MM406: AN INTRODUCTION TO THE FINITE ELEMENT METHOD

GABRIEL R. BARRENECHEA

Abstract. In these notes the general ideas of the finite element method, its analysis, and implementation will be given. This topic constitutes the second half of the class `MM406:Finite Element Methods For Boundary Value Problems And Approximation`.

Contents

1. Notation, stucture and goals of these notes

Partial differential equations (PDE's) are a fundamental tool in applied mathematics and mathematical modelling. They can be found in the modelling of biological systems, population dynamics, micro/macro-economic modelling, game theory, and financial mathematics. They also constitute an important branch of mathematics with applications to different fields such as geometry, fluid and solid mechanics, electromagnetism, dynamical systems, mathematical astronomy and physics, just to name a few.

The goal of this part of the class is to derive, understand, analyse, and implement effective numerical methods that provide "good" approximations to the solution of a specific class of PDEs, namely, elliptic partial differential equations. The course will take a dual approach to this end: we will investigate theoretically the derivation and analysis of the various methods and, in tandem, we will show how to implement these methods on a computer using a high level progamming language such as Matlab.

Seeing an implementation of a method is not only fundamental in understanding and appreciating the method, but it also gives a rewarding feeling if it is successful, something precluded in a purely theoretical approach to the subject. So, in some lectures I will describe the implementaion of the finite element method for basic problems, and showcase its application to more complex problems as well.

Since we are only focussed on elliptic PDEs, in these notes we will use the convention that $u = u(\boldsymbol{x})$ denotes a function depending upon $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$ of a d-dimensional independent variable. We will write

$$\nabla u := \begin{pmatrix} \partial_{x_1} u \\ \vdots \\ \partial_{x_d} u \end{pmatrix}, \tag{1.1}$$

as the *gradient* of u. It is a d dimensional column vector. In addition, for a vector-valued function $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_d)$, we define its *divergence* by

$$\operatorname{div} \boldsymbol{\tau} = \partial_{x_1} \tau_1 + \ldots + \partial_{x_d} \tau_d \,. \tag{1.2}$$

With these ingredients, we define one of the most important differential operators for this class, namely, the *Laplacian*, defined for a smooth function u as follows:

$$\Delta u := \operatorname{div}(\nabla u) = \sum_{i=1}^{d} \partial_{x_i} \partial_{x_i} u \,. \tag{1.3}$$

We will use colour boxes to denote things of specific interest, in particular, we will use red colour boxes to give a summary or to express the main point to an argument in heuristic terms:

Main Idea

We will use a (very ugly) yellow for exercises:

Exercise

We will use an oceanic blue for worked examples:

Worked example

and green for their solutions (although many solutions will be given directly in the lectures).

Solution

Finally, the material of these lecture notes is insprired by many textbookds, most noticeably the following ones:
- Susanne Brenner and L. Ridway Scott: *The Mathematical Theory of Finite Element Methods*, Third Edition, Springer Verlag 2008.
- Alexandre Ern and Jean-Luc Guermond: *Finite Elements I* and *Finite Elements II*. Springer Verlag 2021.
- Claes Johnson: *Numerical Solution of Partial Differential Equations by the Finite Element Method*. Originally published in 1984, reprinted in 2010 by Dover Publications.
- Alfio Quarteroni and Fausto Saleri: *Scientific Computing with Matlab and Octave*, Second Edition. Springer Verlag, 2006.

All of these books can be found in the library. Nevertheless, if you'd like to purchase them (which I can recommend, especially in the case of Claes Johnson's book, as it presents a very good introduction to the method), you can find the Dover printing for approximately £10-15. Also, older versions of the Matlab book by Quarteroni and collaborators can be found for approximately £25 (and also can be downloaded from Springer Link `https://link.springer.com/book/10.1007/978-3-642-45367-0`).

2. Finite elements in 1D

2.1. **A differential equation and its weak form.** Consider the problem of finding a function $u\colon [0,1] \to \mathbb{R}$ such that

$$\text{(D)} \qquad \begin{aligned} -(ku')'(x) + (bu)(x) &= f(x), \qquad x \in (0,1)\,, \\ u(0) &= 0\,, \\ u(1) &= 0\,, \end{aligned}$$

where $k \in C^1[0,1]$, $b \in C[0,1]$ and $f \in C[0,1]$ are given functions[1]. The minus sign in front of the higher-order derivative in (D) is deliberate (the reason for this will become clear later). The function k is assumed to be uniformly positive and b is assumed non-negative; that is,

$$k(x) \geq k_0 > 0, \qquad \text{and} \qquad b(x) \geq 0\,, \quad x \in [0,1]. \tag{2.1}$$

Note that the differential equation in (D) may also be written in the form

$$-(ku'' + k'u') + bu = f \tag{2.2}$$

(i.e., as a second-order ODE), but we prefer not to do this, as (D) is the way that will be make our derivations below more natural. In *very special circumstances* an explicit formula for the solution of (D) can be found, but this is a very rare situation. For example, for the equation $-u'' = e^{x^2}$ an analytical solution is not known. Similarly, it is easy to imagine a problem like (D), even with a simple right-hand side f, but having very complex coefficients k and b, where an analytical solution cannot be computed. So, numerical approximation of the solution is usually necessary.

Main Idea

The idea of the *Finite Element Method* (FEM) is as follows:

Step 1: Rewrite the boundary value problem (D) in a *weak form* only involving first derivatives of u.

Step 2: Approximate this resulting weak form using a finite-dimensional vector space whose elements are functions.

Note that for more general second order equations than (D), the weak formulation can involve the function and its first derivatives; the key point of the weak formulation is that, for second order equations, it no longer contains second derivatives.

Step 1. To derive a weak formulation for our model problem, the first step is to identify a vector space of functions in which the solution u is sought. For this, we define the notion of *piecewise continuous* function.

Definition 1. *A function w is called* piecewise continuous *in an interval $[a,b]$ if there exists a partition $a = p_0 < p_1 < \cdots < p_m = b$ such that $w|_{[p_{i-1},p_i]} \in C[p_{i-1},p_i]$ for all $i = 1,\ldots,m$.*

With this definition, in the one-dimensional case of our model problem the following is an appropriate space for u:

$$V := \{v \in C[0,1] :\ v' \text{ is bounded and piecewise continuous in } [0,1], \text{ and } v(0) = v(1) = 0\}\,. \tag{2.3}$$

Note that the notion of a derivative v' is not defined for arbitrary continuous v. This opens the door to a *weaker* notion of differentiation, that we will study at a later stage.

Now, we multiply equation (D) by an arbitrary function $v \in V$, and integrate over $[0,1]$ to obtain

$$\int_0^1 \left\{(-k(x)u'(x))'v(x) + b(x)u(x)v(x)\right\} \mathrm{d}x = \int_0^1 f(x)v(x)\,\mathrm{d}x\,.$$

We usually drop the (x) and write

$$\int_0^1 \left(-(ku')'v + buv\right) \mathrm{d}x = \int_0^1 fv\,\mathrm{d}x.$$

[1] *Notation:* $C^r[a,b]$ denotes the space of r-times continuously differentiable functions on $[a,b]$.

Integrating the first term by parts and applying the boundary conditions $v(0) = v(1) = 0$ leads to

$$\int_0^1 (ku'v' + buv)\,\mathrm{d}x - \underbrace{[ku'v]_0^1}_{=0} = \int_0^1 fv\,\mathrm{d}x \Rightarrow \int_0^1 (ku'v' + buv)\,\mathrm{d}x = \int_0^1 fv\,\mathrm{d}x\,.$$

Since $v \in V$ was arbitrary, u is a solution of the following problem:
Find $u \in V$ such that

$$\text{(W)} \qquad a(u,v) = L(v) \qquad \forall\, v \in V,$$

where

$$(2.4) \qquad a(u,v) := \int_0^1 (ku'v' + buv)\,\mathrm{d}x \quad \text{and} \quad L(v) := \int_0^1 fv\,\mathrm{d}x.$$

The problem (W) is called the *weak form* of (D).

The forms $a(\cdot,\cdot)$ and $L(\cdot)$ will be used throught this class. We now recall the definition of a bilinear form, and a linear functional.

Definition 2. *Let H be a real vector space. We say that*

- *$F : H \to \mathbb{R}$ is a* **linear functional** *if, for all $u, v \in H$ and all $\alpha \in \mathbb{R}$,*
$$F(u + \alpha v) = F(u) + \alpha F(v)\,.$$
- *$B : H \times H \to \mathbb{R}$ is* **a bilinear form** *if for all $u, v, w \in H$ and all $\alpha \in \mathbb{R}$,*
$$B(u + \alpha v, w) = B(u, w) + \alpha B(v, w)\,,$$
$$B(u, v + \alpha w) = B(u, v) + \alpha B(u, w)\,.$$

Example

Example 1. *Show that $a(\cdot,\cdot)$ defined in (2.4) is bilinear.*

Step 2 now seems easy (in principle): The weak form (W), although very elegant, is as difficult to solve as the differential equation itself. This is due to the fact that V is an infinite-dimensional vector space. This is where the finite element method appears. In fact, we consider a finite-dimensional space $V_h \subseteq V$ (the notation subscript h will become more apparent below), and write the following approximate problem (commonly referred to as *discrete problem*): Find $u_h \in V_h$ such that

$$\text{(W}_h\text{)} \qquad a(u_h, v_h) = L(v_h) \qquad \forall v_h \in V_h.$$

The discrete problem (W$_h$) is the one that practitionners actually solve. For this, the problem needs to be rewritten in a way a computer can understand. The first step in this process is to identify a basis for the space V_h. For the moment, let us suppose the set $\{\phi_1, \ldots, \phi_n\}$ is a basis for V_h, then we can write

$$(2.5) \qquad u_h = \sum_{j=1}^{n} U_j \phi_j\,,$$

for some unknown coefficients $U_j \in \mathbb{R}$. So, using the bilinearity of $a(\cdot,\cdot)$ we see that (W$_h$) is equivalent to

$$(2.6) \qquad \sum_{j=1}^{n} a(\phi_j, v_h)\, U_j = L(v_h), \qquad \forall v_h \in V_h.$$

Also, a is linear with respect to its second argument and L is linear, so (2.6) is equivalent to

$$\sum_{j=1}^{n} a(\phi_j, \phi_i)\; U_j = L(\phi_i), \qquad \forall i = 1, \ldots, n.$$

Thus the vector of coefficients $\mathbf{U} = (U_j)_{j=1}^n \in \mathbb{R}^n$ satisfies the linear system of equations

$$\mathbb{A}\,\mathbf{U} = \mathbf{f}, \tag{2.7}$$

where $\mathbb{A} = (a_{ij})_{i,j=1}^n$ and $\mathbf{f} = (f_i)_{i=1}^n$ are given by

$$a_{ij} := a(\phi_j, \phi_i) \qquad \text{and} \qquad f_i := L(\phi_i).$$

Hence, by solving (2.7) for $\mathbf{U}$ and substituting into (2.5), we obtain the solution of (W_h). The matrix $\mathbb{A}$ is commonly called **stiffness matrix**, and $\mathbf{f}$ is called the **load vector**, for historical reasons.

Main Idea

- Approximate the solution u of (D) by u_h.
- Write u_h in terms of the basis functions.
- The solution can be computed by solving a system of n linear equations (2.7).

2.2. **Practical realisation with piecewise linear finite elements.** We now make the choice of V_h concrete. For this, we choose a very specific (but widely used in practice) finite-dimensional space V_h in (W_h). The first step of its construction is the building of a "*mesh*" on $[0,1]$ with *nodes*

$$0 = x_0 < x_1 < \cdots < x_n < x_{n+1} = 1.$$

The intervals

$$\tau_i := [x_{i-1}, x_i]\,,$$

are called *elements*. We define

$$h_i = x_i - x_{i-1}\,,$$

and set

$$h := \max_{i=1,\dots,n+1} h_i \quad \text{the "}\textit{mesh diameter}\text{"}\,.$$

The mesh is *uniform* when $h_i = h = 1/(n+1)$ for all i; in this case: $x_i = ih$, $i = 0, \dots, n+1$.

With these notations we introduce the following subspace of V:

$$V_h := \left\{ v \in C[0,1] :\ v|_{\tau_i} \text{ is linear and } v(0) = 0 = v(1) \right\}. \tag{2.8}$$

Here "linear" means that v has a straight line graph on each subinterval of the mesh or, equivalently, that v is a linear combination of the functions 1 and x on each subinterval of the mesh.[2]

There are many possible choices for basis of V_h (an infinite number, in fact). The most widely used (and for a good reason!) is the one formed by the so-called *hat functions* $\{\phi_i : i = 1, \dots, n\}$ defined by

$$\phi_i(x) = \begin{cases} \dfrac{x - x_{i-1}}{h_i} & x \in \tau_i, \\ \dfrac{x_{i+1} - x}{h_{i+1}} & x \in \tau_{i+1}, \\ 0 & \text{for all other } x\,. \end{cases}$$

It is immediate to see that

$$\phi_i(x_j) = \delta_{ij} := \begin{cases} 1 & \text{if } i = j\,, \\ 0 & \text{if } i \neq j\,. \end{cases} \tag{2.9}$$

Figure 2.1 illustrates the plots of two hat functions.

[2]The use of the word "linear" is obviously an abuse of language; a more accurate description would be *affine*, but we will stick to linear since it is the widely accepted terminology.

