_F \, dx = \begin{cases} 1 & \text{if } E = F \\ 0 & \text{else.} \end{cases} \tag{16}$$

Here, ν_F is the outer normal vector of T restricted to the edge F. This property is achieved by the following definition

$$\psi_{T,E}(x) := \frac{|E|}{2|T|}(x - P_E)$$

where P_E is the vertex of T opposite to E. The proof of (16) is left as an exercise. Globally, we then define

$$RT_0(\mathcal{T}) := \{v \in H(\operatorname{div},\Omega) : \forall T \in \mathcal{T} \; v|_T \in RT_0(T)\}.$$

This space is called the Raviart–Thomas finite element space. We have seen that it consists of all vector fields that are in $RT_0(T)$ for every triangle T and that are normal-continuous across

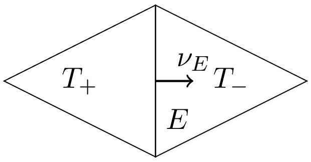


Figure 8: Convention for the edge normal.

each edge. Given any interior edge E, we fix a normal vector. For the two neighbouring triangles T_+ and T_- this vector then points inwards to one of them and outwards to the other one. We use the convention that

$$\nu_E = \nu_{T_+} \quad \text{and} \quad \nu_E = -\nu_{T_-}$$

that is, ν_E is the outward pointing normal to T_+. This is graphically illustrated in Figure 8. If E is a boundary edge, we define $T_- = \emptyset$.
The functions

$$\psi_E(x) = \begin{cases} \psi_{T_+,E}(x) & \text{if } x \in T_+ \\ -\psi_{T_-,E}(x) & \text{if } x \in T_- \\ 0 & \text{else} \end{cases}$$

then form a global basis of $RT_0(\mathcal{T})$.

Lemma 2.75. *The functions $(\psi_E)_{E\in\mathcal{E}}$ form a basis of $RT_0(\mathcal{T})$. They satisfy $\fint_F \psi_E \cdot \nu_F \, dx = \delta_{EF}$.*

Proof. Exercise. □

The Raviart–Thomas space has a canonical interpolation operator, which reads for any sufficiently smooth vector field τ

$$I_{RT}\tau = \sum_{E\in\mathcal{E}} \fint_E \tau \cdot \nu \, ds \psi_E.$$

By construction, it satisfies the conservation property

$$\int_E I_{RT}\tau \cdot \nu_E \, ds = \int_E \tau \cdot \nu \, ds \quad \text{for any } E \in \mathcal{E}.$$

We will see that this operator is not well defined for functions in $H(\operatorname{div}, \Omega)$ but requires further regularity of τ. A sufficient criterion for $I_{RT}\tau$ to exist is for instance $\tau \in [H^1(\Omega)]^2$ because traces along edges are well defined due to the trace theorem. The following result shows H^1 stability of I_{RT}.

Theorem 2.76. *The Raviart–Thomas interpolation is stable with respect to the H^1 norm in the following sense. There exists a constant $C_{I_{RT}}$ only dependent on the shape regularity of $\mathcal{T}$ such that*

$$\|I_{RT}v\|_{H^1(\Omega)} \le C_{I_{RT}} \|v\|_{H^1(\Omega)} \quad \textit{for all } v \in [H^1(\Omega)]^2.$$

Proof. The restriction of $I_{RT}v$ to a triangle K can be written in terms of the basis expansion as follows

$$I_{RT}v|_K = \sum_{E\in\mathcal{E}(K)} \fint_E v\cdot \nu_E \, ds \psi_E.$$

A direct computation for the basis function shows that

$$\|\psi_E\|_{L^2(K)} = \frac{|E|}{2|K|}\|\bullet - P_E\|_{L^2(K)} \leq \frac{|E|}{2|K|}h_K|K|^{1/2} = \frac{|E|}{2|K|^{1/2}}h_K \leq \frac{1}{2\sqrt{c}}h_K$$

where we used that there exists a constant c (only depending on the shape regularity) such that $h_k^2 \leq c|K|$ and the elementary estimate $|E| \leq h_K$. Similarly,

$$\|D\psi_E\|_{L^2(K)} \leq \frac{|E|}{2|K|}\sqrt{2}|K|^{1/2} = \frac{|E|}{2|K|^{1/2}}\sqrt{2} \leq \frac{1}{\sqrt{2c}}.$$

We recall the trace inequality from Theorem 1.27 and compute for the coefficient in front of ψ_E that

$$\begin{aligned} |\fint_E v\cdot\nu_E| = |E|^{-1}|\int_E v\cdot\nu_E| &\leq |E|^{-1/2}\|v\|_{L^2(E)} \\ &\leq |E|^{-1/2}\sqrt{\frac{3|E|}{2|K|}\|v\|^2_{L^2(K)} + \frac{|E|}{2|K|}h_K^2\|Dv\|^2_{L^2(K)}} \\ &= \sqrt{\frac{3}{2|K|}\|v\|^2_{L^2(K)} + \frac{1}{2|K|}h_K^2\|Dv\|^2_{L^2(K)}}. \end{aligned}$$

There exists a constant c (only depending on the shape regularity) such that $h_k^2 \leq c|K|$. We thus infer

$$|\fint_E v\cdot\nu_E| \leq \sqrt{\frac{3}{2h_K^2}\|v\|^2_{L^2(K)} + \frac{1}{2}\|Dv\|^2_{L^2(K)}} \leq C_1(h_K^{-1}\|v\|_{L^2(K)} + \|Dv\|_{L^2(K)})$$

for some universal constant C_1. We use the triangle inequality and compute

$$\begin{aligned} \|I_{RT}v\|_{L^2(K)} \leq \sum_{E\in\mathcal{E}(K)} |\fint_E v\cdot\nu_E \, ds|\|\psi_E\|_{L^2(\Omega)} &\leq 3\frac{1}{2\sqrt{c}}h_K C_1(h_K^{-1}\|v\|_{L^2(K)} + \|Dv\|_{L^2(K)}) \\ &\leq C_2\|v\|_{H^1(K)} \end{aligned}$$

for some constant C_2 (where we used the very rough esimate $h_K \leq$ const for the gradient term). In order to bound the gradient, we observe that $DI_{RT}v = DI_{RT}(v - \fint_K v\,dx)$ for the constant $\fint_K v\,dx$ (component-wise integral mean) because I_{RT} conserves constants (exercise).

We then compute

$$\begin{aligned}\|DI_{RT}v\|_{L^2(K)} &= \|DI_{RT}(v - \fint_K v\,dx)\|_{L^2(\Omega)}\\ &\leq \sum_{E\in\mathcal{E}(K)} |\fint_E (v - \fint_K v\,dx)\cdot\nu_E\,ds|\|D\psi_E\|_{L^2(\Omega)}\\ &\leq 3\frac{1}{\sqrt{2c}}C_1(h_K^{-1}\|(v - \fint_K v\,dx)\|_{L^2(K)} + \|\nabla v\|_{L^2(K)})\\ &\leq C_3\|v\|_{H^1(K)}\end{aligned}$$

for some constant C_3 where we used the Poincaré inequality for the mean-free function $v - \fint_K v\,dx$. Note that the constant in the Poincaré inequality scales like h_K. The claimed bound on the $H^1(\Omega)$ norm follows from using this local argument in the sum expansion

$$\|I_{RT}v\|^2_{H^1(\Omega)} = \sum_{K\in\mathcal{T}} \|I_{RT}v\|^2_{H^1(K)}.$$

□

The following so-called *commuting diagram property* is of particular importance. We denote by $\Pi_0 : L^2(\Omega) \to P_0(\mathcal{T})$ the L^2 projection on piecewise constants. It has the following representation (exercise)

$$(\Pi_0 q)|_T = \fint_T q\,dx \quad \text{for all } q \in L^2(\Omega) \text{ and all } T \in \mathcal{T}.$$

For vector variables, we use the same symbol Π_0 to denote the component-wise L^2 projection on $[P_0(\mathcal{T})]^2$.

