A gentle introduction to the Finite Element Method

Francisco-Javier Sayas

April 27, 2015

An introduction

If you haven't been hiding under a stone during your studies of Engineering, Mathematics or Physics, it is very likely that you have already heard about the Finite Element Method. Maybe you even know some theoretical and practical aspects and have played a bit with FEM software. What you are going to find here is a detailed and mathematically biased introduction to several aspects of the Finite Element Method. This is not however a course on the analysis of the method. It is just a demonstration of how it works, written as applied mathematicians usually write it. There are going to be Mathematics involved, but not lists of theorems and proofs. We are also going from the most particular cases towards useful generalizations, from example to theory.

An aspect where this course differs from most of the many introductory books on finite elements is the fact that I am going to begin directly with the two-dimensional case. I have just sketched the one dimensional case in an appendix. Many people think that the one-dimensional case is a better way of introducing the method, but I have an inner feeling that the method losses richness in that very simple situation, so I prefer going directly to the plane.

The course is divided into five lessons and is thought to be read in that order. We cover the following subjects (but not in this order):

- triangular finite elements,
- finite elements on parallelograms and quadrilaterals,,
- adaptation to curved boundaries (isoparametric finite elements),
- three dimensional finite elements,
- assembly of the finite element method,
- some special techniques such as static condensation or mass lumping,
- eigenvalues of the associated matrices,
- approximation of evolution problems (heat and wave equations).

It is going to be around one hundred pages with many figures. Ideas will be repeated over and over, so that you can read this with ease. These notes have evolved during the decade I have been teaching finite elements to mixed audiences of mathematicians, physicists and engineers. The tone is definitely colloquial. I could just claim that these are my classnotes and that's what I'm like[1]. There's much more than that. First, I believe in doing your best at being entertaining when teaching. At least that's what I try. Behind that there is a deeper philosophical point: take your work (and your life) seriously but, please, don't take yourself too seriously.

I also believe that people should be duly introduced when they meet. All this naming old time mathematicians and scientists only by their last names looks to me too much like the Army. Or worse, high school![2] I think you have already been properly introduced to the great Leonhard Euler, David Hilbert, Carl Friedrich Gauss, Pierre Simon Laplace and George Green[3]. If you haven't so far, consider it done here. This is not about history. It's just good manners. Do you see what I mean by being colloquial?

Anyway, this is not about having fun[4], but since we are at it, let us try to have a good time **while learning**. If you take your time to read these notes with care and try the exercises at the end of each lesson, I can assure that you will have made a significant step in your scientific persona. Enjoy!

These notes were written in its present form during my first year as visiting faculty at the University of Minnesota. They constitute an evolved form of my lecture notes to teach Finite Elements at the graduate level, something I have done for many years in the University of Zaragoza (Spain). The version you are reading now is a revision produced for teaching at the University of Delaware.

[1] To the very common comment *every person has their ways*, the best answer I've heard is *Oh, God, no! We have good manners for that.*

[2] When I was in high school, boys were called by their last names. I was Sayas all over. On the other hand, girls were called by their first names.

[3] You will find here the names of Peter Lejeune Dirichlet, Carl Neumann or Sergei Sobolev, associated to different concepts of PDE theory

[4] Unfortunately too many professional mathematicians advocate fun or beauty as their main motivations to do their job. It is so much better to have a scientific calling than this aristocratic detachment from work...

Lesson 1

Linear triangular elements

1 The model problem

All along this course we will be working with a simple model boundary value problem, which will allow us to put the emphasis on the numerical method rather than on the intricacies of the problem itself. For some of the exercises and in forthcoming lessons we will complicate things a little bit.

In this initial section there is going to be a lot of new stuff. Take your time to read it carefully, because we will be using this material during the entire course.

1.1 The physical domain

The first thing we have to describe is the geometry (the physical setting of the problem). You have a sketch of it in Figure 1.1.

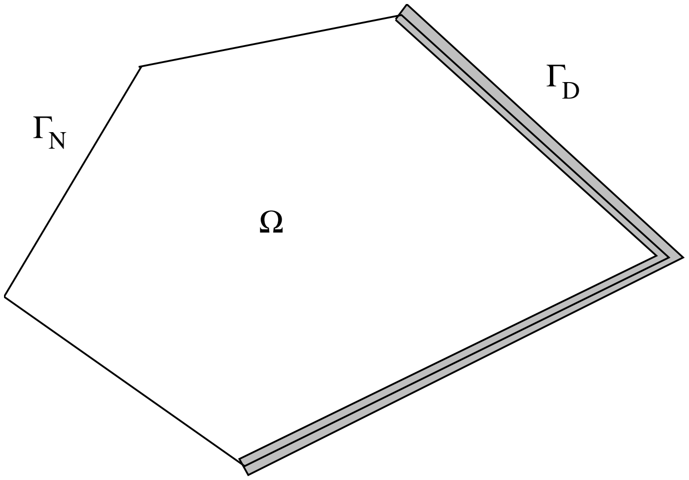


Figure 1.1: The domain Ω and the Dirichlet and Neumann boundaries.