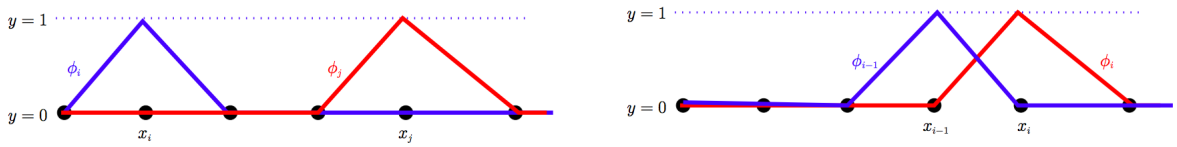

FIGURE 2.1. Two hat functions ϕ_j and ϕ_i (with $j = i - 1$ on the right plot).

Exercise

Exercise 1. *Show the hat functions $\phi_1, \ldots, \phi_n$ are a basis for V_h.*

Remark 1. *The hat functions have many appealing properties. The following two will have a clear impact on the structure of the linear system associated to the discrete problem* (W_h)*:*

$$\phi_i'(x) = \begin{cases} h_i^{-1} & x \in \tau_i \\ -h_{i+1}^{-1} & x \in \tau_{i+1} \\ 0 & \text{elsewhere}\,, \end{cases}$$

and

$$\phi_i \phi_j \equiv 0 \equiv \phi_i' \phi_j' \quad \text{unless} \quad j = i-1, i, i+1\,. \tag{2.10}$$

Remark 2. *Further, we introduce some useful language. For a real function f, we define the* support *of f, and denote it by $\mathrm{spt}(f)$ as the set*

$$\mathrm{spt}(f) = \overline{\{x \in \mathbb{R} : f(x) \neq 0\}}\,, \tag{2.11}$$

where, for an interval I (or a union of intervals) we define its closure *$\overline{I}$ as the smallest closed interval (or union of closed intervals) that contains I (for example, $\overline{(0, 0.5]} = [0, 0.5]$, $\overline{(0,1)} = [0,1]$, $\overline{(0,1) \cup (1,2)} = [0,2]$, and $\overline{(0,1) \cup (2,3]} = [0,1] \cup [2,3]$).*

We now exploit the particular choice of basis functions we have made. We recall that the problem (W_h) is equivalent to the linear system

$$\mathbb{A}\mathbf{U} = \mathbf{f},$$

where $\mathbb{A} = (a_{ij})_{i,j=1}^n$ and $\mathbf{f} = (f_i)_{i=1}^n$ are given by

$$a_{ij} := a\big(\phi_j, \phi_i\big) \quad \text{and} \quad f_i := L(\phi_i).$$

In order to implement the method we need to form the vector $\mathbf{f}$ and matrix $\mathbb{A}$. By the definition of ϕ_i,

$$f_i = \int_0^1 f\phi_i \,\mathrm{d}x = \int_{\tau_i} f\phi_i \,\mathrm{d}x + \int_{\tau_{i+1}} f\phi_i \,\mathrm{d}x\,.$$

That is, f_i is obtained by integrating $f\phi_i$ over the support of ϕ_i.

Moreover,

$$a_{ij} = \int_0^1 (k\phi_j'\phi_i' + b\phi_j\phi_i)\,\mathrm{d}x = \int_{\mathrm{spt}(\phi_j)\cap\mathrm{spt}(\phi_i)} (k\phi_j'\phi_i' + b\phi_j\phi_i)\,\mathrm{d}x\,,$$

thus, thanks to Remark 1, $\mathbb{A}$ takes the form

$$\mathbb{A} = \begin{bmatrix} * & * & & & & & \\ * & * & * & & & & \\ & * & * & * & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & * & * & * \\ & & & & & * & * \end{bmatrix},$$

where $*$ denotes an entry which may be non-zero and the blank spaces mark zero entries. We call such a matrix *tridiagonal*. The non-zero entries are integrals over those elements for which *both* ϕ_i and ϕ_j are non-zero.

To know the matrix $\mathbb{A}$ completely we only need to compute three elements of it per row. We start with the diagonal elements, given by

$$a_{ii} = a(\phi_i, \phi_i) = \int_{\tau_i} (k\phi_i'\phi_i' + b\phi_i\phi_i)\,\mathrm{d}x + \int_{\tau_{i+1}} (k\phi_i'\phi_i' + b\phi_i\phi_i)\,\mathrm{d}x\,.$$

In addition, we notice that the bilinear form $a(\cdot,\cdot)$ is *symmetric*, that is, $a(u,v) = a(v,u)$ for all $u, v \in V$. This implies, in particular, that $a(\phi_j, \phi_i) = a(\phi_i, \phi_j)$, and thus the matrix $\mathbb{A}$ is symmetric, this is $a_{ij} = a_{ji}$ for all $i, j = 1, \ldots, n$. So,

$$a_{i\,i-1} = a_{i-1\,i} = \int_{\tau_i} (k\phi_{i-1}'\phi_i' + b\phi_{i-1}\phi_i)\,\mathrm{d}x\,,$$

where the integral is only computed in $\tau_i = \mathrm{spt}(\phi_i) \cap \mathrm{spt}(\phi_{i-1})$.

It is convenient to introduce the notation

$$a_{g\,\ell}^{\tau_i} = \int_{\tau_i} (k\phi_\ell'\phi_g' + b\phi_\ell\phi_g)\,\mathrm{d}x\,, \quad \text{for} \quad g, \ell = i-1, i\,.$$

The matrix $\mathbb{A}^{\tau_i} = (a_{g\ell}^{\tau_i})$ is called the *element stiffness matrix* for the problem. Then, suppose we compute the quantities

$$a_{i,i-1}^{\tau_i}, \quad a_{i-1,i-1}^{\tau_i}, \quad \text{and} \quad a_{i,i}^{\tau_i}\,, \tag{2.12}$$

for each $i = 1, \ldots, n+1$ (these can be conveniently stored using three vectors in Matlab.) Then the matrix $\mathbb{A}$ can easily be constructed by combining the appropriate components of (2.12):

$$\left.\begin{array}{rcl} a_{i\,i} & = & a_{ii}^{\tau_i} + a_{i,i}^{\tau_{i+1}} \\ a_{i\,i-1} & = & a_{i,i-1}^{\tau_i} \\ a_{i\,i+1} & = & a_{i,i+1}^{\tau_{i+1}} \end{array}\right\} \tag{2.13}$$

If k and b are constant functions, all of (2.12) can be easily found exactly. Otherwise, we may need numerical integration.

In this course, we assume that we can solve linear systems like (2.7) for any size n. The Matlab command \ (backslash) can be used to do this.

Worked example

Example 2. *Suppose $k = 1$, $b = 0$, $f = 1$ and the mesh is uniform (so $h = 1/(n+1)$). Compute entries of the stiffness matrix and load vector.*

Worked example

Example 3. *The ordinary differential equation $-u'' = 1$ on $(0,1)$ subject to $u(0) = u(1) = 0$ has $u(x) = \frac{x(1-x)}{2}$ as exact solution. Use the provided finite element code in 1D to see the finite element approximation of this equation with uniform meshes containing $4, 8$, and 16 elements.*

Worked example

Example 4. *Consider the following two-point boundary value problem:*

$$\begin{aligned} -u'' + u &= f \qquad \text{in } (0,1)\,, \\ u'(0) = 0 \quad &, \quad u'(1) = 0\,. \end{aligned}$$

Find a weak formulation for this problem. Define an appropriate space for the solution and test functions, and write a finite element method for it using piecewise continuous functions. Specify the matrix and right-hand side.

Note that with this small value of n (and relatively large h) the FEM may typically only give a very crude approximation to the true solution u. Later we shall show convergence as $h \to 0$. However for this particular example something intersting happens and you actually get a much more accurate solution than expected. We will be back to this later.

Exercises

Exercise 2. *Repeat the calculations from the worked example 2, but now assuming that the mesh is* not uniform, *that is, $h_i = x_i - x_{i-1}$ does vary with i.*

Exercise 3. *Let $v \in C[0,1]$ be such that $\int_0^1 v^2(x)\,\mathrm{d}x = 0$. Show that $v(x) = 0$ for all $x \in [0,1]$.*

Exercise 4. *Consider the variational formulation for the model problem* (W) *with $k > 0$ a positive constant and $b = 1$. Show that, if it exists, the solution of the weak formulation* (W) *is* unique. *(**Hint:** Suppose that there exist two solutions, u_1, u_2. Consider the weak problem* (W) *for both u_1 and u_2, subtract them, and obtain an equation of the type $a(u_1 - u_2, v) = 0$ for all $v \in V$. Taking $v = u_1 - u_2$ will lead to uniqueness of solutions.)*

Exercise 5. *The goal of this problem is to show that, under appropriate hypotheses, the solution u of* (W) *also soves* (D), *thus, closing the circle. For this we consider $k = b = 1$ (only for simplicity, this might be done in a much more general case), and we suppose that $u \in V$ satisfies* (W), *and,* in addition *we suppose that $u'', f \in C[0,1]$. Integrating by parts (backwards) obtain that u also satisfies*

$$\int_0^1 (-u'' + u)v\,\mathrm{d}x = \int_0^1 fv\,\mathrm{d}x \qquad \forall\, v \in V\,.$$

From this relation, then apply the result from one of the previous exercises to conclude that u satisfies (D).

Exercise 6. *Consider the following two-point boundary value problem: Find u such that*

$$\begin{aligned} -u'' + u' &= 1 \qquad \text{in } (0,1)\,, \\ u(0) = u(1) &= 0\,. \end{aligned}$$

Write a weak formulation for this problem. Next, consider the finite element space V_h defined in (2.8), *and build the corresponding matrix $\mathbb{A}$. Is this matrix symmetric?*

2.3. **Abstract error analysis.** As seen in our computational experiments, as the mesh gets refined, the finite element solution becomes closer and closer to the exact solution. This phenomenon is referred to as *convergence of the discrete solution to the continuous one*. The main role of the mathematical analysis of the finite element method is to justify the observed convergence (by means of rigorous error estimates), and to set the right hypotheses for this to happen.[3]

The results in this section are abstract and can be applied to PDEs in any dimension. We shall use these later in the course as well as in this section. Before we begin, we will introduce a set of spaces and definitions that will be very useful for our purposes.

We start defining, for $D \subset \mathbb{R}^d$ ($d \leq 3$ in typical applications), the space of square integrable functions

$$L^2(D) = \{f : D \to \mathbb{R} \,:\, \int_D |f(\boldsymbol{x})|^2\,\mathrm{d}\boldsymbol{x} < +\infty\}\,. \tag{2.14}$$

In this space, we define the inner product

$$(f,g)_D = \int_D f(\boldsymbol{x})g(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \qquad \forall\, f, g \in L^2(D)\,, \tag{2.15}$$

[3]Interestingly, this process sometimes also provides insight on when things *will not work*. From a mathematical perspective this result is also of interest, as it acts as a red light to make practitioners aware of the fact they should not use certain variants of the method in situations when it will not work.

and associate norm

$$\|f\|_{0,D} = (f,f)_D^{\frac{1}{2}} = \left\{ \int_D |f(\boldsymbol{x})|^2 \, \mathrm{d}\boldsymbol{x} \right\}^{\frac{1}{2}} \qquad \forall\, f \in L^2(D)\,. \tag{2.16}$$

We also define the space of functions with square-integrable derivatives as follows

$$H^1(D) = \{ f \in L^2(D) : \partial_{x_i} f \in L^2(D) \text{ for } i = 1, \dots, d \}\,. \tag{2.17}$$

In $H^1(D)$ we define the inner product

$$(f,g)_{H^1(D)} = \int_D f(\boldsymbol{x}) g(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} + \sum_{i=1}^{d} \int_D \partial_{x_i} f(\boldsymbol{x}) \partial_{x_i} g(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}\,, \tag{2.18}$$

for all $f, g \in H^1(D)$. The norm induced by this inner product is

$$\|f\|_{1,D} = (f,f)_{H^1(D)}^{\frac{1}{2}} = \left\{ \|f\|_{0,D}^2 + \sum_{i=1}^{d} \|\partial_{x_i} f\|_{0,D}^2 \right\}^{\frac{1}{2}}\,. \tag{2.19}$$

Remark 3. *Different ways (sometimes used as the default ones) of writing the definition of $H^1(D)$ and its inner product are the following:*

$$H^1(D) = \{ f \in L^2(D) : \nabla f \in L^2(D)^d \}\,,$$

and

$$(f,g)_{H^1(D)} = \int_D f(\boldsymbol{x}) g(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} + \int_D \nabla f(\boldsymbol{x}) \cdot \nabla g(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}\,.$$

We also want to stress the fact that the requirement for f to be square integrable in D is not spurious. In fact, there are non-integrable functions that belong to $L^2(D)$. One of the most classical examples is the following: the function $f(x) = x^{-1}$ belongs to $L^2(1,\infty)$, but is not integrable in $(1,\infty)$. In the same way, the function $g(x) = x^{-1/2}$ is integrable in $(0,1)$, but doesn't belong to $L^2(0,1)$.

Finally, we define the space of functions in $H^1(D)$ that *vanish* on the boundary of D as

$$H_0^1(D) = \{ f \in H^1(D) : f|_{\partial D} = 0 \}\,, \tag{2.20}$$

where by $f|_{\partial D}$ we denote the *restriction* of f to the boundary of D (also referred to as *trace* of f on ∂D).