Lemma 2.77 (commuting diagram property)**.** *The Raviart–Thomas interpolation* I_{RT} : $[H^1(\Omega)]^2 \to \mathbb{R}T_0(\mathcal{T})$ *satisfies*

$$\operatorname{div} I_{RT}v = \Pi_0 \operatorname{div} v.$$

In other words, the diagram

$$\begin{tikzcd} {[H^1(\Omega)]^2} \arrow[r, "\operatorname{div}"] \arrow[d, "I_{RT}"'] & L^2(\Omega) \arrow[d, "\Pi_0"] \\ RT_0(\mathcal{T}) \arrow[r, "\operatorname{div}"] & P_0(\mathcal{T}) \end{tikzcd}$$

commutes.

Proof. Let $v[H^1(\Omega)]^2$. The divergence theorem shows for any $T \in \mathcal{T}$ with outer unit normal ν that

$$\int_T \operatorname{div} I_{RT}v\,dx = \int_{\partial T} I_{RT}v\cdot\nu\,ds = \sum_{E\in\mathcal{E}(T)} \int_E I_{RT}v\cdot\nu|_E\,ds.$$

For any edge $E \in \mathcal{E}(T)$, the operator I_{RT} conserves the integral of $v \cdot \nu|_E$. Thus

$$\sum_{E\in\mathcal{E}(T)} \int_E I_{RT} v \cdot \nu|_E \, ds = \sum_{E\in\mathcal{E}(T)} \int_E v \cdot \nu|_E \, ds = \int_{\partial T} v \cdot \nu \, ds = \int_T \operatorname{div} v \, ds$$

where we used again the divergence theorem. We combine the above two chains of identities and divide by the area of T to obtain

$$\fint_T \operatorname{div} I_{RT} v \, dx = \fint_T \operatorname{div} v \, ds.$$

The left integrals simply equals $\operatorname{div} I_{RT} v$ because the integrand is constant on T. The assertion follows with the above representation of Π_0 as the piecewise integral mean. □

Let us now turn to the discretization of the mixed Laplacian. The mixed finite element approximation seeks $(\sigma_h, u_h) \in RT_0(\mathcal{T}) \times P_0(\mathcal{T})$ such that

$$\begin{aligned} \int_\Omega \sigma_h \cdot \tau_h \, dx - \int_\Omega \operatorname{div} \tau_h u_h \, dx &= 0 && \text{for all } \tau_h \in RT_0(\mathcal{T}), \\ \int_\Omega \operatorname{div} \sigma_h \, v_h \, dx &= -\int_\Omega f v_h \, dx && \text{for all } v_h \in P_0(\mathcal{T}). \end{aligned}$$

Theorem 2.78. *Given any $f \in L^2(\Omega)$, there is a unique solution $(\sigma_h, u_h) \in RT_0(\mathcal{T}) \times P_0(\mathcal{T})$ to the discrete mixed system. We have the error estimate*

$$\|\sigma - \sigma_h\|_{H(\operatorname{div},\Omega)} + \|u - u_h\|_{L^2(\Omega)} \leq C(\inf_{\tau_h \in RT_0(\mathcal{T})} \|\sigma - \tau_h\|_{H(\operatorname{div},\Omega)} + \inf_{v_h \in P_0(\mathcal{T})} \|u - v_h\|_{L^2(\Omega)})$$

for some constant C.

Proof. It suffices to prove the requirements from Brezzi's splitting theorem. The error estimate then follows from the abstract error estimate from Theorem 2.68. For the proof of coercivity of the form

$$a(\sigma_h, \tau_h)$$

on the kernel Z_h, we first note that any $\tau_h \in Z_h$ satisfies by definition

$$\int_\Omega \operatorname{div} \tau_h v_h \, dx = 0 \quad \text{for all } v_h \in P_0(\mathcal{T}).$$

But since $\operatorname{div} \tau_h \in P_0(\mathcal{T})$, we see that $\operatorname{div} \tau_h = 0$ pointwise in Ω. Therefore

$$a(\tau_h, \tau_h) = \|\tau_h\|^2_{L^2(\Omega)} = \|\tau_h\|^2_{L^2(\Omega)} + \|\operatorname{div} \tau_h\|^2_{L^2(\Omega)} = \|\tau_h\|^2_{H(\operatorname{div},\Omega)},$$

which implies coercivity of a in $RT_0(\mathcal{T}) \subseteq H(\operatorname{div}, \Omega)$.
Let us prove the inf-sup condition for the form

$$b(\tau_h, v_h) := \int_\Omega \operatorname{div} \tau_h \, v_h \, dx.$$

Let any $v_h \in P_0(\mathcal{T})$ be given. In case that Ω is not convex, we increase the domain to a larger convex domain $\hat{\Omega}$ by adding suitable triangles, and we extend v_h by zero to a function $\hat{f} \in L^2(\hat{\Omega})$. On $\hat{\Omega}$ we then solve the weak form of the Dirichlet problem $\Delta\hat{w} = \hat{f}$ for some $\hat{w} \in H^1_0(\hat{\Omega})$. From the H^2 regularity on convex domains (Theorem 1.55) we deduce that $\hat{w} \in H^2(\hat{\Omega})$ with

$$\|\hat{w}\|_{H^2(\Omega)} \leq C_{\text{reg}}\|v_h\|_{L^2(\Omega)}, \quad \nabla\hat{w}|_\Omega \in [H^1(\Omega)]^2, \quad \text{and} \quad \operatorname{div}\nabla\hat{w} = v_h \text{ in } \Omega.$$

Since $\nabla\hat{w}$ in H^1, its Raviart–Thomas interpolation is well defined and satisfies, due to the commuting diagram property,

$$\operatorname{div} I_{RT}\nabla\hat{w} = \Pi_0 \operatorname{div}\nabla\hat{w} = \Pi_0(v_h) = v_h.$$

We furthermore have a bound on the $H(\operatorname{div},\Omega)$ norm

$$\begin{aligned}\|I_{RT}\nabla\hat{w}\|^2_{H(\operatorname{div},\Omega)} = \|I_{RT}\nabla\hat{w}\|^2_{L^2(\Omega)} + \|v_h\|_{L^2(\Omega)} &\leq C_{I_{RT}}\|\nabla\hat{w}\|^2_{H^1(\Omega)} + \|v_h\|^2_{L^2(\Omega)}\\ &\leq (C^2_{I_{RT}}C^2_{\text{reg}} + 1)\|v_h\|^2_{L^2(\Omega)}.\end{aligned}$$

We then compute

$$\begin{aligned}\sup_{\tau_h\in RT_0(\mathcal{T})\setminus\{0\}} \frac{b(\tau,v_h)}{\|\tau\|_{H(\operatorname{div})}\|v_h\|_{L^2(\Omega)}} \geq \frac{b(I_{RT}\nabla\hat{w},v_h)}{\|I_{RT}\nabla\hat{w}\|_{H(\operatorname{div})}\|v_h\|_{L^2(\Omega)}} &= \frac{\|v_h\|^2_{L^2(\Omega)}}{\|I_{RT}\nabla\hat{w}\|_{H(\operatorname{div})}\|v_h\|_{L^2(\Omega)}}\\ &\geq \frac{1}{\sqrt{C^2_{I_{RT}}C^2_{\text{reg}} + 1}}.\end{aligned}$$

This proves the inf-sup condition with a constant that only depends on the shape regularity. □

Remark 2.79. The implementation of the Raviart–Thomas element requires a global enumeration of the edges of the triangulation.

