Lectures in Finite Element Method–1
Winter semester, 2019/2020
Sofia University

Tihomir Ivanov

December 4, 2019

Contents

Chapter 1

1D Finite elements—Introduction

1.1 Piecewise polynomials in a single variable. Interpolation, L_2-projection, a priori error estimates.

In FEM, we look for the best approximation of a certain kind (piecewise polynomial) to the solution of a given differential problem. Before we consider this question, let us discuss the approximation of a given function that is known. The ideas that we shall study are fundamental for FEM and we shall meet them repeatedly during the course.

There exist two main ideas—**interpolation** and finding the **best approximation with respect to a given norm**. We shall discuss them consequently, since they are both important for what follows.

The functions that we shall use for approximation in the present course will be piecewise linear polynomials. On one hand, they are simple enough, so that we can work with them easily. On the other hand, they are sufficiently flexible to approximate functions with complex behaviour.

Let us formulate the following very general problem

Given the function $u \in V$, find the function $u_h \in V_h$, a piecewise polynomial, that is "close" to u.

Let for the time being think of the function u as sufficiently smooth (e.g., infinitely differentiable, $V \equiv C^\infty[a,b]$). We shall specify this question later. The formulated problem is fundamental for the approximation theory. We want to approximate a given (in some sense, complex) function u from the infinite-dimensional space V with something simpler, in an appropriately chosen finite-dimensional subspace $V_h \subset V$. The benefit of working in a finite-dimensional space is that we know the form of all functions in it. They can be represented as $\sum_{i=0}^{n} a_i \varphi_i(x)$, where $\{\varphi_0(x), \ldots, \varphi_n(x)\}$ form a given basis of the space. In other words, every function can be defined by choosing $n+1$ numbers $a_0, \ldots, a_n$ and the question for finding a specific function is reduced to finding $(a_0, \ldots, a_n) \in \mathbb{R}^{n+1}$. Taking into account the latter, we first need to clear out the structure of V_h and choose an appropriate basis to work with.

We shall present the ideas, using the space of piecewise-linear functions. All the ideas can be easily generalized for higher-degree polynomials.

1.1.1 The space of linear polynomials P_1. Interpolation.

In order to discuss the usage of piecewise-linear functions, let us first consider the space of linear polynomials

$$P_1 := \{p(x) = a_0 + a_1 x : (a_0, a_1) \in \mathbb{R}^2\}.$$

The simplest basis of P_1 is $\{1, x\}$. Nevertheless, it is not always the most convenient one. As we know, every line can be uniquely determined by two arbitrary points. Let us, e.g., choose the points (x_0, q_0) and (x_1, q_1). The polynoimal, passing through those points satisfies the linear algebraic system

$$\begin{aligned} p(x_0) = a_0 + a_1 x_0 = q_0, \\ p(x_1) = a_0 + a_1 x_1 = q_1 \end{aligned}$$

or, written in a vector-matrix form,

$$\begin{bmatrix} 1 & x_0 \\ 1 & x_1 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} q_0 \\ q_1 \end{bmatrix}.$$

The latter system has a Vandermonde matrix, which is well known to have poor properties in numerical computations, especially for higher dimensions. It would be much more convenient if we chose the basis $\{\varphi_0(x), \varphi_1(x)\}$, such that the system

$$\begin{bmatrix} \varphi_0(x_0) & \varphi_1(x_0) \\ \varphi_0(x_1) & \varphi_1(x_1) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} q_0 \\ q_1 \end{bmatrix}$$

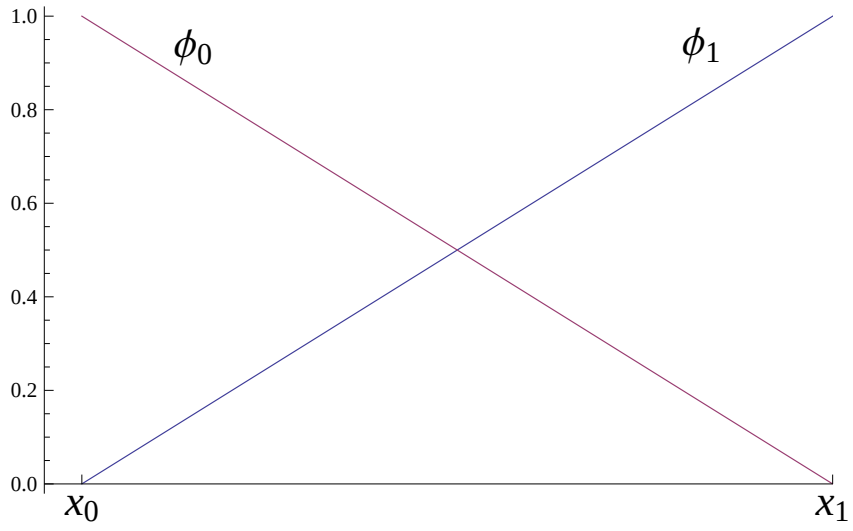
had a diagonal matrix. This gives us the reason to choose the basis functions $\varphi_i(x)$, $i = 0, 1$ in such a way that

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

We shall call such a basis an **interpolation** or **nodal basis**. In this particular case, the nodal basis is given with the Lagrange basis polynomials:

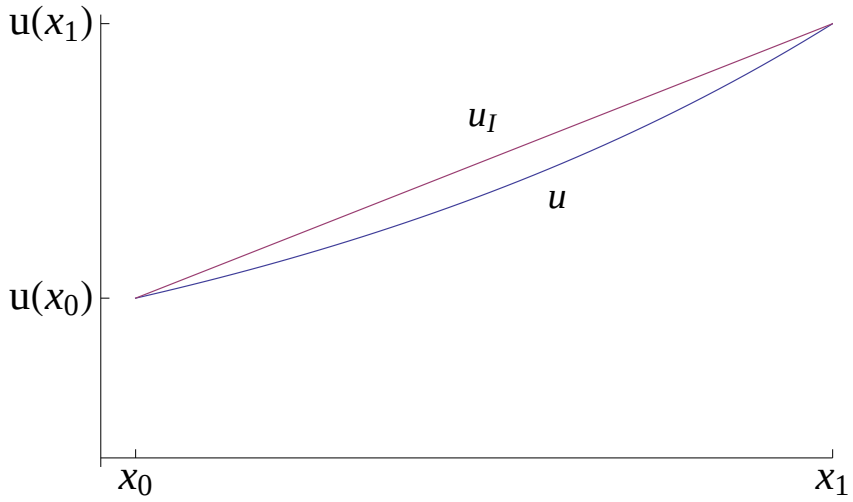
$$\varphi_0(x) = \frac{x - x_1}{x_0 - x_1}, \quad \varphi_1(x) = \frac{x - x_0}{x_1 - x_0}.$$

The graphs of the basis functions have the following form:



Then, if we have a function $u(x)$, which is given, we obtain

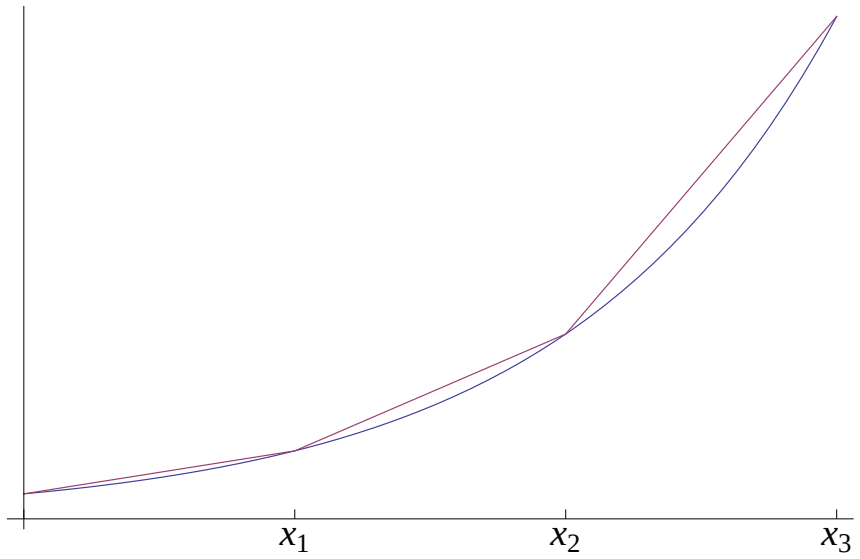
$$u_I(x) = u(x_0)\varphi_0(x) + u(x_1)\varphi_1(x).$$



1.1.2 The space of piecewise-linear polynomials V_h. Interpolation.

It is clear, however, that if want to approximate a given function with complex behaviour over a large interval, this cannot be achieved using a linear function.

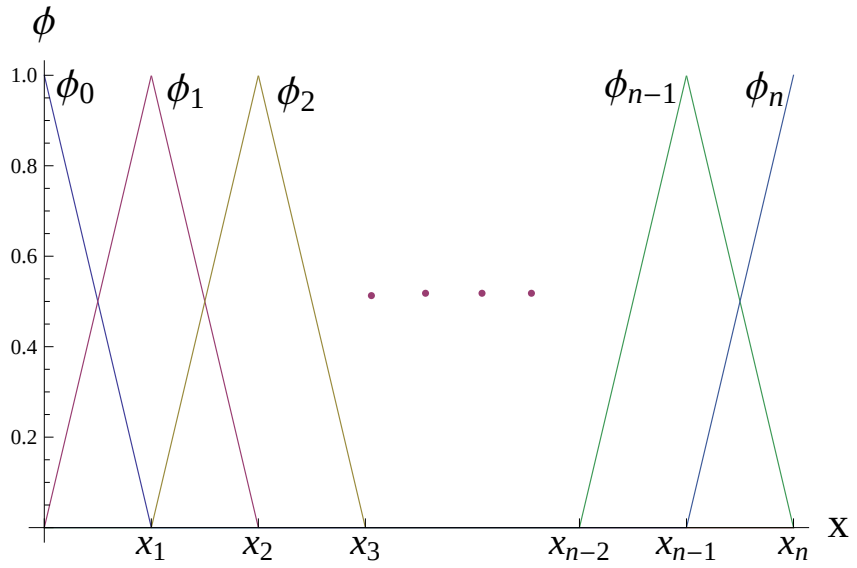
A standard idea in numerical analysis is to divide the interval into subintervals and approximate in each subinterval individually, e.g.:



This leads us to consider the space of piecewise-polynomials (in this case, linear). Let us have the nodes $x_0 < x_1 < \cdots < x_n$ chosen and let $h_i := x_i - x_{i-1}$, $I_i := [x_{i-1}, x_i]$, $i = \overline{1,n}$. We shall call the subintervals I_i **elements**. We define

$$V_h := \{p(x) \in C[a,b] : p(x) \in P_1 \text{ for } x \in I_i, i \in \overline{1,n}\}.$$

The nodal basis of this functional space must look as follows:



Those are the so called "hat"-functions. We can analytically define them as

$$\varphi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{x_i - x_{i-1}}, & x \in I_i, \\ \dfrac{x_{i+1} - x}{x_{i+1} - x_i}, & x \in I_{i+1}, \\ 0, & \textit{otherwise}. \end{cases} \tag{1.1}$$

Obviously, this definition should be trivially modified for the first and the last basis functions.

A very important property of the functions $\varphi_i(x)$ is that they have a finite support, i.e. they have non-zero values only over two elements:

$$\text{supp } \varphi_i(x) = [x_{i-1}, x_i] \cup [x_i, x_{i+1}], \quad i = \overline{1, n-1},$$

except for the first and the last basis function, whose support is the first and the last element, correspondingly.

Thus, the interpolant of the function $u(x)$ at the nodes $x_0, \ldots, x_n$ is

$$u_I(x) = \sum_{i=0}^{n} u(x_i)\varphi_i(x).$$

1.1.3 A priori error estimates for interpolation with functions from P_1

One of the main questions that we shall deal with in this course, is about obtaining a priori error estimates for a given approximation. One of the norms, most widely used for estimating the error, is the L_2-norm. It gives an idea of the "average magnitude" of a given function:

$$\|v\|_{L_2(I)} = \left(\int_I v^2 dx \right)^{1/2}.$$

We shall prove the following estimates

Proposition 1. *The interpolant $u_I \in P_1$ for the function u in the interval $I = [a, b]$ satisfies*

$$\|u - u_I\|_{L_2(I)} \leq Ch^2 \|u''\|_{L_2(I)},$$
$$\|u' - u_I'\|_{L_2(I)} \leq Ch \|u''\|_{L_2(I)}$$

for fixed constants C and $h := b - a$.

Proof. Let us denote the error by

$$e(x) := u(x) - u_I(x).$$

We want to estimate the $L_2(I)$-norm of the error, i.e

$$\sqrt{\int_I e^2(x)dx}.$$

For this purpose, we shall firstly estimate the error at an arbitrary point x. We have

$$\begin{aligned}
e(x) &= e(x_0) + \int_{x_0}^{x} e'(x)dx \\
&= \int_{x_0}^{x} e'(x)dx \\
&\leq \left(\int_{x_0}^{x} 1dx\right)^{1/2} \left(\int_{x_0}^{x} e'^2(x)dx\right)^{1/2} \quad \text{(Schwarz inequality)} \\
&= \sqrt{x - x_0} \left(\int_{x_0}^{x} e'^2(x)dx\right)^{1/2} \\
&\leq \sqrt{h} \left(\int_{x_0}^{x_n} e'^2(x)dx\right)^{1/2} \\
&= \sqrt{h} \|e'\|_{L_2(I)}.
\end{aligned}$$

Therefore,

$$e^2(x) \leq h \|e'\|^2_{L_2(I)}.$$

Integrating both sides of the latter inequality, we obtain

$$\|e\|^2_{L_2(I)} \leq h^2 \|e'\|^2_{L_2(I)},$$

i.e.

$$\|e\|_{L_2(I)} \leq h \|e'\|_{L_2(I)}.$$

We similarly derive

$$\|e'\|_{L_2(I)} \leq h \|e''\|_{L_2(I)}.$$

The difference lies in the fact that the equality $e'(x_0) = 0$ does not hold. Using the Rolle's Theorem, however, we know that there exists a point $\xi \in (x_0, x_1)$, such that $e'(\xi) = 0$. Then, we can write

$$e'(x) = \int_{\xi}^{x} e'' d\tilde{x}$$

and proceed as we did above.

Using the fact that $e'' = u'' - u_I'' = u''$, we conclude the proof. □

Remark 1. We have proved the estimates for $C = 1$. Nevertheless, we leave the constant C, since this is the general form of the estimates we shall obtain.

Remark 2. Let us note the important fact that the error estimate for $\|u - u_I\|_{L_2(I)}$ is of an order, higher with one than the estimate for the derivative of the error.

1.1.4 A priori error estimate for interpolation with functions from V_h

As we shall repeatedly see during the course, working with piecewise polynomials can be reduced directly to working with polynomials in each of the subintervals.

Proposition 2. *The interpolant $u_I \in V_h$ satisfies*

$$\|u - u_I\|_{L_2(I)} \leq Ch^2 \|u''\|_{L_2(I)},$$
$$\|u' - u_I'\|_{L_2(I)} \leq Ch \|u''\|_{L_2(I)}$$

for given constants C and $h := \max h_i$.

Proof. Obtaining the estimates is straightforward by examining the error over each subinterval and applying the previous proposition. We consecutively obtain

$$\begin{aligned}
\|u - u_i\|^2_{L_2(I)} &= \sum_{i=1}^{n} \|u - u_i\|^2_{L_2(I_i)} \\
&\leq \sum_{i=1}^{n} C_i h_i^4 \|u''\|^2_{L_2(I_i)} \\
&\leq Ch^4 \sum_{i=1}^{n} \|u''\|^2_{L_2(I_i)} \\
&= Ch^4 \|u''\|^2_{L_2(I)},
\end{aligned}$$

where $C := \max_{i=\overline{1,n}} C_i$, $h := \max_{i=\overline{1,n}} h_i$.

By taking square roots of both sides, we obtain the first inequality in the proposition. Analogously, we obtain the second one. □

Remark 3. In order to derive the estimate, we have used the crucial fact that we can compute the square of the L_2-norm over I as the sum of the squares of the L_2-norms over all elements. We specifically note this fact, since we will be using it often.

1.1.5 L_2-projection

As we shall see, the interpolation theory is fundamental in the theory of FEM. The idea of FEM itself, however, is much more related to the other main approach to approximating functions—looking for the best approximation with respect to a given norm. In particular, in this section, we shall be interested in obtaining the best approximation with respect to the L_2-norm since it can be computed algorithmically. We, therefore, formulate the following problem:

Given the function $u \in L_2(I)$, find $u_h \in V_h$, such that

$$\|u - u_h\|_{L_2(I)} \to \min_{u_h \in V_h}.$$

Let us remind that the space $L_2(I)$ is defined as

$$L_2(I) := \left\{ u : \int_I u^2 dx < \infty \right\}$$

and is equipped with the norm

$$\|u\|^2_{L_2(I)} := \int_I u^2 dx.$$

A crucial fact about the space $L_2(I)$ is that we can define the L_2-norm via the scalar product

$$(u, v) := \int_I uv dx,$$

i.e. $\|u\|^2_{L_2(I)} := (u, u)$.

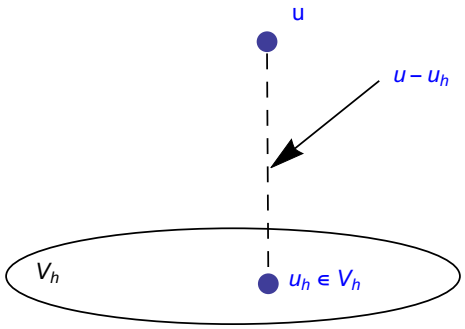
The advantage of having a scalar product, defined in the space, is that it introduces geometry. In particular, we can define the notions of angles, orthogonality, and projections. Using this fact, we can approach the question of finding the best approximation in the following way.

The natural "candidate" for being the best approximation of u from V_h is the orthogonal projection of u, i.e. the function $u_h \in V_h$, such that

$$u - u_h \perp v, \quad \forall v \in V_h,$$

or, which is the same,

$$(u - u_h, v) = 0, \quad \forall v \in V_h. \tag{1.2}$$



Best approximation result. A priori error estimate.

Before we proceed with giving an algorithm for computing u_h, we shall first prove that this is indeed the best approximation and derive an a priori error estimate.

Proposition 3 (Best approximation result)**.** *The orthogonal projection u_h is the best approximation of u from V_h with respect to the $L_2(I)$-norm, i.e.*

$$\|u - u_h\|_{L_2(I)} \leq \|u - v\|_{L_2(I)}, \quad \forall v \in V_h. \tag{1.3}$$

Proof. We have

$$\|u - u_h\|^2_{L_2(I)} = \int_I (u - v + v - u_h)(u - u_h)dx = \int_I (u - v)(u - u_h)dx.$$

In the latter equality, we have used the fact that $v - u_h \in V_h$ and $u - u_h \perp V_h$. Now, we apply the Schwarz inequality and obtain

$$\|u - u_h\|^2_{L_2(I)} \leq \|u - v\|_{L_2(I)} \|u - u_h\|_{L_2(I)}.$$

Dividing both sides to $\|u - u_h\|_{L_2(I)}$, we obtain the proposition. □

Using the latter proposition, we can easily obtain an a priori error estimate for the L_2-projection.

Proposition 4. *For the L_2-projection u_h, the following error estimates hold:*

$$\|u - u_h\|_{L_2(I)} \leq Ch^2 \|u''\|_{L_2(I)},$$
$$\|u' - u_h'\|_{L_2(I)} \leq Ch \|u''\|_{L_2(I)}.$$

Proof. The estimate (1.3) holds true for every $v \in V_h$. We can, therefore, use it with $v \equiv u_I$, for which we already have an error estimate. We obtain

$$\|u - u_h\|_{L_2(I)} \leq \|u - u_I\|_{L_2(I)} \leq Ch^2 \|u''\|_{L_2(I)}.$$

Analogously, we obtain the estimate for the derivative of the error. □

Remark 4. The orthogonality and the best approximation result (and, thus, a priori error estimate) are closely related. We shall prove that the approximation obtained with FEM is "the orthogonal projection" of the unknown function in V_h. We will obtain analogous results to the ones in the present section, but they will be with respect to the so-called energy norm or, more generally, H^1-norm. Therefore, the "natural" a priori error estimates will be with respect to those norms.

Remark 5. In order to use the best approximation result for obtaining a priori error estimates, in general, we shall need error estimates from the theory of interpolation. We shall later formulate, prove, and use a result (the Bramble–Hilbert lemma) that is much more general than the a priori error estimates for 1D linear interpolation that we derived in the previous subsection.

Computing u_h

We are now ready to develop an algorithm for computing u_h. The problem (1.2) is equivalent to the problem for finding $u_h \in V_h$, such that

$$(u_h, v) = (u, v),\ \forall v \in V_h.$$

Using the fact that we work in a finite-dimensional space, the latter will be satisfied if

$$(u_h, \varphi_i) = (u, \varphi_i),\ i = \overline{0, n}.$$

Taking into account that we know the form of $u_h = \sum_{j=0}^{n} q_j \varphi_j(x)$, where $\varphi_j(x)$ are the hat-functions, and the linearity of the scalar product, we finally obtain the following linear algebraic systems for the unknown coefficients:

$$\sum_{j=0}^{n} q_j(\varphi_j, \varphi_i) = (u, \varphi_i), \quad i = \overline{0, n}$$

or, written in a vector-matrix form,

$$\begin{bmatrix} (\varphi_0, \varphi_0) & (\varphi_0, \varphi_1) & \cdots & (\varphi_n, \varphi_0) \\ \vdots & \vdots & \ddots & \vdots \\ (\varphi_0, \varphi_n) & (\varphi_1, \varphi_n) & \cdots & (\varphi_n, \varphi_n) \end{bmatrix} \begin{bmatrix} q_0 \\ \vdots \\ q_n \end{bmatrix} = \begin{bmatrix} (u, \varphi_0) \\ \vdots \\ (u, \varphi_n) \end{bmatrix}. \tag{1.4}$$

This is, actually, the most general form of the system that would be obtained when we look for the orthogonal projection of a given function u in the subspace, spanned by $\varphi_0, \ldots, \varphi_n$, with respect to a scalar product, defined in the space. In our particular case, i.e. looking for the L_2-projection in V_h, we obtain

$$\begin{bmatrix} \int_I \varphi_0^2 dx & \int_I \varphi_1 \varphi_0 dx & \cdots & \int_I \varphi_n \varphi_0 dx \\ \vdots & \vdots & \ddots & \vdots \\ \int_I \varphi_0 \varphi_n dx & \int_I \varphi_1 \varphi_n dx & \cdots & \int_I \varphi_n^2 dx \end{bmatrix} \begin{bmatrix} q_0 \\ \vdots \\ q_n \end{bmatrix} = \begin{bmatrix} \int_I u \varphi_0 dx \\ \vdots \\ \int_I u \varphi_n dx \end{bmatrix}.$$

For short, we write the system as

$$M\mathbf{q} = \mathbf{b},$$

where M is the so called **global mass matrix** and $\mathbf{b}$ is the **global load vector**.

Therefore, in order to compute the unknown coefficients, we need to assemble the matrix M and the vector $\mathbf{b}$ and solve the system. We shall use the fact that the basis functions of V_h have finite support, in order to compute M and $\mathbf{b}$ efficiently. Since φ_i has non-zero values only over I_i and I_{i+1} it makes sense to write M as a

sum of matrices, containing integrals over each element individually, i.e.

$$M = \sum_{i=1}^{n} \begin{bmatrix} \int_{I_i} \varphi_0^2 dx & \int_{I_i} \varphi_1\varphi_0 dx & \cdots & \int_{I_i} \varphi_n\varphi_0 dx \\ \vdots & \vdots & \ddots & \vdots \\ \int_{I_i} \varphi_0\varphi_n dx & \int_{I_i} \varphi_1\varphi_n dx & \cdots & \int_{I_i} \varphi_n^2 dx \end{bmatrix}$$

$$= \begin{bmatrix} \int_{I_0} \varphi_0^2 dx & \int_{I_0} \varphi_1\varphi_0 dx & 0 & \cdots & 0 & 0 \\ \int_{I_0} \varphi_0\varphi_1 dx & \int_{I_0} \varphi_1^2 dx & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \int_{I_1} \varphi_1^2 dx & \int_{I_1} \varphi_1\varphi_2 dx & \cdots & 0 & 0 \\ 0 & \int_{I_1} \varphi_2\varphi_1 dx & \int_{I_1} \varphi_2^2 dx & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

$$+ \cdots + \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \int_{I_n} \varphi_{n-1}^2 dx & \int_{I_n} \varphi_{n-1}\varphi_n dx \\ 0 & 0 & 0 & \cdots & \int_{I_n} \varphi_n\varphi_{n-1} dx & \int_{I_n} \varphi_n^2 dx \end{bmatrix}.$$

Of course, in practice, it would be absolutely impractical to store all those matrices in the sum, having most of their elements equal to zero. Thus, we only compute the 2×2 non-zero blocks and put them "at the correct places" in M. This process is called **assembling the global mass matrix**.