We are thus given a polygon in the plane $\mathbb{R}^2$. We call this polygon Ω. Its boundary is a closed polygonal curve Γ. (There is not much difference if we suppose that there is one or more holes inside Ω, in which case the boundary is composed by more than one polygonal curve).

The boundary of the polygon, Γ is divided into two parts, that cover the whole of Γ and do not overlap:

- the Dirichlet boundary Γ_D,
- the Neumann boundary Γ_N.

You can think in more mechanical terms as follows: the Dirichlet boundary is where displacements are given as data; the Neumann boundary is where normal stresses are given as data.

Each of these two parts is composed by full sides of the polygon. This is not much of a restriction if you admit the angle of 180 degrees as separating two sides, that is, if you want to divide a side of the boundary into parts belonging to Γ_D and Γ_N, you just have to consider that the side is composed of several smaller sides with a connecting angle of 180 degrees.

1.2 The problem, written in strong form

In the domain we will have an elliptic partial differential equation of second order and on the boundary we will impose conditions on the solution: boundary conditions or boundary values. Just to unify notations (you may be used to different ways of writing this), we will always write the Laplace operator, or Laplacian, as follows

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}.$$

By the way, sometimes it will be more convenient to call the space variables (x_1, x_2) rather than (x, y), so expect mixed notations.

The boundary value problem is then

$$\left[\begin{array}{l} -\Delta u + c\,u = f \qquad \text{in } \Omega, \\ u = g_0 \qquad \text{on } \Gamma_D, \\ \partial_n u = g_1 \qquad \text{on } \Gamma_N. \end{array}\right.$$

There are new many things here, so let's go step by step:

- The unknown is a (scalar valued) function u defined on the domain Ω.
- c is a non-negative constant value. In principle we will consider two values $c = 1$ and $c = 0$. The constant c is put there to make clear two different terms when we go on to see the numerical approximation of the problem. By the way, this equation is usually called a **reaction-diffusion equation**. The diffusion term is given by $-\Delta u$ and the reaction term, when $c > 0$, is $c\,u$.
- f is a given function on Ω. It corresponds to source terms in the equation. It can be considered as a surface density of forces.

- There are two functions g_0 and g_1 given on the two different parts of the boundary. They will play very different roles in our formulation. As a general rule, we will demand that g_0 is a continuous function, whereas g_1 will be allowed to be discontinuous.
- The symbol ∂_n denotes the exterior normal derivative, that is,

  $$\partial_n u = \nabla u \cdot \mathbf{n},$$

  where $\mathbf{n}$ is the unit normal vector on points of Γ pointing always outwards and ∇u is, obviously, the gradient of u.

We are not going to worry about regularity issues. If you see a derivative, admit that it exists and go on. We will reach a point where everything is correctly formulated. And that moment we will make hypotheses more precise. If you are a Mathematician and are already getting nervous, calm down and believe that I know what I'm talking about. Being extra rigorous is not what is important at this precise time and place.

1.3 Green's Theorem

The approach to solve this problem above with the Finite Element Method is based upon writing it in a completely different form, which is sometimes called **weak or variational form**. At the beginning it can look confusing to see all this if you are not used to advanced Mathematics in Continuum Mechanics or Physics. We are just going to show here how the formulation is obtained and what it looks like at the end. You might be already bored in search of matrices and something more tangible! Don't rush! If you get familiarized with formulations and with the notations Mathematicians given to frame the finite element method, many doors will be open to you: you will be able to read a large body of literature that would be ununderstandable to you if you stick to what you already know.