Remark 2.80. The space $H(\operatorname{div},\Omega)$ possesses traces in some generalized sense. The trace is defined by duality. Let $v \in H^1(\Omega)$, then we can define the normal trace $\sigma\cdot\nu$ of $\sigma \in H(\operatorname{div},\Omega)$ as a linear functional acting as follows

$$\langle\sigma\cdot\nu, v\rangle := \int_\Omega \sigma\cdot\nabla v\,dx + \int_\Omega v\operatorname{div}\sigma\,dx.$$

This definition is of course inspired by the divergence theorem. Notice carefully that $\sigma\cdot\nu$ need not be an L^2 function over $\partial\Omega$, but in general only exists as a linear functional. It is therefore not possible to restrict it to some part of the boundary or to a single edge of a triangle. This is why the interpolation I_{RT} is not a well-defined operation on $H(\operatorname{div},\Omega)$.

Problem 61. Prove that the mixed form of the Poisson equation satisfies the properties of the Brezzi splitting theorem.

Problem 62. Prove that the local basis functions $\psi_{T,E}$ satisfy the property (16).

Problem 63. Write a routine (Python or pseudocode) that provides a global enumeration of all edges in a given mesh $\mathcal{T}$.

Problem 64. Implement the mixed Raviart–Thomas method for the homogeneous Dirichlet problem of the Laplacian. Use the data from earlier exercises to compute experimental rates of convergence in different norms.

Topic 3: Linear parabolic problems

§12 The heat equation and a numerical scheme *(week 26)*

So far we have seen PDEs that were depending on spatial variables in some domain Ω. We assumed uniform positive definiteness on the diffusion coefficient A (see Theorem 1.41) so that we could use arguments based on coercivity. Such partial differential operators are called *elliptic*. We now introduce an additional time variable $t \in [0,T]$ such that the PDE is posed on the space-time cylinder $\Omega \times [0,T]$. For simplicity we will focus on the *heat equation* as a prototype. It seeks a function $u : \Omega \times [0,T] \to \mathbb{R}$ such that

$$\partial_t u - \Delta u = f \quad \text{in } \Omega \times (0,T] \to \mathbb{R}, \tag{17a}$$

$$u = 0 \quad \text{on } \partial\Omega \times [0,T], \tag{17b}$$

$$u = g \quad \text{on } \Omega \times \{0\}. \tag{17c}$$

Note that $u = u(t,x)$ and that the Laplacian Δ acts with respect to the spatial variable x. Equation (17a) is called the heat equation. It describes the time evolution of heat diffusion in the domain Ω. Condition (17b) is the homogeneous Dirichlet boundary condition on $\partial\Omega$ that holds of all times $t \in [0,T]$. Finally, (17b) is an initial condition on the initial state $u(\cdot,0)$ that should equal a given function $g = g(x)$. The right-hand side $f = f(x,t)$ models (time-dependent) heat sources in the domain Ω.
Before stating a weak formulation for this problem, we give some brief remarks on integration of functions with values in some Banach space X. The construction is analogous to the usual Lebesgue integral.

Definition 3.81 (integral of X-valued functions)**.** Given a Banach space $(X, \|\cdot\|)$, a function $s : [0,T] \to X$ is called a *simple function* if it has the form form $s(t) = \sum_{j=1}^m 1_{A_j}(t) u_j$ with $u_j \in X$ and Lebesgue measurable sets $A_j \subseteq [0,T]$. A function $f : [0,T] \to X$ is said to be *strongly measurable* if it is the limit (a.e. in $[0,T]$) of a sequence of simple functions. The integral of a simple function $s(t) = \sum_{j=1}^m 1_{A_j}(t) u_j$ is defined as

$$\int_0^T s(t)dt := \sum_{j=1}^m \operatorname{meas}(A_j) u_j.$$

A strongly measurable function f is said to be *summable* if there is a sequence $(s_k)_k$ of simple functions such that

$$\int_0^T \|s_k(t) - f(t)\| dt \to 0 \quad \text{as } k \to \infty.$$

For summable f we define the integral

$$\int_0^T f(t)dt := \lim_{k\to\infty} \int_0^T s_k(t)dt.$$

This method of integration is sometimes named after *S. Bochner.* □

We now define, for $1 \leq p < \infty$,

$$L^p(0,T;X) := \{f : [0,T] \to X : f \text{ strongly measurable and } \|f\|_{L^p(0,T;X)} < \infty\}$$

where

$$\|f\|_{L^p(0,T;X)} := \left(\int_0^T \|f(t)\|^p dt\right)^{1/p}.$$

We define a weak derivative as follows.

Definition 3.82. Let $v \in L^2(0,T;X)$. A summable $g : [0,T] \to X$ is called the weak derivative of v if

$$\int_0^T \partial_t \psi(t) v(t)\, dt = -\int_0^T \psi(t) g(t)\, dt \quad \text{for all } \psi \in C_c^\infty(0,T) \text{ (scalar test functions).}$$

and we write $v' = \partial_t v = g$.

In many respects we can operate with Bochner integrals as with ordinary Lebesgue integrals, and this viewpoint will basically be sufficient for our lecture.
We have the following embedding of weakly differentiable functions.

Lemma 3.83. *Let $u \in L^1(0,T;X)$ be summable with $u' \in L^1(0,T;X)$. Then $u \in C([0,T];X)$ and we have*

$$u(t) = u(s) + \int_s^t u'(r) dr \quad \textit{for all } 0 \leq s \leq t \leq T.$$

Proof. We only sketch the idea of the proof. As in prior sections on Sobolev spaces, we can approximate u by some u_ε (through convolution) and see that $u_\varepsilon \to u$ in $L^1(0,T;X)$ as well as $u_\varepsilon \to u'$ on $L^1(s,t;X)$ for compact intervals $[s,t] \subseteq (0,T)$ as $\varepsilon \to 0$. We observe

$$u_\varepsilon(t) = u_\varepsilon(s) + \int_s^t u_\varepsilon'(r) dr$$

and pass to the limit $\varepsilon \to 0$, which shows

$$u(t) = u(s) + \int_s^t u'(r) dr$$

for almost every $0 < s < t < T$. From this representation we see that u is continuous because the integral is continuous as a function of t. □

Remark 3.84. The statement that $u \in L^1(0,T;X)$ is continuous should always be read as: There exists a continuous function in the equivalence class u.

Recall the dual space $H^{-1}(\Omega)$ of $H_0^1(\Omega)$.