The element mass 2×2 matrices can be defined as follows:

$$m_i = \begin{bmatrix} \int_{I_i} \varphi_{i-1}^2 dx & \int_{I_i} \varphi_{i-1}\varphi_i dx \\ \int_{I_i} \varphi_i\varphi_{i-1} dx & \int_{I_i} \varphi_i^2 dx \end{bmatrix} = \frac{h_i}{6} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

The last equality can be easily shown and we leave it as an exercise. We shall compute particular element mass matrices, when we discuss the solution of differential equations with FEM.

We can proceed analogously for the load vector $\mathbf{b}$, by introducing

$$\mathbf{b}_i = \begin{bmatrix} \int_{I_i} \varphi_{i-1} u dx \\ \int_{I_i} \varphi_i u dx \end{bmatrix}.$$

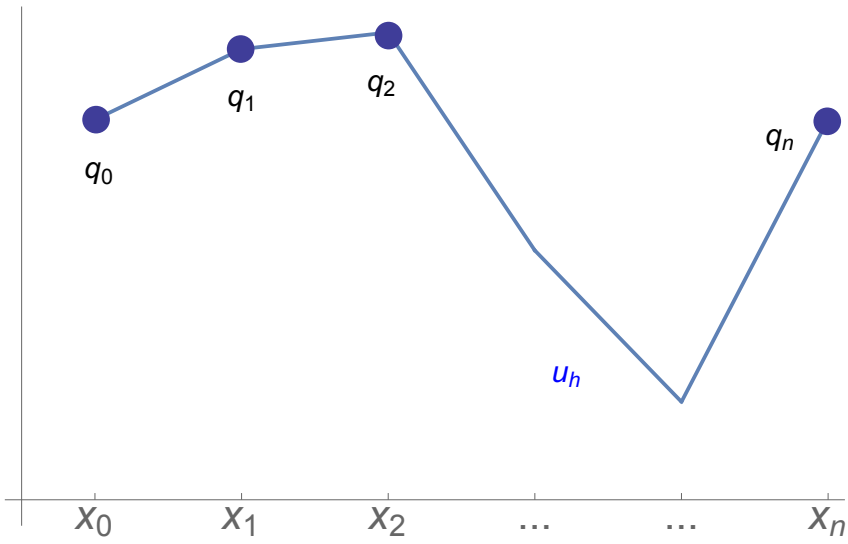
We formulate the following general algorithm for obtaining the best approximation of a given function $u \in L_2(I)$ with respect to the L_2-norm.

ALGORITHM (COMPUTING L_2-PROJECTION):

1. Discretize the interval I by introducing the mesh $x_0 < x_1 < \cdots < x_n$, $I_i := [x_{i-1}, x_i]$, $h_i := x_i - x_{i-1}$.
2. Define the nodal basis $\varphi_i(x)$, $i = \overline{0,n}$, the hat-functions.
3. Implement functions, computing the element mass matrices, M_i, and element load vectors, $\mathbf{b}_i$.
4. Assemble the global mass matrix M and load vector $\mathbf{b}$.
5. Solve the linear algebraic system $M\mathbf{q} = \mathbf{b}$.
6. Using the solution $\mathbf{q}$, obtain

$$u_h = \sum_{i=0}^{n} q_i \varphi_i(x).$$

Let us remark that $q_i = u_h(x_i)$ **are exactly the values of u_h at the nodes**.



In order to get further intuition about what we have discussed so far, before you proceed, see the worked-out examples in Section 1 of the notes from the exercise classes.

1.2 FEM for 1D problems with homogeneous Dirichlet boundary conditions. Variational formulation. Ritz–Galerkin method. A priori error estimate in energy norm. Discretization and derivation of a linear algebraic system. General boundary conditions.

1.2.1 General idea of the method.

We are now ready to apply the ideas, introduced in the previous section, to solving differential equations. We shall present the ideas over a simple example. We consider the differential problem

$$\begin{aligned} -u''(x) &= f, \quad x \in (0, L), \\ u(0) &= u(L) = 0, \end{aligned} \tag{D}$$

where $f \in L_2(I)$ is a given function, $I := [0, L]$. Stated otherwise, we are looking for a function u from the space

$$\mathcal{D} = \{v \in C^2(0, L) \cap C[0, L] : v(0) = v(L) = 0\}$$

that satisfies the differential equation. We shall call such a function a **strong solution** or a **classical solution** of the differential problem (D).

Our approach will be to obtain an integral problem that has the same solution, since working with integrals has certain benefits in complex geometries, discretized by using arbitrarily-shaped objects (e.g., triangles). **In those cases the classical finite-difference methods have serious drawbacks. On the other hand, it makes no difference whether we compute an integral over rectangles (from a rectangular mesh), or over other domains.** We can accomplish this by using the idea we had for the L_2-projection. That is, we shall require that the error (i.e., the difference between the left-hand side and the right-hand side in the differential equation, $-u'' - f$) be orthogonal to "every v". We realize this idea in the following way.

Let us take an arbitrary function v that we shall call a **test function** and take the scalar products of both sides of the differential equation with v. We obtain

$$(-u'', v) = (f, v)$$

or, which is the same,

$$\int_I -u''v dx = \int_I fv dx.$$

Let us consider the integral on the left-hand side in the latter equation and use integration by parts, in order to accomplish two things:

- make the left-hand side "more symmetric";
- decrease the order of the derivatives in the equation and, thus, relax the requirements on the admissible functions u.

Thus, for the left hand-side we obtain

$$\int_I -u''v dx = \int_I u'v' dx - u'(1)v(1) + u'(0)v(0).$$

In order to simplify the expression further, we require that the test functions also satisfy the boundary conditions, i.e. $v(0) = v(1) = 0$. This leads to a very useful simplification, but is also perfectly natural to have the test functions v and the unknown function u be from the same space (and, thus, satisfy the same requirements).

We have, thus, obtained the following integral problem that we shall refer to as the **variational form** or the **weak form**:

Find $u \in V$, such that

$$a(u, v) = F(v), \quad \forall v \in V, \tag{V}$$

where the bilinear form $a(u, v)$ and the linear functional $F(v)$ are defined as follows:

$$a(u, v) = \int_I u'v' dx, \quad F(v) = \int_I fv dx.$$

We shall define the space V to be the largest functional space, for which this problem makes sense, i.e.

$$V = \left\{ v : v' \text{ exists (in a weak sense)}, \int_I v^2 < \infty, \int_I v'^2 < \infty, v(0) = v(L) = 0 \right\} =: H_0^1.$$

Remark 6. Before we proceed, we need to make a couple of remarks:

- The space H_0^1 is an example of a Sobolev space. We shall discuss this question in much more detail later in the course. In particular, we shall introduce the notion of a weak (generalized) derivative;
- We have explicitly required that the functions in this space satisfy the Dirichlet boundary conditions, because we have used this fact in order to obtain the variational formulation. This is a very subtle moment and we shall have a special lecture on the subject;
- Obviously the space $V \equiv H_0^1$ is "larger" than $\mathcal{D}$. On one hand, this is good—if (D) has a solution, it is contained in H_0^1 and is, thus, a solution of (V). On the other hand, the problem (V) in general is not equivalent to (D). However, we shall prove under certain conditions that the problem (V) has a unique solution. Thus, if (D) has a solution, the two problems are indeed equivalent. Otherwise, the solution of (V) will be called a **weak solution**;
- In the variational formulation (V), we require that the unknown function u and the test functions v are from the same space $V \equiv H_0^1$ (Galerkin method). This seems the natural thing to do and we also used it as a justification for imposing the conditions $v(0) = v(1) = 0$ on the test functions. This, however, is not necessarily the case always. When the two spaces are different, we have the so called Petrov–Galerkin method. For the time being, we shall only be interested in the Galerkin method, i.e. the two spaces being equal.

Unfortunately, we cannot solve the problem (V), in general. Therefore, we shall look for an approximate solution in a finite-dimensional subspace of H_0^1 (a problem that can be solved). In particular, we shall look for a piecewise-linear function with zero values on the boundaries, i.e. we shall solve the so-called Ritz–Galerkin problem:

Find $u_h \in V_{h,0} \subset H_0^1$, such that

$$a(u_h, v) = F(v), \quad \forall v \in V_{h,0}, \tag{R.–G.}$$

where

$$V_{h,0} := \{v \in V_h : v(0) = v(L) = 0\} = span(\varphi_1, \ldots, \varphi_{n-1}).$$

Taking into account that we work in the finite-dimensional space $V_{h,0}$, it is sufficient to satisfy

$$a(u_h, \varphi_j) = F(\varphi_j), \quad j = \overline{1, n-1}.$$

Further, using the form of the approximate solution, $u_h(x) = \sum_{i=1}^{n-1} q_i \varphi_i(x)$, we obtain the following linear algebraic system for the coefficients q_i:

$$\begin{bmatrix} a(\varphi_1, \varphi_1) & a(\varphi_1, \varphi_2) & \cdots & a(\varphi_1, \varphi_{n-1}) \\ \vdots & \vdots & \ddots & \vdots \\ a(\varphi_1, \varphi_{n-1}) & a(\varphi_2, \varphi_{n-1}) & \cdots & a(\varphi_{n-1}, \varphi_{n-1}) \end{bmatrix} \begin{bmatrix} q_1 \\ \vdots \\ q_{n-1} \end{bmatrix} = \begin{bmatrix} (f, \varphi_1) \\ \vdots \\ (f, \varphi_{n-1}) \end{bmatrix}. \tag{1.5}$$

We shall write the latter system compactly as

$$M\mathbf{q} = \mathbf{f},$$

where M is the so-called **stiffness matrix**

$$M = \begin{bmatrix} \int_I \varphi_1'^2 dx & \cdots & \int_I \varphi_1' \varphi_{n-1}' dx \\ \vdots & \ddots & \vdots \\ \int_I \varphi_1' \varphi_{n-1}' dx & \cdots & \int_I \varphi_{n-1}'^2 dx \end{bmatrix}.$$

Solving the system with respect to $\mathbf{q}$, we obtain the desired approximate solution.

Before you proceed, see the worked-out examples in Section 2 of the notes from the exercise classes.

1.2.2 A priori error estimate in energy norm

In order to use the idea of substituting the problem (V) with the problem (R.–G.), we need to prove that this method is convergent, i.e. $u_h \to u$ as $h \to 0$. This is directly related to obtaining a priori error estimates for $u - u_h$.

If we compare (1.5) to (1.4), we can note that the two are very much alike. Actually, since the bilinear form $a(\cdot, \cdot)$ is symmetric and positive-definite, we can define the following scalar product and corresponding norm:

$$< u, v >_E := a(u, v) = \int_I u'v' dx, \quad \|u\|_E^2 := a(u, u) = \int_I u'^2 dx.$$

They are called energy scalar product and energy norm, respectively. Then, the matrix of the system (1.5) is a Gram matrix with respect to the energy scalar product. Two consequences follow:

- The Ritz–Galerkin problem has a unique solution;
- One can prove an orthogonality result and a best approximation result (with respect to the energy norm), from which an a priori error estimate and convergence follow.

Remark 7. In general, the bilinear form needs not be symmetric and it does not define a scalar product. We shall discuss the more general case later in the course.

Proposition 5 (Galerkin orthogonality)**.** *The finite element approximation u_h satisfies*

$$a(u - u_h, v),\ \forall v \in V_{h,0}.$$

Proof. For the solution u of the variational problem (V), we have

$$a(u, v) = F(v),\ \forall v \in H_0^1 \supset V_{h,0}.$$

For the solution u_h of the Ritz–Galerkin problem, we have

$$a(u_h, v) = F(v),\ \forall v \in V_{h,0}.$$

Therefore, subtracting the latter two equations and using the linearity of $a(\cdot, \cdot)$, we obtain

$$a(u - u_h, v) = 0,\ \forall v \in V_{h,0}.$$

□

Remark 8. The obtained result is very general. It obviously holds true for every bilinear form $a(\cdot, \cdot)$ and every functional $F(\cdot)$.

Remark 9. Since for our model problem the bilinear form $a(\cdot, \cdot)$ defines a scalar product, we have proven that u_h is the orthogonal projection of u in $V_{h,0}$ with respect to the energy scalar product $< \cdot, \cdot >_E$:

$$< u - u_h, v >_E = 0,\ \forall v \in V_{h,0}.$$

In general, when the bilinear form is not symmetric, the Galerkin "orthogonality" is not really a orthogonality with respect to some scalar product. This result, however, will still allow us to obtain a priori error estimates.

Now, as before, the best approximation result follows directly:

Proposition 6 (Best approximation result)**.** *The finite element solution u_h is the best approximation of u with respect to the energy norm, i.e.*

$$\|u - u_h\|_E \leq \|u - v\|_E,\ \forall v \in V_{h,0}.$$

Proof. We obtain consecutively

$$\begin{aligned}
\|u - u_h\|^2 &= \int_I (u - v + v - u_h)'(u - u_h)' dx \\
&= \int_I (u - v)'(u - u_h)' dx \qquad \text{(Galerkin orthogonality)} \\
&\leq \|u - v\|_E \|u - u_h\|_E \qquad \text{(Schwarz inequality).}
\end{aligned}$$

□

Proposition 7. *We are now ready to derive an a priori error estimate. We have*

$$\|u - u_h\|_E \leq \|u - u_I\|_E = \|(u - u_I)'\|_{L_2} \leq Ch\|u''\|_{L_2}.$$

Two questions immediately follow from this error estimate.

1. We have proven convergence for the derivatives, $u_h' \to u'$ as $h \to 0$. Do we have convergence for the approximate solution?

2. Can we prove second order convergence for $u_h \to u$ in L_2-norm that holds for the L_2-projection?

We can answer the first question affirmatively, since $u(0) = u_h(0) = 0$ and we have convergence for the derivatives. therefore as $h \to 0$, the following hods:

$$u_h(x) = u_h(0) + \int_0^x u_h' dx \to u(0) + \int_0^x u' dx = u(x).$$

Concerning the second question, it is not so trivial. We shall prove in a few lectures that under certain conditions (but not in the most general case) second-order convergence can be shown to hold in L_2-norm.

1.2.3 FEM in 1D with more general boundary conditions

We shall now consider one more example, with which we shall illustrate the case of Robin boundary conditions.

$$\begin{aligned} -(au')' &= f, \quad x \in (0, L), \\ au'(0) &= \varkappa_0(u(0) - g_0), \\ -au'(L) &= \varkappa_L(u(L) - g_L), \end{aligned} \tag{D}$$

where $a(x) > 0$ and $f(x)$ are given function, $\varkappa_0 > 0$, $\varkappa_L > 0$, g_0, g_L are given constant parameters.

We derive the corresponding variational problem by multiplying both sides with an arbitrary test function v and integrating over $I := [0, L]$. For the left-hand side, lhs, using integration by parts, we obtain

$$lhs = -\int_I (au')' v dx = \int au'v'dx - a(L)u'(L)v(L) + a(0)u'(0)v(0).$$

Here comes the difference with the case of Dirichlet boundary conditions. In this case, we know something about u' and, therefore, can include the boundary conditions in the variational form. We obtain

$$lhs = \int_I au'v'dx + \varkappa_L(u(L) - g_L)v(L) + \varkappa_0(u(0) - g_0)v(0).$$

By leaving on the left-hand side only the terms that include the unknown function, we obtain the following variational problem.

Find $u \in H^1$, such that

$$a(u, v) = F(v), \ \forall v \in H^1, \tag{V}$$

where

$$a(u,v) := \int au'v'dx + \varkappa_L u(L)v(L) + \varkappa_0 u(0)v(0),$$
$$F(v) := \int_I fvdx + g_0 v(0) + g_L v(L).$$

Since in the derivation of the variational form, we didn't impose any explicit conditions on the test functions v, we can work in the space H^1. As we shall discuss in more detail later in the course, the fact that we have included the boundary conditions in the variational form means that the solution will automatically satisfy them. Therefore, **if we have Dirichlet boundary conditions, we impose them in the functional space (e.g., we work in H_0^1), but if we have Neumann or Robin boundary conditions, we need not care about them and solve the problem in H^1.**

Further, we approximate the variational problem (V) by searching an approximate solution in the subspace $V_h := span(\varphi_0, \ldots, \varphi_n)$. We obtain the following problem.

Find $u_h \in V_h \subset H^1$, such that

$$a(u_h, v) = F(v), \ \forall v \in V_h. \qquad \text{(R.–G.)}$$

Rewriting the latter as a linear algebraic system with respect to the unknown coefficients in the representation $u_h = \sum_{i=0}^{n} q_i \varphi_i(x)$, we obtain

$$M\mathbf{q} = \mathbf{f},$$

where

$$M_{ij} := \int_I a\varphi_i'\varphi_j'dx + \varkappa_L\varphi_j(L)\varphi_i(L) + \varkappa_0\varphi_j(0)\varphi_i(0),$$
$$f_i := \int_I f\varphi_i(x) + \varkappa_L g_L\varphi_i(L) + \varkappa_0 g_0\varphi_i(0).$$

Let us note that the red terms give non-zero contributions only for $i = j = 0$ and $i = j = n$.

Since we have seen that the general approach is basically the same, no matter what the specific problem is, we formulate the following algorithm.

ALGORITHM (GENERAL STEPS OF 1D FEM):

1. Obtain the variational formulation of the original differential problem. Determine the functional space, where (V) is solved, carefully, keeping in mind that homogeneous Dirichlet conditions need to be explicitly included in the definition of the functional space.
2. Discretize the domain by introducing the mesh $x_0 < \cdots < x_n$ and define the finite dimensional subspace V_h.
3. Formulate the Ritz–Galerkin problem and obtain the linear algebraic system $M\mathbf{q} = \mathbf{b}$,
4. Compute M and $\mathbf{b}$. In practice, this is achieved by computing local element matrices/vectors and then assembling the global ones.
5. Solve the linear algebraic system to obtain $\mathbf{q}$.
6. The approximate solution is

$$u_h = \sum_{i=0}^{n} q_i \varphi_i(x).$$

It is easy to obtain an a priori error estimate in this case as well. Since the bilinear form $a(\cdot,\cdot)$ is again symmetric, we can define the energy scalar product and corresponding norm in the following way:

$$< u, v >_E := \int_I au'v'dx + \varkappa_L u(L)v(L) + \varkappa_0 u(0)v(0), \quad \|u\|_E^2 := \int_I au'^2dx + \varkappa_L u^2(L) + \varkappa_0 u^2(0).$$

Obviously, the Galerkin orthogonality holds for the above-defined energy scalar product and, thus, we shall proceed with establishing the best approximation result.

Proposition 8 (Best approximation result)**.** $\|u - u_h\|_E \leq \|u - v\|$ *for all* $v \in V_h$.

Proof. We have

$$\|u - u_h\|^2 = \int_I a(u - u_h)'(u - v + v - u_h)'dx + \varkappa_L(u(L) - u_h(L))(u(L) - v(L) + v(L) - u_h(L))$$
$$+ \varkappa_0(u(0) - u_h(0))(u(0) - v(0) + v(0) - u_h(0)).$$

From the Galerking orthogonality, it follows that the red terms sum up to zero and, therefore,

$$\|u - u_h\|^2 =< u - u_h, u - v >_E \leq \|u - u_h\|_E \|u - v\|_E.$$

□

Now, the a priori error estimate follows easily. We obtain consecutively

$$\begin{aligned}
\|u - u_h\|_E^2 &\leq \|u - u_I\|_E^2 \\
&= \int_I a(u - u_I)'^2dx + \cancelto{0}{\varkappa_L(u(L) - u_I(L))^2} + \cancelto{0}{\varkappa_0(u(0) - u_I(0))^2} \\
&\leq \max_I a \|(u - u_I)'\|_{L_2}^2 \\
&\leq Ch^2 \|u''\|_{L_2}^2.
\end{aligned}$$

1.3 A priori error estimates in H^1 and L_2 norms. Nitsche's trick.

1.3.1 Error estimates in H^1-norm. Coercivity and continuity of the bilinear form.

As we can see from the above discussions, it is very natural to obtain a priori error estimates in the energy norm. Unfortunately, this has some drawbacks. First, the bilinear form does not define a scalar product and, thus, a norm, if it is not symmetric. Furthermore, the energy norm is problem-specific, which means that concrete computations should be carried for each individual problem.

Instead of using the energy norm, we shall further show that error estimates in the H^1-norm can be obtained just as easily, but do not have the same drawbacks. Furthermore, they are in some sense equivalent to the estimates in energy norm. The H^1-norm is defined as

$$\|u\|^2_{H^1} := \int_I (u^2 + u'^2) dx.$$

It is a natural measure for the magnitude of an element u from H^1, taking into account that we require that the elements in H^1 satisfy

$$\int_I u^2 dx < \infty, \ \int_I u'^2 dx < \infty.$$

We shall introduce the ideas, using the first model problem, which we re-state here for convenience:

$$\begin{aligned} -u''(x) &= f, \quad x \in (0, L), \\ u(0) &= u(L) = 0. \end{aligned} \tag{D}$$

The corresponding variational problem is:
Find $u \in H^1_0$, such that

$$a(u, v) = F(v), \quad \forall v \in H^1_0, \tag{V}$$

where

$$a(u, v) = \int_I u'v' dx, \quad F(v) = \int_I fv dx.$$

It turns out that if the bilinear form is coercive in H^1, then we can obtain an error estimate in H^1-norm, directly from the error estimate in energy norm.

Definition 1. The bilinear form $a(\cdot, \cdot)$ is said to be **coercive** in the functional space V if

$$a(u, u) \geq \alpha \|u\|^2_V$$

for some constant α that is not dependent on u.

Even though we still need to check that the bilinear form in (V) is indeed coercive in H^1, let us, for the time being, assume that it is. Then, we can straightforwardly derive an error estimate in H^1-norm, using what we know for the error in energy norm. We obtain consecutively

$$\|u - u_h\|^2_{H^1} \leq Ca(u - u_h, u - u_h) = C\|u - u_h\|^2_E \leq Ch^2 \|u''\|^2_{L_2}.$$

Therefore, an equivalent a priori error estimate holds in H^1-norm:

$$\|u - u_h\|_{H^1} \leq Ch\|u''\|_{L_2}. \tag{1.6}$$

Actually, the estimate (1.6) has a further benefit. If we have convergence in H^1 norm, i.e. if we can show that

$$\|(u - u_h)\|_{H^1} \to 0,$$

then it follows that both

$$\int_I (u - u_h)^2 dx \to 0 \text{ and } \int_I (u - u_h)'^2 dx \to 0.$$

Stated otherwise, the following holds true:

$$\|u\|_{H^1}^2 = \|u\|_{L_2}^2 + \|u'\|_{L_2}^2.$$

We have, thus, already proven the following error estimate for our model problem.

Proposition 9. *For the finite element solution u_h of the variational problem (V), the following a priori error estimate in H^1-norm holds:*

$$\|u - u_h\|_{H^1} \leq Ch\|u''\|_{L_2}.$$

We, nevertheless, are interested in obtaining error estimates even in the case when the bilinear form is not symmetric and, thus, we cannot define the energy norm like we have done so far. It turns out that the coercivity allows us to directly obtain a "best approximation"[1] result. Indeed, we will prove the following.

Proposition 10. *For the finite element solution $u_h \in V_h$ of (V),*

$$\|u - u_h\|_{H^1} \leq C\|u - v\|_{H^1}, \ \forall v \in V_h,$$

holds true.

Proof. Using the coercivity of the bilinear form and the Galerkin orthogonality (that is a general result and, thus, not problem-dependent) we consecutively obtain:

$$\begin{aligned}
\|u - u_h\|_{H^1}^2 &\leq Ca(u - u_h, u - v + v - u_h) \\
&= Ca(u - v, u - u_h) \\
&= C\int_I (u - v)'(u - u_h)' dx \leq C\sqrt{\int_I (u - v)'^2 dx}\sqrt{\int_I (u - u_h)'^2 dx} \\
&\leq C\|u - v\|_{H^1}\|u - u_h\|_{H^1}.
\end{aligned}$$

In the latter inequality, we have used the obvious fact that

$$\int_I (u - v)'^2 dx \leq \int_I \left[(u - v)^2 + (u - v)'^2\right] dx.$$

□

[1] The inverted commas are used, because we won't really obtain a best approximation result, but a similar one, which will be sufficient for our purposes.

Remark 10. Let us note that in the latter proof, we have used the specific problem in the row that is coloured in red. If we change this row with the following abstract condition on the bilinear form (we shall call the bilinear forms that satisfy this condition **continuous** in H^1):

$$a(u,v) \leq C\|u\|_{H^1}\|v\|_{H^1},$$

we shall again obtain the statement of the proposition. Therefore, the statement of the proposition holds for all bilinear forms that are coercive and continuous in H^1.