Remark 4. *We also remind here the definition and basic properties of a norm: Let V be a real vector space, equipped with a norm $\|\cdot\|_V$ i.e., $\|v\|_V$ is a real number, and we have the properties, for all $v, w \in V$ and $\alpha \in \mathbb{R}$:*

$$\|\alpha v\|_V = |\alpha| \|v\|_V\,, \quad \|v+w\|_V \le \|v\|_V + \|w\|_V\,,$$

and

$$\|v\|_V \ge 0 \quad \textit{with equality if and only if} \quad v = 0\,.$$

We finally recall a set of inequalities that will be very useful in the rest of the class.

Fundamental inequalities

Cauchy's inequality Let $a, b \in \mathbb{R}$, then

$$2ab \le a^2 + b^2\,. \tag{2.21}$$

Cauchy-Schwarz's inequality Let $u, v \in L^2(D)$ then

$$(u,v)_D \le \|u\|_{0,D} \|v\|_{0,D}\,. \tag{2.22}$$

Discrete Cauchy-Schwarz's inequality Let $\boldsymbol{a} = (a_1, \dots, a_n)\,, \boldsymbol{b} = (b_1, \dots b_n) \in \mathbb{R}^n$ then

$$\boldsymbol{a} \cdot \boldsymbol{b} \le \left\{ \sum_{i=1}^{n} a_i^2 \right\}^{\frac{1}{2}} \left\{ \sum_{i=1}^{n} b_i^2 \right\}^{\frac{1}{2}} =: \|\boldsymbol{a}\|_2 \|\boldsymbol{b}\|_2\,. \tag{2.23}$$

Example

Example 5. *Prove the above inequalities.*

We have already encountered a variational formulation in (W). In there, two main ingredients were needed. We will consider a *generic* Hilber space V, and recall that a *bilinear form* on V is a mapping $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ which is linear in each of its arguments. In addition, we recall that a linear mapping $L : V \to \mathbb{R}$ is called a *linear functional* on V.

Lets return to the abstract problem introduced in Section 2.1:

Continuous Problem

Find $u \in V$ such that

$$a(u,v) = L(v) \quad \text{for all} \quad v \in V\,, \tag{2.24}$$

where a is a bilinear form and L is a linear functional.

For problems of the type (2.24), we need V to be a *Hilbert space*, which means that, in addition, all Cauchy sequences in V converge to an element that belongs to V. $H^1(D)$, $H^1_0(D)$, and $L^2(D)$ are examples of Hilbert spaces.

We introduce the following hypotheses on $a(\cdot,\cdot)$ and $L(\cdot)$. For all $v, w \in V$:

$$\text{(A1)} \qquad a \text{ is } \textit{bounded}\text{: } |a(v,w)| \le \gamma \|v\|_V \|w\|_V,$$

$$\text{(A2)} \qquad a \text{ is } V\textit{-elliptic}\text{: } \quad a(v,v) \ge \alpha \|v\|_V^2,$$

$$\text{(A3)} \qquad L \text{ is } \textit{bounded}\text{: } \quad |L(v)| \le \lambda \|v\|_V,$$

where γ, α and λ are non-negative constants independent of v and w, and $\alpha > 0$.

The importance of these hypotheses on the bilinear form and linear functional is that, under those assumptions, the following fundamental well-posedness result can be proven. This Theorem, known as *Lax-Milgram's Lemma*, has had an immense impact on the theory of the finite element method. We will not prove all of it, since the first part of its proof requires techniques on functional analysis and/or metric spaces that some of you might not be acquainted with, but very nice proofs can be found in the books by Brenner & Scott, and by Ern & Guermond.

Theorem 1. *Suppose $a(\cdot,\cdot)$ is a bilinear form and $L(\cdot)$ is a linear functional and they satisfy* (A1)–(A3). *Then, there exists a unique solution u of* (2.24). *In addition, this solution satisfies the following* stability *estimate:*

$$\|u\|_V \le \frac{\lambda}{\alpha}\,. \tag{2.25}$$

Proof To prove uniqueness, assume that (2.24) has two solutions, $u_1, u_2 \in V$. Then

$$a(u_1, v) = L(v) = a(u_2, v)\,,$$

for all $v \in V$, and so by linearity, $a(u_1 - u_2, v) = 0$ for all $v \in V$. Since V is a vector space, $u_1 - u_2 \in V$ and thus $a(u_1 - u_2, u_1 - u_2) = 0$. By (A2) this implies

$$\alpha \|u_1 - u_2\|_V^2 \le a(u_1 - u_2, u_1 - u_2) = 0$$

and so $\|u_1 - u_2\|_V = 0$ which implies $u_1 = u_2$ (since $\|\cdot\|_V$ is a norm). To get (2.25) we use (A2) and (A3) to obtain

$$\alpha \|u\|_V^2 \ \le a(u,u) = L(u) \le \lambda \|u\|_V\,,$$

from which the result follows. (Divide each side by $\alpha\|u\|_V$ when $u \neq 0$, and note that (2.25) is trivially true when $u = 0$.) □

To analyse the error in a general way, we suppose that V_h is a finite-dimensional subspace of V, and consider the approximate problem:

Approximate Problem

Find $u_h \in V_h$ such that

$$a(u_h, v_h) = L(v_h) \quad \text{for all} \quad v_h \in V_h\,. \tag{2.26}$$

The following result appears as just applying Theorem 1 to (2.26). We will give a proof though since it is very interesting how, now we are operating with finite-dimensional objects, we are able to prove existence of solutions at the same time as uniqueness.

Theorem 2. *Suppose $a(\cdot,\cdot)$ is a bilinear form and $L(\cdot)$ is a linear functional and they satisfy* (A1)–(A3). *Then, the discrete problem problem* (2.26) *has a unique solution $u_h \in V_h$, and it satisfies*

$$\|u_h\|_V \leq \frac{\lambda}{\alpha}\,. \tag{2.27}$$

Remark 5. *The estimate* (2.27) *is a "discrete stability" result. The numerical solution cannot get indefinitely large as $h \to 0$ (this is, as the mesh gets refined, and the linear system involves more unknowns), and its bound is the same as that for the true solution.*

Proof We consider (2.26) and introduce a basis $\phi_1, \ldots, \phi_n$ of V_h. Then

$$u_h \in V_h \qquad \Longleftrightarrow \qquad u_h = \sum_{j=1}^{n} U_j \phi_j \qquad \text{for some } \boldsymbol{U} = (U_j)_{j=1}^n \in \mathbb{R}^n\,,$$

and (following the argument done previously), (2.26) is equivalent to

$$\mathbb{A}\mathbf{U} = \mathbf{f}, \tag{2.28}$$

where

$$a_{ij} := a(\phi_j, \phi_i) \quad \text{and} \quad f_i := L(\phi_i).$$

The existence and uniqueness of solutions of (2.26) then reduces to showing that the matrix $\mathbb{A}$ is non-singular. For this, let suppose $\boldsymbol{V} = (V_i)_{i=1}^n \in \mathbb{R}^n$ satisfies the equation $\mathbb{A}\boldsymbol{V} = \mathbf{0}$ and let us denote $v_h = \sum_{j=1}^n V_j \phi_j \in V_h$. Then,

$$0 = \boldsymbol{V}^T(A\boldsymbol{V}) = \sum_{i=1}^{n} V_i (A\boldsymbol{V})_i = \sum_{i=1}^{n}\sum_{j=1}^{n} V_i a\big(\phi_j, \phi_i\big)\, V_j$$

$$= a\left(\underbrace{\sum_{j=1}^{n} V_j \phi_j}_{=: v_h}, \sum_{i=1}^{n} V_i \phi_i \right) = a(v_h, v_h) \geq \alpha \|v_h\|_V^2 \geq 0,$$

where the second to last inequality comes from (A3). This chain of inequalities implies $v_h = 0$. Hence, since $\{\phi_1, \ldots, \phi_n\}$ is a linearly independent set, we have $\boldsymbol{V} = \mathbf{0}$. Therefore, $\mathbb{A}$ is a square matrix such that 0 is **not** one of its eigenvalues, which means that $\mathbb{A}$ is non-singular (i.e., invertible). Thus, (2.28), and hence (2.26), has a unique solution.

The bound (2.27) is proved analogously: we apply first (A2), then (W_h), and finally (A3), to obtain

$$\alpha \|u_h\|_V^2 \leq a(u_h, u_h) = L(u_h) \leq \lambda \|u_h\|_V,$$

giving (2.27). □

Now we have proven that both the continuous and discrete problems have a unique solution, we turn our attention to estimate the error. This will be done in two main steps. The first step has a geometric interpretation in the theory of Hilbert spaces, which justifies the use of the word *orthogonality* in its name.

Theorem 3 (Galerkin orthogonality)**.** *Let $a(\cdot,\cdot)$ be a bilinear form and let $L(\cdot)$ be a linear functional and suppose (A1)–(A3) are satisfied. Let $u \in V$ be the unique solution of (2.24) and let $u_h \in V_h$ be the unique solution of (2.26). Then*

$$a(u - u_h, v_h) = 0 \quad \forall\, v_h \in V_h. \tag{2.29}$$

Proof By (2.24) and the fact $V_h \subset V$ we have

$$a(u, v_h) = L(v_h) \quad \forall\, v_h \in V_h. \tag{2.30}$$

Also, u_h satisfies(2.26), which means

$$a(u_h, v_h) = L(v_h) \quad \forall\, v_h \in V_h. \tag{2.31}$$

Hence, subtracting the above two equalities and using that $a(\cdot,\cdot)$ is bilinear we obtain

$$a(u - u_h, v_h) = a(u, v_h) - a(u_h, v_h) = L(v_h) - L(v_h) = 0\,, \tag{2.32}$$

for all $v_h \in V_h$, which finishes the proof. □

As noticeable consequence of the Galerkin orthogonoality is the following results, known as Céa's Lemma. This result, in essence, shows that the error in the finite element method becomes smaller, if the space V_h becomes richer. [4]

Theorem 4 (Céa's lemma)**.** *Under the conditions of Theorem 3 we have*

$$\|u - u_h\|_V \le \frac{\gamma}{\alpha} \|u - v_h\|_V \qquad \text{for all } v_h \in V_h\,.$$

Hence

$$\|u - u_h\|_V \le \frac{\gamma}{\alpha} \min_{v_h \in V_h} \|u - v_h\|_V\,. \tag{2.33}$$

Remark 6. *We see that the error $u - u_h$ can never be any greater in norm than the error in best approximation of u by elements of V_h, multiplied by the constant γ/α (which is independent of h).*

Proof Using Theorem 3, we have, for any $v_h \in V_h$,

$$\begin{aligned}
\alpha \|u - u_h\|_V^2 &\le a(u - u_h, u - u_h) && \text{[by (A2)]}\\
&= a(u - u_h, u - v_h) + \underbrace{a(u - u_h, \underbrace{v_h - u_h}_{\in V_h})}_{=0 \text{ by Galerkin orthogonality}}\\
&\le \gamma \|u - u_h\|_V \|u - v_h\|_V\,. && \text{[by (A1)]}
\end{aligned}$$

If $u \neq u_h$, divide by $\alpha \|u - u_h\|_V$ to get

$$\|u - u_h\|_V \le \frac{\gamma}{\alpha} \|u - v_h\|_V\,.$$

The result is trivially true for $u = u_h$. □

Main Idea

If the finite element space V_h is chosen with care, then the finite element error $\|u - u_h\|_V$ will get smaller as the dimension of the space V_h increases. In addition, the rate will be the same rate as the *interpolation rate* $\min_{v_h \in V_h} \|u - v_h\|_V$.

In the next two examples (covered in the lectures), we will explore some more in details properties of the finite element method in which the bilinear form $a(\cdot,\cdot)$ is supposed to be, in addition, symmetric.

[4] This result was first proven by Jean Céa in his PhD thesis in 1964. It has become a cornerstone in the analysis of the finite element method. The deep importance of it is that it converts the problem of approximating a PDE to a problem of approximating a function by a piecewise polynomial (in principle, a simpler problem), like the ones you dealt with in the first part of this class.

Example

Example 6. *Let $a : V \times V \to \mathbb{R}$ be a bilinear form satisfying (A1) and (A2). In addition, let us suppose that $a(\cdot,\cdot)$ is* symmetric*, this is, for all $u, v \in V$,*

$$a(u,v) = a(v,u)\,.$$

Prove that $a(\cdot,\cdot)$ is an inner product, and thus that

$$\|v\|_a := a(v,v)^{\frac{1}{2}} \qquad \forall\, v \in V\,,$$

is a norm on V.

Example

Example 7. *(An error estimate in the energy norm). Prove that, under the same conditions as in Cea's Lemma, but assuming that the bilinear form $a(\cdot,\cdot)$ is also symmetric, the following refined version of Cea's Lemma holds:*

$$\|u - u_h\|_a = \min_{v_h \in V_h} \|u - v_h\|_a\,.$$

Main Idea

In the case of a symmetric bilinear form, the FEM actually computes the **best approximation** to u in the energy norm!

We can use Theorem 4 to obtain a rate of convergence and estimate the error $u - u_h$. We do this by choosing a particular $v_h \in V_h$ so that the right hand side of (2.33) is (a) easily estimated and (b) a good estimate of the error. To see this in action we return to the 1D case.

Exercises

Exercise 7. *Prove the following inequalities:*

$$ab \le \frac{a^2}{2} + \frac{b^2}{2} \qquad , \qquad ab \le a^2 + \frac{1}{4}b^2\,,.$$

for all $a, b \in \mathbb{R}$.