The most important theorem in this process or reformulating the problem is Green's Theorem, one of the most popular results of Vector Calculus. Sometimes it is also called Green's First Formula (there's a popular second one and a less known third one). The theorem states that

$$\int_\Omega (\Delta u)\, v + \int_\Omega \nabla u \cdot \nabla v = \int_\Gamma (\partial_n u)\, v.$$

Note that there are two types of integrals in this formula. Both integrals in the left-hand side are domain integrals in Ω, whereas the integral in the right-hand side is a line integral on the boundary Γ. By the way, the result is also true in three dimensions. In that case, domain integrals are volume integrals and boundary integrals are surface integrals. The dot between the gradients denotes simply the Euclidean product of vectors, so

$$\nabla u \cdot \nabla v = \frac{\partial u}{\partial x_1}\frac{\partial v}{\partial x_1} + \frac{\partial u}{\partial x_2}\frac{\partial v}{\partial x_2}$$

Remark. This theorem is in fact a simple consequence of the Divergence Theorem:

$$\int_\Omega (\operatorname{div}\mathbf{p})\,v + \int_\Omega \mathbf{p}\cdot\nabla v = \int_\Gamma (\mathbf{p}\cdot\mathbf{n})\,v.$$

Here $\operatorname{div}\mathbf{p}$ is the divergence of the vector field $\mathbf{p}$, that is, if $\mathbf{p} = (p_1, p_2)$

$$\operatorname{div}\mathbf{p} = \frac{\partial p_1}{\partial x_1} + \frac{\partial p_2}{\partial x_2}.$$

If you take $\mathbf{p} = \nabla u$ you obtain Green's Theorem. □

1.4 The problem, written in weak form

The departure point for the weak or variational formulation is Green's Theorem. Here it is again

$$\int_\Omega (\Delta u)\,v + \int_\Omega \nabla u\cdot\nabla v = \int_\Gamma (\partial_n u)\,v = \int_{\Gamma_D} (\partial_n u)\,v + \int_{\Gamma_N} (\partial_n u)\,v.$$

Note that we have broken the integral on Γ as the sum of the integrals over the two sub-boundaries, the Dirichlet and the Neumann boundary. You may be wondering what v is in this context. In fact, it is nothing but a test. Wait for comments on this as the section progresses.

Now we substitute what we know in this formula: we know that $\Delta u = c\,u - f$ in Ω and that $\partial_n u = g_1$ on Γ_N. Therefore, after some reordering

$$\int_\Omega \nabla u\cdot\nabla v + c\int_\Omega u\,v = \int_\Omega f\,v + \int_{\Gamma_N} g_1\,v + \int_{\Gamma_D} (\partial_n u)\,v.$$

Note now that I have written all occurrences of u on the left hand side of the equation except for one I have left on the right. In fact we don't know the value of $\partial_n u$ on that part of the boundary. So what we will do is impose that v cancels in that part, that is,

$$v = 0 \qquad \text{on } \Gamma_D.$$

Therefore

$$\int_\Omega \nabla u\cdot\nabla v + c\int_\Omega u\,v = \int_\Omega f\,v + \int_{\Gamma_N} g_1\,v, \qquad \text{if } v = 0 \text{ on } \Gamma_D.$$

Notice now three things:

- We have not imposed the Dirichlet boundary condition ($u = g_0$ on Γ_D) yet. Nevertheless, we have imposed a similar one to the function v, *but in a homogeneous way.*

- As written now, data (f and g_1) are in the right-hand side and coefficients of the equation (the only one we have is c) are in the left-hand side.

- The expression on the left-hand side is linear in both u and v. It is a bilinear form of the variables u and v. The expression on the right-hand side is linear in v.

Without specifying spaces where u and v are, the weak formulation can be written as follows:

$$\left[\begin{array}{l} \text{find } u \text{ such that} \\ u = g_0 \qquad \text{on } \Gamma_D, \\ \displaystyle\int_\Omega \nabla u \cdot \nabla v + c \int_\Omega u\, v = \int_\Omega f\, v + \int_{\Gamma_N} g_1\, v \qquad \text{for all } v \text{ such that } v = 0 \text{ on } \Gamma_D. \end{array}\right.$$

Note how the two boundary conditions appear in very different places of this formulation:

- The Dirichlet condition (given displacements) is imposed apart from the formulation and involves imposing it homogeneously to the testing function v. It is called an **essential boundary condition**.

- The Neumann condition (given normal stresses) appears inside the formulation. It is called a **natural boundary condition**.

Being essential or natural is not inherently tied to the boundary condition: it is related to the role of the boundary condition in the formulation. So when you hear (or say) essential boundary condition, you mean a boundary condition that is imposed apart from the formulation, whereas a natural boundary condition appears inside the formulation. *For this weak formulation of a second order elliptic equation we have*

Dirichlet=essential *Neumann=natural*

What is v? At this point, you might (you should) be wondering what v is in the formulation. In the jargon of weak formulations, v is called a test function. It tests the equation that is satisfied by u. The main idea is that instead of looking at the equation as something satisfied point-by-point in the domain Ω, you have an averaged version of the equation. Then v plays the role of a weight function, something you use to average the equation. In many contexts (books on mechanics, engineering or physics) v is called a virtual displacement (or virtual work, or virtual whatever is pertinent), emphasizing the fact that v is not the unknown of the system, but something that only exists virtually to write down the problem. The weak formulation is, in that context, a principle of virtual displacements (principle of virtual work, etc). □