1.3.2 Poincaré inequality in H_0^1

We are now ready to deal with the question of showing coercivity in H_0^1. This is obviously needed for obtaining the "good" a priori estimates, in H^1-norm. A crucial result from the theory of Sobolev spaces that will allow us to do so is the so-called Poincaré inequality.

Proposition 11 (Poincaré inequality)**.** *For every function $u \in H_0^1(I)$, the following inequality is valid:*

$$\int_0^1 u^2 dx \leq C \int_0^1 u'^2 dx.$$

Proof. We shall first estimate $u(x)$ and then square and integrate both sides. We have

$$\begin{aligned} u(x) &= \cancelto{0}{u(0)} + \int_0^x u'(x)dx \\ &\leq \sqrt{\int_0^x 1 d\tilde{x}}\sqrt{\int_0^x u'^2 d\tilde{x}} \\ &\leq \sqrt{x}\sqrt{\int_0^1 u'^2 d\tilde{x}}. \end{aligned}$$

Therefore,

$$u^2(x) \leq x \int_0^1 u'^2 dx$$

and, integrating both sides, we finally obtain

$$\int_0^1 u^2(x)dx \leq \frac{1}{2}\int_0^1 u'^2 dx.$$

□

Remark 11. Note that the Poincaré inequality holds only in H_0^1. We shall later formulate a similar result in H^1, a so-called Poincaré–Friedrichs inequality.

Remark 12. We have proven the Poincaré inequality with $C = 1/2$. One can derive a better result for the constant, $C = 1/6$ if the right boundary condition is also used (see the second set of additional problems). Nevertheless, the constant is not important for our purposes and, thus, we shall not care about obtaining sharper estimates for the constant.

Since we are heading towards the main part of the course, i.e. dealing with 2D problems, let us take this opportunity and prove the Poincaré inequality in 2D, as well.....

Now, we are ready to prove that the bilinear form in the model problem (V) is coercive. We have

$$\|u\|_{H_1}^2 = \int (u^2 + u'^2)dx \leq C \int u'^2 dx + \int u'^2 dx = Ca(u,u).$$

Therefore, all the results, we obtained based on the coercivity of $a(\cdot,\cdot)$, are valid.

1.3.3 Error estimates in L_2-norm. Nitsche's trick.

As we have shown, the best approximation to u from V_h is the L_2-projection and the convergence is of second order as $h \to 0$ (w.r.t. the L_2-norm). The following natural question arises—if we are not interested in the approximation of the derivative, can we show a second-order convergence in L_2-norm for the FEM solution. It turns out that this is the case under the assumption of full regularity for the exact solution:

$$\|u''\| \leq C\|f\|_{L_2}.$$

This result, however, cannot be shown to hold true, solely based on interpolation theory (as we did for the error estimate in energy norm and H_1-norm). It is obtained by using the so-called Nitsche's trick. That is, we consider the following dual problem to (V):

Find $\hat{(u)} \in H_0^1$, such that

$$a(v,\hat{u}) = (u - u_h, v), \ \forall v \in H_0^1,$$

i.e. we solve the same variational problem, but for the right-hand side we choose the error $u - u_h$.

Then, since the variational equality holds for every v, then, in particular, it holds for $v = u - u_h$ and we obtain

$$\begin{aligned}
\|u - u_h\|^2 &= a(u - u_h, \hat{u}) \\
&= a(u - u_h, \hat{u} - \hat{u}_I) \text{ (Galerkin orthogonality)} \\
&\leq C\|u - u_h\|_{H^1}\|\hat{u} - \hat{u}_I\|_{H^1} \text{ (coercivity)} \\
&\leq Ch\|u''\|_{L_2} h\|\hat{u}''\|_{L_2} \text{ (Error estimates for } u - u_h \text{ and } \hat{u} - \hat{u}_I) \\
&\leq Ch^2\|u''\|_{L_2}\|u - u_h\|_{L_2} \text{ (Full regularity assumption).}
\end{aligned}$$

Remark 13. Let us note that for the particular model problem that we considered, the requirement for full regularity is satisfied, because $-u'' = f$.

Chapter 2

2D Finite elements

2.1 Piecewise linear polynomials in two variables.

Defining piecewise polynomials is directly related to the domain discretization. Therefore, we shall begin our study with this question.

2.1.1 Triangulation

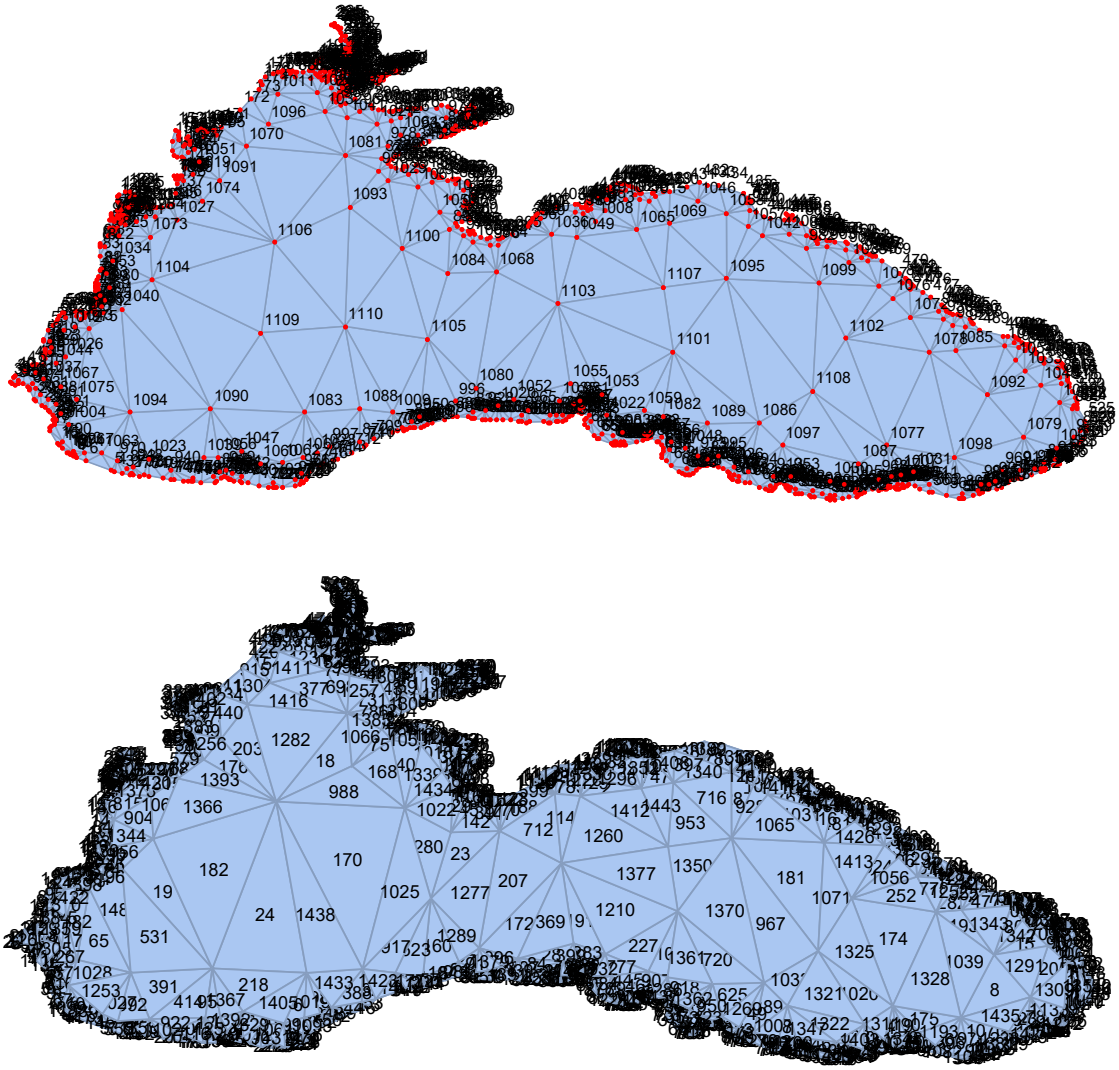
Let $\Omega \subset \mathbb{R}^2$ be a bounded region with polygonal[1] boundary $\partial\Omega$.

Definition 2. A **triangulation** $\mathcal{K}$ in Ω is a set of triangles τ, such that $\Omega = \cup_{\tau\in\mathcal{K}}\tau$ and the intersection of any two triangles is either a common edge, or a common vertex, or empty.

Let us consider as an example the following discretization of the Black Sea.

As the triangulation is generated, to each of its nodes (i.e., each vertex of a triangle) and to each element (i.e., each triangle) a number is assigned, as shown below (we shall further use the notation $\mathbf{N}_1, \ldots, \mathbf{N}_N$ for the mesh nodes):

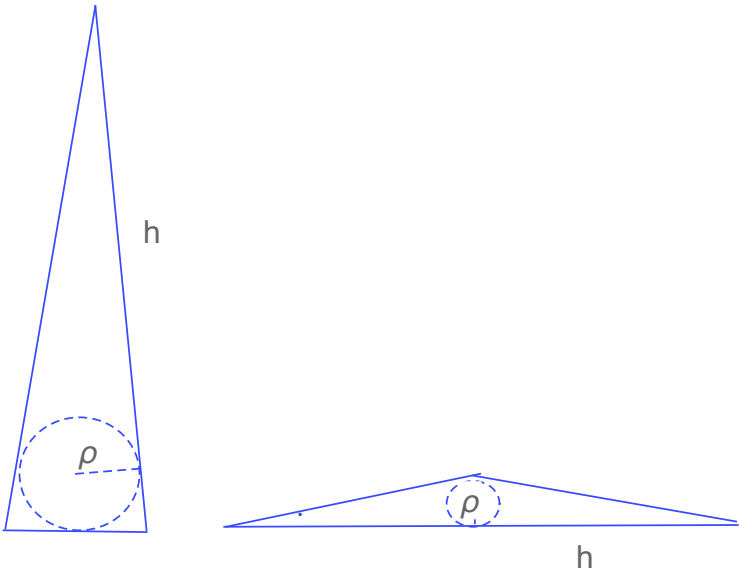
[1] The boundary can certainly be (and in most practical situations will be) non-polygonal. This will not change the algorithm of FEM. The discrete problem is still solved over the introduced mesh. It, however, gives an approximation of the real domain. Therefore, this fact must be taken into account in the error estimates. We shall deal with it in the second part of the course.



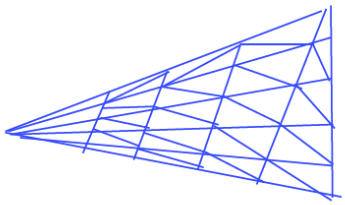
In order to derive error estimates, we are interested in measuring the size of the mesh. Thus, we need some characteristic length h_k for each element τ_k. We define h_k to be the length of the largest side of the respective triangle and $h := \max h_k$ is a global measure for the size of the triangulation. Further, we define

$$\beta_k = \frac{h_k}{\rho_k},$$

the so-called chunkiness parameter, where ρ_k is the radius of the inscribed circle. It is a measure of the quality of triangulation. As we shall see, we need to control the magnitude of β_k, in order to obtain good approximations. Stated otherwise, we do not want "too extreme" shapes in the triangulation—too wide or too narrow (in both cases β_k is large):

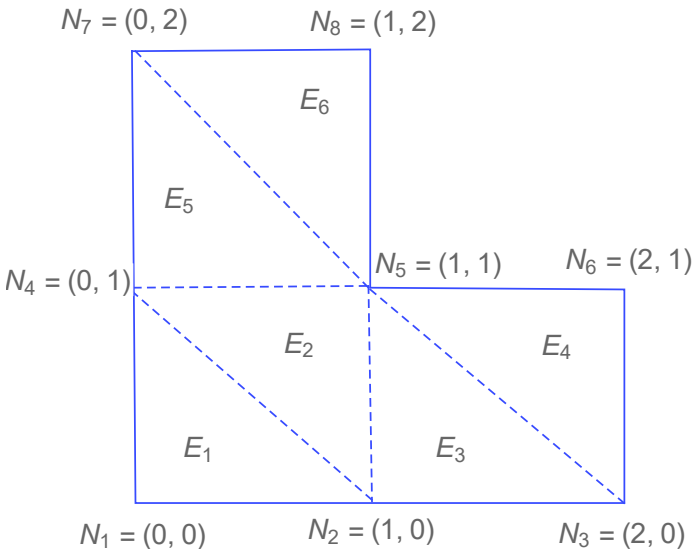


An example for a bad triangulation is the following:



Obviously, as we decrease h, then near the left vertex of the triangular domain, β_k becomes larger and larger.

Usually, we represent and store a triangulation in the memory by two matrices—of the nodes and the elements, respectively. E.g., the following triangulation:

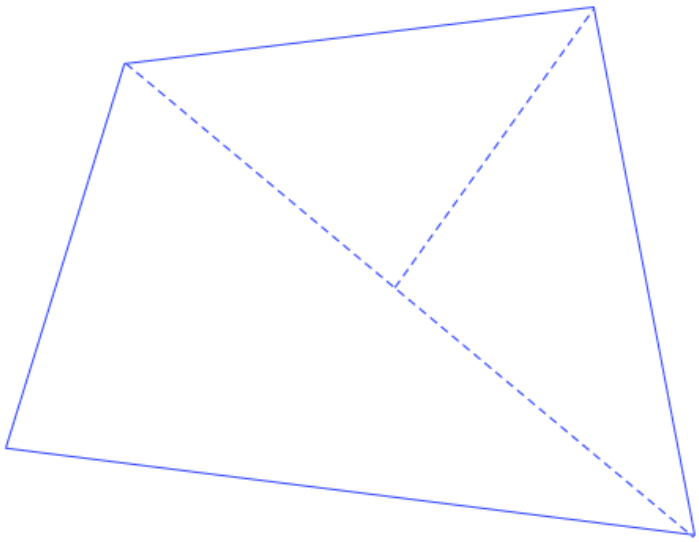


could be stored in the memory in the following way:

$$nodes = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 2 & 0 \\ 0 & 1 \\ 1 & 1 \\ 2 & 1 \\ 0 & 2 \\ 1 & 2 \end{bmatrix}, \quad elements = \begin{bmatrix} 1 & 2 & 4 \\ 2 & 5 & 4 \\ 2 & 3 & 5 \\ 3 & 6 & 5 \\ 4 & 5 & 7 \\ 5 & 8 & 7 \end{bmatrix}.$$

Having numbered each node and each element, the order of the rows in the two matrices correspond to nodes (their coordinates) and elements (the indices of the nodes that define the corresponding element). Also, information about the boundaries must be given in a similar way.

Now, that we have given a few examples of triangulations, let us give an example of a triangular discretization that is **not** a triangulation:



The problem of this discretization is that one cannot ensure continuity of the piecewise linear function—there is one linear function on the left of the main diagonal, but two different ones on the right.

There exist various packages that generate triangulations for given regions. Nevertheless, constructing a discretization with good quality is still a difficult problem, especially in 3D domains with complex geometries. In this course, we shall assume that the triangulation is given and will not deal with the subject of generating a mesh.

Before we continue, let us just make the following remark.

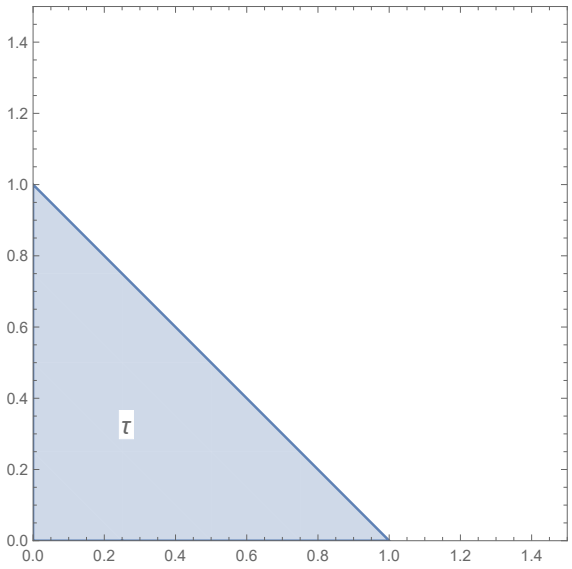
Remark 14. The region Ω can be discretized using different shapes, e.g. quadrilaterals, curvilinear triangles, etc. We shall deal with the possibilities in the second part of the course. The algorithm of FEM is, however, unchanged no matter what elements are used.

2.1.2 2D linear polynomials

The space of linear polynomials in 2D over a domain Ω is defined as

$$P_1(\Omega) := \{a_0 + a_1 x_1 + a_2 x_2, (x_1, x_2) \in \Omega, a_0, a_1, a_2 \in \mathbb{R}\}.$$

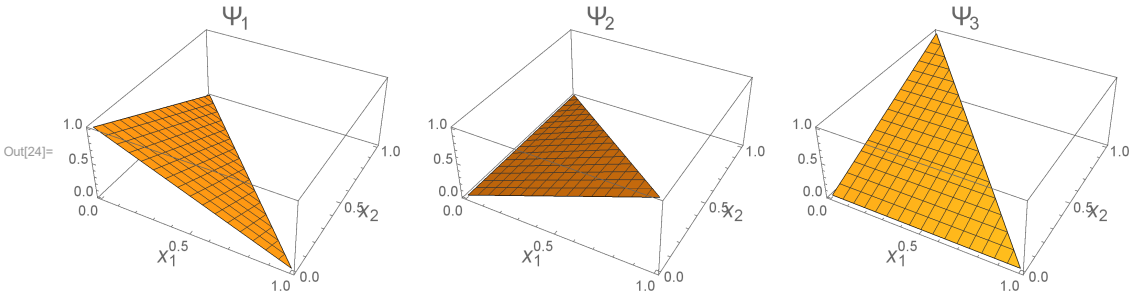
For our purposes, we shall be interested in Ω being a triangle. Without loss of generality, let us consider the **reference triangle**:



We can transform any arbitrary triangle into the reference triangle with a simple linear change of variables. We shall do this often, as it will be our general approach when we need to compute something (integrals, in particular, over an arbitrary element τ).

Over the reference triangle, we introduce the following nodal basis:

$$\Psi_1(x_1, x_2) = 1 - x_1 - x_2,$$
$$\Psi_2(x_1, x_2) = x_1,$$
$$\Psi_3(x_1, x_2) = x_2.$$

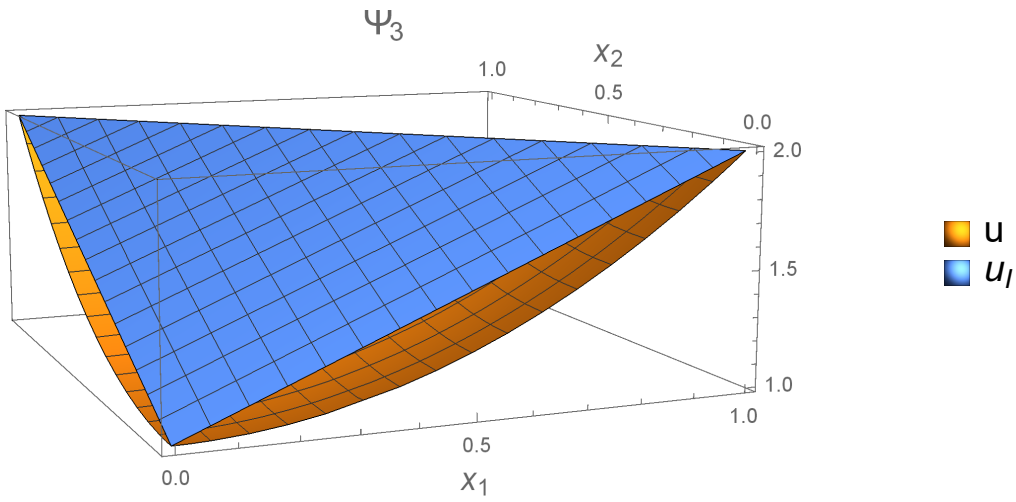


We shall call those three functions **shape functions** for the reference linear triangular element. Then, we can write the interpolant u_I of a given function u as

$$u_I = q_1 \Psi_1 + q_2 \Psi_2 + q_3 \Psi_3,$$

where q_1, q_2, q_3 are, as usual, the values at the three nodes ($q_i = u(\mathbf{N}_i)$, $i = 1, 2, 3$).

An example for the linear interpolant of a given function is depicted below:

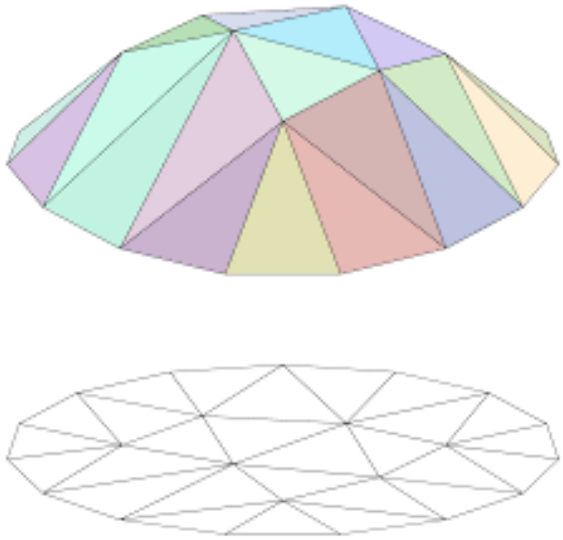


2.1.3 2D piecewise linear polynomials

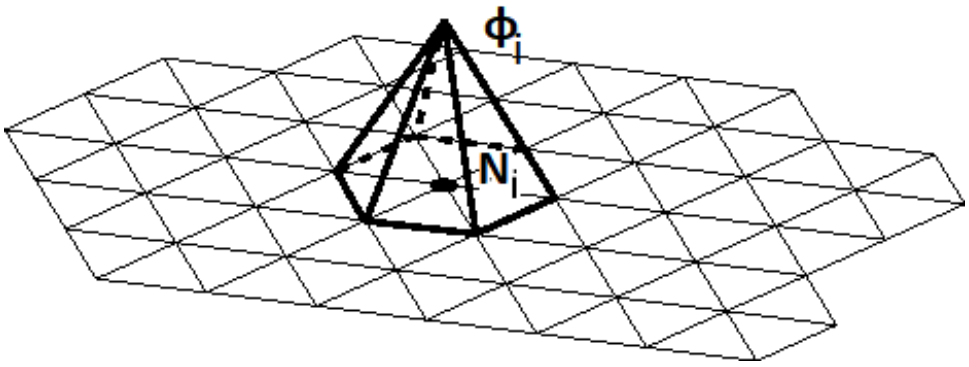
Given a triangulation $\mathcal{K}$ of Ω, we define the space of continuous piecewise linear polynomials

$$V_h(\mathcal{K}) := \{v \in C(\Omega) : v|_\tau \in P_1(\tau), \forall \tau \in \mathcal{K}\}.$$

An example graph of a piecewise linear function is given below:



We can again introduce a nodal basis $\{\varphi_k\}_{k=1}^N$, such that $V_h = span\{\varphi_1, \ldots, \varphi_N\}$. It is a little bit harder to define analytically the hat-functions in 2D and, thus, we will only illustrate one basis function graphically:



Nevertheless, a computer program can be easily implemented that computes those functions (see the exercise classes). It is important, however, to note that

$$\text{supp } \varphi_k = \cup\{\tau \in \mathcal{K} : \mathbf{N}_k \text{ is a vertex of } \tau\}.$$

Stated otherwise, over each element the only basis functions that are different than zero are the ones that correspond to the vertices of the triangle.