Exercise 8. *Prove that the matrix $\mathbb{A}$ associated to the finite element method for the problem*

$$-u'' = f \quad \text{in } (0,1) \quad , \quad u(0) = u(1) = 0\,,$$

using the finite element space V_h is invertible.

Exercise 9 ($b = 0$ is also acceptable)**.** *(1) Prove the following inequality, known as* Poincaré inequality*:*

$$\|v\|_{0,(0,1)} \le \|v'\|_{0,(0,1)} \qquad \forall\, v \in H_0^1(0,1)\,. \tag{2.34}$$

(2) Then, show that the weak problem: find $u \in H_0^1(0,1)$ such that

$$\int_0^1 u'v'\,\mathrm{d}x = \int_0^1 fv\,\mathrm{d}x \qquad \forall\, v \in H_0^1(0,1)\,,$$

has a unique solution $u \in H_0^1(0,1)$.

2.4. **Application of the abstract theory in 1D.** Here we apply the theory of the previous section to the differential equation (D) and its weak form (W) on the space V of Section 2.1, approximated using piecewise linear FEM.

Here we shall assume b and k to be positive constants. Thus:

$$a(u,v) = \int_0^1 \left(ku'v' + buv\right)\,\mathrm{d}x \quad \text{and} \quad L(v) := \int_0^1 fv\,\mathrm{d}x\,,$$

and the space V is given in this case by $H_0^1(0,1)$.

Example

Example 8. *Show that $a(\cdot,\cdot)$ satisfies* (A1) *and* (A2) *with $V = H_0^1(0,1)$.*

Let u_h be the FEM solution. Then Cea's Lemma implies that

$$\|u - u_h\|_{1,(0,1)} \le C\,\|u - v_h\|_{1,(0,1)} \qquad \forall v_h \in V_h\,, \tag{2.35}$$

where $C > 0$ depends only on the values of k and b, but not on h.

The left-hand side of (2.35) is a measure of the error $u - u_h$, where u is the unknown exact solution. To prove $u_h \to u$ as $h \to 0$, we make a clever choice for v_h so that the right-hand side can be estimated in terms of some power of h. For this, at least when u is smooth, we typically make use of the *Lagrange interpolant* of u, denoted by $\Pi_h u$. For any $v \in C[0,1]$ with $v(0) = 0 = v(1)$, let

$$\Pi_h v := \sum_{i=1}^{n} v(x_i)\,\phi_i \quad \in V_h\,.$$

Then $\Pi_h v(x_j) = v(x_j)$ for each $j = 1,\ldots,n$. Also, $\Pi_h : C[0,1] \to V_h$ is a linear mapping.

To analyse the error we define the following space

$$H^2(0,1) = \{v \in H^1(0,1) : v'' \in L^2(0,1)\}\,. \tag{2.36}$$

With the help of this space, we now state the following *interpolation estimate*.

Lemma 1. *Suppose $v \in H^2(0,1)$ with $v(0) = 0 = v(1)$. Then, there exists a constant $C > 0$, independent of h, such that*

$$\|(v - \Pi_h v)'\|_{0,(0,1)} \le \frac{h}{\sqrt{2}}\,\|v''\|_{0,(0,1)}\,, \tag{2.37}$$

$$\|v - \Pi_h v\|_{0,(0,1)} \le \frac{h^2}{\sqrt{2}}\,\|v''\|_{0,(0,1)}\,. \tag{2.38}$$

(2.39)

So we can use $v_h = \Pi_h u$ in (2.35) and then Lemma 1 with $v = u$ to get:

Corollary 1. *Suppose u solves* (W) *and suppose $u \in H^2(0,1)$. Then, there exists $C > 0$, independent of h, such that*

$$\|u - u_h\|_{1,(0,1)} \le C\,h\,\|u''\|_{0,(0,1)}\,. \tag{2.40}$$

Main Idea

Thus, provided u has two continuous derivatives, $u_h \to u$ with $\mathrm{O}(h)$ in the norm of $H^1(0,1)$. (Note that $\Pi_h u$ is used for theoretical reasons only. In general, we do not know u and so we do not know $\Pi_h u$ either.)

Worked example

Example 9. *Consider the problem*

$$-u'' = f \quad \text{on} \quad [0,1], \quad u(0) = 0 = u(1)\,.$$

Use the provided Matlab code to compute the finite element solution using $h = 1/11, 1/21, 1/41$, *for each of the cases (a)* $f(x) = -2$ *and (b)* $f(x) = 9\pi^2 \sin(3\pi x)$. *For each case, compute the error* $\|u - u_h\|_{0,(0,1)}$ *and* $\|u - u_h\|_{1,(0,1)}$.
Explain the observed behaviour of the error as n *increases in each case.*

Worked example

Example 10. *Consider the problem from the previous example with*

$$f(x) = \exp(x^2)\,.$$

The solution is not known analytically. Find, $u(1/2)$, *correct to 8 significant figures.*

2.5. **Some further interesting properties of the finite element method.** One problem that poses itself in some instances is related to the preservation of positivity of the finite element method. For example, if a partial differential equation models concentration of a certain substance, the its solution should belong to the interval $[0,1]$ (as concentrations usually do). It is natural to ask ourselves the following question: if the solution of the continuous problem belongs to a certain interval, does the solution of the finite element method belong to the same interval? The answer is unfortunately negative in most interesting cases. To explain this, we will introduce the concept of matrices of non-negative type, and then apply it to a finite element method in one space dimension as an example.

We will consider the finite element method: Find $u_h \in V_h = \operatorname{span}\big(\{\phi_1, \ldots, \phi_n\}\big)$ such that

$$a(u_h, v_h) = L(v_h) \qquad \forall\, v_h \in V_h\,. \tag{2.41}$$

For the moment we will not assume that the bilinear form $a(\cdot,\cdot)$, linear functional $L(\cdot)$, or the finite element space V_h are the ones from the model problems in the last sections. We will only assume that the matrix $\mathbb{A} = (a_{ij})_{i,j=1}^n$ associate to it, that is, $a_{ij} = a(\phi_j, \phi_i)$, is invertible.

Definition 3 (Matrix of non-negative type)**.** *The matrix* $\mathbb{A} \in \mathbb{R}^{n\times n}$ *will be said to be of* **non-negative type** *if*

$$a_{ij} \leq 0 \qquad \forall\, i \neq j, 1 \leq i, j \leq n\,, \tag{2.42}$$

$$\sum_{j=1}^{n} a_{ij} \geq 0 \qquad \forall\, 1 \leq i \leq n\,. \tag{2.43}$$

The main result concerning this property is the following theorem. Its origins can be traced back to very early works by Philippe G. Ciarlet in the late 1960s.

Theorem 5. *Let* $\mathbb{A}$ *be the matrix associated to the finite element problem* (2.41). *Then, if the matrix* $\mathbb{A}$ *is of non-negative type we have the following result:*
- If $f \geq 0$, *then the solution* u_h *cannot attain a local minimum value inside the interval* $(0,1)$;
- If $f \leq 0$, *then the solution* u_h *cannot attain a local maximum value inside the interval* $(0,1)$.

The proof of this problem will be covered in the lectures, but it is interesting to notice one of its main consequences. Imagine that f is positive or zero, then this means that the solution remains positive in the interval $[0,1]$. In fact, if it were negative at some point inside $(0,1)$, then forcibly it would have a local minimum, which contradicts the main result. So, in this particular case, the finite element method produces a positive approximate solution if the right-hand side is positive. This property is referred to as *preservation of positivity* and has been a very active area of develpment

in the last couple of decades, especially for more complex problems where the quantities of interest **must** have a sign (e.g., the previous example related to concentration).

So, going back to our model problem (D) approximated by the finite element method with the space V_h defined in (2.8) (and its standard 'hat' basis functions), we recall that the matrix is tridiagonal and given by

$$\mathbb{A} = \frac{1}{h}\begin{bmatrix} 2 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 2 \end{bmatrix}, \tag{2.44}$$

and then we see that:

- $\mathbb{A}$ is invertible;
- for every $i = 1, \dots, n$ and all $j \neq i$, then a_{ij} is either -1, or 0. In any case, $a_{ij} \leq 0$, which proves (2.42); and
- we see that

$$\sum_{j=1}^{n} a_{ij} = \begin{cases} 2 - 1 = 1 > 0 & \text{if } i = 1 \text{ or } i = n\,, \\ -1 + 2 - 1 = 0 & \text{if } i \in \{2, \dots, n-1\}\,, \end{cases} \tag{2.45}$$

  which proves (2.43).

So, for our model problem (D) with piecewise linear finite element methods then whenever the right-hand side f is positive, then the solution is positive.

Worked example

Example 11. *Compute the matrix $\mathbb{A}$ associated to the following reaction-diffusion problem*

$$-\varepsilon u'' + u = f \quad \text{in } (0,1) \quad , \quad u(0) = u(1) = 0\,. \tag{2.46}$$

Compute the largest value for h for which the matrix associated to this problem is of non-negative type.
In addition, for $\varepsilon = 10^{-3}$ check whether the results in the Matlab code reproduce this fact.

Exercises

Exercise 10. *Consider the systems of linear equations arising from a finite element method:*

$$\mathbb{A}\mathbf{U}_1 = \mathbb{F}_1 \qquad \text{and} \qquad \mathbb{A}\mathbf{U}_2 = \mathbb{F}_2\,,$$

that is, consider the finite element method with two different right-hand sides. Let us assume that $\mathbb{F}_i = (f_j^i)_{j=1}^n$ for $i = 1, 2$, and suppose that for all $j = 1, \dots, n$ we have $f_j^1 \geq f_j^2$. Then, prove that each component of $\mathbf{U}_1 - \mathbf{U}_2$ is positive.

Exercise 11. *Consider the following reaction-diffusion problem*

$$b(x)u - \big(k(x)u'\big)' = f \quad \text{in } (0,1) \quad , \quad u(0) = u(1) = 0\,. \tag{2.47}$$

Let us assume that the coefficients b and k are C^1 and satisfy $0 < k_0 \leq k(x)$ and $0 \leq b(x) \leq b_1$ for all $x \in (0,1)$. Find a condition on $h, k(\cdot)$ and $b(\cdot)$ that needs to be satisfied so the matrix $\mathbb{A}$ associated to the finite element method is of non-negative type.

3. Finite elements in 2D

3.1. **A partial differential equation and its weak form.** Let $\Omega \subset \mathbb{R}^2$ be a bounded open domain with a polygonal boundary Γ (i.e., Γ is composed of straight-line segments). Consider the following partial differential equation (PDE)[5] on Ω:

$$-\operatorname{div}(\mathcal{D}\nabla u) = f, \tag{3.1}$$

where the unknown function $u = u(\boldsymbol{x})$, for $\boldsymbol{x} \in \Omega$, has to be found subject to some boundary conditions (specified later). The data $f = f(\boldsymbol{x})$ is assumed to be continuous, $f \in C(\overline{\Omega})$, and $\mathcal{D} = \mathcal{D}(\boldsymbol{x})$ is assumed to be a 2×2 symmetric matrix for all $\boldsymbol{x} \in \Omega$. We will specify more conditions on $\mathcal{D}$ later on.[6] In most cases this matrix is called the *diffusion* matrix, but it can also stand for other physical characteristics (such as porosity, or conductivity).

Example

Example 12 (Canonical example)**.** *The simplest case, namely $\mathcal{D} = I_2$ (the 2×2 identity matrix) leads to the Poisson equation. In fact,* $\operatorname{div}(\mathcal{D}\nabla u) = \operatorname{div}(\nabla u) = \Delta u$.

Example

Example 13 (Aeronautical engineering)**.** *A diffusion equation in flow around an aeroplane wing takes the form* (3.1) *with*

$$\mathcal{D} := \begin{bmatrix} (1 - M^2) & 0 \\ 0 & 1 \end{bmatrix}.$$

Example

Example 14 (Heat conduction)**.** *Let Ω be a conducting body with heat sources $f(\boldsymbol{x})$, $\boldsymbol{x} \in \Omega$, as in Figure 3.1. Let q_i be the heat flow in direction x_i. Then, if $u = u(\boldsymbol{x})$ denotes the temperature at $\boldsymbol{x} \in \Omega$,* Fourier's law *states that*

$$q_i = -\, d_{i1}\frac{\partial u}{\partial x} - d_{i2}\frac{\partial u}{\partial y} = -(\mathcal{D}\nabla u)_i. \tag{3.2}$$

Here, $\mathcal{D} = (d_{ij})$ is the conductivity matrix*. Also, the* law of conservation of energy *states that in the stationary (time-independent) case*

$$\operatorname{div} \mathbf{q} = f. \tag{3.3}$$

That is, the heat flow is balanced by the heat sources. Substituting (3.2) *into* (3.3)*, we have again a problem of the form* (3.1)*.*

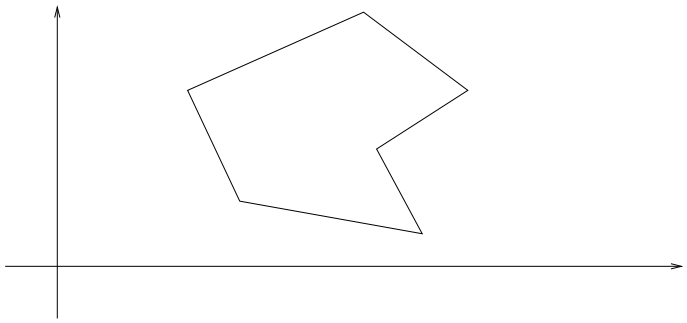

Figure 3.1. A polygonal heat-conducting body

[5] Later we will generalise this to $-\operatorname{div}(\mathcal{D}\nabla u) + bu = f$

[6] In particular we will assume that $\mathcal{D}$ is uniformly positive definite.