1.5 Delimiting spaces

We have reached a point where we should be a little more specific on where we are looking for u and where v belongs. The first space we need is the space of square-integrable functions

$$L^2(\Omega) = \Big\{ f : \Omega \to \mathbb{R} \, : \, \int_\Omega |f|^2 < \infty \Big\}.$$

A fully precise definition of this space requires either the introduction of the Lebesgue integral or applying some limiting ideas. If you know what this is all about, good for you! If you don't, go on: for most functions you know you will always be able to check whether they belong to this space or not by computing or estimating the integral and seeing if it is finite or not.

The second space is one of the wide family of Sobolev spaces:

$$H^1(\Omega) = \Big\{ u \in L^2(\Omega) \,:\, \tfrac{\partial u}{\partial x_1}, \tfrac{\partial u}{\partial x_2} \in L^2(\Omega) \Big\}.$$

There is a norm related to this space

$$\|u\|_{1,\Omega} = \Big(\int_\Omega |\nabla u|^2 + \int_\Omega |u|^2 \Big)^{1/2} = \Big(\int_\Omega \Big| \frac{\partial u}{\partial x_1} \Big|^2 + \int_\Omega \Big| \frac{\partial u}{\partial x_2} \Big|^2 + \int_\Omega |u|^2 \Big)^{1/2}.$$

Sometimes this norm is called the energy norm and functions that have this norm finite (that is, functions in $H^1(\Omega)$) are called functions of finite energy. The concept of energy is however related to the particular problem, so it's better to get used to have the space and its norm clearly written down and think of belonging to this space as a type of admissibility condition.

A particular subset of this space will be of interest for us:

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

Note that $H^1_{\Gamma_D}(\Omega)$ is a subspace of $H^1(\Omega)$, that is, linear combinations of elements of $H^1_{\Gamma_D}(\Omega)$ belong to the same space.

The Mathematics behind. An even half-trained Mathematician should be wondering what do we mean by the partial derivatives in the definition of $H^1(\Omega)$, since one cannot think of taking the gradient of an arbitrary function of $L^2(\Omega)$, or at least to taking the gradient and finding something reasonable. What we mean by restriction to Γ_D in the definition of $H^1_{\Gamma_D}(\Omega)$ is not clear either, since elements or $L^2(\Omega)$ are not really functions, but classes of functions, where values of the function on particular points or even on lines are not relevant. To make this completely precise there are several ways:

- Define a weak derivative for elements of $L^2(\Omega)$ and what we understand by saying that that derivative is again in $L^2(\Omega)$. Then you move to give a meaning to that restriction of a function in $H^1(\Omega)$ to one part of its boundary.

- Go the whole nine yards and take time to browse a book on distribution theory and Sobolev spaces. It takes a while but you end up with a pretty good intuition of what this all is about.

- Take a shortcut. You first consider the space of functions

$$\mathcal{C}^1(\overline{\Omega}) = \Big\{ u \in \mathcal{C}(\overline{\Omega}) \,:\, \tfrac{\partial u}{\partial x_1}, \tfrac{\partial u}{\partial x_2} \in \mathcal{C}(\overline{\Omega}) \Big\},$$

> which is simple to define, and then you close it with the norm $\|\cdot\|_{1,\Omega}$. To do that you have to know what closing or completing a space is (it's something similar to what you do to define real numbers from rational numbers). Then you have to prove that restricting to Γ_D still makes sense after this completion procedure.

My recommendation at this point is to simply go on. If you are a Mathematician you can take later on some time with a good simple book on elliptic PDEs and will see that it is not that complicated. If you are a Physicist or an Engineer you will probably not need to understand all the details of this. There's going to be a very important result in the next section that you will have to remember and that's almost all. Nevertheless, if you keep on doing research related to finite elements, you should really know something more about this. In due time you will have to find any of the dozens of books on Partial Differential Equations for Scientists and Engineers, and read the details, which will however not be given in the excruciating detail of PDE books for Mathematicians. But this is only an opinion. □

1.6 The weak form again

With the spaces defined above we can finally write our problem in a proper and fully rigorous way:

$$\left[\begin{array}{l} \text{find } u \in H^1(\Omega) \text{ such that} \\ u = g_0 \quad \text{on } \Gamma_D, \\ \displaystyle\int_\Omega \nabla u \cdot \nabla v + c \int_\Omega u\, v = \int_\Omega f\, v + \int_{\Gamma_N} g_1\, v \qquad \forall v \in H^1_{\Gamma_D}(\Omega). \end{array}\right.$$

Let me recall that the condition on the general test function $v \in H^1_{\Gamma_D}(\Omega)$ is the same as

$$v \in H^1(\Omega) \quad \text{such that } v = 0 \text{ on } \Gamma_D,$$

that is, v is in the same space as the unknown u but satisfies a homogeneous version of the essential boundary condition.