Lemma 3.85. *If $u \in L^2(0,T;H_0^1(\Omega))$ with $u' \in L^2(0,T;H^{-1}(\Omega))$ be given. Then $u \in C([0,T];L^2(\Omega))$ and furthermore*

$$\partial_t \|u(t)\|^2_{L^2(\Omega)} = 2\langle u'(t), u(t)\rangle.$$

Proof. The proof again works by regularizing u and showing convergence of the regularization u_ε in $C([0,T];L^2(\Omega)$. The claimed formula for the derivative of the squared norm is easily verified for smooth functions u_ε and remains true after taking limits. The details can be found in [Eva10, §5.9]. □

Let us derive a weak formulation for (17). We interpret Δ as the weak Laplacian on $H^1_0(\Omega)$. If we assume $f \in L^2(\Omega \times [0,T])$ for the right-hand side, the solution $u(\cdot,t)$ belongs to $H^1_0(\Omega)$ for all $t \in [0,T]$. We thus have $u \in L^2([0,T];H^1_0(\Omega))$. From (17a), we see that $\partial_t u$ equals $\Delta u + f$ at all times, which is an element of $H^{-1}(\Omega)$.

Definition 3.86. Let $\Omega \subseteq \mathbb{R}^2$ be an open, bounded, connected Lipschitz polygon and $0 < T < \infty$. Let $f \in L^2(\Omega \times [0,T])$ and $g \in L^2(\Omega)$ be given. A function $u \in L^2([0,T];H^1_0(\Omega))$ with $\partial_t u \in L^2([0,T];H^{-1}(\Omega))$ is said to be a *solution to the initial/boundary-value problem* of the heat equation if it satisfies

$$\langle \partial_t u(\cdot,t), v\rangle + \int_\Omega \nabla u(\cdot,t) \cdot \nabla v \, dx = \int_\Omega f(\cdot,t) v \, dx \quad \text{for all } v \in H^1_0(\Omega) \text{ and a.e. } t \in [0,T]$$

and

$$u(\cdot,0) = g.$$

From Lemma 3.83 we see that posing a condition on $u(\cdot,0)$ is meaningful.

Numerical methods

We will later prove that there exists a unique solution to the heat equation. Let us first define a numerical method, so that we can start to do actual computations. The idea is to discretize the spatial derivatives with a finite element method. The time derivative is discretized by difference quotients. Ignoring for the moment the time discretization, we use finite elements (with respect to a triangulation $\mathcal{T}$ of Ω) in space and obtain the so-called *semidiscrete* equation: Seek $\tilde{u}_h \in S^1_0(\mathcal{T}) \times [0,T]$ such that

$$\langle \partial_t \tilde{u}_h(\cdot,t), v_h\rangle + \int_\Omega \nabla \tilde{u}_h(\cdot,t) \cdot \nabla v_h \, dx = \int_\Omega f(\cdot,t) v_h \, dx \quad \text{for all } v_h \in S^1_0(\mathcal{T}_h) \text{ and a.e. } t \in [0,T] \tag{18}$$

and

$$\tilde{u}_h(\cdot,0) = g_h$$

where g_h is a suitable approximation to g, for instance the L^2 projection to $S^1_0(\mathcal{T})$ or some (quasi-)interpolation.

The equation is called semidiscrete because the dependence on time has not been resolved by a numerical method, yet. In order to obtain an actual numerical method, we need to discretize the semidiscrete problem (18) in time. To this end, we approximate the time derivative by difference quotients.

Definition 3.87. Given a time step size Δt and a sequence $(U_j)_{j=0,\dots,J}$ of elements of some vector space, we define

$$\partial_t^+ U_j := \frac{U_{j+1} - U_j}{\Delta t}, \quad (j = 0, \dots, J-1) \quad \textit{(forward difference quotient)}$$

and

$$\partial_t^- U_j := \frac{U_j - U_{j-1}}{\Delta t}, \quad (j = 1, \dots, J) \quad \textit{(backward difference quotient).}$$

Let now the interval $[0, T]$ be uniformly subdivided by the time step size $\Delta t = T/J$ as

$$t_0 = 0,\ t_1 = \Delta t,\ \dots,\ t_j = j\Delta t,\ \dots,\ t_J = T.$$

By Taylor expansion one derives the following approximation property.

Lemma 3.88. *Given $u \in C^2([0,T])$, we have for ∂_t^+ and ∂_t^- that*

$$|\partial_t^\pm u(t_j) - \partial_t u(t_j)| \leq \frac{\Delta t}{2} \|\partial_{tt}^2 u\|_{C([0,T])}.$$

Proof. Problem 66. □

We introduce a uniform time step size $\Delta t = 1/J$. If we replace the time derivative in (18) by the backward difference quotient ∂_t^-, we arrive at fully discrete problem. We denote by $(u_h^k)_{k=0}^J$ the sequence of spatial unknowns in $S_0^1(\mathcal{T})$ and obtain the equations

$$\langle \partial_t^- u_h^k, v_h \rangle + \int_\Omega \nabla u_h^k \cdot \nabla v_h \, dx = \int_\Omega f(\cdot, t_k) v_h \, dx \quad \text{for all } v_h \in S_0^1(\mathcal{T}) \text{ and all } k = 1, \dots, J. \tag{19}$$

The initial condition is $u_h^0 = g_h$ where g_h is some approximation to g in $S_0^1(\mathcal{T})$. In matrix-vector notation for the coefficient vector x^k of u_h^k this reads as

$$M^\top \partial_t^- x^k + A^\top x^k = b_k$$

where M is the mass matrix, A is the stiffness matrix, and b_k is the right-hand side vector for f at t_k. We use the definition of ∂_t^- and the fact that the x^0 is known and obtain the following numerical scheme

$$(M^\top + \Delta t A^\top) x_k = \Delta t b_k + M^\top x^{k-1} \quad \text{for } j = 1, \dots, J.$$

This method is called the *implicit Euler method* or *backward Euler method.* We note that in each time step we have to solve a linear system.

We can perform an analogous derivation for the forward difference quotient ∂_t^+ and obtain the following scheme

$$M^\top x_k = \Delta t b_k + (M^\top - \Delta t A^\top) x_{k-1} \quad \text{for } j = 1, \dots, J$$

called the *explicit Euler method* or *forward Euler method.* Here, a system with the mass matrix M has to be solved in each step. It turns out that this can be done much more efficiently in comparison with the implicit Euler method. The reason is that the mass matrix can be suitably approximated by the diagonal matrix $\tilde{M}$ whose entries M_{jj} equals the row sum of M. This procedure is referred to as *mass lumping*. Each time step in the lumped scheme is then very cheap because no linear system has to be solved. This explains why the scheme is called *explicit*. It will, however, turn out that the time step size needs to be chosen much smaller for the explicit method than for the implicit method in order to obtain reasonable approximations.

Problem 65. Prove that the L^2 inner product on finite element spaces is represented by the *mass matrix*

$$M_{jk} = \int_\Omega \varphi_j \varphi_k \, dx.$$

Prove that the local mass matrix is given by

$$M_{jk} = \frac{|T|}{12} \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$$

(see also Problem 23). Implement an assembling routine for M in Python.

Problem 66. Prove the approximation properties of the difference quotients stated in Lemma 3.88.

Problem 67.

- Implement the backward and the forward Euler method for the heat equation. Approximate g by the finite element interpolation $g_h = I_h g$.
- Let the initial value $u_0(x) = \sin(\pi x_1)\sin(\pi x_2)$ on the unit square $\Omega = (0,1)^2$ be given. Prove that the solution to the heat equation with $f = 0$ is given by

  $$u(t,x) = \sin(\pi x_1)\sin(\pi x_2)\exp(-2\pi^2 t).$$

- Take these data (and $T = 1$) and compute experimental convergence rates with respect to h and Δt. Use the following two choices for the norm:

  $$\max_{k=0,\dots,N} \|\nabla u - \nabla u_h^k\|_{L^2(\Omega)}$$

  and

  $$\sqrt{\int_\Omega \|\nabla u(t) - \nabla u_h(t)\|^2_{L^2(\Omega)} \, dt}.$$

  The last integral can be approximated by the midpoint rule in space and the Simpson rule in time.