To determine the solution u of (3.1) uniquely, we need to specify boundary conditions; for example, we may specify the *homogeneous Dirichlet conditions*:

$$u(\boldsymbol{x}) = 0, \qquad \boldsymbol{x} \in \Gamma. \tag{3.4}$$

Main Idea

To solve (3.1),(3.4) numerically, we use the same two-stage process as in Section 2:

Step 1: Write the PDE in weak form.
Step 2: Approximate this weak form.

To get the weak form, we need the two-dimensional version of *integration by parts*:

Theorem 6 (Green's formula)**.** *Let $v, w\colon \Omega \to \mathbb{R}$ be sufficiently smooth, and let the components of $\mathcal{D}$ also be sufficiently smooth. Then,*

$$\int_\Omega \operatorname{div}(\mathcal{D}\nabla w)\, v \,\mathrm{d}\boldsymbol{x} = \int_\Gamma (\mathcal{D}\nabla w)\cdot \boldsymbol{n}\, v \,\mathrm{d}S - \int_\Omega (\mathcal{D}\nabla w)\cdot\nabla v \;\mathrm{d}\boldsymbol{x},$$

where $\boldsymbol{n}$ denotes the unit outward normal to Ω on Γ.

Proof. Let $\mathbf{F} := (\mathcal{D}\nabla w)\, v$. Then $\operatorname{div}\mathbf{F} = \operatorname{div}(\mathcal{D}\nabla w)\, v + (\mathcal{D}\nabla w)\cdot\nabla v$. We insert this into the divergence theorem:

$$\int_\Omega [\operatorname{div}(\mathcal{D}\nabla w)\, v + (\mathcal{D}\nabla w)\cdot\nabla v] \;\mathrm{d}\boldsymbol{x} = \int_\Gamma (\mathcal{D}\nabla w)\cdot\boldsymbol{n}\, v \,\mathrm{d}S,$$

which is the desired result. □

We return to the problem (3.1),(3.4):

$$\begin{aligned} -\operatorname{div}(\mathcal{D}\nabla u) &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \Gamma. \end{aligned} \tag{D}$$

To obtain the weak form, let V be a space of suitable functions on Ω satisfying $v = 0$ on Γ (more details below). We multiply (D) by $v \in V$, integrate and use Green's formula to get

$$\int_\Omega (\mathcal{D}\nabla u)\cdot\nabla v \,\mathrm{d}\boldsymbol{x} - \underbrace{\int_\Gamma (\mathcal{D}\nabla u)\cdot\boldsymbol{n}\, v \;\mathrm{d}S}_{=0 \text{ since } v|_\Gamma = 0} = \int_\Omega f v \,\mathrm{d}\boldsymbol{x}.$$

Hence the weak form of (D) in two dimensions is: Find $u \in V$ such that

$$a(u, v) = L(v) \qquad \forall\, v \in V, \tag{W}$$

where

$$a(u,v) := \int_\Omega (\mathcal{D}\nabla u)\cdot\nabla v \,\mathrm{d}\boldsymbol{x} \quad \text{and} \quad L(v) := \int_\Omega f v \,\mathrm{d}\boldsymbol{x}. \tag{3.5}$$

Worked example

Example 15. *Compute the weak form with $\mathcal{D} = k\boldsymbol{I}$ ($k : \mathbb{R}^2 \to \mathbb{R}$ is a scalar function and I is the 2×2 identity)*

Looking at the weak formulation (W) the natural requirements for V become clear. First, functions in V need to vanish on the boundary. In addition, the products uv and $\nabla u\cdot\nabla v$ need to be integrable for every u and v in V. So, a natural choice for V is the following:

$$V = \left\{ v\colon \Omega\to\mathbb{R} : \int_\Omega |v|^2 \;\mathrm{d}\boldsymbol{x} < \infty\,, \int_\Omega |\nabla v|^2 \;\mathrm{d}\boldsymbol{x} < \infty\,, \text{and } v = 0 \text{ on } \Gamma \right\} = H_0^1(\Omega)\,.$$

The abstract approximation of (W) is the same as before: Choose $V_h \subset V$ and seek $u_h \in V_h$ such that

$$(\mathrm{W}_h) \qquad a(u_h, v_h) = L(v_h) \qquad \forall\, v_h \in V_h.$$

3.2. **Piecewise-linear finite elements in 2D.** Let Ω be a region with a polygonal boundary Γ. We decompose Ω into a *mesh* of triangular *elements*.

The vertices of the triangles are called *nodes*. Denote the collection of elements by $\mathscr{T}$ and a typical element including its boundary by $K \in \mathscr{T}$. The set of nodes of the mesh is denoted by $\mathscr{N}$. We will only work with *admissible* triangulations, for which we require:

- $\overline{\Omega} := \Omega \cup \Gamma = \bigcup_{K \in \mathscr{T}} K$.
- If $K, K' \in \mathscr{T}$, $K \neq K'$, then either
  (1) $K \cap K' =$ a whole edge of two triangles, or
  (2) $K \cap K' =$ a node, or
  (3) $K \cap K' = \emptyset$.

In Figure 3.2 we present an example of an admissible and of a non-admissible mesh.

The *mesh diameter* h of $\mathscr{T}$ is defined as

$$h := \max_{K \in \mathscr{T}} h_K,$$

where h_K is the *diameter* of K (i.e., the length of the longest side of K).

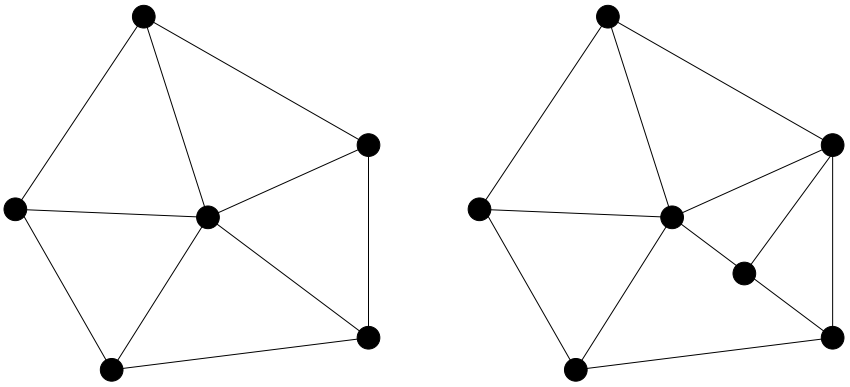

FIGURE 3.2. Two meshes, the one on the left being admissible and the one on the right not being admissible (the dark dots denotes the nodes of the triangulation).

Remark 7. *If Ω had curved rather than straight edges, we could approximate it well by sufficiently fine triangles or use curvi-linear triangles. We will not go into the latter type of elements in this class, but generating meshes with more general geometries for the elements has been a very active area of research in the last few years. Nowadays, the mesh generation process can be carried out automatically, and there are several* mesh generators *like gmsh* `https://gmsh.info/`*, or Triangle* `https://www.cs.cmu.edu/~quake/triangle.html`.

The next step is to build the finite element space. The first step for this is to define appropriate basis functions. For any $\boldsymbol{x}_i \in \mathscr{N}$, define $\phi_i \colon \Omega \to \mathbb{R}$ such that

$$(3.6) \qquad \phi_i(\boldsymbol{x}_j) = \delta_{ij} \qquad (\text{where } \delta_{ij} \text{ is defined as in (2.9)}),$$

and such that

$$(3.7) \qquad \phi_i(\boldsymbol{x}) \text{ is a polynomial of degree} \leq 1 \text{ in every } K \in \mathscr{T}.$$

We stress that by 'polynomial of degree ≤ 1' we mean that the restriction of $\phi_i(\boldsymbol{x})$ to K is a linear combination of the functions 1, x, and y. Equation (3.7) implies that the graph of ϕ_i is *planar* on each $K \in \mathscr{T}$ and that it is determined uniquely by its values at the three nodes of K(three points determine a plane; a three-legged table never wobbles). Equation (3.6) implies that ϕ_i is non-zero

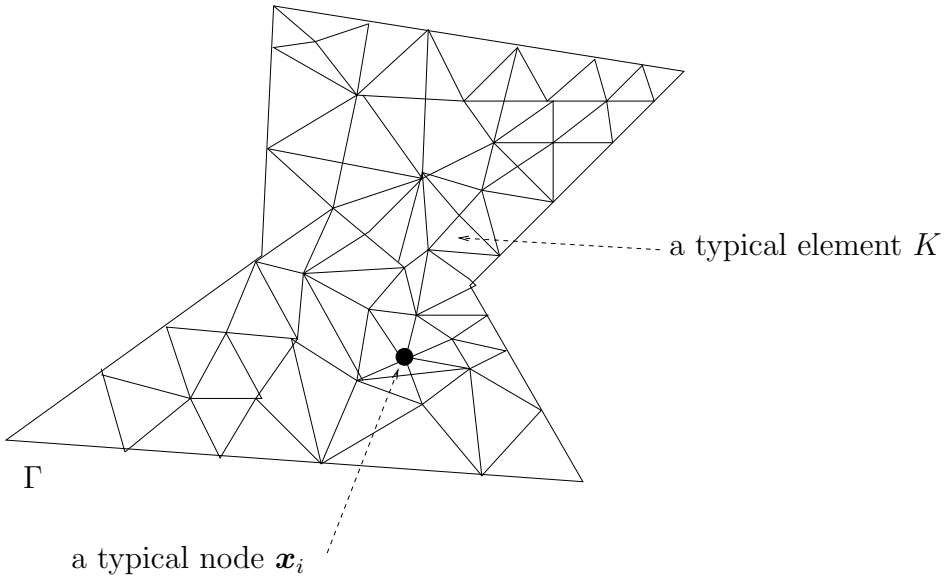

FIGURE 3.3. A typical mesh on a more complicated domain Ω

only on triangles containing node $\boldsymbol{x}_i$ (see Figure 3.4). A three-dimensional graph of a typical ϕ_i is given in Figure 3.4.

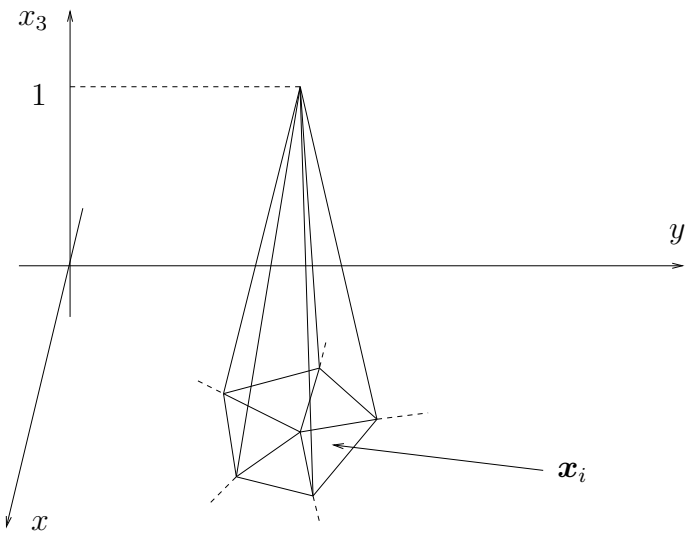

FIGURE 3.4. Three-dimensional graph of a typical ϕ_i

We now introduce the space

$$V_h = \left\{ v_h \in C\left(\bar{\Omega}\right) : \ v_h|_K \in \mathbb{P}_1 \ \forall K \in \mathscr{T} \text{ and } v_h|_\Gamma = 0 \right\}.$$

A suitable basis for V_h is

$$\{\phi_i : \boldsymbol{x}_i \in \mathscr{N}_0\},$$

where

$$\mathscr{N}_0 = \{\boldsymbol{x}_i \in \mathscr{N} \ : \ \boldsymbol{x}_i \notin \Gamma\}.$$

Remark 8. *We notice that the boundary nodes are excluded to ensure the homogeneous boundary conditions are satisfied. We will come back to this later.*

Consider the problem (W). Using the particular finite-dimensional subspace $V_h \subset V$, the problem (W_h) reads: Find $u_h \in V_h$ such that

(W_h) $$a(u_h, v_h) = L(v_h) \qquad \forall v_h \in V_h.$$

Writing

$$u_h = \sum_{\boldsymbol{x}_j \in \mathscr{N}_0} U_j \phi_j,$$

and following exactly the same steps as in the last section, we get to

$$u_h \text{ solves } (\mathrm{W}_h) \qquad \Longleftrightarrow \qquad \sum_{\boldsymbol{x}_j \in \mathscr{N}_0} \underbrace{a(\phi_j, \phi_i)}_{=: a_{ij}} U_j = \underbrace{L(\phi_i)}_{=: f_i} \qquad \forall \boldsymbol{x}_i \in \mathscr{N}_0,$$

or equivalently

$$\mathbb{A}\mathbf{U} = \mathbf{f}. \tag{3.8}$$

In practice, the nodes in $\mathscr{N}_0$ are given an ordering, say $\{\boldsymbol{x}_j : j = 1, \ldots, n\}$, and so (3.8) is a linear system in $\mathbb{R}^n$ with

$$a_{ij} := \int_\Omega \big(\mathcal{D}\nabla\phi_j\big) \cdot \nabla\phi_i \, \mathrm{d}\boldsymbol{x}\,, \tag{3.9}$$

and

$$f_i := \int_\Omega f\phi_i \, \mathrm{d}\boldsymbol{x}. \tag{3.10}$$

We call $\mathbb{A}$ the *stiffness matrix* and $\mathbf{f}$ the *load vector*.[7] Note that $a_{ij} = 0$ except when $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are common nodes of at least one element $K \in \mathscr{T}$ (see Figure 3.5).