2.1.4 L_2-projection

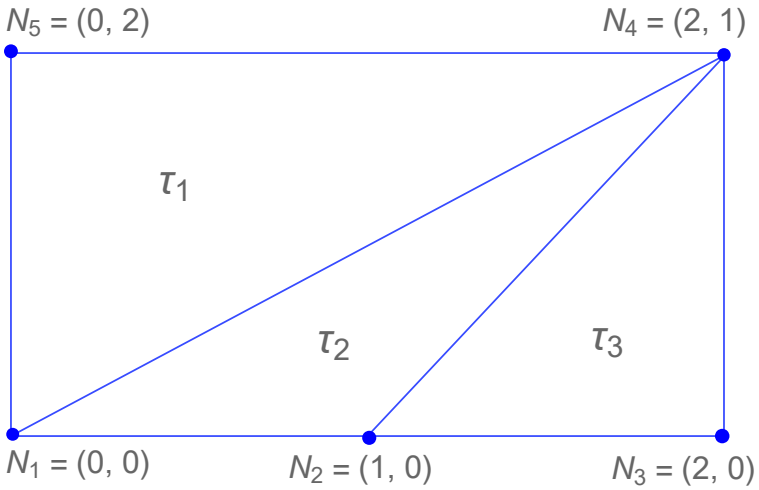
As we have already seen, the important question from the point of view of FEM is how to obtain an orthogonal projection of a given function onto V_h. Thus, we shall again (as we did for the 1D case) consider the question of obtaining the L_2-projection. I.e., given a function $u \in L_2(\Omega)$, we look for $u_h \in V_h$, such that

$$(u - u_h, v) = 0, \ \forall v \in V_h.$$

Obviously, a linear algebraic system, analogous to (1.4) is the general form of the system for the unknown nodal values. In particular, for the 2D case, we have

$$\begin{bmatrix} \iint_\Omega \varphi_1^2 d\Omega & \iint_\Omega \varphi_2\varphi_1 d\Omega & \cdots & \iint_\Omega \varphi_N\varphi_1 d\Omega \\ \vdots & \vdots & \ddots & \vdots \\ \iint_\Omega \varphi_1\varphi_N d\Omega & \iint_\Omega \varphi_2\varphi_N d\Omega & \cdots & \iint_\Omega \varphi_N^2 d\Omega \end{bmatrix} \begin{bmatrix} q_1 \\ \vdots \\ q_N \end{bmatrix} = \begin{bmatrix} \iint_\Omega u\varphi_1 d\Omega \\ \vdots \\ \iint_\Omega u\varphi_N d\Omega \end{bmatrix}.$$

Of course, we shall again compute the mass matrix M and the load vector $\mathbf{b}$ by using element-wise computations. We shall illustrate the idea on the basis of the following example. Consider the following triangulation of the rectangular domain Ω:



For the thus-introduced mesh, we have the following mass matrix:

$$M = \iint_{\Omega} \begin{bmatrix} \varphi_1^2 & \varphi_2\varphi_1 & \cdots & \varphi_5\varphi_1 \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1\varphi_5 & \varphi_2\varphi_5 & \cdots & \varphi_5^2 \end{bmatrix} d\Omega.$$

Taking into account that

$$\text{supp } \varphi_1 = \tau_1 \cup \tau_2, \text{supp } \varphi_2 = \tau_2 \cup \tau_3, \text{supp } \varphi_3 = \tau_3, \text{supp } \varphi_4 = \tau_1 \cup \tau_2 \cup \tau_3, \text{supp } \varphi_5 = \tau_1,$$

we can write M as a sum of three element matrices:

$$\begin{aligned} M = &\iint_{\tau_1} \begin{bmatrix} \varphi_1^2 & 0 & 0 & \varphi_1\varphi_4 & \varphi_1\varphi_5 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \varphi_4\varphi_1 & 0 & 0 & \varphi_4^2 & \varphi_4\varphi_5 \\ \varphi_5\varphi_1 & 0 & 0 & \varphi_5\varphi_4 & \varphi_5^2 \end{bmatrix} d\Omega + \iint_{\tau_2} \begin{bmatrix} \varphi_1^2 & \varphi_1\varphi_2 & 0 & \varphi_1\varphi_4 & 0 \\ \varphi_2\varphi_1 & \varphi_2^2 & 0 & \varphi_2\varphi_4 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \varphi_4\varphi_1 & \varphi_4\varphi_2 & 0 & \varphi_4^2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} d\Omega \\ &+ \iint_{\tau_3} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \varphi_2^2 & \varphi_2\varphi_3 & \varphi_2\varphi_4 & 0 \\ 0 & \varphi_3\varphi_3^2 & \varphi_3\varphi_4 & \varphi_3\varphi_4 & 0 \\ 0 & \varphi_4\varphi_2 & \varphi_4\varphi_3 & \varphi_4^2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} d\Omega \\ &=: M_{\tau_1} + M_{\tau_2} + M_{\tau_3}. \end{aligned}$$

Each global element matrix has 3×3 non-zero elements and, therefore, we only need to compute those 3×3 (local) element mass matrices and, when assemble the global mass matrix, to add them at the correct places. It can be shown that

$$M_\tau = \frac{1}{12} \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix} |\tau|.$$

For our example, we have, therefore

$$M_{\tau_1} = \frac{1}{12}\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}, \quad M_{\tau_2} = \frac{1}{24}\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}, \quad M_{\tau_3} = \frac{1}{24}\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}.$$

Taking into account that the matrix, representing the elements in the triangulation is

$$elements = \begin{bmatrix} 1 & 4 & 5 \\ 1 & 2 & 4 \\ 2 & 3 & 4 \end{bmatrix},$$

we can assemble M in the following way (colours show the contributions from each element matrix):

$$M = \frac{1}{12}\begin{bmatrix} 0+2+1 & 0+1/2 & 0 & 0+1+1/2 & 0+1 \\ 0+1/2 & 0+1+1 & 0+1/2 & 0+1/2+1/2 & 0 \\ 0 & 0+1/2 & 0+1 & 0+1/2 & 0 \\ 0+1+1/2 & 0+1/2+1/2 & 0+1/2 & 0+2+1+1 & 0+1 \\ 0+1 & 0 & 0 & 0+1 & 0+2 \end{bmatrix}.$$

Analogously, we can assemble the global load vector.

We can formulate the following general algorithm for computing the L_2-projection.

ALGORITHM (L_2-PROJECTION IN 2D):

1. Compute(assemble) M and $\mathbf{b}$. For this purpose, iterate for k over the rows of the matrix *elements*:
   1.1 Compute the element mass matrix m_{τ_k} and the local load vector $\mathbf{b}_{\tau_k}$;
   1.2 Add m_{τ_k} and $\mathbf{b}_{\tau_k}$ to the corresponding positions (that are written in the k-th row of *elements*) in M and $\mathbf{b}$;
2. Solve the linear algebraic system $M\mathbf{q} = \mathbf{b}$;
3. The L_2-projection of u is, thus,

$$u_h = \sum_{i=1}^{N} q_i \varphi_i(x_1, x_2).$$

Remark 15. The algorithm we formulated is not specific for 2D. It only requires that the discretization is stored in the memory as descried above. Furthermore, it allows for arbitrary elements (not only linear triangular elements).

2.1.5 A priori error estimates

In order to generalize the a priori error estimates that we know for the 1D case, we need to first introduce some additional notation. Let

$$\alpha = (\alpha_1, \alpha_2)$$

be a multiindex and let $|\alpha| := \alpha_1 + \alpha_2$. We denote

$$D^\alpha u := \frac{\partial^{|\alpha|} u}{\partial^{\alpha_1} x_1 \partial^{\alpha_2} x_2}.$$

We define the Sobolev norm in the space $H^k(\Omega)$:

$$\|u\|^2_{H^k} = \int_\Omega \sum_{|\alpha|\leq k} |D^\alpha u|^2 d\Omega.$$

We also define the Sobolev semi-norms:

$$|u|^2_{H^k} = \int_\Omega \sum_{|\alpha|=k} |D^\alpha u|^2 d\Omega.$$

We shall write explicitly several norms and semi-norms:

- In 1D:

$$\begin{aligned}
&\|u\|^2_{H^0} \equiv \|u\|^2_{L_2} = \int_I u^2 dx; \\
&\|u\|^2_{H^1} = \int_I (u^2 + u'^2) dx, \qquad |u|^2_{H^1} = \int_I u'^2 dx; \\
&\|u\|^2_{H^2} = \int_I (u^2 + u'^2 + u''^2) dx, \quad |u|^2_{H^2} = \int_I u''^2 dx.
\end{aligned}$$

- In 2D:

$$\begin{aligned}
&\|u\|^2_{H^0} \equiv \|u\|^2_{L_2} = \int_\Omega u^2 dx; \\
&\|u\|^2_{H^1} = \iint_\Omega \left[u^2 + \left(\frac{\partial u}{\partial x_1}\right)^2 + \left(\frac{\partial u}{\partial x_2}\right)^2 \right] d\Omega, \\
&\quad |u|^2_{H^1} = \iint_\Omega \left[\left(\frac{\partial u}{\partial x_1}\right)^2 + \left(\frac{\partial u}{\partial x_2}\right)^2 \right] d\Omega; \\
&\|u\|^2_{H^2} = \iint_\Omega \left[u^2 + \left(\frac{\partial u}{\partial x_1}\right)^2 + \left(\frac{\partial u}{\partial x_2}\right)^2 + \left(\frac{\partial^2 u}{\partial x_1^2}\right)^2 + \left(\frac{\partial^2 u}{\partial x_1 \partial x_2}\right)^2 + \left(\frac{\partial^2 u}{\partial x_2^2}\right)^2 \right] d\Omega, \\
&\quad |u|^2_{H^2} = \iint_\Omega \left[\left(\frac{\partial^2 u}{\partial x_1^2}\right)^2 + \left(\frac{\partial^2 u}{\partial x_1 \partial x_2}\right)^2 + \left(\frac{\partial^2 u}{\partial x_2^2}\right)^2 \right] d\Omega.
\end{aligned}$$

The following proposition is a generalization of the result we proved for the linear interpolant in 1D.

Proposition 12. *The following a priori error estimate holds for the linear interpolant $u_I \in P_1$ of a given function u:*

$$\begin{aligned}
&\|u - u_I\|_{L_2} \leq Ch^2 |u|_{H^2}, \\
&\|\nabla(u - u_I)\|_{L_2} \leq Ch |u|_{H^2}.
\end{aligned} \tag{2.1}$$

We shall not prove this result at the time being since we will prove a more general one later in the course. Using those estimates, however, we are ready to derive error estimates for the L_2-projection u_h, as well. First, we can prove a best approximation result, using the orthogonality in L_2-norm.

Proposition 13. *The L_2-projection in V_h (the function u_h) is the best approximation of u with respect to the L_2-norm, i.e.*

$$\|u - u_h\|_{L_2(\Omega)} \leq \|u - v\|_{L_2(\Omega)}, \forall v \in V_h.$$

Proof. We have

$$\begin{aligned}\|u - u_h\|^2_{L_2(\Omega)} &= \iint_\Omega (u - u_h)(u - v + \cancelto{0}{v - u_h})d\Omega \\ &\leq \|u - u_h\|_{L_2(\Omega)}\|u - v\|_{L_2(\Omega)}.\end{aligned}$$

□

Then, the a priori error estimate for the L_2-projection follows in the usual way.

Proposition 14. *The following a priori error estimates hold for the L_2-projection in V_h of a given function u:*

$$\begin{aligned}\|u - u_h\|_{L_2(\Omega)} &\leq Ch^2|u|_{H^2(\Omega)}, \\ \|\nabla(u - u_h)\|_{L_2(\Omega)} &\leq Ch|u|_{H^2(\Omega)}.\end{aligned}$$

Proof. We shall prove the first inequality. The second one follows similarly. We consequtively obtain, using (2.1):

$$\begin{aligned}\|u - u_h\|^2_{L_2(\Omega)} &\leq \|u - u_I\|^2_{L_2(\Omega)} \\ &= \sum_{\tau\in\mathcal{K}} \|u - u_I\|^2_{L_2(\tau)} \\ &\leq \sum_{\tau\in\mathcal{K}} Ch_\tau^4|u|^2_{H^2(\tau)} \\ &\leq Ch^4 \sum_{\tau\in\mathcal{K}} |u|^2_{H^2(\tau)} \\ &= Ch^4|u|^2_{H^2(\Omega)}\end{aligned}$$

□

Before you proceed, see the worked-out examples in Section 3 of the notes from the exercise classes.

2.2 FEM for 2D stationary problems

2.2.1 Preliminaries from vector calculus

In order to generalize the ideas that we presented for the 1D case, we shall need several notions from vector calculus.

The analog of the ordinary differential operator (actually, one of the analogs) for higher-dimensional problems is the ∇ operator. We define it in Cartesian coordinates as follows:

$$\nabla := \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y} \right).$$

The latter operator can be applied to different objects (scalars, vectors).

If we apply it to the scalar function u, we obtain

$$\nabla u = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y} \right) u = \left(\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y} \right),$$

the so-called **gradient vector**. Let us remind the following important facts about the gradient vector:

- it points in the direction of fastest increase of the function u;
- the projection of ∇u onto a direction $\mathbf{d}$ (unit vector) is the directional derivative in that direction, i.e.

$$\frac{\partial u}{\partial \mathbf{d}} = \nabla u \cdot \mathbf{d}.$$

The ∇ operator can be applied to the vector function $\mathbf{j} = (j_x, j_y)$ in two different ways: $\nabla \cdot \mathbf{j}$ or $\nabla \times \mathbf{j}$. Both of them have important physical interpretations. For the time being, we shall be interested only in the first one, i.e. we shall consider the divergence

$$\operatorname{div} \mathbf{j} := \nabla \cdot \mathbf{j} = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y} \right) \cdot (j_x, j_y) = \frac{\partial j_x}{\partial x} + \frac{\partial j_y}{\partial y}.$$

The divergence can be interpreted physically as the net outflow per unit area at a given point.

Having this in mind, the following fundamental theorem for the integral calculus of vector functions can be justified on an intuitive level:

Proposition 15 (Divergence theorem)**.** *Let Ω be a compact subset of $\mathbb{R}^n$ that has a piecewise smooth boundary $\partial\Omega$. If $\boldsymbol{F}$ is continuously differentiable function, defined in an open region that contains Ω, then*

$$\int_\Omega \nabla \cdot \boldsymbol{F} d\Omega = \int_{\partial\Omega} \boldsymbol{F} \cdot \boldsymbol{n} ds,$$

where $\boldsymbol{n}$ is the outer normal.

The physical intuition behind this theorem is the following. The outflux at each point needs to result in an influx at a "neighboring" point. Therefore, the sum of all outflows inside Ω must be equal to what goes out of the boundary $\partial\Omega$.

In particular, we will be interested in using this theorem for $\mathbf{F} = \mathbf{j}v$, which will give us an analog to the Integration by parts theorem for higher-dimensional problems. We shall formulate it in 2D. Taking into account that

$$\iint_\Omega \nabla \cdot (\mathbf{j}v) d\Omega = \iint_\Omega (\nabla \cdot \mathbf{j}v + \mathbf{j} \cdot \nabla v) d\Omega,$$

we obtain the following integration by parts formula.

Proposition 16 (Integration by parts formula)**.** *Under the assumptions of the Divergence theorem, the following relation holds true:*

$$\iint_\Omega \nabla \cdot \boldsymbol{j} v d\Omega = -\iint_\Omega \boldsymbol{j} \cdot \nabla v d\Omega + \int_{\partial\Omega} (\boldsymbol{j} \cdot \boldsymbol{n}) v ds.$$

2.2.2 FEM for 2D stationary problems with homogeneous Dirichlet boundary conditions

We begin presenting the ideas over the following relatively simple example:

$$\begin{aligned} -\Delta u &= f, \text{ in } \Omega, \\ u &= 0, \text{ on } \partial\Omega, \end{aligned} \tag{D}$$

where $\Delta u = \nabla \cdot (\nabla u)$ is the Laplace operator (or Laplacian). The considered differential equation is the so-called Poisson equation.

In order to obtain the variational form of the differential problem, we take the scalar products of both sides of the differential equation with a test function v and obtain

$$(-\Delta u, v) = (f, v).$$

For the left-hand side, using the Integration by parts formula, we obtain consecutively

$$\begin{aligned} (-\Delta u, v) &= \iint_\Omega -\nabla \cdot (\nabla u) v d\Omega \\ &= \iint_\Omega \nabla u \cdot \nabla v d\Omega - \cancelto{0,\text{ if we require } v \in H_0^1(\Omega)}{\int_{\partial\Omega} (\nabla u \cdot \mathbf{n}) v ds}. \end{aligned}$$

Thus, we have obtained the following variational problem:

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

$$a(u, v) = F(v), \ \forall v \in H_0^1(\Omega), \tag{V}$$

where

$$a(u, v) := \iint_\Omega \nabla u \cdot \nabla v d\Omega, \quad F(v) := \iint_\Omega f v d\Omega.$$

Discretizing the domain Ω, by introducing the triangulation $\mathcal{K}$, we formulate the corresponding approximate problem:

Find $u_h \in V_{h,0}(\mathcal{K})$, such that

$$a(u_h, v) = F(v), \quad \forall v \in V_{h,0}(\mathcal{K}). \tag{R.–G.}$$

Remark 16. We can now understand why it is important from a theoretical point of view to have Ω with polygonal boundary. In the opposite case $\mathcal{K}$ might not be the same as Ω and, therefore, $V_{h,0}(\mathcal{K})$ might not be a subet of $H_0^1(\Omega)$. This would be one of the "variational crimes" that we shall discuss in the second part of the course. In those cases, we will need to work a little bit harder, in order to show convergence of the method.

On the other hand, from an algorithmic point of view, nothing would change if Ω has an arbitrary boundary. We would still need to solve the linear algebraic system that follows below.

Using the fact that $V_{h,0} = span\{\varphi_1, \ldots, \varphi_{N_{inter}}\}$, where $\varphi_1, \ldots, \varphi_{N_{inter}}$ are the hat-functions, corresponding to the interior nodes of the triangulation, and $u_h(x) = \sum_{j=1}^{N_{inter}} q_j \varphi_j(x)$, we obtain the equivalent problem:

Find $(q_1, \ldots, q_{N_{inter}}) \in \mathbb{R}^{N_{inter}}$, such that

$$\sum_{j=1}^{N_{inter}} a(\varphi_j, \varphi_i) q_j = F(\varphi_i), \quad i = \overline{1, N_{inter}}.$$

Substituting with the specific bilinear form, we obtain the following linear algebraic system with respect to the unknown coefficients, $M\mathbf{q} = \mathbf{b}$:

$$\begin{bmatrix} \iint_\Omega \nabla\varphi_1 \cdot \nabla\varphi_1 d\Omega & \cdots & \iint_\Omega \nabla\varphi_1 \cdot \nabla\varphi_{N_{inter}} d\Omega \\ \vdots & \ddots & \vdots \\ \iint_\Omega \nabla\varphi_{N_{inter}} \cdot \nabla\varphi_1 d\Omega & \cdots & \iint_\Omega \nabla\varphi_{N_{inter}} \cdot \nabla\varphi_{N_{inter}} d\Omega \end{bmatrix} \begin{bmatrix} q_1 \\ \vdots \\ q_{N_{inter}} \end{bmatrix} = \begin{bmatrix} \iint_\Omega f\varphi_1 d\Omega \\ \vdots \\ \iint_\Omega f\varphi_{N_{inter}} d\Omega \end{bmatrix}.$$

Here, $M = M^1$ is the stiffness matrix.

2.2.3 A priori error estimate

Of course, the Galerkin orthogonality holds:

$$a(u - u_h, v) = 0, \ \forall v \in V_{h,0}(\mathcal{K})$$

and, furthermore, the bilinear form defines a scalar product:

$$< u, v >_E := a(u, v) = \iint_\Omega \nabla u \cdot \nabla v d\Omega$$

and the corresponding energy norm

$$\|u\|_E^2 := a(u, u) = \iint_\Omega (\nabla u)^2 d\Omega.$$

Thus, we can proceed in the usual way and obtain a Best approximation result in energy norm and, subsequently, an energy-norm error estimate:

$$\|u - u_h\|_E \leq Ch|u|_{H^2}.$$

In this section, however, we shall derive the more general H^1-norm error estimate (which in this case is equivalent to the energy-norm estimate). As we have discussed, the important properties that we would like to be satisfied by the bilinear form $a(\cdot, \cdot)$ are coercivity and continuity in H^1. Let us check if they are satisfied for the particular problem.

First, **coercivity** in H^1 means

$$a(u, u) \geq C\|u\|_{H^1}^2.$$

We have

$$\begin{aligned} \|u\|_{H^1}^2 &= \iint_\Omega [u^2 + (\nabla u)^2] d\Omega \\ &= \|u\|_{L_2}^2 + \|\nabla u\|_{L_2}^2 \\ &\leq C\|\nabla u\|_{L_2}^2 + \|\nabla u\|_{L_2}^2 \qquad \text{(using Poincaré inequality)} \\ &= Ca(u, u). \end{aligned}$$

Thus, we have established coercivity of $a(\cdot,\cdot)$ in H^1.

Furthermore, we can show that it is also **continuous** in H^1, i.e.

$$a(u,v) \le C\|u\|_{H^1}\|v\|_{H^1}.$$

Indeed, we consecutively obtain

$$\begin{aligned} a(u,v) &= \iint_\Omega \nabla u \cdot \nabla v d\Omega \\ &\le \sqrt{\iint_\Omega (\nabla u)^2 d\Omega}\sqrt{\iint_\Omega (\nabla v)^2 d\Omega} \qquad \text{(using Schwarz inequality)} \\ &\le \sqrt{\iint_\Omega \big(u^2 + (\nabla u)^2\big) d\Omega}\sqrt{\iint_\Omega \big(v^2 + (\nabla v)^2\big) d\Omega} \\ &= \|u\|_{H^1}\|v\|_{H^1}. \end{aligned}$$

Having established coercivity and continuity, we can obtain an upper estimate of $\|u-u_h\|_{H^1}$ in the usual way.

Proposition 17. *For the FEM solution u_h, the following result is valid:*

$$\|u-u_h\|_{H^1} \le C\|u-v\|_{H^1}, \quad \forall v \in V_{h,0}.$$

Proof. Considering an arbitrary function $v \in V_{h,0}$, we get

$$\begin{aligned} \|u-u_h\|^2_{H^1} &\le Ca(u-u_h, u-v+v-u_h) && \text{(using coercivity)} \\ &= Ca(u-u_h, u-v) && \text{(using Galerkin's orthogonality)} \\ &\le C\|u-u_h\|_{H^1}\|u-v\|_{H^1}. && \text{(using continuity).} \end{aligned}$$

Dividing both sides to $\|u-u_h\|_{H^1}$, we obtain the desired result. □

Using the latter result with $v = u_I$, we obtain the following.

Proposition 18. *For the FEM solution u_h, the following a priori error estimate holds:*

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

Proof. Using Proposition 17, we have

$$\begin{aligned} \|u-u_h\|^2_{H^1(\Omega)} &\le C\|u-u_I\|^2_{H^1(\Omega)} \\ &= C(\|u-u_I\|^2_{L_2(\Omega)} + \|\nabla(u-u_I)\|^2_{L_2(\Omega)}) \\ &= C\sum_{\tau\in\mathcal{K}} \left(\|u-u_I\|^2_{L_2(\tau)} + \|\nabla(u-u_I)\|^2_{L_2(\tau)}\right). \end{aligned}$$

Now, from (2.1), it follows that for sufficiently small h

$$\|u-u_h\|^2_{H^1(\Omega)} \le C\sum_{\tau\in\mathcal{K}} h_\tau^2 |u|^2_{H^2(\tau)} \le Ch^2|u|^2_{H^2(\Omega)}.$$

□

Remark 17. Basically, in the latter proof we have used the main approach underlying FEM—we have decomposed the piecewise-linear problem into a sum of linear problems over each element and, at the end, we assembled back the global problem over Ω.

2.2.4 FEM for stationary 2D problems with more general boundary conditions

Let us now consider a problem with the more general Robin boundary conditions. We shall also include in the problem formulation a (possibly) variable coefficient $\alpha = \alpha(x, y) \geq \alpha_{min} > 0$:

$$\begin{aligned} -\nabla \cdot (\alpha \nabla u) &= f, \quad x \in \Omega, \\ -\mathbf{n} \cdot (\alpha \nabla u) &= \varkappa(u - g_0) - g_N, \quad x \in \partial\Omega, \end{aligned} \tag{D}$$

where $\mathbf{n}$ is the outer normal to the boundary.

In order to obtain the variational form, we multiply both sides with a test function v and integrate. For the left-hand side, we obtain

$$\begin{aligned} -\iint_\Omega \nabla \cdot (\alpha \nabla u) v d\Omega &= \iint_\Omega \alpha \nabla u \cdot \nabla v d\Omega - \int_{\partial\Omega} \mathbf{n} \cdot (\alpha \nabla u) v ds \\ &= \iint_\Omega \alpha \nabla u \cdot \nabla v d\Omega + \int_{\partial\Omega} [\varkappa(u - g_0) - g_N] v ds. \end{aligned}$$

In the latter inequality, we have used the Robin boundary condition. By leaving on the left-hand side only the terms that contain the unknown function u, we obtain the variational problem:

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

$$a(u, v) = F(v), \ \forall v \in H^1(\Omega), \tag{V}$$

where

$$a(u, v) := \iint_\Omega \alpha \nabla u \cdot \nabla v d\Omega + \int_{\partial\Omega} \varkappa u v ds, \quad F(v) := \iint_\Omega f v d\Omega + \int_{\partial\Omega} (\varkappa g_0 + g_N) v ds.$$

The corresponding Ritz–Galerkin problem is:

Find $u_h \in V_h(\mathcal{K})$, such that

$$a(u_h, v) = F(v), \quad \forall v \in V_h. \tag{R.–G.}$$

Here, $V_h = span\{\varphi_1, \ldots, \varphi_N\}$, where we have used the hat-functions, corresponding to all nodes in the triangulation $\mathcal{K}$.