  Remark: We interpret $u_h(t)$ to be piecewise affine in time (a polygonal line through the points u_h^k).

§13 Error analysis *(week 27)*

We now perform an error analysis for the implicit Euler method. The proof follows a general guideline for the error analysis of any time-stepping scheme. The two building blocks are:

- *Stability:* The sequence of discrete approximations stays bounded (uniformly in Δt) in norms that we expect to be bounded for the exact solution. This is a reasonable requirement for convergence with respect to these norms.
- *Consistency:* Usually, the exact (or semidiscrete) solution will not satisfy the recursion rule of the numerical method. Consistency means that the resulting error terms converge to zero for $\Delta t \to 0$.

Lemma 3.89 (stability)**.** *The iterates u_h^k of the backward Euler scheme satisfy*

$$\max_{k=1,\dots,J} \|u_h^k\|_{L^2(\Omega)}^2 + \Delta t \sum_{k=1}^{J} \|\nabla u_h^k\|_{L^2(\Omega)}^2 \leq 2\|u_h^0\|_{L^2(\Omega)}^2 + 2\Delta t \sum_{k=1}^{J} \|f(\cdot, t_k)\|_{H^{-1}(\Omega)}^2.$$

Proof. From the definition of ∂_t^- we obtain the following identity

$$u_h^k = \frac{1}{2}(u_h^k + u_h^{k-1}) + \frac{1}{2}\Delta t \partial_t^- u_h^k.$$

A straightforward computation then yields

$$\begin{aligned}\langle \partial_t^- u_h^k, u_h^k\rangle = \int_\Omega \partial_t^- u_h^k u_h^k \,dx &= \int_\Omega \partial_t^- u_h^k(\frac{1}{2}(u_h^k + u_h^{k-1}))\,dx + \int_\Omega \partial_t^- u_h^k(\frac{1}{2}\Delta t \partial_t^- u_h^k)\,dx \\ &= \frac{1}{2\Delta t}(\|u_h^k\|_{L^2(\Omega)}^2 - \|u_h^{k-1}\|_{L^2(\Omega)}^2) + \frac{1}{2}\Delta t \|\partial_t^- u_h^k\|_{L^2(\Omega)}^2.\end{aligned}$$

We use $v_h := u_h^k$ as a test function at step k of (19) and obtain

$$\begin{aligned}\frac{1}{2\Delta t}(\|u_h^k\|_{L^2(\Omega)}^2 - \|u_h^{k-1}\|_{L^2(\Omega)}^2) + \frac{1}{2}\Delta t \|\partial_t^- u_h^k\|_{L^2(\Omega)}^2 + \|\nabla u_h^k\|_{L^2(\Omega)}^2 \\ = \int_\Omega f(\cdot, t_k) u_h^k \,dx \leq \|f(\cdot, t_k)\|_{H^{-1}(\Omega)} \|\nabla u_h^k\|_{L^2(\Omega)} \\ \leq \frac{1}{2}\|f(\cdot, t_k)\|_{H^{-1}(\Omega)}^2 + \frac{1}{2}\|\nabla u_h^k\|_{L^2(\Omega)}^2.\end{aligned}$$

After re-arranging the gradient terms and estimating the norm of ∂_t^- from below by 0, multiplication of the estimate by $2\Delta t$ and summation over k results in

$$\sum_{k=1}^{K}(\|u_h^k\|_{L^2(\Omega)}^2 - \|u_h^{k-1}\|_{L^2(\Omega)}^2) + \Delta t \sum_{k=1}^{K} \|\nabla u_h^k\|_{L^2(\Omega)}^2 \leq \Delta t \sum_{k=1}^{K} \|f(\cdot, t_k)\|_{H^{-1}(\Omega)}^2$$

for any $K \leq J$. The first term is a telescoping sum and equals $(\|u_h^K\|_{L^2(\Omega)}^2 - \|u_h^0\|_{L^2(\Omega)}^2)$. Increasing the right-hand side by replacing K by J, we thus see that

$$\max_{k=1,\dots,J} \|u_h^k\|_{L^2(\Omega)}^2 \leq \|u_h^0\|_{L^2(\Omega)}^2 + \Delta t \sum_{k=1}^{J} \|f(\cdot, t_k)\|_{H^{-1}(\Omega)}^2.$$

The combination with the foregoing estimate then implies the assertion. □

The solution of the spatial finite element method defines a map $G_h : H_0^1(\Omega) \to S_0^1(\mathcal{T})$, called the *Galerkin projection.* The Galerkin orthogonality reads

$$\int_\Omega \nabla(u - G_h u) \cdot \nabla v_h \, dx = 0 \quad \text{for all } v_h \in S_0^1(\mathcal{T}).$$

On convex domains, we have from elliptic regularity (Theorem 1.55 and Corollary 1.54) that

$$\|\nabla(u - G_h u)\|_{L^2(\Omega)} \leq Ch \|D^2 u\|_{L^2(\Omega)}. \tag{20a}$$

Furthermore, Theorem 1.56 implies

$$\|u - G_h u\|_{L^2(\Omega)} \leq Ch^2 \|D^2 u\|_{L^2(\Omega)}. \tag{20b}$$

Lemma 3.90 (consistency)**.** *Assume that the domain Ω is convex and that the solution u to the heat equation additionally satisfies*

$$u \in C^1([0,T]; H^2(\Omega)) \ \cap \ C^2([0,T]; L^2(\Omega)).$$

Then, the Galerkin projection $z_h^k := G_h u(\cdot, t_k)$ of the exact solution $u(\cdot, t_k)$ at t_k satisfies, for all $k = 1, \dots, J$,

$$\langle \partial_t^- z_h^k, v_h \rangle + \int_\Omega \nabla z_h^k \cdot \nabla v_h \, dx = \int_\Omega f(\cdot, t_k) v_h \, dx + \mathcal{C}_k(v_h)$$

where the $\mathcal{C}_k \in H^{-1}(\Omega)$ are linear functionals satisfying

$$\Delta t \sum_{k=1}^J \|\mathcal{C}_k\|_{H^{-1}(\Omega)}^2 \leq C(h^4 \int_0^T \|D^2 \partial_t u(\cdot, s)\|_{L^2(\Omega)}^2 \, ds + (\Delta t)^2 \int_0^T \|\partial_{tt} u(\cdot, s)\|_{L^2(\Omega)}^2 \, ds).$$

Proof. We have the Galerkin orthogonality

$$\int_\Omega \nabla(z_h^k - u(\cdot, t_k)) \cdot \nabla v_h \, dx = 0.$$

This and the solution property of $u(\cdot, t_k)$ show

$$\begin{aligned}
&\langle \partial_t^- z_h^k, v_h \rangle + \int_\Omega \nabla z_h^k \cdot \nabla v_h \, dx \\
&= \langle \partial_t^- z_h^k - \partial_t u(\cdot, t_k), v_h \rangle + \langle \partial_t u(\cdot, t_k), v_h \rangle + \int_\Omega \nabla u(\cdot, t_k) \cdot \nabla v_h \, dx \\
&= \mathcal{C}_k + \int_\Omega f(\cdot, t_k) v_h \, dx
\end{aligned}$$

where $\mathcal{C}_k$ is defined by

$$\mathcal{C}_k(v) := \int_\Omega (\partial_t^- z_h^k - \partial_t u(\cdot, t_k)) v dx \quad \text{for any } v \in H_0^1(\Omega).$$