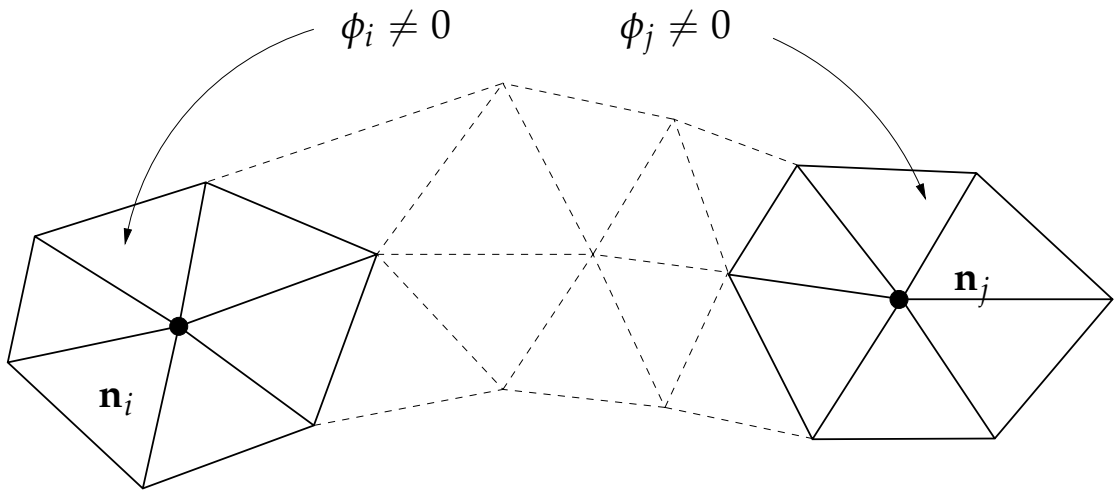

FIGURE 3.5. The matrix entries a_{ij} vanish unless $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are common nodes of at least one element $K \in \mathscr{T}$.

The stiffness matrix $\mathbb{A}$ is again sparse. In fact, the number of non-zero elements in row i equals the number of nodes connected to $\boldsymbol{x}_i$ by zero or one edge, which is typically small compared to the total number of nodes.

3.3. **System assembly.** One of the key reasons for the popularity of the finite element method is the elegant way of assembling the system (3.8). Instead of integrating over all of Ω in (3.9), we write the integral as

$$a_{ij} = \sum_{K \in \mathscr{T}} \underbrace{\int_K (\mathcal{D}\nabla\phi_j) \cdot \nabla\phi_i \, \mathrm{d}\boldsymbol{x}}_{=: a_{ij}^K}\,. \tag{3.11}$$

[7]These terms originate from the application in elasticity where finite elements were first introduced.

Definition 4. *The matrix $\mathbb{A}^K$ is called the* element stiffness matrix *for element $K \in \mathscr{T}$.* [8]

We note that $a_{ij}^K = 0$ unless $\boldsymbol{x}_i, \boldsymbol{x}_j$ are both nodes of element K. So $\mathbb{A}^K$ (which is an $n \times n$ matrix) can in practice be stored as a 3×3 matrix $\widetilde{\mathbb{A}}^K = (\tilde{a}_{pq}^K)_{p,q=1}^3$ with rows and columns corresponding to the 3 nodes of K. The sum in (3.11) can be taken over those K which contain both $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ (*not* over all the triangles in $\mathscr{T}$).

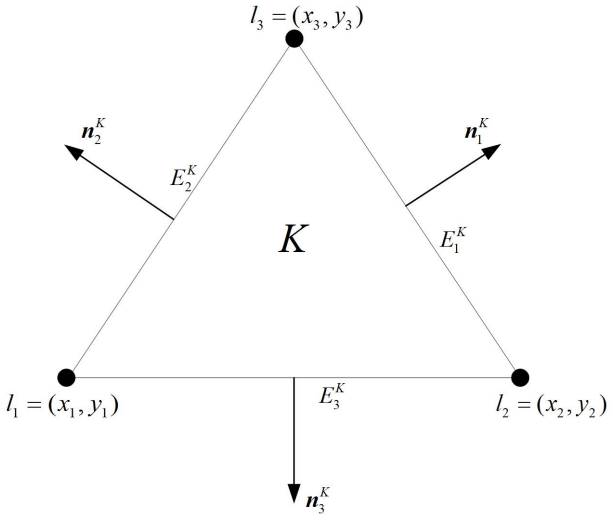

FIGURE 3.6. Anti-clockwise numbering of the nodes of a triangle K, and normal and tangent vectors.

Now lets consider any $K \in \mathscr{T}$ and introduce a *local anti-clockwise numbering* $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \boldsymbol{\ell}_3$ of the nodes of K as in Figure 3.6. With respect to this numbering, we define the *local basis functions* $\phi_1^K, \phi_2^K, \phi_3^K$ as the only $\mathbb{P}_1$ functions that satisfy

$$\phi_p^K(\boldsymbol{\ell}_q) = \delta_{pq} \qquad \text{for } p, q = 1, 2, 3\,;$$

that is, the restriction to K of the basis functions of the space V_h.

For the moment, we only need to compute the gradient of the basis functions to assemble the local matrix. Let then $\boldsymbol{n}_1^K, \boldsymbol{n}_2^K, \boldsymbol{n}_3^K$ be the 3 normal vectors associated to the edges E_1^K, E_2^K, E_3^K of K (see Figure 3.6). The first thing to notice is that the gradient of these 3 local basis functions is a constant vector in $\mathbb{R}^2$. Then, for $p = 1, 2, 3$, the gradient of ϕ_p^K can be decomposed as

$$\nabla\phi_p^K = \left(\nabla\phi_p^K \cdot \boldsymbol{t}_p^K\right) \boldsymbol{t}_p^K + \left(\nabla\phi_p^K \cdot \boldsymbol{n}_p^K\right) \boldsymbol{n}_p^K\,,$$

where $\boldsymbol{t}_p^K$ is the *unit tangent vector* to E_p^K, that is,

$$\boldsymbol{t}_1^K = \frac{1}{|\boldsymbol{\ell}_3 - \boldsymbol{\ell}_2|}(\boldsymbol{\ell}_3 - \boldsymbol{\ell}_2) \quad , \quad \boldsymbol{t}_2^K = \frac{1}{|\boldsymbol{\ell}_1 - \boldsymbol{\ell}_3|}(\boldsymbol{\ell}_1 - \boldsymbol{\ell}_3) \quad , \quad \boldsymbol{t}_3^K = \frac{1}{|\boldsymbol{\ell}_2 - \boldsymbol{\ell}_1|}(\boldsymbol{\ell}_2 - \boldsymbol{\ell}_1)\,. \tag{3.12}$$ [9]

Next, since the local basis function ϕ_p^K vanishes on the edge E_p^K (as it vanishes at its two end points), then the derivative of ϕ_p^K in the direction of $\boldsymbol{t}_p^K$ is zero. Hence, denoting by $|K|$ the area of the element K, and by $|E_1^K| = |\boldsymbol{\ell}_3 - \boldsymbol{\ell}_2|$ (and analogously for $|E_2^K|$ and $|E_3^K|$), $\nabla\phi_p^K$ is given by

$$\nabla\phi_p^K = -\frac{|E_p^K|}{2|K|}\boldsymbol{n}_p^K\,. \tag{3.13}$$

Example

Example 16. *Prove formula* (3.13).

[8] This is the analogue of the local matrices introduced in §1.2

[9] Notice that for this decomposition we have used that the normal and tangent vectors are actually orthogonal.

Remark 9. *There is even a simpler expression to compute $\nabla\phi_p^K$, realising that the normal vector to an edge outer to K is the rotation of its tangent vector by an angle of $\frac{3\pi}{2}$. So, we can see that*

$$\boldsymbol{n}_1^K = \begin{pmatrix} \cos\left(\frac{3\pi}{2}\right) & -\sin\left(\frac{3\pi}{2}\right) \\ \sin\left(\frac{3\pi}{2}\right) & \cos\left(\frac{3\pi}{2}\right) \end{pmatrix} \boldsymbol{t}_1^K = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} \frac{x^3-x^2}{|E_1^K|} \\ \frac{y^3-y^2}{|E_1^K|} \end{pmatrix} = \frac{1}{|E_1^K|} \begin{pmatrix} (y^3-y^2) \\ -(x^3-x^2) \end{pmatrix}, \tag{3.14}$$

and analogous expressions for $\boldsymbol{n}_2^K$ and $\boldsymbol{n}_3^K$. This expression for the normal vector is very useful to implement the finite element method in an efficient way.

To compute the area of the element K we will use a classical formula, known as *Heron's formula*. If we denote by s the number

$$s = \frac{|E_1^K| + |E_2^K| + |E_3^K|}{2}, \tag{3.15}$$

then, Heron's formula states that the area of the triangle K is given by

$$|K| = \sqrt{s(s-|E_1^K|)(s-|E_2^K|)(s-|E_3^K|)}\,. \tag{3.16}$$

Its proof will not be shown in this class, but it is a very good exercise in trigonometry.

Thus, with these results, the local matrix $\mathbb{A}^K$ is given by

$$a_{ij}^K = \tilde{a}_{pq}^K = \frac{|E_q^K|\,|E_p^K|}{4|K|^2} \int_K \left(\mathcal{D}\boldsymbol{n}_q^K\right) \cdot \boldsymbol{n}_p^K \,\mathrm{d}\boldsymbol{x}\,, \tag{3.17}$$

where p and q are the local numbers of the nodes $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, respectively.

In general, the coefficient matrix $\mathcal{D}$ in (3.17) depends on $\boldsymbol{x}$ and numerical quadrature is needed to compute the integrals in (3.11). When $\mathcal{D} = I$, everything can be done exactly, and is what will be done in the next section.

3.4. **Special case: Poisson's equation.** Consider Poisson's equation (i.e., set $\mathcal{D} = I$ in (3.1)). Since $\operatorname{div}\nabla u = \Delta u$, this corresponds to

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

The weak form of (D) is again (W) in Section 3.1, but with $\mathcal{D} = I$. Therefore (3.11) becomes

$$a_{ij} = \sum_{K\in\mathcal{T}} \int_K \nabla\phi_j \cdot \nabla\phi_i \,\mathrm{d}\boldsymbol{x} = \sum_{K\in\mathcal{T}:\boldsymbol{x}_i,\boldsymbol{x}_j\in K} \underbrace{\int_K \nabla\phi_j \cdot \nabla\phi_i \,\mathrm{d}\boldsymbol{x}}_{a_{ij}^K}. \tag{3.18}$$

Here

$$\begin{aligned} a_{ij}^K &= |K|\,\Big(\nabla\phi_j \cdot \nabla\phi_i\Big)|_K && [\nabla\phi_i, \nabla\phi_j \text{ are constant vectors on } K] \\ &= |K|\,\nabla\phi_q^K \cdot \nabla\phi_p^K && [\text{where locally on } K\text{: } \boldsymbol{\ell}_p = \boldsymbol{x}_i \text{ and } \boldsymbol{\ell}_q = \boldsymbol{x}_j] \\ &= \frac{|E_q^K|\,|E_p^K|}{4|K|}\,\boldsymbol{n}_q^K \cdot \boldsymbol{n}_p^K && [\text{using (3.13)}] \\ &=: \widetilde{a}_{pq}^K. \end{aligned}$$

Here $\boldsymbol{\ell}_p, \boldsymbol{\ell}_q$ are the local nodes on K corresponding to the global nodes $\boldsymbol{x}_i, \boldsymbol{x}_j$; $\mathbb{A}^K$ is stored as the 3×3 matrix $\widetilde{\mathbb{A}}^K$ (using the local numbering).

Main (algorithmic) Idea

We have that the $n \times n$ matrix $\mathbb{A}$ and the 3×3 matrix $\tilde{\mathbb{A}}^K$ are intimately linked. This is at the heart of the practical implementation of the finite element method. A typical finite element code (as the one we will deal with in a later lecture), *performs a loop over the elements, computes the local matrices, and sends the entries of the local matrices to the right positions in the global one.*
More precisely, since two nodes belong (simultaneously) to at most two elements, then the global matrix element a_{ij} is the sum of at most two contributions coming from the local matrices of these elements.

Worked example

Example 17. *Let us suppose that for a given triangulation we have that the nodes $\boldsymbol{x}_{10}$ and $\boldsymbol{x}_{12}$ belong to the two elements K and K'. Also, let us suppose that for K and K' we have the following local numbering:*

(3.19)

	local	*global*
for K	*1*	*10*
	2	*15*
	3	*1*

,

	local	*global*
for K'	*1*	*10*
	2	*30*
	3	*15*

.