Following the same path as usual, we obtain the linear system $M\mathbf{q} = \mathbf{b}$, where

$$M = \begin{bmatrix} \iint_\Omega \alpha \nabla \varphi_1 \cdot \nabla \varphi_1 d\Omega & \cdots & \iint_\Omega \alpha \nabla \varphi_1 \cdot \nabla \varphi_N d\Omega \\ \vdots & \ddots & \vdots \\ \iint_\Omega \alpha \nabla \varphi_N \cdot \nabla \varphi_1 d\Omega & \cdots & \iint_\Omega \alpha \nabla \varphi_N \cdot \nabla \varphi_N d\Omega \end{bmatrix} + \begin{bmatrix} \int_{\partial\Omega} \varkappa \varphi_1 \varphi_1 ds & \cdots & \int_{\partial\Omega} \varkappa \varphi_1 \varphi_N ds \\ \vdots & \ddots & \vdots \\ \int_{\partial\Omega} \varkappa \varphi_N \varphi_1 ds & \cdots & \int_{\partial\Omega} \varkappa \varphi_N \varphi_N ds \end{bmatrix}$$

and

$$\mathbf{b} = \begin{bmatrix} \iint_\Omega f \varphi_1 d\Omega \\ \vdots \\ \iint_\Omega f \varphi_N d\Omega \end{bmatrix} + \begin{bmatrix} \int_{\partial\Omega} (\varkappa g_0 + g_N) \varphi_1 ds \\ \vdots \\ \int_{\partial\Omega} (\varkappa g_0 + g_N) \varphi_N ds \end{bmatrix}.$$

Note that we have decomposed the matrix M into a sum of the stiffness matrix and a matrix, corresponding to contributions from the boundary conditions. We have proceeded analogously with the load vector $\mathbf{b}$. This has certain benefits from an algorithmic point of view that we shall discuss later.

Since in this case the variational problem is solved over H^1 and the Poincaré inequality does not hold, we still don't have the machinery to show coercivity of the bilinear form and, thus, derive an a priori error estimate in H^1-norm. For the time being, we will confine ourselves to presenting only an energy-norm error estimate. It is based on the usual Best approximation result. Let us just remark that for the variational problem (V) the bilinear form again defines an energy scalar product

$$< u, v >_E := a(u,v) = \iint_\Omega \alpha \nabla u \cdot \nabla v d\Omega + \int_{\partial\Omega} \varkappa uv ds$$

and we can define the energy norm

$$\|u\|_E^2 := a(u,u) = \iint_\Omega \alpha (\nabla u)^2 d\Omega + \int_{\partial\Omega} \varkappa u^2 ds.$$

Proposition 19. *For the FEM solution u_h, the following holds true:*

$$\|u - u_h\|_E \leq \|u - v\|_E, \quad \forall v \in V_h.$$

Proof. We obtain consecutively

$$\begin{aligned}
\|u-u_h\|_E^2 &= a(u-u_h, u-v+\cancelto{0}{v-u_h}), \text{ Galerkin orthogonality} \\
&= \iint_\Omega \big(\sqrt{\alpha}\nabla(u-u_h)\big)\big(\sqrt{a}\nabla(u-v)\big)d\Omega + \int_{\partial\Omega}\big(\sqrt{\varkappa}(u-u_h)\big)\big(\sqrt{\varkappa}(u-v)\big)ds \\
&\leq \sqrt{\iint_\Omega \alpha\big[\nabla(u-u_h)\big]^2 d\Omega}\sqrt{\iint_\Omega \alpha\big[\nabla(u-v)\big]^2 d\Omega} \\
&+ \sqrt{\int_{\partial\Omega}\varkappa(u-u_h)^2 ds}\sqrt{\int_{\partial\Omega}\varkappa(u-v)^2 ds} \\
&=: \sqrt{A}\sqrt{B} + \sqrt{C}\sqrt{D}.
\end{aligned}$$

We are now left to prove that

$$\sqrt{A}\sqrt{B} + \sqrt{C}\sqrt{D} \leq \sqrt{A+C}\sqrt{B+D} = \|u-u_h\|_E\|u-v\|_E.$$

If we square both sides of the inequality, we obtain

$$AB + CD + 2\sqrt{ABCD} \leq (A+C)(B+D)$$

or, which is the same

$$2\sqrt{ABCD} \leq AD + BC.$$

The latter is obviously satisfied due to the inequality between the geometric mean and the arithmetic mean. Thus, we have established the statement of the proposition. □

Therefore,

$$\|u - u_h\|_E \leq \|u - u_I\|_E \leq Ch|u|_{H^2}$$

Chapter 3

Practical aspects of FEM. Element-wise computations.

From our discussion so far, we should be aware that the application of FEM for a given differential problem follows the next 5 steps:

1. Write the differential problem in its weak form (V);
2. Subdivide the region Ω into triangles (or, e.g., quadrilaterals);
3. Define the test space V as the span of appropriately chosen basis functions (in what we have discussed so far, hat function);
4. Derive the Ritz–Galerkin problem over the finite-dimensional space V_h and the corresponding linear algebraic system $M\mathbf{q} = \mathbf{b}$;
5. Assemble M and $\mathbf{b}$ and solve the system.

All the computations are in step 5. But the first four steps determine whether the method will be efficient and accurate for our specific problem. In particular, the discretization and the choice of the finite-dimensional space V_h matter a lot.

We have already covered the basic ideas, underlying the first four steps. Now, we need to deal with the question of computing M and $\mathbf{b}$ in practice. We shall work one element at a time for reasons that we explained when we considered assembling the mass matrix for the L_2-projection.

We shall follow the same approach we have embraced from the beginning of the course—first, we shall obtain some intuition over a 1D example, and then generalize in 2D.

3.1 Element-wise computations in 1D

We shall present the ideas over a much more general example than we have done so far:

$$\begin{aligned} -(\alpha u')' + \beta u &= f, \quad x \in (0,1) \\ u'(0) = u'(1) &= 0, \end{aligned} \tag{D}$$

where $\alpha(x) \geq \alpha_{min} > 0$ and $\beta(x)$ are sufficiently smooth known functions.

Basically, any linear second-order boundary-value problem for ODE can be put in the above form (the so-called divergence form). What is important to note is that we have imposed homogeneous Neumann boundary conditions. This is the easiest problem from computational point of view and, thus, we begin with it. Note that here we don't discuss the well-posedness of the problem. We are only interested in the formal computation of the global matrix and load vector of the resulting system and considering Neumann boundary conditions is a good place to start from.

In order to discuss the computational aspects of FEM, we need to first derive the linear algebraic system. Here, we shall only sketch the derivation and focus on what happens after that. First, we derive the weak form in the usual way and obtain the following:

Find $u \in H^1$, such that

$$a(u,v) = F(v), \quad \forall v \in H^1, \tag{V}$$

where

$$a(u,v) = \int_I (\alpha u'v' + \beta uv)dx, \quad F(v) = \int_I fvdx.$$

Introducing a (not necessarily uniform) mesh $0 = x_0 < \cdots < x_n = 1$ and denoting $h_i := x_i - x_{i-1}$, $I_i := [x_{i-1}, x_i]$, we define $V_h := span\{\varphi_0(x), \ldots, \varphi_n(x)\}$. I.e., we have the full basis, which consists of hat-functions, corresponding to all nodes from the mesh.

Thus, the corresponding Ritz–Galerking problem becomes:

Find $u \in V_h$, such that

$$a(u_h,v) = F(v), \quad \forall v \in V_h. \tag{R.–G.}$$

Finally, we derive the linear algebraic system $M\mathbf{q} = \mathbf{b}$, where

$$M = \begin{bmatrix} \int_I (\alpha\varphi_0'^2 + \beta\varphi_0^2)dx & \cdots & \int_I (\alpha\varphi_0'\varphi_n' + \beta\varphi_0\varphi_n)dx \\ \vdots & \ddots & \vdots \\ \int_I (\alpha\varphi_n'\varphi_0' + \beta\varphi_n\varphi_0)dx & \cdots & \int_I (\alpha\varphi_n'^2 + \beta\varphi_n^2)dx \end{bmatrix}$$

is the sum of the mass matrix M^0 and the stiffness matrix M^1:

$$M^0 = \begin{bmatrix} \int_I \beta\varphi_0^2 dx & \cdots & \int_I \beta\varphi_0\varphi_n dx \\ \vdots & \ddots & \vdots \\ \int_I \beta\varphi_n\varphi_0 dx & \cdots & \int_I \beta\varphi_n^2 dx \end{bmatrix}, \; M^1 = \begin{bmatrix} \int_I \alpha{\varphi_0'}^2 dx & \cdots & \int_I \alpha\varphi_0'\varphi_n' dx \\ \vdots & \ddots & \vdots \\ \int_I \alpha\varphi_n'\varphi_0' dx & \cdots & \int_I \alpha{\varphi_n'}^2 dx \end{bmatrix}.$$

As we have already discussed, in order to take advantage of the finite support of the basis functions, we decompose those matrices as sums over the elements:

$$M^0 = \sum_{i=0}^{n} M^0_{i,glob} := \sum_{i=0}^{n} \begin{bmatrix} \int_{I_i} \beta\varphi_0^2 dx & \cdots & \int_{I_i} \beta\varphi_0\varphi_n dx \\ \vdots & \ddots & \vdots \\ \int_{I_i} \beta\varphi_n\varphi_0 dx & \cdots & \int_{I_i} \beta\varphi_n^2 dx \end{bmatrix},$$

$$M^1 = \sum_{i=0}^{n} M^1_{i,glob} := \sum_{i=0}^{n} \begin{bmatrix} \int_{I_i} \alpha{\varphi_0'}^2 dx & \cdots & \int_{I_i} \alpha\varphi_0'\varphi_n' dx \\ \vdots & \ddots & \vdots \\ \int_{I_i} \alpha\varphi_n'\varphi_0' dx & \cdots & \int_{I_i} \alpha{\varphi_n'}^2 dx \end{bmatrix}.$$

Let us, first, show how, using the latter, to assemble M^0. Assembling M^1 follows the same path. We know that the i-th global element matrix $M^0_{i,glob}$ has only 2×2 non-zero elements,

$$M^0_{i,glob} = \begin{bmatrix} 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots\cdots & \cdots & \cdots\cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \int_{I_i} \beta\varphi_{i-1}^2 dx & \int_{I_i} \beta\varphi_{i-1}\varphi_i dx & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \int_{I_i} \beta\varphi_i\varphi_{i-1} dx & \int_{I_i} \beta\varphi_i^2 dx & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots\cdots & \cdots & \cdots\cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix},$$

because of the finite support of the hat-functions:

We, therefore, need to compute only the 2×2 (local) element mass matrices:

$$M^0_i = \begin{bmatrix} \int_{I_i} \beta\varphi_{i-1}^2 dx & \int_{I_i} \beta\varphi_{i-1}\varphi_i dx \\ \int_{I_i} \beta\varphi_i\varphi_{i-1} dx & \int_{I_i} \beta\varphi_i^2 dx \end{bmatrix}. \tag{3.1}$$

We can do this directly, because we have analytic definitions for $\varphi_{i-1}(x)$ and $\varphi_i(x)$ in I_i (see (1.1)). Nevertheless, this approach would cause us problems in 2D, because we haven't given an analytic definition for the hat functions in 2D. Furthermore, the integrals will need to be computed over arbitrary triangles, which might cause difficulties. Thus, we shall present here the general idea. That is, we shall first make a change of variables that transforms I_i into the **reference element** $I := [0, 1]$. Over the reference element, all computations will become straight-forward.

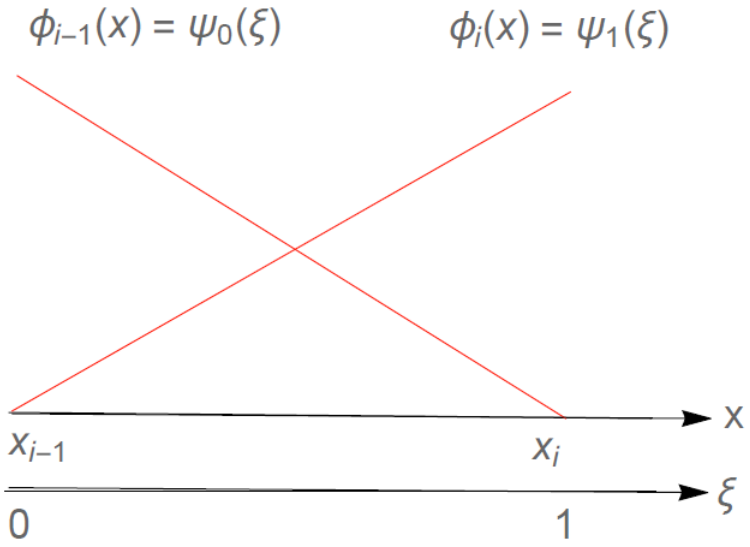
It is obvious that the following change of variables will do the work:

$$x = x_{i-1} + \xi h_i.$$

Then, the interval $x \in [x_{i-1}, x_i]$ is transformed into $\xi \in [0, 1]$. Furthermore, we have

$$\begin{aligned} &\varphi_{i-1}(x) = \varphi_{i-1}(x_{i-1} + \xi h_i) =: \psi_0(\xi) = 1 - \xi, \\ &\varphi_i(x) = \varphi_{i-1}(x_i + \xi h_i) =: \psi_1(\xi) = \xi, \\ &dx = h_i d\xi. \end{aligned}$$

The choice of the functions ψ_0 and ψ_1 is correct because they are linear for $\xi \in [0, 1]$ and satisfy the same conditions as φ_{i-1}, φ_i, respectively, in the nodes x_{i-1} and x_i. The functions ψ_0 and ψ_1 are called **shape functions**. We have, obviously (and for obvious reasons!), obtained the nodal basis of P_1.



Changing the variable in (3.1), we obtain

$$\begin{aligned} M_i^0 &= h_i \int_0^1 \beta(x_{i-1} + \xi h_i) \begin{bmatrix} \psi_0^2 & \psi_0\psi_1 \\ \psi_1\psi_0 & \psi_1^2 \end{bmatrix} d\xi \\ &= h_i \int_0^1 \beta(x_{i-1} + \xi h_i) \begin{bmatrix} 1 - 2\xi + \xi^2 & \xi - \xi^2 \\ \xi - \xi^2 & \xi^2 \end{bmatrix} d\xi. \end{aligned} \tag{3.2}$$

Analogously, we obtain for the element stiffness matrix

$$\begin{aligned} M_i^1 &= \frac{1}{h_i} \int_0^1 \alpha(x_{i-1} + \xi h_i) \begin{bmatrix} \psi_0'^2 & \psi_0'\psi_1' \\ \psi_1'\psi_0' & \psi_1'^2 \end{bmatrix} d\xi \\ &= \frac{1}{h_i} \int_0^1 \alpha(x_{i-1} + \xi h_i) \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \end{aligned} \tag{3.3}$$

In the latter, we have used the fact that

$$\frac{d\varphi_j}{dx} = \frac{d\varphi_j}{d\xi}\frac{d\xi}{dx} = \frac{d\varphi_j}{d\xi}\frac{1}{h_i}, \quad j = i-1, i.$$

Remark 18. We can also write the following

$$\begin{aligned} M_i^0 &= h_i \int_0^1 \beta(x_{i-1} + \xi h_i) \mathbf{\Psi}^T \mathbf{\Psi} d\xi, \\ M_i^1 &= \frac{1}{h_i} \int_0^1 \beta(x_{i-1} + \xi h_i) \mathbf{\Psi}'^T \mathbf{\Psi}' d\xi, \end{aligned}$$

where $\mathbf{\Psi} = (\psi_0, \psi_1)$.

This vector form is very useful, because it can be shown to hold for any shape functions (e.g., if we work with piecewise quadratic functions, we would obtain $\mathbf{\Psi} = (2\xi^2 - 3\xi + 1, -4\xi^2 + 4\xi, 2\xi^2 - \xi)$ for the shape functions). This is connected to one of the benefits of FEM—we can derive results that are independent on the specific choice of elements and reuse them in different particular cases.

Let us consider two specific examples, in order to illustrate the ideas.

Example. Let $\alpha \equiv 1, \beta \equiv 1$ hold. Then, for the element mass and stiffness matrices from (3.2) and (3.3) we obtain, correspondingly,

$$M_i^0 = \frac{h_i}{6}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \quad M_i^1 = \frac{1}{h_i}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

Example. Now, let α and β be arbitrary functions. Then, the integrals cannot be computed analytically and we need to use quadrature formulas. For this example, we shall use Simpson's rule:

$$\int_a^b F(x)dx \approx \frac{b-a}{6}\left[F(a) + 4F\left(\frac{a+b}{2}\right) + F(b)\right].$$

Then, from (3.2), we obtain

$$M_i^0 \approx \frac{h_i}{6}\left(\beta(x_{i-1})\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + 4\beta(x_{i-1}+h_i/2)\begin{bmatrix} 1/4 & 1/4 \\ 1/4 & 1/4 \end{bmatrix} + \beta(x_{i-1})\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}\right).$$

For the element stiffness matrix, using (3.3), we have

$$M_i^1 \approx \frac{1}{6h_i}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\big(\alpha(x_{i-1}) + 4\alpha(x_{i-1}+h_i/2) + \alpha(x_i)\big).$$

Let us note that when computing the integrals approximately, **we need to choose a quadrature formula, having at least the same order of convergence as the FEM approximation**. We shall get back to this question in much more detail later in the course.

Having obtained the element mass and stiffness matrices, we can assemble the global matrices in the usual way. We formulate the following algorithm for computing M.

ALGORITHM (ASSEMBLING GLOBAL MATRICES AND LOAD VECTORS):

1. Implement functions that compute the local element matrices and load vectors (those functions are specific for each particular problem);
2. Implement a function for assembling the global matrix M and the global load vector $\mathbf{b}$. For this purpose, iterate over the elements and for each element:
   2.1. compute the element matrices and vectors, using the functions in 1;
   2.2. add the results at the correct places in M and $\mathbf{b}$.

3.2 Element-wise computations in 2D

3.2.1 Preliminaries from Multivariable Calculus

We shall first remind a few fundamental results from multivariable calculus, concerning change of variables. Consider the following change of variables

$$x = x(\xi, \eta),$$
$$y = y(\xi, \eta).$$

For what follows, let us think that the change is linear (this case will be of interest for us), even though the results are valid in the general case.

Proposition 20 (Chain rule)**.** *For sufficiently smooth functions, the following holds true:*

$$\frac{\partial u}{\partial x} = \frac{\partial u}{\partial \xi}\frac{\partial \xi}{\partial x} + \frac{\partial u}{\partial \eta}\frac{\partial \eta}{\partial x},$$
$$\frac{\partial u}{\partial y} = \frac{\partial u}{\partial \xi}\frac{\partial \xi}{\partial y} + \frac{\partial u}{\partial \eta}\frac{\partial \eta}{\partial y}.$$

Proof. We shall prove the first equality. The second is analogous. We linearize u and obtain (using standard notation)

$$\Delta u = \frac{\partial u}{\partial \xi}\Delta\xi + \frac{\partial u}{\partial \eta}\Delta\eta + O(|2|).$$

We divide both sides to Δx and let $\Delta x \to 0$ to obtain the first equation. □

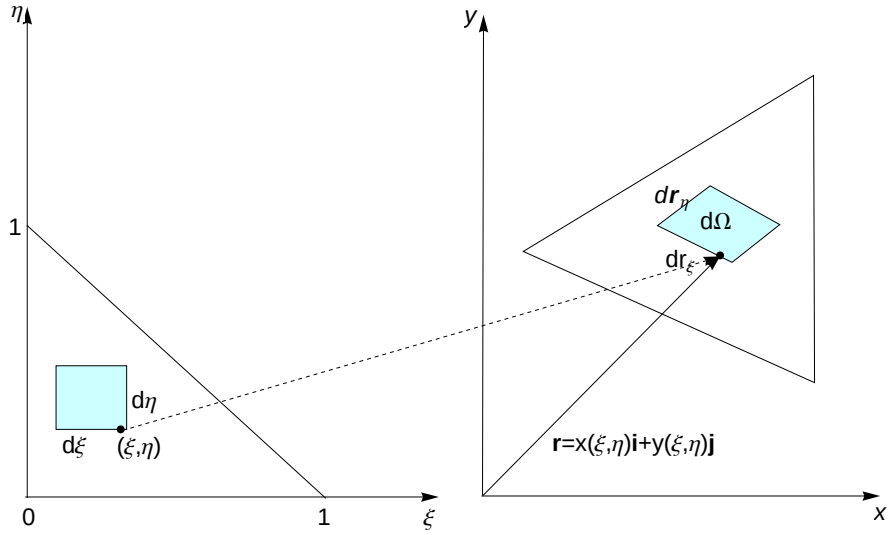
Proposition 21. *Let E be the image of a given region Ω under the transformation. Let*

$$J = det \begin{bmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial y}{\partial \xi} \\ \frac{\partial x}{\partial \eta} & \frac{\partial y}{\partial \eta} \end{bmatrix}.$$

Then,

$$\iint_\Omega u d\Omega = \iint_E u(x(\xi,\eta), y(\xi,\eta))|J| d\xi d\eta.$$

Proof. The only real challenge in proving the statement is to find the relation between $d\Omega$ and $d\xi d\eta$. Consider an infinitesimal rectangle, defined by $d\xi$ and $d\eta$, which is the image of $d\Omega$ under the transformation with respect to the new variables. Taking into account that linear change of variables preserves parallelism, $d\Omega$ must be a parallelogram.



We can parameterize the boundary of the parallelogram as

$$\mathbf{r} = x(\xi, \eta)\mathbf{i} + y(\xi, \eta)\mathbf{j},$$

where ξ, η lie on the boundary of the rectangle in $\xi - \eta$. Therefore, for the sides of the parallelogram, we obtain

$$d\mathbf{r}_\xi = \left[\frac{\partial x}{\partial \xi}\mathbf{i} + \frac{\partial y}{\partial \xi}\mathbf{j}\right] d\xi$$

and

$$d\mathbf{r}_\eta = \left[\frac{\partial x}{\partial \eta}\mathbf{i} + \frac{\partial y}{\partial \eta}\mathbf{j}\right] d\eta.$$

Therefore,

$$d\Omega = |d\mathbf{r}_\xi \times d\mathbf{r}_\eta| = |J| d\xi d\eta.$$

□

3.2.2 Model problem

Now, we are ready to discuss how the linear algebraic system is assembled in 2D on the basis of the following example.

$$\begin{aligned} -\Delta u &= f, \text{ in } \Omega, \\ \frac{\partial u}{\partial \mathbf{n}} &= 0, \text{ on } \partial\Omega. \end{aligned} \tag{D}$$

The corresponding Ritz–Galerkin problem, obtained in the usual way, is (Check!):

Find $u_h \in V_h$, such that

$$a(u_h, v) = F(v), \ \forall v \in V_h, \tag{R.–G.}$$

where

$$a(u, v) = \iint_\Omega \nabla u \cdot \nabla v d\Omega, \quad F(v) = \iint_\Omega f v d\Omega.$$

We are, therefore, interested in solving the linear algebraic system $M^1\mathbf{q} = \mathbf{b}$, where

$$M^1 = \iint_\Omega \begin{bmatrix} \nabla\varphi_1 \cdot \nabla\varphi_1 & \cdots & \nabla\varphi_1 \cdot \nabla\varphi_N \\ \vdots & \ddots & \vdots \\ \nabla\varphi_N \cdot \nabla\varphi_1 & \cdots & \nabla\varphi_N \cdot \nabla\varphi_N \end{bmatrix} d\Omega, \quad \mathbf{b} = \iint_\Omega \begin{bmatrix} f\varphi_1 \\ \vdots \\ f\varphi_N \end{bmatrix} d\Omega.$$

3.2.3 Computing the element matrices

In order to implement FEM, derived in the previous section, in practice, we need to assemble the stiffness matrix and the load vector. We do this, as usual, working one element at a time. We have

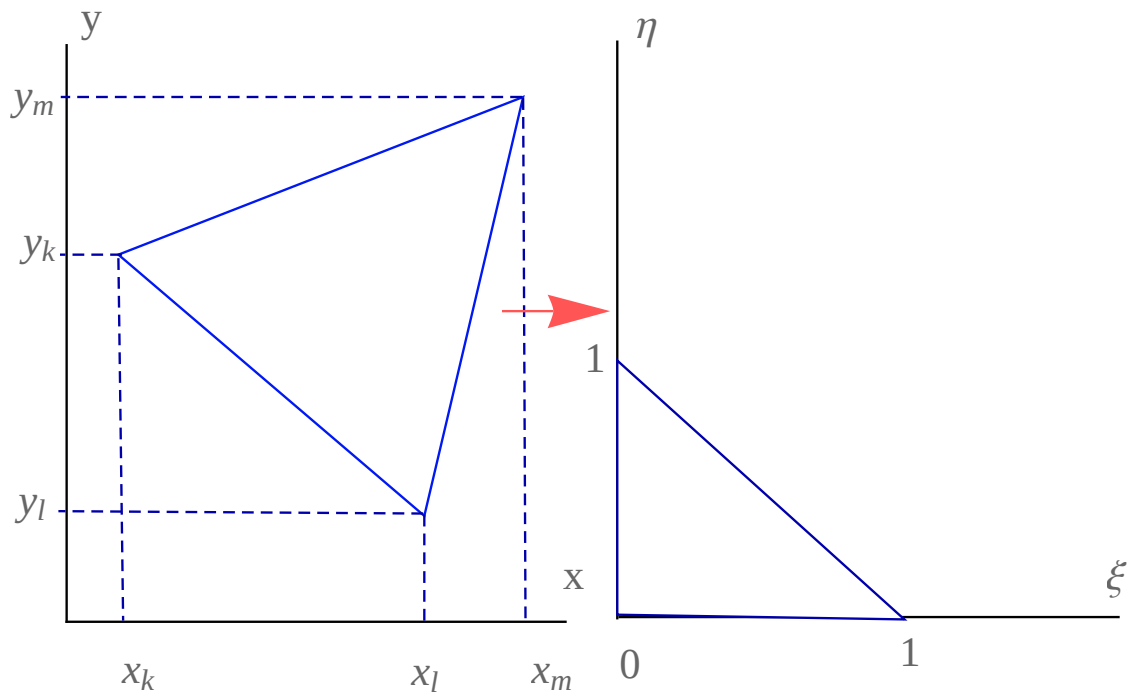
$$\begin{aligned} M^1 &= \iint_\Omega \begin{bmatrix} \nabla\varphi_1 \cdot \nabla\varphi_1 & \cdots & \nabla\varphi_1 \cdot \nabla\varphi_N \\ \vdots & \ddots & \vdots \\ \nabla\varphi_N \cdot \nabla\varphi_1 & \cdots & \nabla\varphi_N \cdot \nabla\varphi_N \end{bmatrix} d\Omega \\ &= \sum_{\tau \in \mathcal{K}} \iint_\tau \begin{bmatrix} \nabla\varphi_1 \cdot \nabla\varphi_1 & \cdots & \nabla\varphi_1 \cdot \nabla\varphi_N \\ \vdots & \ddots & \vdots \\ \nabla\varphi_N \cdot \nabla\varphi_1 & \cdots & \nabla\varphi_N \cdot \nabla\varphi_N \end{bmatrix} d\Omega. \end{aligned}$$

Therefore, we are interested in computing the following 3×3 matrices (containing the only non-zero elements in the global element matrices):

$$M_\tau^1 = \iint_\tau \begin{bmatrix} \nabla\varphi_k\cdot\nabla\varphi_k & \nabla\varphi_k\cdot\nabla\varphi_l & \nabla\varphi_k\cdot\nabla\varphi_m \\ \nabla\varphi_l\cdot\nabla\varphi_k & \nabla\varphi_l\cdot\nabla\varphi_l & \nabla\varphi_l\cdot\nabla\varphi_m \\ \nabla\varphi_m\cdot\nabla\varphi_k & \nabla\varphi_m\cdot\nabla\varphi_l & \nabla\varphi_m\cdot\nabla\varphi_m \end{bmatrix} d\Omega,$$

where the element τ is defined by nodes k, l, m with coordinates (x_k, y_k), (x_l, y_l), (x_m, y_m).