To estimate the H^{-1} norm of $\mathcal{C}_k$, we split the consistency term on the right-hand side as follows

$$\mathcal{C}_k(v) = \int_\Omega \partial_t^-(z_h^k - u(\cdot, t_k))v dx + \int_\Omega (\partial_t^- u(\cdot, t_k) - \partial_t u(\cdot, t_k))v dx. \tag{21}$$

By the fundamental theorem of calculus (see also Lemma 3.83) the first term on the right-hand side of (21) equals

$$\begin{aligned}\int_\Omega \partial_t^-(z_h^k - u(\cdot, t_k))v dx &= \frac{1}{\Delta t}\int_{t_{k-1}}^{t_k}\int_\Omega (G_h\partial_t u(\cdot, s) - \partial_t u(\cdot, s))v \, dx \, ds \\ &\leq \frac{1}{\Delta t}\int_{t_{k-1}}^{t_k} \|G_h\partial_t u(\cdot, s) - \partial_t u(\cdot, s)\|_{L^2(\Omega)} \, ds \|v\|_{L^2(\Omega)}.\end{aligned}$$

Thus we obtain with the error bound (20a) on the Galerkin projection and Hölder's inequality that

$$\begin{aligned}\int_\Omega \partial_t^-(z_h^k - u(\cdot, t_k))v dx &\leq C\frac{h^2}{\Delta t}\int_{t_{k-1}}^{t_k} \|D^2\partial_t u(\cdot, s)\|_{L^2(\Omega)} \, ds \|v\|_{L^2(\Omega)} \\ &\leq C\frac{h^2}{\sqrt{\Delta t}}\sqrt{\int_{t_{k-1}}^{t_k} \|D^2\partial_t u(\cdot, s)\|^2_{L^2(\Omega)} \, ds}\|v\|_{L^2(\Omega)}.\end{aligned}$$

For the difference in the second term on the right-hand side of (21), we obtain through Taylor's formula

$$\begin{aligned}\partial_t^- u(\cdot, t_k) - \partial_t u(\cdot, t_k) &= \frac{1}{\Delta t}(u(\cdot, t_k) - u(\cdot, t_{k-1})) - \partial_t u(\cdot, t_k) \\ &= -\frac{1}{\Delta t}\int_{t_{k-1}}^{t_k} (s - t_{k-1})\partial_{tt} u(\cdot, s) \, ds.\end{aligned}$$

Thus, with Hölder's inequality,

$$\begin{aligned}\int_\Omega (\partial_t^- u(\cdot, t_k) - \partial_t u(\cdot, t_k))v_h dx &= -\frac{1}{\Delta t}\int_{t_{k-1}}^{t_k} (s - t_{k-1})\int_\Omega \partial_{tt} u(\cdot, s)v_h \, dx \, ds \\ &\leq \int_{t_{k-1}}^{t_k} \|\partial_{tt} u(\cdot, s)\|_{L^2(\Omega)} \, ds \|v_h\|_{L^2(\Omega)} \\ &\leq \sqrt{\Delta t}\sqrt{\int_{t_{k-1}}^{t_k} \|\partial_{tt} u(\cdot, s)\|^2_{L^2(\Omega)} \, ds}\|v_h\|_{L^2(\Omega)}.\end{aligned}$$

Combining the foregoing estimates with (21) results in

$$\mathcal{C}_k(v_h) \leq \left(C\frac{h^2}{\sqrt{\Delta t}}\sqrt{\int_{t_{k-1}}^{t_k} \|D^2\partial_t u(\cdot, s)\|^2_{L^2(\Omega)} \, ds} + \sqrt{\Delta t}\sqrt{\int_{t_{k-1}}^{t_k} \|\partial_{tt} u(\cdot, s)\|^2_{L^2(\Omega)} \, ds}\right)\|v_h\|_{L^2(\Omega)}.$$

This implies a bound on $\|\mathcal{C}_k\|_{H^{-1}}$. Taking squares and summing over k implies the stated bound on the sum of $\|\mathcal{C}_k\|^2_{H^{-1}}$. □

Theorem 3.91 (error estimate for the implicit Euler method)**.** *Under the assumption of Lemma 3.90, the implicit Euler method with initial value $u_h^0 = I_h g$ (nodal interpolation) has the approximation order*

$$\sqrt{\Delta t \sum_{k=1}^{J} \|\nabla(u(\cdot,t_k) - u_h^k)\|_{L^2(\Omega)}^2} \leq \mathcal{O}(h + \Delta t)$$

and

$$\max_{k=1,\dots,J} \|u(\cdot,t_k) - u_h^k\|_{L^2(\Omega)} \leq \mathcal{O}(h^2 + \Delta t).$$

Proof. We introduce the Galerkin projections $z_k := G_h u(\cdot,t_k)$ and split the error as follows

$$\Delta t \sum_{k=1}^{J} \|\nabla(u(\cdot,t_k) - u_h^k)\|_{L^2(\Omega)}^2 \leq 2\Delta t \sum_{k=1}^{J} \|\nabla(u(\cdot,t_k) - z_k)\|_{L^2(\Omega)}^2 + 2\Delta t \sum_{k=1}^{J} \|\nabla(z_k - u_h^k)\|_{L^2(\Omega)}^2.$$

The first term is estimated by the approximation property (20a) of the Galerkin projection and $\Delta T = 1/J$ as follows

$$2\Delta t \sum_{k=1}^{J} \|\nabla(u(\cdot,t_k) - z_k)\|_{L^2(\Omega)}^2 \leq Ch^2 \Delta t \sum_{k=1}^{J} \|D^2 u(\cdot,t_k)\|_{L^2(\Omega)}^2 \leq Ch^2 \|u\|_{C^0([0,T];H^2(\Omega))}.$$

For the second term, observe that Lemma 3.90 implies that $z_k - u_h^k$ is the sequence of an implicit Euler scheme with right-hand side $\mathcal{C}_k$. Therefore, the stability of Lemma 3.89 shows

$$\begin{aligned}\Delta t \sum_{k=1}^{J} \|\nabla u_h^k - z_k\|_{L^2(\Omega)}^2 &\leq 2\|I_h g - z_0\|_{L^2(\Omega)} + 2\Delta t \sum_{k=1}^{J} \|\mathcal{C}_k\|_{H^{-1}(\Omega)} \\ &\leq 2\|g - G_h g\|_{L^2(\Omega)} \\ &\quad + Ch^4 \int_0^T \|D^2 \partial_t u(\cdot,s)\|_{L^2(\Omega)}^2 \, ds + C(\Delta t)^2 \int_0^T \|\partial_{tt} u(\cdot,s)\|_{L^2(\Omega)}^2 \, ds.\end{aligned}$$

Since we have

$$\|I_h g - z_0\|_{L^2(\Omega)} \leq \|I_h g - g\|_{L^2(\Omega)} + \|g - z_0\|_{L^2(\Omega)} \leq Ch^2 \|D^2 g\|_{L^2(\Omega)}$$

(note that $g \in H^2(\Omega)$), the error estimate for the norm involving the gradient is established. The error estimate for the maximal L^2 error is shown analogously and left as an exercise. □

We have seen in Theorem 3.91 that the choice $h \approx \Delta t$ yields a balanced error bound for the discrete L^2-H^1 norm. This is the case for the implicit Euler method. The explicit Euler method satisfies a similar error bound under more restrictive assumptions. A stability analysis of the explicit Euler method shows that stability is achieved under the additional condition

$$\frac{\Delta t}{h^2} \leq c$$

for some global constant c. This means that we have to chose the time step much smaller, namely of order $(\Delta t)^2$, which is not rewarded by the error estimate. In spite of the low computational costs in each time step, this makes the explicit method rather unattractive. In contrast, the implicit scheme is *unconditionally stable.*

Problem 68. Prove the error estimate

$$\max_{k=1,\ldots,J} \|u(\cdot, t_k) - u_h^k\|_{L^2(\Omega)} \leq \mathcal{O}(h^2 + \Delta t).$$

from Theorem 3.91.