Then, we have that

(3.20)
$$a_{10\,15} = \tilde{a}^K_{12} + \tilde{a}^{K'}_{13} \,.$$

Worked example

Example 18. *We approximation the Poisson equation given by*

$$\begin{aligned} -\Delta u &= f \qquad \text{in } \Omega = (0,3) \times (0,2), \\ u &= 0 \qquad \text{on } \Gamma \,, \end{aligned}$$

with the mesh *on Ω shown in Figure 3.7. In this example we will assemble the 2×2 matrix $\mathbb{A}$.*

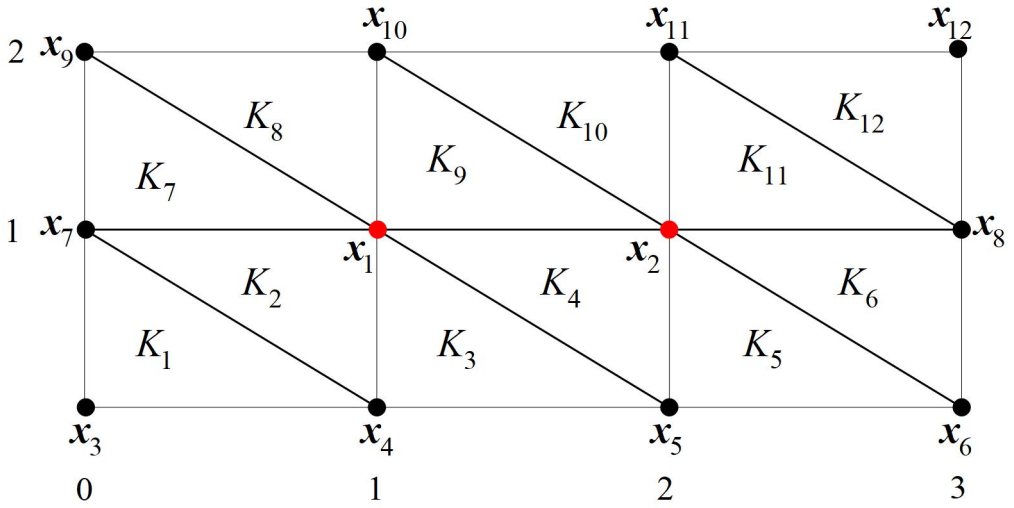

FIGURE 3.7. The mesh used to solve Poisson's equation on $\Omega = (0,3) \times (0,2)$.

Now consider the right-hand side (load) vector in system (3.8), given by

$$f_i = \int_\Omega f\phi_i \, \mathrm{d}\boldsymbol{x} = \sum_{K\in\mathscr{T}\,:\,\boldsymbol{x}_i\in K} \underbrace{\int_K f\phi_i \, \mathrm{d}\boldsymbol{x}}_{=:f_i^K} . \tag{3.21}$$

The vector $\mathbf{f}^K$ is called the *element load vector*. If K has local nodes $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \boldsymbol{\ell}_3$, then $\mathbf{f}^K$ can be represented by the 3-dimensional vector

$$\tilde{f}_p^K = \int_K f(\boldsymbol{x})\phi_p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \qquad p = 1, 2, 3. \tag{3.22}$$

Computing this often requires quadrature. Here is a simple quadrature rule that is sufficient for our purposes.

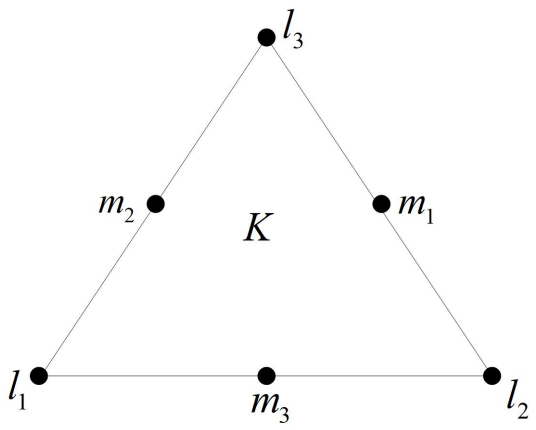

FIGURE 3.8. Quadrature points in the element K

Lemma 2. *Let K be any triangle with nodes $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \boldsymbol{\ell}_3$, and denote the mid-points of the edges of K by $\boldsymbol{m}_1, \boldsymbol{m}_2$, and $\boldsymbol{m}_3$ (see Figure 3.8). Then the quadrature rule*

$$\int_K g \, \mathrm{d}\boldsymbol{x} \approx \frac{|K|}{3} \sum_{s=1}^{3} g(\boldsymbol{m}_s) , \tag{3.23}$$

is exact for quadratic functions g on K.

Using this lemma we can approximate $\tilde{f}_p$ to get (since $\phi^p(\boldsymbol{x}^q) = \delta_{pq}$),

$$\tilde{f}_p^K \approx \frac{|K|}{3} \sum_{s=1}^{3} f(\boldsymbol{m}_s)\phi_p(\boldsymbol{m}_s) = \frac{|K|}{3} \frac{f(\boldsymbol{m}_q) + f(\boldsymbol{m}_r)}{2} . \tag{3.24}$$

So

$$\tilde{\mathbf{f}}^K \approx \frac{|K|}{6} \begin{bmatrix} f(\boldsymbol{m}_2) + f(\boldsymbol{m}_3) \\ f(\boldsymbol{m}_1) + f(\boldsymbol{m}_3) \\ f(\boldsymbol{m}_1) + f(\boldsymbol{m}_2) \end{bmatrix} ,$$

and, by Lemma 2, equality holds in (3.24) if $f(\boldsymbol{x})$ is a piecewise linear function. Once these local right-hand sides are computed, then the global right-hand side is assembled following the formula (3.22).

Worked example

Example 19. *Return to Example 17 of solving $-\Delta u = f$ on the mesh given in Figure 3.7 and assume $f(\boldsymbol{x}) = 1$. Compute the right-hand side of the system, and thus find the approximate solution to the Poisson problem.*

Exercise

Exercise 12. *Using properties of the cross product of vectors, prove the following expression for the area of a triangle K with corners (oriented counterclockwise) $(x_1, y_1), (x_2, y_2), (x_3, y_3)$:*

$$|K| = \frac{1}{2} det \begin{pmatrix} 1 & 1 & 1 \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{pmatrix} . \tag{3.25}$$

This formula stresses the importance of the counterclockwise *orientation of the local numbering of the nodes in the triangle K. In fact, if this convention is not respected, then a negative area will be obtained, which is unrealistic.*

Exercise 13. *Let K be the triangle having the following nodes: $(0,0), (0.5,0), (0.5,1)$ (you may assume that is the local ordering of the nodes). Compute the local matrix $\tilde{\mathbb{A}}^K = \left(\int_K \nabla\phi_q^K \cdot \nabla\phi_p^K \, \mathrm{d}\boldsymbol{x}\right)_{p,q=1}^3$.*

Exercise 14. *Consider the mesh from Example 18. Consider also the partial differential equation*

$$-\operatorname{div}(\mathcal{D}\nabla u) = f \qquad \text{in } \Omega\,, \tag{3.26}$$

and $u = 0$ on Γ, where the diffusion matrix is given by

$$\mathcal{D} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} .$$

Compute the element a_{12} of the system matrix.

Exercise 15. *Consider the mesh from Example 18. Consider also the partial differential equation*

$$-\Delta u + u = f \qquad \text{in } \Omega\,, \tag{3.27}$$

and $u = 0$ on Γ. Compute the element a_{21} of the system matrix. To compute the extra term arising from the reaction term $\int_\Omega uv \, \mathrm{d}\boldsymbol{x}$, the quadrature formula (3.23) will be useful.

3.5. **Application of the abstract error analysis in two space dimensions.** In this section we will detail the proof of the existence, uniqueness, and error estimates for the two-dimensional problem:

$$\begin{aligned} -\operatorname{div}(\mathcal{D}\nabla u) + bu &= f \qquad \text{in } \Omega\,, \\ u &= 0 \qquad \text{on } \Gamma\,. \end{aligned} \tag{3.28}$$

Here, $\Omega \subseteq \mathbb{R}^2$ is a bounded open set with boundary Γ. As before, we will suppose that the matrix $\mathcal{D}$ is symmetric and uniformly positive-definite, and that $b \geq 0$. By this, we mean that there exists a constant $\lambda_{\min} > 0$ such that

$$(s\ t)\mathcal{D}(\boldsymbol{x}) \begin{pmatrix} s \\ t \end{pmatrix} \geq \lambda_{\min} |(s\,,t)|_2^2 \qquad \forall\,(s,t) \in \mathbb{R}^2 \text{ and all } \boldsymbol{x} \in \Omega\,, \tag{3.29}$$

where $|\mathbf{c}|_2 = |(c_1, c_2)|_2 = (c_1^2 + c_2^2)^{\frac{1}{2}}$ is the Euclidean norm on $\mathbb{R}^2$.

Remark 10. *The notation $\lambda_{\min}$ for the constant refers to the fact that it is, in fact, the minimum eigenvalue of $\mathcal{D}$, and that the main assumption is that this value is positive.*

We recall the weak formulation of (3.28): Find $u \in H_0^1(\Omega)$ such that

$$\text{(W)} \qquad a(u,v) = L(v) \qquad \forall\, v \in H_0^1(\Omega),$$

where

$$a(u,v) := \int_\Omega \{(\mathcal{D}\nabla u)\cdot\nabla v + buv\}\ \mathrm{d}\boldsymbol{x} \quad \text{and} \quad L(v) := \int_\Omega fv\ \mathrm{d}\boldsymbol{x}. \tag{3.30}$$

Theorem 7. *Let $f \in L^2(\Omega)$. Then, the problem* (W) *has a unique solution $u \in H_0^1(\Omega)$.*

The proof of this result will be presented later. We present first two results that will be very useful.

Lemma 3 (Poincaré inequality)**.** *Let us suppose that Ω is bounded. Then, there exists $C_\Omega > 0$, depending only on the geometry of Ω, such that*

$$\|v\|_{0,\Omega} \le C_\Omega \left\{\|\partial_x v\|_{0,\Omega}^2 + \|\partial_y v\|_{0,\Omega}^2\right\}^{\frac{1}{2}} \qquad \forall\, v \in H_0^1(\Omega)\,. \tag{3.31}$$

Proof. Let us first assume that $\Omega = (0,d)\times(0,c)$ where c and d are positive numbers. Then, for all $\boldsymbol{x} = (x,y) \in \Omega$, by the fundamental theorem of calculus we have

$$\begin{aligned}
|v(\boldsymbol{x})|^2 &= \left|\underbrace{v(0,y)}_{=0} + \int_0^x \frac{\partial v}{\partial x}(t,y)\ \mathrm{d}t\right|^2 \\
&\le \left(\int_0^x \left|\frac{\partial v}{\partial x}(t,y)\right|\ \mathrm{d}t\right)^2 \\
&\le \left(\int_0^x 1\ \mathrm{d}t\right)\left(\int_0^x \left|\frac{\partial v}{\partial x}(t,y)\right|^2\ \mathrm{d}t\right) \quad \text{by Hölder in } L^2(0,x) \\
&\le d\int_0^d \left|\frac{\partial v}{\partial x}(t,y)\right|^2\ \mathrm{d}t. \qquad (3.32)
\end{aligned}$$

Integrating (3.32) over Ω, we get

$$\begin{aligned}
\int_\Omega |v(\boldsymbol{x})|^2\ d\boldsymbol{x} &\le \int_0^c\int_0^d d\int_0^d \left|\frac{\partial v}{\partial x}(t,y)\right|^2\ \mathrm{d}t\ \mathrm{d}x\ \mathrm{d}y \\
&= d^2\int_\Omega \left|\frac{\partial v}{\partial x}(t,y)\right|^2\ \mathrm{d}t\,\mathrm{d}y \\
&\le d^2\int_\Omega \left(|\partial_x v|^2 + |\partial_y v|^2\right)\ \mathrm{d}\boldsymbol{x}\,,
\end{aligned}$$

which proves the result in the case that Ω is a rectangle.

The general case can be reduced to the simple one in a rectangle: Since Ω is bounded, possibly up to a translation, there exists $d, c \in \mathbb{R}$, such that

$$\Omega \subset [0,d]\times[0,c] =: \hat{\Omega}.$$

Any function $v \in H_0^1(\Omega)$ can be extended by zero to a function $\hat{v} \in H_0^1(\hat{\Omega})$ as follows

$$\hat{v}(\boldsymbol{x}) := \begin{cases} v(\boldsymbol{x}) & \text{for } \boldsymbol{x} \in \Omega, \\ 0 & \text{for } \boldsymbol{x} \in \hat{\Omega}\setminus\Omega. \end{cases}$$

Then, we see that $\|v\|_{0,\Omega} = \|\hat{v}\|_{0,\hat{\Omega}}$ and $|v|_{1,\Omega} = |\hat{v}|_{1,\hat{\Omega}}$, and using the Poincaré inequality in $\hat{\Omega}$ we get the result. □

One of the main consequences of Poincaré's inequality is that in $H_0^1(\Omega)$ there is more than one possible definition for norm, as the following result shows.

Corollary 2. *[Equivalence of norms in $H^1_0(\Omega)$] Let $|\cdot|_{1,\Omega}$ be the $H^1(\Omega)$-seminorm defined by*

$$|v|_{1,\Omega} := \left\{ \sum_{i=1}^{2} \|\partial_{x_i} v\|^2_{0,\Omega} \right\}^{\frac{1}{2}} = \|\nabla v\|_{0,\Omega} \,. \tag{3.33}$$

Then, the following holds

$$\frac{1}{\sqrt{C_\Omega^2 + 1}} \|v\|_{1,\Omega} \leq |v|_{1,\Omega} \leq \|v\|_{1,\Omega} \qquad \forall\, v \in H^1_0(\Omega) \,. \tag{3.34}$$

As a consequence, this seminorm is a norm in $H^1_0(\Omega)$, equivalent to the standard $H^1(\Omega)$ norm.