We shall make those computations by making a transformation to the standard element:



This has serious benefits. First, the shape functions are well-known and do not depend on the coordinates of the vertices. Furthermore, quadrature formulas can be easily found for the standard element.

The change in question is

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x_l - x_k & x_m - x_k \\ y_l - y_k & y_m - y_k \end{bmatrix} \begin{bmatrix} \xi \\ \eta \end{bmatrix} + \begin{bmatrix} x_k \\ y_k \end{bmatrix}. \tag{3.4}$$

Exercise. Derive the change of variables, by taking into account the correspondence between points with (x, y)-coordinates and (ξ, η)-coordinates.

Based on Section 3.2.1, we need to make some computations, in order to be ready to make the change for the gradient and the integral. By differentiating (3.4) with respect to ξ and η, we correspondingly obtain

$$\begin{bmatrix} \frac{\partial x}{\partial \xi} \\ \frac{\partial y}{\partial \xi} \end{bmatrix} = \begin{bmatrix} x_l - x_k \\ y_l - y_k \end{bmatrix}, \quad \begin{bmatrix} \frac{\partial x}{\partial \eta} \\ \frac{\partial y}{\partial \eta} \end{bmatrix} = \begin{bmatrix} x_m - x_k \\ y_m - y_k \end{bmatrix}.$$

Therefore

$$J = \begin{bmatrix} \dfrac{\partial x}{\partial \xi} & \dfrac{\partial y}{\partial \xi} \\ \dfrac{\partial x}{\partial \eta} & \dfrac{\partial y}{\partial \eta} \end{bmatrix} = \begin{bmatrix} x_l - x_k & y_l - y_k \\ x_m - x_k & y_m - y_k \end{bmatrix}.$$

The latter is needed for the change under the integral.

We also need to be able to make the following change:

$$\begin{bmatrix} \dfrac{\partial}{\partial x} \\ \dfrac{\partial}{\partial y} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial \xi}{\partial x} & \dfrac{\partial \eta}{\partial x} \\ \dfrac{\partial \xi}{\partial y} & \dfrac{\partial \eta}{\partial y} \end{bmatrix} \begin{bmatrix} \dfrac{\partial}{\partial \xi} \\ \dfrac{\partial}{\partial \eta} \end{bmatrix}.$$

By differentiating (3.4) with respect to x and y and then using Cramer's rule, we obtain

$$\begin{bmatrix} \dfrac{\partial}{\partial x} \\ \dfrac{\partial}{\partial y} \end{bmatrix} = \frac{1}{\det J} \underbrace{\begin{bmatrix} y_m - y_k & y_k - y_l \\ x_k - x_m & x_l - x_k \end{bmatrix}}_{B} \begin{bmatrix} \dfrac{\partial}{\partial \xi} \\ \dfrac{\partial}{\partial \eta} \end{bmatrix}.$$

Having the above results in mind, we are ready to compute the elements of M_τ^1. All of them have the form

$$\iint_\Omega \nabla_{(x,y)} \varphi_\alpha \cdot \nabla_{(x,y)} \varphi_\beta d\Omega,$$

where α, β are among k, l, m. Making a change to the standard element, we obtain consecutively

$$\begin{aligned} \iint_\Omega \nabla_{(x,y)} \varphi_\alpha \cdot \nabla_{(x,y)} \varphi_\beta d\Omega &= \iint_\Omega \left(\begin{bmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{bmatrix} \varphi_\alpha \right) \cdot \left(\begin{bmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \end{bmatrix} \varphi_\beta \right) d\Omega \\ &= \iint_E \left(\frac{1}{\det J} B \begin{bmatrix} \frac{\partial}{\partial \xi} \\ \frac{\partial}{\partial \eta} \end{bmatrix} \psi_\alpha \right) \cdot \left(\frac{1}{\det J} B \begin{bmatrix} \frac{\partial}{\partial \xi} \\ \frac{\partial}{\partial \eta} \end{bmatrix} \psi_\beta \right) |\det J| d\xi d\eta \\ &= \frac{1}{|\det J|} \iint_E \left(B \nabla_{(\xi,\eta)} \psi_\alpha \right) \cdot \left(B \nabla_{(\xi,\eta)} \psi_\beta \right) d\xi d\eta \\ &= \frac{1}{|\det J|} \iint_E (\nabla_{(\xi,\eta)} \psi_\alpha)^T B^T B \nabla_{(\xi,\eta)} \psi_\beta d\xi d\eta, \end{aligned}$$

where ψ_α, ψ_β must be interpreted in the following way: $\varphi_k \to \psi_0$, $\varphi_l \to \psi_1$, $\varphi_m \to \psi_2$.

Therefore, the matrix M_τ^1 is

$$M_\tau^1 = \frac{1}{|\det J|} \iint_E (\nabla \mathbf{\Psi})^T B^T B \nabla \mathbf{\Psi} d\xi d\eta,$$

where $\mathbf{\Psi} = (\psi_0, \psi_1, \psi_2)$ and

$$\nabla \mathbf{\Psi} = \begin{bmatrix} \dfrac{\partial}{\partial \xi} \\ \dfrac{\partial}{\partial \eta} \end{bmatrix} (\psi_0, \psi_1, \psi_2) = \begin{bmatrix} \dfrac{\partial \psi_0}{\partial \xi} & \dfrac{\partial \psi_1}{\partial \xi} & \dfrac{\partial \psi_2}{\partial \xi} \\ \dfrac{\partial \psi_0}{\partial \eta} & \dfrac{\partial \psi_1}{\partial \eta} & \dfrac{\partial \psi_2}{\partial \eta} \end{bmatrix} = \begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}.$$

For the load vector, we obtain

$$\mathbf{b}_\tau = \iint_\tau f(x,y) \begin{bmatrix} \varphi_k \\ \varphi_l \\ \varphi_m \end{bmatrix} d\Omega = \iint_E f(x(\xi,\eta), y(\xi,\eta)) \mathbf{\Psi}^T |det\ J| d\xi d\eta.$$

As a separate example, we shall show that we can analogously compute an element mass matrix. The following equality obviously holds:

$$\iint_\Omega \varphi_\alpha \varphi_\beta d\Omega = \iint_E \psi_\alpha \psi_\beta |\det J| d\xi d\eta,$$

and, therefore, the local mass matrix is

$$\begin{aligned} M^0_\tau &= |\det J| \iint_E \begin{bmatrix} \psi_0^2 & \psi_0\psi_1 & \psi_0\psi_1 \\ \psi_1\psi_0 & \psi_1^2 & \psi_1\psi_2 \\ \psi_2\psi_0 & \psi_2\psi_1 & \psi_2^2 \end{bmatrix} d\xi d\eta \\ &= |\det J| \iint_E \mathbf{\Psi}^T \mathbf{\Psi} d\xi d\eta. \end{aligned}$$

Exercise. Make the necessary computations to show that the latter result agrees with what we stated in Section 2.1.4 for M_τ.

3.2.4 Element-wise computations for more general boundary conditions

For more general boundary conditions, there might be additional contributions to the left-hand side matrix and the load vector that include integrals over the boundary. Consider the following example

$$\begin{aligned} -\nabla \cdot [(x^2+y^2+1)\nabla u] &= f,\ x \in \Omega, \\ -(x^2+y^2+1)\frac{\partial u}{\partial \mathbf{n}} &= u,\ x \in \partial\Omega. \end{aligned} \tag{D}$$

The corresponding variational problem becomes (Check!):
Find $u \in H^1(\Omega)$, such that

$$a(u,v) = F(v), \tag{V}$$

where

$$a(u,v) = \iint_\Omega (x^2+y^2+1)\nabla u \cdot \nabla v d\Omega + \int_{\partial\Omega} uv ds, \quad F(v) = \iint_\Omega f v d\Omega.$$

Approximating the latter variational problem in the finite-dimensional subspace V_h, we obtain eventually the linear algebraic system

$$(M^1 + \Gamma)\mathbf{q} = \mathbf{b}.$$

Here, M^1 and $\mathbf{b}$ are the stiffness matrix and load vector. We have already shown in the previous section how to assemble those. Let us consider the matrix

$$\Gamma = \int_{\partial\Omega} \begin{bmatrix} \varphi_1^2 & \varphi_2\varphi_1 & \cdots & \varphi_N\varphi_1 \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1\varphi_N & \varphi_2\varphi_N & \cdots & \varphi_N^2 \end{bmatrix} ds.$$

Following our usual procedure, we can write it as a sum over all boundary segments γ in the triangulation:

$$\Gamma = \sum_{\gamma} \int_{\gamma} \begin{bmatrix} \varphi_1^2 & \varphi_2\varphi_1 & \cdots & \varphi_N\varphi_1 \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_1\varphi_N & \varphi_2\varphi_N & \cdots & \varphi_N^2 \end{bmatrix} ds.$$

For the local matrix that corresponds to the boundary segment $\gamma(lm)$, defined by nodes $l = (x_l, y_l), m = (x_m, y_m)$, we are interested only in the 2×2 non-zero elements

$$\Gamma_{lm} = \int_{\gamma(lm)} \begin{bmatrix} \varphi_l^2 & \varphi_l\varphi_m \\ \varphi_m\varphi_l & \varphi_m^2 \end{bmatrix} ds.$$

We make a change of variables, so that the segment, defined by nodes l, m is transformed into the standard 1D element, $\xi \in [0, 1]$, over which the basis functions φ_l, φ_m are transformed into $\psi_{\gamma,0} = 1 - \xi$ and $\psi_{\gamma,1} = \xi$. The change of variables is obviously

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x_l \\ y_l \end{bmatrix} + \xi \begin{bmatrix} x_m - x_l \\ y_m - y_l \end{bmatrix}$$

and, therefore,

$$ds = \sqrt{(x_m - x_l)^2 + (y_m - y_l)^2} d\xi.$$

For the local matrix, we finally obtain

$$\int_{\gamma(lm)} \begin{bmatrix} \varphi_l^2 & \varphi_l\varphi_m \\ \varphi_m\varphi_l & \varphi_m^2 \end{bmatrix} ds = \int_0^1 \begin{bmatrix} \psi_{\gamma,0}^2 & \psi_{\gamma,0}\psi_{\gamma,1} \\ \psi_{\gamma,1}\psi_{\gamma,0} & \psi_{\gamma,1}^2 \end{bmatrix} \sqrt{(x_m - x_l)^2 + (y_m - y_l)^2} d\xi$$

3.2.5 Quadrature formulae for the standard triangular element

In order to compute the local element matrices and load vectors, we need to compute integrals over the standard triangle. Here, we summarize several quadrature formulas:

Nodes	Weights	Exact for polynomials of degree
(1/2,1/2)	1/2	1
(0,0),(1,0),(0,1)	1/6, 1/6, 1/6	1
(1/2,0),(1/2,1/2),(0,1/2)	1/6, 1/6, 1/6	2
(0,0), (1,0), (0,1) (1/2,0),(1/2,1/2),(0,1/2) (1/3,1/3)	3/120, 3/120, 3/120 8/120,8/120,8/120 27/120	3

Example. If we compute approximately the matrix

$$M = \iint_E \begin{bmatrix} \psi_0^2 & \psi_0\psi_1 & \psi_0\psi_1 \\ \psi_1\psi_0 & \psi_1^2 & \psi_1\psi_2 \\ \psi_2\psi_0 & \psi_2\psi_1 & \psi_2^2 \end{bmatrix} d\xi d\eta,$$

using the second formula, we obtain

$$M \approx \frac{1}{6}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \frac{1}{6}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \frac{1}{6}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Exercise. Which formula would guarantee exact computation of the integral in the last example?

3.3 Imposing Dirichlet boundary conditions

When we have Dirichlet boundary conditions, there is a certain inconvenience, when we try to assemble the linear algebraic system. Usually, we write the matrix of the system as a sum of element matrices. The latter are computed by making a change to a standard element, in order to unify the computations. However, when Dirichlet boundary conditions are imposed, the variational problem is solved in H_0^1, i.e. the basis functions, corresponding to boundary nodes, are excluded. Let us reconsider the 1D example from Section 1.2.1. We need to solve the system

$$\int_I \begin{bmatrix} \varphi_1'^2 & \cdots & \varphi_1'\varphi_{n-1}' \\ \vdots & \ddots & \vdots \\ \varphi_1'\varphi_{n-1}' & \cdots & \varphi_{n-1}'^2 \end{bmatrix} dx \begin{bmatrix} q_1 \\ \vdots \\ q_{n-1} \end{bmatrix} = \int_I f \begin{bmatrix} \varphi_1 \\ \vdots \\ f, \varphi_{n-1} \end{bmatrix} dx.$$

If we start computing the element matrices as usual, then the matrix that corresponds to element I_1 is different than the rest, because there is only one non-zero element ($\int_I \varphi_1'^2 dx$). The same holds for the element matrix for I_n.

In order to be able to use our general approach, in practice we assemble the matrix as usual (as if we had Neumann boundary conditions) and obtain an $(n+1)\times(n+1)$ linear system:

$$\int_I \begin{bmatrix} \varphi_0'^2 & \cdots & \varphi_0'\varphi_n' \\ \vdots & \ddots & \vdots \\ \varphi_0'\varphi_n' & \cdots & \varphi_n'^2 \end{bmatrix} dx \begin{bmatrix} q_0 \\ \vdots \\ q_n \end{bmatrix} = \int_I f \begin{bmatrix} \varphi_0 \\ \vdots \\ \varphi_n \end{bmatrix} dx.$$

Let $\tilde{m}_{ij}$ and $\tilde{b}_i$, $i = \overline{0,n}$, $j = \overline{0,n}$, denote the elements of the so-obtained matrix and right-hand side vector, correspondingly.

Only then, we impose the boundary conditions by changing the first and last rows of the system as well as the first and last column as follows:

$$\begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & \tilde{m}_{11} & \cdots & \tilde{m}_{1,n-1} & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \tilde{m}_{n-1,1} & \cdots & \tilde{m}_{n-1,n-1} & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} q_0 \\ \vdots \\ q_n \end{bmatrix} = \begin{bmatrix} 0 \\ \tilde{b}_1 \\ \vdots \\ \tilde{b}_{n-1} \\ 0 \end{bmatrix}.$$

The change in the columns is made, so that the structure of the matrix is kept symmetric.

If the boundary conditions are not homogeneous, but

$$u(0) = u_{left},\ u(1) = u_{right},$$

then the system becomes

$$\begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & \tilde{m}_{11} & \cdots & \tilde{m}_{1,n-1} & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \tilde{m}_{n-1,1} & \cdots & \tilde{m}_{n-1,n-1} & 0 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} q_0 \\ \vdots \\ q_n \end{bmatrix} = \begin{bmatrix} u_{left} \\ \tilde{b}_1 - \tilde{m}_{10}u_{left} - \tilde{m}_{1n}u_{right} \\ \vdots \\ \tilde{b}_{n-1} - \tilde{m}_{n-1,0}u_{left} - \tilde{m}_{n-1,n}u_{right} \\ u_{right} \end{bmatrix}.$$

We can proceed analogously in the 2D case, by changing the rows and columns of the matrix, that correspond to the boundary nodes, after we have assembled as if there were Neumann boundary conditions.

Chapter 4

FEM for time-dependent problems

Until now, we have only considered stationary problems. This is, in some sense, natural, since FEM is usually a method for spatial discretization. We shall now see what can it give us for time-dependent problems. We shall discuss the ideas on the basis of two examples—the linear 1D diffusion equation and the 2D wave equation. Thus, by the end of this section we will have covered the main linear second-order linear PDEs that serve as a foundation for many mathematical models, used in practice. We shall postpone the transport equation until the second part of the course, since there are certain problems, concerning solving it numerically.

Furthermore, we shall only consider homogeneous Dirichlet boundary conditions in our examples, since our focus will be on the time-dependent part of the problem. We already know what to do in cases when the boundary conditions are different.

4.1 FEM for the 1D linear diffusion/heat equation

We consider the following differential problem over the domain $x \in I := [0, L]$, $y \in J := [0, T]$:

$$\begin{aligned} &\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = f, \\ &u(0,t) = u(L,t) = 0, \\ &u(x,0) = u_0(x). \end{aligned} \tag{D}$$

We proceed in the usual way to obtain the variational formulation. We multiply both sides with a test function v and integrate **with respect to the spatial variable**.

For the left hand-side we obtain

$$\begin{aligned} lhs &= \int_I \left(\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} \right) v dx \\ &= \int_I \frac{\partial u}{\partial t} v dx + \int_I \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} dx - \cancelto{0,\ \text{if } \forall t \in J, v(\cdot,t) \in H_0^1(I)}{\frac{\partial u}{\partial x} v \Big|_0^L}. \end{aligned}$$

Then, the variational problem becomes:

For every $t \in J$, find $u(\cdot, t) \in H_0^1(I)$, such that

$$\left(\frac{\partial u}{\partial t}, v \right) + a(u, v) = F(v), \forall v \in H_0^1, \tag{V}$$

where

$$a(u,v) := \int_I \frac{\partial u}{\partial x}\frac{\partial v}{\partial x}dx, \quad F(v) = \int_I fvdx.$$

Remark 19. In the variational equation, we treat t as something fixed and then we want to solve the variational equation for every fixed t. Thus, in some sense, we have decomposed the problem into a spatial and a time-dependent problem.

Remark 20. Note that the left-hand side is not symmetric anymore.

Further, **we discretize the spatial domain** in the usual way (i.e., **we obtain a semi-discretization of the numerical domain $\boldsymbol{I \times J}$**) and look for an approximate solution of the form

$$u_h = \sum_{j=1}^{n-1} q_j(t)\varphi_j(x),$$

which is a piecewise-linear polynomial for every fixed t. In the latter, $\varphi_j(x)$, $j = \overline{1, n-1}$ are the hat-functions, corresponding to the internal nodes of the discretization of I. Then, we can approximate the variational problem (V) with the following one:[1]

For every $t \in J$, find the function u_h, such that $V_{h,0} \ni u_h(\cdot, t) = \sum_{j=1}^{n-1} q_j(t)\varphi_j(\cdot)$, and

$$\left(\frac{\partial u_h}{\partial t}, v\right) + a(u_h, v) = F(v), \ \forall v \in V_{h,0}. \qquad \text{(R.–G.)}$$

This leads to the equivalent semi-discrete problem:

For every $t \in J$, find $(q_1(t), \ldots, q_{n-1}(t)) \in \mathbb{R}^{n-1}$, such that

$$\sum_{j=1}^{n-1} (\varphi_j, \varphi_i) \frac{dq_j}{dt}(t) + \sum_{j=1}^{n-1} a(\varphi_j, \varphi_i) q_j(t) = F(\varphi_i), \quad i = \overline{1, n-1}.$$

If we denote $\mathbf{q}(t) = (q_1(t), \ldots, q_{n-1}(t))^T$, we can write the problem compactly as

$$M^0 \frac{d\mathbf{q}}{dt} + M^1 \mathbf{q} = \mathbf{b}, \qquad (4.1)$$

where M^0 and M^1 are the usual mass and stiffness matrices, respectively, and $\mathbf{b}$ is the usual load vector. The differential problem can be closed with

$$\mathbf{q}(0) = \mathbf{q}_0,$$

where $\mathbf{q}_0$ are the coefficients in a piecewise linear approximation $u_{h,0}(x)$ of $u_0(x)$. For instance, we can choose $u_{h,0}(x)$ to be the piecewise-linear interpolant or the L_2-projection of $u_0(x)$ onto $V_{h,0}$. We shall discuss this in a little more detail when we discuss stability and convergence.

Thus, using FEM for the time-dependent problem, we have obtained a spatial discretization and we have reduced the original PDE boundary-value problem to an initial-value problem for a system of ODEs. In this

[1] It would be more correct to call the obtained semi-discrete problem a Faedo–Galerkin formulation, but we shall continue to use the abbreviation R.–G.

case the ODEs are linear, since the original problem is also linear. A remarkable fact is that in the ODE problem the same mass and stiffness matrices take part.

We shall give two examples, concerning how we can solve numerically the arising ODE system (4.1). For simplicity, we shall assume that f does not depend on time and, therefore $\mathbf{b}$ is constant.

Example. Let us introduce the uniform time discretization

$$0 = t_0 < \cdots < t_m = L$$

and let τ be the time discretization step. If we use the backward Euler method, we obtain

$$M^0 \frac{\mathbf{Q}_{i+1} - \mathbf{Q}_i}{\tau} + M^1 \mathbf{Q}_{i+1} = \mathbf{b},$$

where $\mathbf{Q}_i$ is the approximate solution at time t_i, $i = \overline{0, m-1}$. We start with $\mathbf{Q}_0 = \mathbf{q}_0$ and then solve the resulting system on each time layer.

Example. Using the improved Euler method, we obtain the following Crank–Nicolson scheme. The difference with respect to the previous example is that the linear algebraic system that needs to be solved on each time layer is

$$M^0 \frac{\mathbf{Q}_{i+1} - \mathbf{Q}_i}{\tau} + M^1 \frac{\mathbf{Q}_{i+1} + \mathbf{Q}_i}{2} = \mathbf{b}.$$

In both examples, we have approximated the ODE system and obtained a linear algebraic system that needs to be solved.

Actually, we can use any numerical method for first-order ODEs (e.g., Euler methods, Runge–Kutta methods, Adams-like methods, etc.), if we put the problem in the appropriate form.

Inverting the mass matrix, however, destroys its sparse structure. One possible solution to this problem that is done from practical considerations, is the so-called **mass lumping**. This means to add all off-diagonal elements of the mass matrix to the diagonal one and, thus, obtain a diagonal approximation to the original matrix. Then, inverting it, will also result in a diagonal matrix.

4.2 Stability and convergence for the semi-discrete problem

One key concept, concerning the numerical solution of time-dependent problems is the one of stability. We need to understand what is the effect of some errors that are introduced in the computations (e.g. round-off errors, approximation errors, etc.) on the final result. We consider the effect of the non-exact representation of the initial data and right-hand side for the example problem from the previous section (the boundary conditions were homogeneous there and can, therefore, be represented exactly). Let us remark that an error can be obtained at any time-layer (not only in the initial data), but we can use the same results with this time-layer as "initial".