§14 Existence and uniqueness for the heat equation *(week 28)*

We now prove existence of weak solutions in a constructive procedure. We follow the following roadmap. In a first step, we discretize the PDE in space with a finite-dimensional Galerkin method. At this stage, we are not interested in actual numerical computations but rather use the Galerkin method as a tool from analysis. The space-discretized PDE can then be interpreted as a system of ordinary differential equations, which is solvable by known arguments. In a second step, we derive so-called *energy estimates* stating that certain norms of the space-discrete solutions are uniformly bounded with respect to the dimension of our Galerkin subspace. In the third step, we pass to the limit and see that the Galerkin solutions weakly converge to some limit, which is then proven to satisfy the heat equation.
In order to define Galerkin approximations, let $\mathcal{T}_0$ be a triangulation of Ω. We consider the shape-regular sequence $(\mathcal{T}_j)_j$ of triangulations resulting from j red refinements. The straight-forward space-discrete Galerkin method (already introduced in (18)) is to find $u_j \in S^1_0(\mathcal{T}_j) \times [0,T]$ such that

$$\langle \partial_t u_j(\cdot,t), v_j\rangle + \int_\Omega \nabla u_j(\cdot,t)\cdot\nabla v_j\,dx = \int_\Omega f(\cdot,t)v_j\,dx \quad \text{for all } v_j \in S^1_0(\mathcal{T}_j) \text{ and all } t\in[0,T]$$

and

$$u_j(\cdot,0) = \Pi^{(j)}g.$$

Where $\Pi^{(j)}g$ is the L^2 projection of g to the finite element space $S^1_0(\mathcal{T}_j)$.

Lemma 3.92. *For each $j = 0,1,\dots$ there exists a unique (semidiscrete) Galerkin approximation u_j to the heat equation.*

Proof. Letting $x_j(t)$ denote the coefficient vector of the spatial part of u_j, we see that the Galerkin equation is equivalent to

$$M^\top \partial_t x_j(t) + A^\top x_j(t) = b(t)$$

where M is the mass matrix, A is the stiffness matrix, and $b(t)$ is the right-hand side vector. This is a linear ODE system which is complemented by the initial condition $x_j(0) = y_j$ where y_j are the coefficients of $\Pi^{(j)}g$. Thus, it is uniquely solvable by standard results on ODEs. □

We now turn to the announced energy estimates.

Theorem 3.93 (energy estimates)**.** *There exists a constant $C > 0$ (independent of f, g, j, u_j) such that, for all $j = 0,1,\dots,$*

$$\max_{0\le t\le T}\|u_j(\cdot,t)\|_{L^2(\Omega)} + \|\nabla u_j\|_{L^2([0,T];L^2(\Omega))} + \|\partial_t u\|_{L^2([0,T];H^{-1}(\Omega))}$$
$$\le C(\|f\|_{L^2([0,T];L^2(\Omega))} + \|g\|_{L^2(\Omega)}.)$$

Proof. An application of the chain rule shows the identity

$$\langle \partial_t u_j, u_j \rangle = \partial_t(\frac{1}{2}\|u_j\|^2_{L^2(\Omega)}).$$

We use the test function $v_j = u_j(t)$ in the Galerkin equation, multiply by 2 and obtain with the Cauchy and Young inequalities

$$\partial_t(\|u_j\|^2_{L^2(\Omega)}) + 2\|\nabla u_j\|^2_{L^2(\Omega)} = 2\int_\Omega f u_j \, dx \le \|f\|^2_{L^2(\Omega)} + \|u_j\|^2_{L^2(\Omega)} \tag{22}$$

for almost every $t \in [0, T]$. This is a differential inequality bounding the growth of $\|u_j(\cdot, t)\|^2_{L^2(\Omega)}$ by the quantity itself and $\|f(\cdot, t)\|^2_{L^2(\Omega)}$. Gronwall's lemma thus implies the bound

$$\|u_j(\cdot, t)\|^2_{L^2(\Omega)} \le \exp(t)\left(\|u_j(\cdot, 0)\|^2_{L^2(\Omega)} + \int_0^t \|f(\cdot, s)\|^2_{L^2(\Omega)} ds\right).$$

Note that $u_j(\cdot, 0)$ equals the L^2 projection of g, whence $\|u_j(\cdot, 0)\|_{L^2(\Omega)} \le \|g\|_{L^2(\Omega)}$. We then have (with $C_1 = \exp(T)$)

$$\max_{0\le t\le T} \|u_j(\cdot, t)\|^2_{L^2(\Omega)} \le C_1(\|f\|^2_{L^2([0,T];L^2(\Omega))} + \|g\|^2_{L^2(\Omega)}) \le C_1(\|f\|_{L^2([0,T];L^2(\Omega))} + \|g\|_{L^2(\Omega)})^2$$

and thus the bound for the first term on the left-hand side of the assertion.
We now integrate (22) with respect to time and deduce

$$\frac{1}{2}\|\nabla u_j\|^2_{L^2([0,T];\Omega)} \le \|f\|^2_{L^2([0,T];\Omega)} + \int_0^T \|u_j(\cdot, s)\|^2_{L^2(\Omega)} ds + \|u_j(\cdot, 0)\|^2_{L^2(\Omega)} - \|u_j(\cdot, T)\|^2_{L^2(\Omega)}.$$

With the bound on the maximum on $\|u_j(\cdot, s)\|_{L^2(\Omega)}$ just shown we thus find

$$\begin{aligned}\frac{1}{2}\|\nabla u_j\|^2_{L^2([0,T];\Omega)} &\le \|f\|^2_{L^2([0,T];\Omega)} + (T+2)\max_{0\le t\le T} \|u_j(\cdot, t)\|^2_{L^2(\Omega)} \\ &\le (1 + C_1(T+2))(\|f\|_{L^2([0,T];\Omega)} + \|g\|_{L^2(\Omega)})^2\end{aligned}$$

which implies the bound on the second term on the left-hand side of the asserted estimate. In order to bound the third term on the left-hand side of the assertion including the negative norm, let $v \in H^1_0(\Omega)$ with $\|\nabla v\|_{L^2(\Omega)} = 1$ be arbitrary. We denote by $\Pi^{(j)}v \in S^1_0(\mathcal{T}_j)$ the L^2 projection of v to the finite element space and use it to test the Galerkin equation. We obtain

$$\langle \partial_t u_j(\cdot, t), \Pi^{(j)}v\rangle + \int_\Omega \nabla u_j(\cdot, t) \cdot \nabla \Pi^{(j)} v \, dx = \int_\Omega f(\cdot, t)\Pi^{(j)} v \, dx.$$