Remark 11.

- *A simple inspection of the definition of the seminorm shows that for the function $v(\boldsymbol{x}) = 1$ the seminorm vanishes. That explains why we use the term* seminorm *instead of norm. Looking at this more carefully, we realise that, since* 1 *does not vanish at the boundary, then $v(\boldsymbol{x}) = 1$ does not belong to $H^1_0(\Omega)$. This is one further explanation on why $|\cdot|_{1,\Omega}$ is a norm on $H^1_0(\Omega)$.*
- *Following from the last point, we can consider the space $H^1_0(\Omega)$ provided with both the norm $\|\cdot\|_{1,\Omega}$, and the seminorm $|\cdot|_{1,\Omega}$. For simplicity, we will always consider (and always have considered) $H^1(\Omega)$ (and* **any** *subspace of it, such as $H^1_0(\Omega)$, or $H^1_{\Gamma_D}(\Omega)$ defined later), with the full norm $\|\cdot\|_{1,\Omega}$.*

The *maximum norm* of a continuous function g in Ω, defined as

$$\|g\|_\infty = \max\{|g(\boldsymbol{x})| : \boldsymbol{x} \in \Omega\} \,, \tag{3.35}$$

will be used extensively in what follows.

Proof of Theorem 7 : The proof will be carried out for the simpler case $\mathcal{D} = I$. The more general case (and further remarks) will be explained in the lectures. We first prove that $a(\cdot,\cdot)$ is continuous. Let us first recall that the discrete Hölder's inequality (2.23) in $\mathbb{R}^2$ implies that for any $\mathbf{a}, \mathbf{b} \in \mathbb{R}^2$,

$$\mathbf{a} \cdot \mathbf{b} \leq |\mathbf{a}|_2 |\mathbf{b}|_2 \,. \tag{3.36}$$

Then

$$\begin{aligned}
|a(v,w)| &\leq \left| \int_\Omega (\nabla v \cdot \nabla w + bvw) \, \mathrm{d}\boldsymbol{x} \right| \\
&\leq \int_\Omega \big(|\nabla v \cdot \nabla w| + |bvw| \big) \, \mathrm{d}\boldsymbol{x} \\
&\leq \max\{1, \|b\|_\infty\} \int_\Omega \left(|\nabla v|_2 \, |\nabla w|_2 + |v||w| \right) \, \mathrm{d}\boldsymbol{x} \\
&\leq \max\{1, \|b\|_\infty\} \int_\Omega \left(|\nabla v|_2^2 + |v|_2^2 \right)^{\frac{1}{2}} \left(|\nabla w|^2 + |w|_2^2 \right)^{\frac{1}{2}} \, \mathrm{d}\boldsymbol{x} \qquad \text{using} \quad (3.36) \\
&\leq \max\{1, \|b\|_\infty\} \left(\int_\Omega |\nabla v|_2^2 + |v|^2 \, \mathrm{d}\boldsymbol{x} \right)^{\frac{1}{2}} \left(\int_\Omega |\nabla w|_2^2 + |w|^2 \, \mathrm{d}\boldsymbol{x} \right)^{\frac{1}{2}} \quad \text{using Hölder's inequality} \\
&= \max\{1, \|b\|_\infty\} \, \|v\|_{1,\Omega} \, \|w\|_{1,\Omega} \,,
\end{aligned}$$

which shows that $a(\cdot,\cdot)$ is bounded with $\gamma = \max\{1, \|b\|_\infty\}$.

To show that $a(\cdot,\cdot)$ is elliptic we use the Poincaré inequality. More precisely, let $v \in H^1_0(\Omega)$. Since $b(\boldsymbol{x}) \geq 0$, from Corollary 2 we have

$$\begin{aligned}
a(v,v) &= \int_\Omega \left(|\nabla v(\boldsymbol{x})|_2^2 + b|v(\boldsymbol{x})|^2 \right) \mathrm{d}\boldsymbol{x} \\
&\geq \int_\Omega |\nabla v(\boldsymbol{x})|_2^2 \, \mathrm{d}\boldsymbol{x} \\
&= |v|^2_{1,\Omega} \\
&\geq \frac{1}{1 + C_\Omega^2} \|v\|^2_{1,\Omega} \,.
\end{aligned}$$

Thus, $a(\cdot,\cdot)$ is elliptic with $\alpha = \frac{1}{1+C_\Omega^2}$.

Finally, to show that $L(\cdot)$ is continuous, we use Hölder's inequality in $L^2(\Omega)$ to get to

$$|L(v)| \leq \|f\|_{0,\Omega}\,\|v\|_{0,\Omega} \leq \|f\|_{0,\Omega}\,\|v\|_{1,\Omega}\,,$$

so (A3) is satisfied with $\lambda := \|f\|_{0,\Omega}$.

Finally, since we have shown the hypotheses of Lax-Milgram's Lemma, then we conclude that the problem (W) has a unique solution $u \in H_0^1(\Omega)$. □

Worked example

Example 20. *Let $\mathcal{D} = (d_{ij})_{i,j=1}^2$ be a symmetric matrix function. Suppose that $\mathcal{D}$ satisfies (3.29), and we will suppose that $\|d_{ij}\|_\infty \leq M$ for $i,j = 1,2$. We consider the Neumann problem: Find u such that*

$$\begin{aligned} -\operatorname{div}\big(\mathcal{D}\nabla u\big) + u &= f \qquad \text{in } \Omega\,, \\ \big(\mathcal{D}\nabla u\big)\cdot \boldsymbol{n} &= 0 \qquad \text{on } \Gamma\,, \end{aligned} \tag{3.37}$$

where $\boldsymbol{n}$ is the unit outward normal vector to Γ. Find a weak formulation for (3.37), and prove it has a unique solution.

3.6. **Interpolation error, and final error estimate.** To estimate the error, we need to introduce some notations and some assumptions in the triangulation. More precisely, for a triangle K from now on we will denote by h_K the diameter of K (that is, the length of its longest edge). For a triangulation $\mathscr{T}$ made up of triangles we denote $h := \max\{h_K : K \in \mathscr{T}\}$, and consequently we will denote $\mathscr{T}$ by $\mathscr{T}_h$ from now on. In addition, we will assume that in the finite element method we use a *family of conforming and shape-regular triangulations* $\{\mathscr{T}_h\}_{h>0}$. That is, the family $\{\mathscr{T}_h\}_{h>0}$ satisfies

(1) $h \to 0$;
(2) every triangulation $\mathscr{T}_h$ is admissible;
(3) let, for every triangle K, ρ_K be the dyameter of the circle inscribed in K; then, there exists a constant $\sigma > 0$ such that, for every $h > 0$, and every $K \in \mathscr{T}_h$ we have

$$h_K \leq \sigma\,\rho_K\,. \tag{3.38}$$

Remark 12. *The condition (3.38) is the really distinguishing feature of a shape-regular triangulation. It has a very strong implication in geometrical terms. In fact, if the family of meshes $\{\mathscr{T}_h\}_{h>0}$ is shape-regular, then there exists a constant $\theta_0 > 0$ such that*

$$\theta \geq \theta_0 \quad \text{for every internal angle } \theta \text{ of every element } K \in \mathscr{T}_h\,, \tag{3.39}$$

and for all $h > 0$. This condition is called the minimal angle condition*, and is fundamental in the proof of the interpolation estimates presented below.*

Over the family of triangulations, we define the finite element spaces

$$V_h = \{v_h \in C^0(\bar\Omega) : v_h|_K \in \mathbb{P}_1\ \forall\, K \in \mathscr{T}_h\,,\ v_h|_\Gamma = 0\}\,. \tag{3.40}$$

As before, we denote the standard basis functions for this space $\{\phi_i : \boldsymbol{x}_i \in \mathcal{N}_0\}$, and they will be used in the definition of the Lagrange interpolation operator onto V_h. In fact, for any continuous function v, we define the *Lagrange interpolation of v* on the mesh $\mathscr{T}_h$ by

$$\Pi_h v(\boldsymbol{x}) := \sum_{\boldsymbol{x}_j \in \mathcal{N}_0} v(\boldsymbol{x}_j)\phi_j(\boldsymbol{x})\,,$$

that is, the only function on V_h that coincides with v at the nodes $\boldsymbol{x}_j$ of the mesh $\mathscr{T}_h$. That is, $\Pi_h v$ is linear on each element $K \in \mathscr{T}_h$ and $\Pi_h v(\boldsymbol{x}_j) = v(\boldsymbol{x}_j)$ for all $\boldsymbol{x}_j \in \mathcal{N}_0$. In this section, we will estimate the interpolation error $v - \Pi_h v$ in the L^2 and H^1 norms.

One final component for our analysis is the definition of regular functions H^2. For a domain $D \subseteq \mathbb{R}^2$ we define the space

$$H^2(D) = \{v \in H^1(D) : \partial_{xx} v, \partial_{xy} v, \partial_{yy} v \in L^2(D)\}\,, \tag{3.41}$$

that is, the space of functions with all derivatives up to second order that belong to $L^2(D)$. This space is provided with the norm

$$\|v\|_{2,D} := \left\{\|v\|_{1,D}^2 + \|\partial_{xx} v\|_{0,D}^2 + 2\|\partial_{xy} v\|_{0,D}^2 + \|\partial_{yy} v\|_{0,D}^2\right\}^{\frac{1}{2}}\,. \tag{3.42}$$

With the above definitions, notations and assumptions, the following interpolation result can be proven. Its proof is, unfortunately, outwith the scope of this class, but very nice proofs can be found in the books by Ern & Guermond, and by Brenner & Scott.

Theorem 8 (Interpolation estimate)**.** *Suppose $v \in H^2(\Omega)$ and $\{\mathscr{T}_h\}_{h>0}$ is a shape-regular family of meshes. Then, there exist constants $C_1, C_2 > 0$, independent of h, such that*

(i) $\|v - \Pi_h v\|_{0,K} \leq C_1 h_K^2 \, \|v\|_{2,K}$ *and*
(ii) $\|v - \Pi_h v\|_{1,K} \leq C_2 h_K \, \|v\|_{2,K}$,

for every $K \in \mathscr{T}_h$, and all $h > 0$, where C_1, C_2 are two constants independent of h and v.

The final result in this part of the class is the application of the interpolation estimate in combination with Cea's Lemma in order to give an error estimate for the model problem (3.28). We will just detail the proof in the simple case $\mathcal{D} = I$ and $b = 0$, but the argument can be extended naturally to more complicated situations.

Theorem 9 (Poisson's equation with homogeneous Dirichlet conditions)**.** *Consider the problem*

$$\begin{aligned} -\Delta u &= f \quad \text{on } \Omega, \\ u &= 0 \quad \text{on } \Gamma\,, \end{aligned} \tag{3.43}$$

and let us assume that its solution belongs to the space $H^2(\Omega)$. Then, there exists $C > 0$, independent of h, such that

$$\|u - u_h\|_{1,\Omega} \leq Ch \, \|u\|_{2,\Omega}\,, \tag{3.44}$$

for all $h > 0$.

Proof. The weak form of (3.43) is given by: Find $u \in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}\boldsymbol{x} = \int_\Omega f v \, \mathrm{d}\boldsymbol{x} \qquad \forall v \in H_0^1(\Omega).$$

It follows from Theorem 7 that the hypotheses of Lax-Milgram's Lemma hold with $V = H_0^1(\Omega)$, $\gamma = 1$, and $\alpha = \frac{1}{1+C_\Omega^2}$ where C_Ω is the Poincaré constant. Next, we approximate this problem in the space V_h. Then by Céa's Lemma and the Interpolation Estimate we get the following estimate

$$\begin{aligned} \|u - u_h\|_{1,\Omega} &\leq \frac{1}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|_{1,\Omega} \\ &\leq \frac{1}{\alpha} \, \|u - \Pi_h u\|_{1,\Omega} \\ &\leq C\, h \, \|u\|_{2,\Omega}\,, \end{aligned} \tag{3.45}$$

which finishes the proof. □

Exercise

Exercise 16. *Let us suppose that* $\Gamma = \Gamma_N \cup \Gamma_D$, *where* $\Gamma_N \cap \Gamma_D = \emptyset$. *Let us consider the following problem with* mixed boundary conditions*:*

$$\begin{aligned} -\Delta u &= f && \text{in } \Omega\,, \\ u &= 0 && \text{on } \Gamma_D\,, \\ \partial_{\mathbf{n}} u &= 0 && \text{on } \Gamma_D\,. \end{aligned}$$

Give a weak formulation of the above problem using the space

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

and prove that the weak problem has a unique solution.
Hint : The Poincaré inequality from Lemma 3 also holds in the space $H^1_{\Gamma_D}(\Omega)$**.**

Exercise 17. *For the boundary value problem*

$$\begin{aligned} -\operatorname{div}\left(\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \nabla u\right) &= f && \text{in } \Omega\,, \\ u &= 0 && \text{on } \Gamma\,, \end{aligned}$$

find a weak formulation, and prove it has a unique solution.

Exercise 18. *Prove the minimum angle condition.*

Email address: gabriel.barrenechea@strath.ac.uk