In this section, $\|\cdot\|$ will mean the L_2-norm.

First, we shall prove that the following stability result holds true.

Proposition 22. *For the FEM solution, it holds that*

$$\|u_h(\cdot,t)\| \leq \|u_h(\cdot,0)\| + \int_0^t \|f(\cdot,s)\|ds,$$

i.e. it is stable with respect to initial conditions and right-hand side.

Remark 21. Before we prove this result, let us first clarify what it means. Let us consider the solution $\tilde{u}_h$ of the semi-discrete problem with slightly perturbed initial data $\tilde{u}_h(x,0)$ and right-hand side $\tilde{f}$. Then, the difference between the two solutions $u_h - \tilde{u}_h$ satisfies the problem with initial data $u_h(x,0) - \tilde{u}_h(x,0)$ and right-hand side $f - \tilde{f}$ and is, therefore, bounded by

$$\|u_h(\cdot,t) - \tilde{u}_h(\cdot,t)\| \leq \|u_h(\cdot,0) - \tilde{u}_h(\cdot,0)\| + \int_0^t \|f(\cdot,s) - \tilde{f}(\cdot,s)\|ds.$$

In this sense, small perturbations in the initial data and right-hand side (e.g. from round-off errors) have small effect on the solution and it is stable.

Now, we are ready to continue with the proof of the stability estimate.

Proof. Since (R.–G.) holds for every $v \in V_{h,0}$, it holds, in particular, for $v = u_h$, i.e.

$$\int_I (\dot{u}_h u_h + (u_h')^2)dx = \int_I f u_h dx \leq \|f\|\|u_h\|. \tag{4.2}$$

Let us work a little with the left-hand side. We obtain consecutively

$$\begin{aligned}\int_I (\dot{u}_h u_h + (u_h')^2)dx &= \int_I \left[\frac{1}{2}\frac{\partial (u_h)^2}{\partial t} + (u_h')^2\right] \\ &= \frac{1}{2}\frac{\partial}{\partial t}(\|u_h\|^2) + \|u_h'\|^2 \\ &\geq \|u_h\|\frac{\partial \|u_h\|}{\partial t}.\end{aligned}$$

From the latter inequality and (4.2), we obtain

$$\frac{\partial \|u_h\|}{\partial t} \leq \|f\|.$$

Now, using the triangle inequality, the statement of the proposition follows directly. □

For the classical finite difference methods, it is well-known that

$$\text{consistency} + \text{stability} = \text{convergence}.$$

It turns out that we can show the same here.

Proposition 23. *For the FEM solution, it holds that for every $t \in J$,*

$$\|u(t) - u_h(t)\| \leq Ch^2 \left(\|u_0''\| + \int_0^t \|\dot{u}''(\cdot,s)\|ds\right).$$

Proof. As we mentioned, convergence is a result of consistency and stability. We shall incorporate this idea by decomposing the error into two parts:

$$u - u_h = \underbrace{(u - R_h u)}_{consistency} + \underbrace{(R_h u - u_h)}_{stability} =: \rho + \theta,$$

where $R_h u$ is the so-called Ritz projection of u onto $V_{h,0}$, i.e.

$$((u - R_h u)', v') = 0, \ \forall v \in V_{h,0}.$$

Our reasoning behind the above representation of the error is the following. We shall show that we can approximate u with something (the Ritz projection) from $V_{h,0}$ (consistency of (R.–G) with (V)) and then we shall show that, starting close to the Ritz projection, we stay close to it (stability) and, therefore, stay close to u. Stating the same thing in a formal way, we shall bound separately ρ and θ by using approximation theory and our stability result, respectively.

First, for the Ritz projection, taking into account that it is actually the solution of the model problem in Section 1.2.1, we have the following result (see Section 1.3.3) for sufficiently regular problems:

$$\begin{aligned}
\|\rho(\cdot, t)\| &\leq Ch^2 \|u''(\cdot, t)\| \\
&= Ch^2 \left\| u''(\cdot, 0) + \int_0^t \dot{u}''(\cdot, s) ds \right\| \\
&\leq Ch^2 \left(\|u_0''(\cdot)\| + \int_0^t \|\dot{u}''(\cdot, s)\| ds \right).
\end{aligned}$$

Considering θ, let us note that it is a solution of (R.–G.) with right-hand side $f = -\dot{\rho}$ (Check!). Thus, it satisfies the stability estimate

$$\begin{aligned}
\|\theta(\cdot, t)\| &\leq \int_0^t \|\dot{\rho}(\cdot, s)\| ds + \cancelto{0,\text{ assuming that } u_h(\cdot,0) = R_h u_0(\cdot)}{\|\theta(\cdot, 0)\|} \\
&= \int_0^t \frac{\partial}{\partial t} \|\rho(\cdot, s) ds\| \\
&\leq Ch^2 \int_0^t \|\dot{u}''(\cdot, s)\| ds.
\end{aligned}$$

□

Remark 22. From the above proof, it follows that the best initial condition for the semi-discrete problem is the Ritz projection of $u_0(x)$. In order to guarantee second-order accuracy, we need to either use this initial condition or an initial condition that approximates the Ritz projection with second-order accuracy.

Remark 23. We can use what we know about the FEM solution of the corresponding stationary problem (i.e. the Ritz projection), in order to study the convergence for the time-dependent problem. Therefore, we shall concentrate on studying the error for the stationary problems.

Remark 24. We have derived stability and convergence results for the semi-discrete solution (for which FEM is responsible). In order to obtain convergence for the fully-discrete solution, we must also ensure that the numerical method, used for time-discretization, is stable and has an appropriate convergence rate.

We can also derive similar stability results in other norms, e.g. H^1-norm.

4.3 FEM for the 2D linear wave equation

Let us now consider a 2D example, as well.

$$\begin{aligned} \frac{\partial^2 u}{\partial t^2} - c^2 \Delta u &= f, \quad \text{in } \Omega \times J, \\ u &= 0, \text{ on } \partial\Omega \times J, \\ u &= u_0, \text{ in } \Omega \text{ for } t = 0, \\ \frac{\partial u}{\partial t} &= v_0, \text{ in } \Omega \text{ for } t = 0. \end{aligned} \tag{D}$$

Since there is nothing really new, we shall only sketch the FEM formulation.

Using our standard procedure, we obtain the following variational problem:

For each $t \in J$, find $u(\cdot, t) \in H_0^1(\Omega)$, such that

$$\left(\frac{\partial^2 u}{\partial t^2}, v \right) + a(u, v) = F(v), \ \forall v \in H_0^1(\Omega), \tag{V}$$

where

$$a(u, v) = c^2 \iint_\Omega \nabla u \cdot \nabla v d\Omega, \ F(v) = \iint_\Omega f v d\Omega.$$

We approximate (V) with the problem

For each $t \in J$, find $u_h = \sum_{j=1}^{N_{inter}} q_j(t) \varphi_j(\mathbf{x}) \in V_{h,0}(\mathcal{K})$, such that

$$\left(\frac{\partial^2 u_h}{\partial t^2}, v \right) + a(u_h, v) = F(v), \ \forall v \in V_{h,0}(\mathcal{K}), \tag{V}$$

where $\varphi_1(\mathbf{x}), \ldots, \varphi_{N_{inter}}(\mathbf{x})$ correspond to the interior nodes of the triangulation $\mathcal{K}$ and $V_{h,0}(\mathcal{K}) = span(\varphi_1(\mathbf{x}), \ldots, \varphi_{N_{inter}}(\mathbf{x}))$.

Using the ansatz form of u_h, we obtain the following semi-discrete problem—an ODE system for the unknown coefficients $(q_1(t), \ldots, q_{N_{inter}}(t))^T =: \mathbf{q}(t)$.

For each $t \in J$, find $\mathbf{q}(t)$, such that

$$M^0 \frac{d^2 \mathbf{q}}{dt^2} + M^1 \mathbf{q} = \mathbf{b}.$$

The mass matrix and the load vector have their usual forms. The elements of the stiffness matrix are

$$M^1_{ij} = c^2 \iint_\Omega \nabla \varphi_i \cdot \nabla \varphi_j d\Omega.$$

We can reduce the latter system to the following first-order ODE system:

$$\begin{aligned} \frac{d\mathbf{q}}{dt} &= \boldsymbol{\xi}, \\ M^0 \frac{d\boldsymbol{\xi}}{dt} &= -M^1 \mathbf{q} + \mathbf{b}. \end{aligned}$$

Chapter 5

Обща теория на МКЕ за елиптични задачи

Абстрактната вариационна постановка, която разглеждаме, е следната.

Нека V е Хилбертово пространство с въведено скаларно произведение $(\cdot,\cdot)$ и породена от него норма $\|\cdot\|$. Нека $a : V \times V \to \mathbb{R}$ е билинейна форма, която е

(i) коерцитивна във V, т.е. $\exists \alpha > 0$:

$$a(u,u) \geq \alpha \|u\|^2, \ \forall u \in V;$$

(ii) непрекъсната във V, т.е. $\exists C_0$:

$$a(u,v) \leq C_0 \|u\| \|v\|, \ \forall u, v \in V.$$

Нека $F : V \to \mathbb{R}$ е линейна форма (линеен функционал), който е

(iii) непрекъснат във V, т.е. $\exists \Lambda$:

$$F(v) \leq \Lambda \|v\|, \forall v \in V.$$

Търсим $u \in V$:

$$a(u,v) = F(v), \ \forall v \in V.$$

5.1 Съществуване и единственост на решението на вариационната задача

Proposition 24 (Riesz Representation Theorem)**.** *Нека V е Хилбертово пространство с въведено скаларно произведение $(\cdot,\cdot)$. Тогава всяка непрекъсната линейна форма $F : V \to \mathbb{R}$ може да се представи по единствен начин във вида*

$$F(\cdot) = (v, \cdot)$$

за някое фиксирано $v \in V$.

Proof. Единствеността се доказва непосредствено. Да допуснем, че има две такива представяния:

$$F(\cdot) = (u_1, \cdot) \;\; F(\cdot) = (u_2, \cdot).$$

Тогава $(u_1 - u_2, v) = 0$ за всяко $v \in V$ и следователно $u_1 - u_2 = 0$.

Остава да докажем съществуването на такова представяне. Ако $F(\cdot) \equiv 0$, твърдението очевидно е изпълнено с $v = 0$. Иначе, $\mathcal{N}(F) \neq V$ и следователно има ненулеви елементи в $\mathcal{N}^\perp$. Нека изберем б.о.о. u_0 така, че $F(u_0) = 1$. Тогава

$$F(u - F(u)u_0) = 0, \;\; \forall u \in V,$$

т.е. $u - F(u)u_0 =: w \in \mathcal{N}$ и u може да се разложи по следния начин на компоненти в $\mathcal{N}$ и $\mathcal{N}^\perp$:

$$u = w + F(u)u_0.$$

Следователно, като умножим двете страни с $u_0 \in \mathcal{N}^\perp$, получаваме

$$F(u) = \frac{(u, u_0)}{(u_0, u_0)}.$$

Така доказахме теоремата с $v = \frac{u_0}{(u_0, u_0)}$. □

Proposition 25 (Lax–Milgram Theorem)**.** *Нека V е Хилбертово пространство с въведено скаларно произведение $(\cdot, \cdot)$ и нека $a(\cdot, \cdot)$ е билинейна форма, която е непрекъсната и коерцитивна във V. Нека $F(\cdot)$ е линейна форма, която е непрекъсната във V. Тогава съществува единствено решение $u \in V$ на вариационната задача*

$$a(u, v) = F(v), \;\; \forall v \in V.$$

При това е в сила следната априорна оценка за устойчивост:

$$\|u\|_V \leq \frac{\Lambda}{\alpha}.$$

Идея на доказателството. Ще покажем основната идея на едно от доказателствата за съществуване и единственост, като ще пропуснем някои детайли, които изискват по-задълбочени знания от теорията на Хилбертовите пространства.

Тъй като $F(\cdot)$ е непрекъсната линейна форма, съществува $b \in V$ такова, че

$$F(v) = (b, v), \;\; \forall v \in V.$$

Освен това $a(u, \cdot)$ е непрекъсната линейна форма във V и следователно съществува $w =: Au \in V$, така че

$$a(u, v) = (Au, v), \;\; \forall v \in V.$$

Може да се докаже, че операторът A е линеен и непрекъснат. Тогава вместо вариационната задача можем да разглеждаме линейното операторно уравнение

$$Au = b.$$

За да докажем, че последното има решение е достатъчно да докажем $Range(A) \equiv V$. Последното се получава като следствие от факта, че $\mathcal{N}(A) = \{0\}$. Действително,

да допуснем, че $z \in \mathcal{N}(A)$. Тогава от коерцитивността на билинейната форма имаме

$$\alpha\|z\|^2 \leq a(z,z) = (Az,z) \leq C_0\|Az\|\|z\|$$

и следователно

$$\|z\| \leq \frac{C_0}{\alpha}\|Az\| = 0 \Longrightarrow z = 0.$$

Остава да докажем единствеността. Нека $Au_1 = b$ и $Au_2 = b$, следователно $A(u_1 - u_2) = 0$. Отново използваме коерцитивността и получаваме

$$\alpha\|u_1 - u_2\|^2 \leq a(u_1 - u_2, u_1 - u_2) \leq \|A(u_1 - u_2)\|\|u_1 - u_2\| = 0.$$

□

Сега ще илюстрираме на базата на няколко примера как теоремата на Lax–Milgram може да се използва за доказване на съществуване и единственост на решението на дадена елиптична задача.

Example. Разглеждаме уравнението на Поасон с гранично условие на Дирихле:

$$-\Delta u = f,\ \mathbf{x} \in \Omega,$$
$$u = 0,\ \mathbf{x} \in \partial\Omega.$$

Съответната вариационна задача е с

$$a(u,v) = (\nabla u, \nabla v),\ F(v) = (f,v),$$

като Хилбертовото пространство е $V = H_0^1(\Omega)$. Бихме могли да нормираме това пространство, използвайки енергетична норма или H^1-норма. Разбира се, ще използваме тази норма, която ще ни даде резултата по-лесно. Знаем, че обикновено резултатите излизат почти непосредствено в енергетичната норма и затова ще използваме нея:

$$\|u\|_E^2 =< u, u >_E = a(u,u).$$

Ще докажем, че за вариационната задача са в сила условията на теоремата на Lax–Milgram:

(i) Коерцитивност на билинейната форма в E-норма следва директно от дефиницията:

$$a(u,u) \geq \alpha\|u\|_E$$

за $\alpha = 1$. Последното нестрого неравенство всъщност е винаги равенство.

(ii) Непрекъснатост на билинейната форма в E-норма следва веднага от неравенството на К.-Б.-Ш.:

$$a(u,v) =< u, v >_E \leq C_0\|u\|_E\|v\|_E,$$

където $C_0 = 1$.

(iii) Непрекъснатостта на линейната форма получаваме, като използваме неравенството на Poincaré (да забележим, че сме в просртранството H_0^1):

$$F(v) = (f,v) \leq \|f\|_{L_2}\|v\|_{L_2} \leq \underbrace{C\|f\|_{L_2}}_{\Lambda}\|v\|_E.$$

Следователно условията на теоремата са изпълнени и вариационната задача има единствено решение.

Remark 25. Да обърнем внимание, че в случаите, когато билинейната форма дефинира скаларно произведение, съществуването и единствеността следват директно и от теоремата на Riesz. Действително, от нея следва, че съществува единствен елемент $u \in V$ такъв, че

$$F(v) =< u, v >_E = a(u, v), \ \forall v \in V.$$

Сега ще приведем пример, в който граничните условия не са на Дирихле. За тази цел ще ни бъде необходимо обобщение на неравенството на Poincaré, което да е валидно в H^1. Ще приведем две неравенства, които са изпълнени за всяко $u \in H^1(\Omega)$:

$$\int_{\partial\Omega} u^2 ds \leq C\|u\|^2_{H^1(\Omega)},$$
$$\iint_\Omega u^2 d\Omega \leq C\left\{\iint_\Omega |\nabla u|^2 d\Omega + \int_{\partial\Omega} u^2 ds\right\} \quad \text{(Неравенство на Friedrichs).}$$

Доказателството на тези две неравенства в 1D е поставено като задача във втория списък от допълнителни задачи.

В по-общ вид могат да се запишат следните еквивалентни неравенства

$$\begin{aligned}\|u\|^2_{L_2(\partial\Omega)} &\leq C\|u\|^2_{H^1(\Omega)},\\ \|u\|^2_{L_2(\Omega)} &\leq C\left(\|\nabla u\|^2_{L_2(\Omega)} + \|u\|^2_{L_2(\partial\Omega)}\right),\\ \|u\|^2_{L_2(\partial\Omega)} &\leq C\left(\|u\|^2_{L_2(\Omega)} + \|\nabla u\|^2_{L_2(\Omega)}\right).\end{aligned}$$

Последните неравенства са от т.нар. неравенства за следата (Trace inequalities), тъй като дават връзка между поведението на функцията по границата на областта (нейната следа върху границата) и поведението ѝ в областта.

Example. Разглеждаме уравнението на Лаплас със смесени гранични условия:

$$\begin{aligned}-\Delta u &= 0, \ \mathbf{x} \in \Omega,\\ u &= 0, \ \mathbf{x} \in \Gamma_D,\\ \mathbf{n}\cdot\nabla u &= g_N, \ \mathbf{x} \in \Gamma_N,\end{aligned}$$

където $g_N \in L_2(\Gamma_N)$.

Съответната вариационна задача се получава за

$$a(u, v) = (\nabla u, \nabla v), \quad F(v) = (g_N, v)_{L_2(\Gamma_N)}$$

в Хилбертовото пространство

$$V = \{v \in H^1(\Omega) : v|_{\Gamma_D} = 0\}.$$

Коерцитивността и непрекъснатостта на билинейната форма в енергетична норма следват непосредствено, тъй като отново билинейната форма дефинира енергетичното скаларно произведение.

Ще покажем непрекъснатостта на линейната форма. Получаваме последователно

$$\begin{aligned}F^2(v) &= (g_N, V)^2_{L_2(\Gamma_N)}\\ &\leq \|g_N\|^2_{L_2(\Gamma_N)}\|v\|^2_{L_2(\Gamma_N)} \qquad \text{(К.-Б.-Ш.)}\\ &\leq C\|g_N\|^2_{L_2(\Gamma_N)}(\|v\|^2_{L_2(\Omega)} + \|\nabla v\|^2_{L_2(\Omega)}) \qquad \text{(Friedrichs)}\\ &\leq C\|g_N\|^2_{L_2(\Gamma_N)}\|v\|^2_E \qquad \text{(Poincaré).}\end{aligned}$$

Сега ще дадем пример, в който ще покажем условията на теоремата на Lax–Milgram в H^1-норма.

Example. Разглеждаме стационарното уравнение на топлопроводността

$$\begin{aligned} -\Delta u + pu &= f,\ \mathbf{x} \in \Omega, \\ \mathbf{n} \cdot \nabla u &= 0, \mathbf{x} \in \partial\Omega, \end{aligned}$$

където $p \in L_2(\Omega),\ p(\mathbf{x}) \geq p_0 > 0$.

Съответната вариационна задача се получава за

$$a(u,v) = (\nabla u, \nabla v) + (pu, v),\ F(v) = (f, v)$$

в Хилбертовото пространство $V = H^1(\Omega)$. Да проверим условията на теоремата.

- Коерцитивност на билинейната форма. Получаваме последователно

$$\begin{aligned} a(u,u) &= (\nabla u, \nabla u) + (pu, u) \\ &\geq \|u\|^2_{L_2(\Omega)} + p_0 \|\nabla u\|^2_{L_2(\Omega)} \\ &\geq \underbrace{\min\{1, p_0\}}_{\alpha} \left(\|u\|^2_{L_2(\Omega)} + \|\nabla u\|^2_{L_2(\Omega)}\right) \\ &= \alpha \|u\|^2_{H^1(\Omega)}. \end{aligned}$$

- Непрекъснатост на билинейната форма:

$$\begin{aligned} a(u,v) &= (\nabla u, \nabla v) + (pu, v) \\ &\leq \|\nabla u\|_{L_2(\Omega)} \|\nabla v\|_{L_2(\Omega)} + \|p\|_{L_\infty(\Omega)} \|u\|_{L_2(\Omega)} \|v\|_{L_2(\Omega)} \\ &\leq \max\{1, \|p\|_{L_\infty(\Omega)}\} \left(\|\nabla u\|_{L_2(\Omega)} \|\nabla v\|_{L_2(\Omega)} + \|u\|_{L_2(\Omega)} \|v\|_{L_2(\Omega)}\right) \\ &\leq C_0 \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)}. \end{aligned}$$

  Последното неравенство се получава подобно на доказателството на Твърдение 19.

- Непрекъснатост на линейната форма:

$$\begin{aligned} F^2(v) = (f,v)^2 &\leq \|f\|^2_{L_2(\Omega)} \|v\|^2_{L_2(\Omega)} \\ &\leq \|f\|^2_{L_2(\Omega)} \left(\|v\|^2_{L_2(\Omega)} + \|\nabla v\|^2_{L_2(\Omega)}\right) \\ &= \Lambda^2 \|v\|^2_{H^1(\Omega)} \end{aligned}$$

  за $\Lambda = \|f\|^2_{L_2(\Omega)}$.

5.2 Априорни оценки на грешката за МКЕ за абстрактната вариационна задача

Съответната абстрактна формулировка на МКЕ за абстрактната вариационна задача е следната:

Търсим $u_h \in V_h \subset V$, където пространството V_h е крайномерно, така, че

$$a(u_h, v) = F(v),\ \forall v \in V_h.$$

Основният резултат, на който се основават априорните оценки на грешката, както знаем, е ортогоналността по Гальоркин. Ще приведем резултата тук отново.

Proposition 26 (Galerkin orthogonality)**.** *В сила е*

$$a(u - u_h, v) = 0, \ \forall v \in V_h.$$

Proof. От вариационната задача и задачата на Ритц–Гальоркин знаем, че

$$a(u, v) = F(v), \ \forall v \in V,$$
$$a(u_h, v) = F(v), \ \forall v \in V_h.$$

Вадейки почленно двете вариационни тъждества, получаваме твърдението. □

Тогава можем да оценим грешката на решението, получено по МКЕ, с грешката на апроксимация за кой да е друг елемент на крайномерното подпространство V_h, благодарение на следния резултат.

Proposition 27 (Лема на Céa)**.** *За всяко $v \in V_h$ е в сила оценката*

$$\|u - u_h\|_V \leq \frac{C_0}{\alpha} \|u - v\|_V.$$

Proof. За да докажем твърдението, ще използваме коерцитивността и непрекъснатостта на билинейната форма. Получаваме последователно

$$\begin{aligned} \|u - u_h\|_V^2 &\leq \frac{1}{\alpha} a(u - u_h, u - v + v - u_h) \\ &= \frac{1}{\alpha} a(u - u_h, u - v) \\ &\leq \frac{C_0}{\alpha} \|u - u_h\|_V \|u - v\|_V. \end{aligned}$$

□

Като директно следствие получаваме, че можем да оценим грешката на решението по МКЕ с грешката на интерполант от V_h:

$$\|u - u_h\|_V \leq \frac{C_0}{\alpha} \|u - u_I\|_V.$$

И така, за да можем да направим оценка на грешката, използвайки последното, ние трябва да имаме подходящ инструмент от теория на апроксимациите. Много силен резултат е лемата на Bramble–Hilbert.

Proposition 28 (Лема на Bramble–Hilbert)**.** *Нека τ е референтна единична област в $\mathbb{R}^n$. Нека $q(u)$ е функционал в $H^{k+1}(\tau)$, за който:*

(i) $q(u + v) \leq q(u) + q(v)$ *(sublinearity);*

(ii) $|q(u)| \leq C\|u\|_{H^{k+1}(\tau)}$ *(ограниченост);*

(iii) $q(u) = 0$, *ако* $u \in P_k$.

Тогава съществува константа C_B *такава, че*

$$|q(u)| \leq C_B |u|_{H^{k+1}(\tau)}.$$

Както видяхме, за да приложим лемата на Céa, ни е необходима оценка за $\|u - u_I\|_{H^1(\Omega)}$. Ще приложим стандартния подход на МКЕ, като запишем $\|u - u_I\|^2_{H^1(\Omega)} = \sum_{\tau \in \mathcal{K}} \|u - u_I\|^2_{H^1(\tau)}$, а разстоянията в сумата оценим, като направим смяна към стандартния елемент. И така, първо ще започнем с оценяването на $\|\nabla(u - u_I)\|_{L_2(E)}$ и $\|u - u_I\|_{L_2(E)}$, като за целта ще използваме лемата на Bramble–Hilbert.