Note that the term with angular brackets on the left-hand side is nothing but the L^2 product because the solution to the ODE system is the FEM function with coefficients $\partial_t x_j$. In the L^2 inner products (without gradients) in the above identity, the projection property of the

L^2 projection shows that we can replace the function $\Pi^{(j)}v$ by v. After rearranging terms we therefore have

$$\langle \partial_t u_j(\cdot,t), v\rangle = \int_\Omega f(\cdot,t)\Pi^{(j)}v\,dx - \int_\Omega \nabla u_j(\cdot,t)\cdot\nabla\Pi^{(j)}v\,dx.$$

With the Cauchy-Schwarz inequality and the nonexpansivity of $\Pi^{(j)}$ with respect to the L^2 norm we then obtain the bound

$$|\langle \partial_t u_j, v\rangle| \le \|f\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + \|\nabla u_j\|_{L^2(\Omega)}\|\nabla\Pi^{(j)}v\|_{L^2(\Omega)}.$$

We have seen in Problem 57 that the L^2 projection is H^1 stable on sequences of red refined meshes so that there is some constant C_3 with $\|\nabla\Pi^{(j)}v\|_{L^2(\Omega)} \le C_3\|\nabla v\|_{L^2(\Omega}$. We use this bound and the Friedrichs inequality for $\|v\|_{L^2(\Omega)}$ in the above estimate and get

$$|\langle \partial_t u_j, v\rangle| \le C_4(\|f\|_{L^2(\Omega)} + \|\nabla u_j\|_{L^2(\Omega)})$$

for some constant C_4 because $\|\nabla v\|_{L^2(\Omega)}$. Taking the supremum over such v and integrating we obtain

$$\int_0^T \|\partial_t u_j\|_{H^{-1}(\Omega)} \le C_4(\|f\|^2_{L^2([0,T];\Omega)} + \|\nabla u_j\|^2_{L^2([0,T];\Omega)}.).$$

We can now use the (already established) bound on $\|\nabla u_j\|_{L^2([0,T];\Omega)}$ to prove the estimate on the third term on the left-hand side of the asserted inequality. □

Theorem 3.94 (existence of a weak solution)**.** *There exists a weak solution to the initial/boundary value problem of the heat equation.*

Proof. The bounds from the energy estimates show that the sequence $(u_j)_j$ of Galerkin solutions is bounded in $L^2([0,T];H^1_0(\Omega))$ and $(\partial_t u_j)_j$ is bounded in $L^2([0,T];H^{-1}(\Omega))$. Since these spaces are reflexive, there is a subsequence (which we do not relabel with an additional index) and some $u \in L^2([0,T];H^1_0(\Omega))$ and $v \in L^2([0,T];H^{-1}(\Omega))$ such that

$$\begin{cases} u_j \rightharpoonup u & \text{weakly in } L^2([0,T];H^1_0(\Omega)) \\ \partial_t u_j \rightharpoonup v & \text{weakly in } L^2([0,T];H^{-1}(\Omega)). \end{cases}$$

It is then an exercise (cf. the arguments in the proof of completeness of Sobolev spaces) to prove that $v = \partial_t u$. The plan of the proof is to show that u satisfies the heat equation and the initial condition. We momentarily fix an index m and choose a finite element test function $v_m \in L^2([0,T];S^1_0(\mathcal{T}_m))$. Recall that $S^1_0(\mathcal{T}_m) \subseteq S^1_0(\mathcal{T}_j)$ for any $j \ge m$ and so v_m is an admissible test function on all finer triangulations. We therefore obtain from the Galerkin equation and integration with respect to time that

$$\int_0^T \langle \partial_t u_j, v_m\rangle dt + \int_0^T\int_\Omega \nabla u_j\cdot\nabla v_m\,dxdt = \int_0^T\int_\Omega f v_m\,dxdt \quad \text{for all } j \ge m. \tag{23}$$

By the above weak convergence results, we can pass to the limit $j \to \infty$ to get

$$\int_0^T \langle \partial_t u, v_m\rangle dt + \int_0^T\int_\Omega \nabla u\cdot\nabla v_m\,dxdt = \int_0^T\int_\Omega f v_m\,dxdt. \tag{24}$$

This identity is valid for all $m \in \mathbb{N}$. Since the finite element functions on a sequence of red refined meshes are dense in $H^1_0(\Omega)$ (see Problem 58), the identity even holds for all $v \in L^2([0,T]; H^1_0(\Omega))$. In particular, the weak heat equation is satisfied for almost every t and all test functions $v \in H^1_0(\Omega)$.

We now proceed by showing that u satisfies the initial condition $u(\cdot, 0) = g$. We consider (23) with $v_m \in C^1([0,T]; S^1_0(\mathcal{T}_m))$ with $v_m(T) = 0$ being a test function for the Galerkin equation with $j \geq m$. Integration by parts (with respect to time) reveals

$$\int_0^T -\langle \partial_t v_m, u_j \rangle dt + \int_0^T \int_\Omega \nabla u_j \cdot \nabla v_m \, dxdt = \int_0^T \int_\Omega f v_m \, dxdt + \int_\Omega u_j(\cdot, 0) v_m(\cdot, 0) \, dx.$$

Letting $j \to \infty$ and observing that $u_j(\cdot, 0) = \Pi^{(j)} g \to g$ in $L^2(\Omega)$, we use the weak convergence relations and find

$$\int_0^T -\langle \partial_t v, u \rangle dt + \int_0^T \int_\Omega \nabla u \cdot \nabla v \, dxdt = \int_0^T \int_\Omega f v \, dxdt + \int_\Omega g v(\cdot, 0) \, dx$$

where we again used density of the finite element functions v_m. On the other hand, a similar argument for (24) results in

$$\int_0^T -\langle \partial_t v, u \rangle dt + \int_0^T \int_\Omega \nabla u \cdot \nabla v \, dxdt = \int_0^T \int_\Omega f v \, dxdt + \int_\Omega u(\cdot, 0) v(\cdot, 0) \, dx.$$

Comparing these two formulas then leads to $u(0) = g$ because the test function v was arbitrary. □

Theorem 3.95 (uniqueness). *The weak solution to the initial/boundary value problem of the heat equation is unique.*

Proof. The difference e of two weak solutions satisfies the heat equation with right-hand side $f = 0$ and initial values $g = 0$. We then have (cf. Lemma 3.85) that

$$\partial_t (\frac{1}{2} \|e\|^2_{L^2(\Omega)}) + \|\nabla e\|^2_{L^2(\Omega)} = \langle \partial_t e, e \rangle + \int_\Omega \nabla e \cdot \nabla e \, dx = 0$$

for almost all t. In particular, we have $\partial_t (\|e\|^2_{L^2(\Omega)}) \leq 0$ and the norm of e is nonincreasing. The initial condition thus shows that $e = 0$. □

References

[Bar16] Sören Bartels. *Numerical approximation of partial differential equations.*, volume 64 of *Texts Appl. Math.* Springer, Cham, 2016. Online im Netz der Uni Jena verfügbar.

[Bra13] Dietrich Braess. *Finite Elemente.* Texts Appl. Math. Springer, 5. edition, 2013. Online im Netz der Uni Jena verfügbar.

[Dob10] Manfred Dobrowolski. *Angewandte Funktionalanalysis. Funktionalanalysis, Sobolev-Räume und elliptische Differentialgleichungen.* Berlin: Springer, 2nd revised and extended ed. edition, 2010.

[Eva10] Lawrence C. Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics.* American Mathematical Society, Providence, RI, 2. edition, 2010.