Proposition 29. *Нека* u_I *е интерполант на* u, *който е по части полином от* P_k. *Тогава*

$$\|u - u_I\|_{L_2(E)} \leq C|u|_{H^{k+1}(E)}$$

и

$$\|\nabla(u - u_I)\|_{L_2(E)} \leq C|u|_{H^{k+1}(E)}.$$

Proof. Ще докажем първото неравенство. Второто се проверява аналогично. Разглеждаме

$$q(u) := \left\{ \iint_E (u - u_I)^2 d\xi d\eta \right\}^{1/2}.$$

Ще проверим, че условията на лемата на Bramble–Hilbert са изпълнени:

- Sublinearity:

$$\begin{aligned} q(u+v) &= \left\{ \iint_E (u - u_I + v - v_I)^2 d\xi d\eta \right\}^{1/2} \\ &= \|u - u_I + v - v_I\|_{L_2(E)} \\ &\leq \|u - u_I\|_{L_2(E)} + \|v - v_I\|_{L_2(E)} \\ &= q(u) + q(v); \end{aligned}$$

- Ограниченост:

$$\begin{aligned} q(u) &= \|u - u_I\|_{L_2(E)} \\ &\leq \|u\|_{L_2(E)} + \|u_I\|_{L_2(E)} \\ &\leq C\|u\|_{H^{k+1}}. \end{aligned}$$

В последното неравенство използвахме факта, че $\|u\|_{L_2(E)} \leq \|u\|_{H^{k+1}(E)}$, както и

$$\begin{aligned} u_I(x) &= \sum_{i=1}^{k+1} u(P_i)\varphi_i(x) \\ &\leq \max_{x \in E} |u(x)| \sum_{i=1}^{k+1} \varphi_i(x) \\ &\leq C\|u\|_{H^{k+1}(E)} \qquad \text{от лемата на Соболев (вж. Параграф 5.4)} \end{aligned}$$

където $\varphi_i(x)$ образуват интерполационния базис, отговарящ на възлите P_i, $i = \overline{1, k+1}$.

- Очевидно $q(u) = 0$, ако $u \in P_k$.

Тогава условията на лемата на Bramble–Hilbert са изпълнени и твърдението е доказано. □

Сега вече сме готови да докажем основния резултат от тази тема.

Proposition 30. *Нека u_I, по части полином от P_k, е интерполант на функцията u, която е достатъчно регулярна. Тогава са в сила следните априорни оценки:*

$$\|u - u_I\|_{L_2(\Omega)} \leq Ch^{k+1}|u|_{H^{k+1}(\Omega)},$$
$$\|\nabla(u - u_I)\|_{L_2(\Omega)} \leq C\beta h^k|u|_{H^{k+1}(\Omega)},$$

където β е горна граница на h_τ/ρ_τ, а $h := \max_{\tau\in\mathcal{K}} h_\tau$.

Proof. Както отбелязахме, класическата идея в МКЕ е да сведем задачата, която е поставена върху областта Ω, до задача върху елементите, след което да "асемблираме" резултатите. Имаме

$$\|\nabla(u - u_I)\|_{L_2(\Omega)} = \left(\sum_{\tau\in\mathcal{K}} \int_\tau |\nabla(u - u_I)|^2 d\tau\right)^{1/2} = \left(\sum_{\tau\in\mathcal{K}} \|\nabla(u - u_I)\|^2_{L_2(\tau)}\right)^{1/2}.$$

Достатъчно е да оценим $Q_\tau(u) := \|\nabla(u - u_I)\|_{L_2(\Omega)}$. За тази цел ще използваме лемата на Bramble–Hilbert и следователно трябва да направим трансформация към стандартния триъгълен елемент. Нека елементът τ е определен от възлите $(x_k, y_k), (x_l, y_l), (x_m, y_m)$. Тогава смяната на променливите, която трансформира τ в стандартния триъгълен елемент, е

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} + B \begin{bmatrix} \xi \\ \eta \end{bmatrix},$$

където

$$B = \begin{bmatrix} x_l - x_k & x_m - x_k \\ y_l - y_k & y_m - y_k \end{bmatrix},$$

виж параграф 3.2. Освен това, $\nabla_x u = B^{-T}\nabla_\xi u$ и $\nabla_\xi u = B^T \nabla_x u$. Да обърнем внимание, че всички елементи на матрицата B са по модул по малки от h_τ. Нека означим още $J := det\ B$.

Тогава

$$\begin{aligned} Q_\tau(u) &= \left\{\int_E \|B^{-T}\nabla_\xi(u - u_I)\|_2^2 |J| d\xi d\eta\right\}^{1/2} \\ &\leq \underbrace{\|B^{-T}\||J|^{1/2}}_{\text{Тук влиза геометрията}} \underbrace{\left\{\int_E |\nabla_\xi(u - u_I)|^2 d\xi d\eta\right\}^{1/2}}_{\text{Ще използваме Bramble–Hlibert}}. \end{aligned} \tag{5.1}$$

И така, ще оценим всеки от множителите в последното поотделно.

- Знаем, че детерминантата може да бъде интерпретирана геометрично като лице. Действително, за лицето на τ имаме

$$|\tau| = \int_\tau d\tau = \int_E |J| d\xi d\eta = \frac{1}{2}|J|.$$

Следователно

$$|J| = \rho_\tau(h_k + h_l + h_m)$$

и

$$2\rho_\tau h_\tau \leq |J| \leq 3\rho_\tau h_\tau.$$

- За нормата на B^{-T} имаме по дефиниция

$$\|B^{-1}\| := \sup_{\zeta \in \mathbb{R}^2} \frac{\|B^{-1}\zeta\|}{\|\zeta\|}. \tag{5.2}$$

От друга страна, за произволно ζ е изпълнено

$$\begin{aligned}
\|B^{-1}\zeta\|^2 &= \left\| \frac{1}{J} \begin{bmatrix} y_m - y_k & x_k - x_m \\ y_k - y_l & x_l - x_k \end{bmatrix} \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix} \right\|^2 \\
&\leq \frac{2}{|J|^2}(\zeta_1 h_\tau + \zeta_2 h_\tau)^2 \\
&\leq \frac{4h_\tau^2(\zeta_1^2 + \zeta_2^2)}{|J|^2} \\
&= \frac{4h_\tau^2\|\zeta\|^2}{|J|^2}
\end{aligned}$$

Замествайки последното в (5.2), получаваме окончателно

$$\|B^{-1}\| \leq \frac{1}{\rho_\tau}. \tag{5.3}$$

- Да отбележим, че по подобен начин се получава и $\|B\| \leq h$, което ще използваме малко по-късно.

- За последния множител в $Q_\tau(u)$ от предходното твърдение имаме

$$\|\nabla_\xi(u - u_I)\|_{L_2(E)} \leq C|u|_{H^{k+1}(E)} \leq Ch^{k+1}|J|^{-1/2}|u|_{H^{k+1}(\tau)}. \tag{5.4}$$

Последното се основава на следното наблюдение

$$\|\nabla_\xi u\|_{L_2(E)} = \|B^T\nabla_x u\|_{L_2(E)} \leq \|B^T\|\|\nabla_x u\|_{L_2(E)} \leq Ch_\tau\|\nabla_x u\|_{L_2(\tau)}|J|^{-1/2}.$$

С други думи за всяко прилагане на диференциалния оператор при смяната "се появява" B^T, т.е. "излиза" по едно h. Тъй като в $|u|_{H^{k+1}}$ участват производните от ред $k+1$, получаваме оценката (5.4).

И така, замествайки (5.3) и (5.4) в (5.1), получаваме

$$\|\nabla(u - u_I)\|_{L_2(\tau)} \leq C\frac{1}{\rho_\tau}|J|^{1/2}h_\tau^{k+1}|J|^{-1/2}|u|_{H^{k+1}(\tau)}.$$

С това доказахме второто неравенство от твърдението.

По подобен начин се получава и първото неравенство. Единствената разлика е, че при смяната от τ към стандартния елемент не излиза B^{-T}, поради което получаваме и оценка, която е с един ред по-висока. □

5.3 Изследване на МКЕ за общата елиптична задача

Разглеждаме общата елиптична задача при гранични условия на Дирихле,

$$\begin{aligned} \nabla \cdot (-K(x)\nabla u + \mathbf{b}(x)u) + q(x)u = f(x),\, x \in \Omega \subset \mathbb{R}^d, \\ u(x) = 0,\, x \in \partial\Omega, \end{aligned} \tag{D}$$

където $d = 1, 2, 3$ е размерността на задачата, а $K(x) \in \mathbb{R}^{d\times d}$ е положително-определена матрица, т.е. $\boldsymbol{\xi}^T K(x)\boldsymbol{\xi} \geq k_0 \boldsymbol{\xi}^T \boldsymbol{\xi}$ за всяко $\boldsymbol{\xi} \in \mathbb{R}^d$ и някоя положителна константа k_0.

Като умножим двете страни скаларно с тестова функция v и интегрираме по части лявата страна, получаваме следната вариационна задача:

Да се намери $u \in H_0^1(\Omega)$ така, че

$$a(u, v) = F(v), \forall v \in H_0^1(\Omega), \tag{V}$$

където

$$\begin{aligned} a(u, v) &= \iint_\Omega \Big[\big(K(x)\nabla u - \mathbf{b}(x)u\big) \cdot \nabla v + q(x)uv \Big] d\Omega, \\ F(v) &= \iint_\Omega f v d\Omega. \end{aligned}$$

Proposition 31. *За вариационната задача (V), при предположения*

(i) $K(x)$ *е положително-определена матрица, т.е.* $\boldsymbol{\xi}^T K(x)\boldsymbol{\xi} \geq k_0 \boldsymbol{\xi}^T \boldsymbol{\xi}$ *за всяко* $\boldsymbol{\xi} \in \mathbb{R}^d$ *и някоя положителна константа* k_0*,*

(ii) $q(x) + \frac{1}{2}\nabla \cdot \boldsymbol{b}(x) \geq 0$*, за всяко* $x \in \Omega$

винаги съществува, при това единствено, решение. Приближеното решение във вид на по части полином от P_k*, получено по МКЕ, изпълнява следната априорна оценка на грешката:*

$$\|u - u_h\|_{H^1(\Omega)} \leq C\beta h^k |u|_{H^{k+1}(\Omega)}.$$

При това, ако точното решение изпълнява условие за пълна (елиптична) регулярност, т.е.

$$|u|_{H^2(\Omega)} \leq C\|f\|_{L_2(\Omega)},\ \forall f,$$

то е в сила

$$\|u - u_h\|_{L_2(\Omega)} \leq C\beta h^{k+1} |u|_{H^{k+1}(\Omega)}.$$

Proof. Непрекъснатостта на билинейната форма и дясната страна се проверяват лесно, като се използва неравенството на Коши–Буняковски–Шварц. Тук ще проверим, че при направените предположения, билинейната форма е коерцитивна. Действително, имаме

$$\begin{aligned} a(u, u) &= \iint_\Omega [(K(x)\nabla u - \mathbf{b}(x)u) \cdot \nabla u + q(x)u^2] d\Omega, \\ &\geq \iint_\Omega [k_0(\nabla u)^2 - (\mathbf{b}(x)u) \cdot \nabla u + q(x)u^2] d\Omega. \end{aligned}$$

Използвайки известното тъждество

$$-u\mathbf{b}\cdot\nabla u = -\frac{1}{2}\nabla\cdot(\mathbf{b}u^2) + \frac{1}{2}u^2\nabla\cdot\mathbf{b},$$

получаваме

$$\begin{aligned}\iint_\Omega -u\mathbf{b}\cdot\nabla u &= -\frac{1}{2}\iint_\Omega \nabla\cdot(\mathbf{b}u^2)d\Omega + \frac{1}{2}\iint_\Omega u^2\nabla\cdot\mathbf{b}d\Omega \\ &= \cancelto{0}{-\frac{1}{2}\int_{\partial\Omega}(\mathbf{b}u^2)\cdot\mathbf{n}ds} + \frac{1}{2}\iint_\Omega u^2\nabla\cdot\mathbf{b}d\Omega.\end{aligned}$$

Тогава

$$\begin{aligned}a(u,u) &\geq \iint_\Omega \left[k_0(\nabla u)^2 + \left(q(x) + \frac{1}{2}\nabla\cdot\mathbf{b}\right)u^2\right]d\Omega \\ &\geq k_0\|\nabla u\|^2_{L_2(\Omega)} \\ &\geq C\|u\|^2_{H^1(\Omega)}.\end{aligned}$$

За последното неравенство е използвано неравенството на Poincaré.

И така, от теоремата на Lax–Milgram следва съществуване и единственост на решението на вариационната задача. При това е в сила лемата на Céa, т.е.

$$\|u-u_h\|_{H^1(\Omega)} \leq \frac{C_0}{\alpha}\|u-u_I\|_{H^1(\Omega)},\ \forall v\in V_h,$$

откъдето получаваме априорната оценка на грешката в H^1-норма, като използваме Твърдение 30.

За да докажем оценката в L_2-норма, ще приложим трика на Nitsche. Разглеждаме дуалната задача на (V):

$$a(v,\hat{u}) = (u-u_h, v),\ \forall v\in V.$$

В частност за $v = u - u_h$ имаме

$$\begin{aligned}\|u-u_h\|^2_{L_2(\Omega)} &= a(u-u_h,\hat{u}) \\ &= a(u-u_h,\hat{u}-\hat{u}_I) \\ &\leq C\|u-u_h\|_{H^1(\Omega)}\|\hat{u}-\hat{u}_I\|_{H^1(\Omega)} \\ &\leq C\beta h^k|u|_{H^{k+1}(\Omega)}h|\hat{u}|_{H^2(\Omega)}.\end{aligned}$$

При предположение за пълна регулярност е изпълнено $|\hat{u}|_{H^2(\Omega)} \leq C\|u-u_h\|_{L_2(\Omega)}$ и твърдението е доказано. □

Remark 26. За да бъде ефективна оценката, е **необходимо триангулацията да се построява по такъв начин, че β да се контролира**. Това обикновено се постига чрез контролиране на минималните ъгли на всеки от триъгълниците. Това обяснява и казаното за триангулацията в параграф 2.1.1.

5.4 Избрани теми от теорията на Соболевите и Хилбертовите пространства

Тук ще приведем някои обзорни сведения от теорията на Соболевите и Хилбертовите пространства, на които се основават някои от теоретичните резултати, разгледани дотук.

По-долу V ще бъде дадено линейно нормирано пространстно с норма $\|\cdot\|$ освен ако не е казано нещо друго.

Definition 3. Казваме, че редицата $\varphi_0, \varphi_1, \varphi_2, \ldots$ е редица на Коши (фундаментална редица), ако за всяко $\varepsilon > 0$ съществува $n \in \mathbb{N}$ така, че $\|\varphi_k - \varphi_l\| < \varepsilon$ за $k, l > n$.

Definition 4. Казваме, че линейното нормирано пространство V е Банахово (пълно метрично) пространство, ако една редица е сходяща тогава и само тогава, когато е редица на Коши.

Example. Пространството $\mathbb{R}^n$, нормирано със стандартната Евклидова норма, е Банахово пространство.

Example. Може да се покаже, че простанството $C(\Omega)$ от непрекъснатите в областта Ω функции, нормирано с равномерната норма, е Банахово.

Example. Пространството $\mathbb{Q}$ не е Банахово, тъй като в него има "дупки". Поточно можем да намерим редици на Коши (от рационални числа), които не са сходящи в $\mathbb{Q}$ (сходящи са например в $\mathbb{R}$). Такъв пример е редицата от приближения на $\sqrt{2}$ до първия, втория и т.н. знак след десетичната точка.

В пълните пространства има "хубава" теория за съществуване, единственост, сходимост и т.н. Пример за такъв резултат е теоремата за неподвижната точка от курса "Числени методи на анализа", която се обобщава в теоремата на Банах за произволно Банахово пространство.

Особено добра структура имат пространствата, в която е въведена геометрия (посредством скаларно произведение).

Definition 5. Казваме, че линейното пространството V е предхилбертово пространство, ако в него е въведено скаларно произведение.

Най-основният резултат в предхилбертовите пространства е неравенството на Коши–Буняковски–Шварц.

Proposition 32 (Неравенство на Коши–Буняковски-–Шварц)**.** *За всеки два елемента u, v на предхилбертовото пространство V е в сила*

$$|(u, v)| \leq \|u\|\|v\|.$$

Proof. За всяко реално α е в сила

$$(u + \alpha v, u + \alpha v) = (u, u) + 2\alpha(u, v) + \alpha^2(v, v) \geq 0.$$

Следователно дискримантата на получения квадратен тричлен е неположителна, т.е.

$$(u, v)^2 \leq (u, u)(v, v).$$

□

Definition 6. Казваме, че предхилбертовото пространство H е Хилбертово пространство, ако H е Банахово.

За нашите цели най-важните Банахови и Хилбертови пространства са функционалните пространства от интегруеми функции.

Definition 7. Дефинираме пространството $L_p(\Omega) := \{v : \Omega \to \mathbb{R} : \|v\|_{L_p(\Omega)} < \infty\}$, където

$$\|v\|_{L_p(\Omega)} = \left\{ \int_\Omega |v|^p d\Omega \right\}^{1/p}, \ p < \infty, \qquad \|v\|_{L_\infty(\Omega)} = \sup_{x \in \Omega} |v(x)|.$$

Може да се покаже, че пространствата L_p са Банахови за всяко p, но единствено за $p = 2$ съответното пространство L_2 е Хилбертово.

В Хилбертовите пространства L_p функциите по принцип не са достатъчно регулярни, за да говорим за производни в класическия смисъл. По тази причина се въвеждат т.нар. обобщени (слаби) производни. Тук ще уточним това понятие. Когато обобщаваме дадено понятие в математиката, трябва да изпълним две неща:

- действително дефиницията да обобщава класическото понятие (т.е. то да е частен случай на обобщението);
- да се запазват "хубавите" свойства от гледна точка на теорията, която разглеждаме.

Definition 8. Нека $u \in L^1_{loc}(\Omega)$ е дадена функция. Казваме, че функцията $\partial_{x_i} u = \frac{\partial u}{\partial x_i} \in L^1_{loc}(\Omega)$ е обобщена (слаба) производна на u, ако

$$\int_\Omega \partial_{x_i} u \varphi d\Omega = -\int_\Omega u \partial_{x_i} \varphi d\Omega, \quad \forall \varphi \in C_0^\infty(\Omega).$$

Може да се покаже, че формулата за интегриране по части, както и основните свойства на диференциалния оператор се запазват при обобщената производна. Това се основава на факта, че интегруемите функции могат да се апроксимират с гладки такива, но този въпрос излиза извън рамките на настоящия курс.

Example. Да разгледаме функцията

$$u = \begin{cases} x, & x \geq 0, \\ 0, & x \leq 0, \end{cases}$$

дефинирана в интервала $[-1, 1]$. Пресмятаме

$$\int_{-1}^1 u\varphi' dx = \int_{-1}^0 u\varphi' dx + \int_0^1 u\varphi' dx = \int_0^1 x\varphi'(x) dx = -\int_0^1 \varphi dx.$$

Следователно, ако функцията u има слаба производна, то това трябва да е такава функция u', за която

$$\int_{-1}^1 u'\varphi dx = \int_0^1 \varphi dx. \ \forall \varphi \in C_0^\infty.$$

Такава функция действително съществува, например функцията на Heaviside:

$$u' = H(x) = \begin{cases} 0, & x \leq 0, \\ 1, & x > 0. \end{cases}$$

Remark 27. Да обърнем внимание, че слабата производна е определена с точност до стойността ѝ върху множество с мярка 0. Това не е проблем, тъй като в пространствата L_p такива функции се идентифицират.

Remark 28. Слабата производна съвпада с класическата производна в точките, където u е диференцируема в класическия смисъл.

Remark 29. Наличието на "чупки" в графиката на функцията, както виждаме, не е проблем за съществуването на слаба производна. Следователно функциите-колибка, които са основни за целите на настоящия курс за действително H^1-функции.

Example. Нека разгледаме функцията на Heaviside $H(x)$. Ако тя има слаба производна, това трябва да е функция H' такава, че за всяко $\varphi \in C_0^\infty$ е в сила

$$\int_{-1}^{1} H'\varphi dx = -\int_{-1}^{1} H\varphi' dx = \varphi(0).$$

Горното интегрално тъждество се изпълнява от δ-функцията на Дирак, но тя не е достатъчно регулярна, за да удовлетворява приведената дефиниция.

Функции, в които има такива "скокове" не са диференцируеми и в слаб смисъл.

Както видяхме в рамките на курса, важно условие за функциите, с които работим в МКЕ, е те да изпълняват определени изисквания за съществуване на производни (в слаб смисъл), за да имат съответните вариационни задачи смисъл. Ето защо естествените пространства, в които разглеждаме задачите, са Соболевите пространства. Това са подпространства на L_p, в които функциите изпълняват определени условия за съществуване на слаби производни.

Definition 9. Соболевото пространство $W_k^p(\Omega)$ се състои от тези функции, които заедно със слабите си производни $D^\alpha u$ от ред $|\alpha| \leq k$ са $L_p(\Omega)$-функции. Казано иначе,

$$\|u\|_{W_k^p} < \infty,$$

където

$$\|u\|_{W_k^p(\Omega)} := \left(\sum_{|\alpha| \leq k} \|D^\alpha u\|^p_{L_p(\Omega)} \right)^{1/p}, \ 1 \leq p < \infty.$$

Пространствата, които се получават при $p = 2$ са Хилбертови пространства. За да се подчертае този факт, обикновено се използва означението $W_k^2 =: H^k$. Така например нормата в пространството $H^k(I)$ се поражда от скаларното произведение

$$< u, v >:= \int_I (uv + u'v' + \cdots + u^{(k)}v^{(k)})dx.$$

В сила е следният резултат (Лема/Теорема на Соболев за влаганията), който ще приведем без доказателство.

Proposition 33 (Лема на Соболев)**.** *Нека $n = 1, 2, 3$ е размерността на дадена задача. Ако $u \in H^k(\Omega)$ и $k > n/2$, тогава u е непрекъсната и*

$$\max_{x \in \Omega} |u(x_1, \ldots, x_n)| \leq C\|u\|_{H^k(\Omega)}.$$

5.4.1 МКЕ и граничните условия

Остана да коментираме въпроса за граничните условия и по-точно въпроса защо някои гранични условия налагаме експлицитно във вариационната задача (ще ги наричаме **главни гранични условия**), а за други (ще ги наричаме **естествени гранични условия**) – не.

Proposition 34. *Пространството*

$$V := \{v \in H^1(\Omega) : \nabla v \cdot \boldsymbol{n} = 0 \text{ } \textit{върху } \partial\Omega\}$$

не е Банахово.

Proof. Достатъчно е да дадем един пример. Дефинираме редицата

$$v_i(x) := \begin{cases} x, & x \leq 1 - \frac{1}{i}, \\ 1 - \frac{1}{i}, & x \geq 1 - \frac{1}{i}. \end{cases}$$

Очевидно това е редица на Коши но тя не е сходяща във V, тъй като производната на граничната функция при $x = 1$ е 1 (направете чертеж!). □

Proposition 35. *Пространството H_0^1 е Банахово.*

Proof. Ще докажем в 1D. За всеки два елемента v_n, v_m на H_0^1 е в сила

$$\begin{aligned} |v_n(x) - v_m(x)|^2 &= \left| \int_0^x (v_n - v_m)' dx \right|^2 \\ &\leq \int_0^x |(v_n - v_m)'|^2 dx \\ &\leq \|v_n - v_m\|_{H^1}^2. \end{aligned}$$

Нека $\{v_i\}$ е редица на Коши в H_0^1. Тогава

$$\|v_n - v_m\|_{H^1} \to 0$$

и съществува граница на тази редица в H^1. От друга страна, показахме, че

$$\|v_n - v_m\|_{H^1} \geq \max_{x \in I} |v_n(x) - v_m(x)| = \|v_n - v_m\|_{C[a,b]}$$

и следователно редицата е редица на Коши в $C[a, b]$ и нейната граница е непрекъсната функция v такава, че $v(0) = 0$. □

И така, ако в H^1 наложим експлицитно гранични условия, в които участват производни, то полученото пространство няма да бъде Хилбертово и цялата теория, която развихме, не би била в сила. Оказва се обаче, че експлицитното налагане на тези условия не е и необходимо. Фактът, че сме ги използвали при формулирането на вариационното уравнение, означава, че те ще бъдат автоматично изпълнени от решението на вариационната задача. Ще илюстрираме идеята на базата на следния пример.

Exercise. Коя е диференциалната задача, която, при предположение за достатъчна регулярност на решението, съответства на следната вариационна задача:

Да се намери $u \in H^1(I)$ така, че

$$a(u,v) = F(v),\ \forall v \in H^1(I),$$

където

$$a(u,v) = -\int_I u'v'dx, \quad F(v) = \int_I fvdx.$$

Упътване. Интегрирайте по части $a(u,v)$ и използвайте факта, че интегралното тъждество е изпълнено за всяко v. Изберете последователно тестови функции, които се нулират на двете граници, на лявата граница, на дясната граница. $\square$