The data are in the following spaces

$$f \in L^2(\Omega), \qquad g_1 \in L^2(\Gamma_N), \qquad g_0 \in H^{1/2}(\Gamma_D).$$

We have already spoken of the first of these spaces. The space $L^2(\Gamma_N)$ follows essentially the same idea, with line integrals on Γ_N instead of domain integrals on Ω. The last space looks more mysterious: it is simply the space of restrictions to Γ_D of functions of $H^1(\Omega)$, that is, $g_0 \in H^{1/2}(\Gamma_D)$ means that there exists at least a function $u_0 \in H^1(\Omega)$ such that $u_0 = g_0$ on Γ_D. In fact, all other functions satisfying this condition (in particular our solution u) belong to

$$u_0 + H^1_{\Gamma_D}(\Omega) = \{u_0 + v \,:\, v \in H^1_{\Gamma_D}(\Omega)\} = \{w \in H^1(\Omega) \,:\, w = g_0 \quad \text{on } \Gamma_D\}$$

(can you see why?). Unlike $H^1_{\Gamma_D}(\Omega)$, this set is not a subspace of $H^1(\Omega)$. The only exception is the trivial case, when $g_0 = 0$, since the set becomes $H^1_{\Gamma_D}(\Omega)$.

The fact that g_0 is in $H^{1/2}(\Gamma_D)$ simply means that we are not looking for the solution in the empty set. I cannot give you here a simple and convincing explanation on the name of this space. Sorry for that.

2 The space of continuous linear finite elements

It's taken a while, but we are there! *Numerics start here.* We are now going to discretize all the elements appearing in this problem: the physical domain, the function spaces and the variational/weak formulation.

We are going to do it step by step. At the end of this section you will have the simplest example of a space of finite element functions (or simply finite elements). Many Mathematicians call these elements Courant elements, because Richard Courant introduced them several decades ago with theoretical more than numerical intentions. In the jargon of the business we call them triangular Lagrange finite elements of order one, or simply linear finite elements, or for short (because using initials and short names helps speaking faster and looking more dynamic) $\mathbb{P}_1$ elements.

2.1 Linear functions on a triangle

First of all, let us think for a moment about linear functions. A linear function (or, more properly, affine function) of two variables is the same as a polynomial function of degree at most one

$$p(x_1, x_2) = a_0 + a_1\, x_1 + a_2\, x_2.$$

The set of these functions is denoted $\mathbb{P}_1$. Everybody knows that a linear function is uniquely determined by its values on three different non-aligned points, that is, on the vertices of a (non-degenerate) triangle.

Let us then take an arbitrary non-degenerate triangle, that we call K. You might prefer calling the triangle T, as many people do. However, later on (in Lesson 3) the triangle will stop being a triangle and will become something else, maybe a quadrilateral, and then the meaning of the initial T will be lost. We draw it as in Figure 1.2, marking its three vertices. With this we mean that *a function*

$$p \in \mathbb{P}_1 = \{a_0 + a_1\, x_1 + a_2\, x_2 \; : \; a_0, a_1, a_2 \in \mathbb{R}\}$$

is uniquely determined by its values on these points. Uniquely determined means two things: (a) there is only one function with given values on the vertices; (b) there is in fact one function, that is, the values on the vertices are arbitrary. We can take any values we want and will have an element of $\mathbb{P}_1$ with these values on the vertices. Graphically it is just hanging a flat (linear) function from three non-aligned points.

Thus, a function $p \in \mathbb{P}_1$ can be determined

- either from its three defining coefficients (a_0, a_1, a_2)
- or from its values on the three vertices of a triangle K.

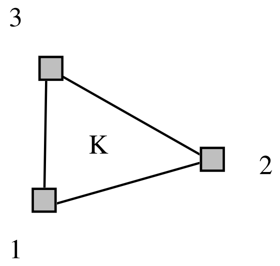


Figure 1.2: A triangle and its three vertices.

Both possibilities state that the space $\mathbb{P}_1$ is a vector space of dimension three. While the first choice (coefficients) gives us a simple expression of the function, the second is more useful for many tasks, in particular for drawing the function. The three values of the function on the vertices will be called the **local degrees of freedom**.

There is another important property that will be extremely useful in the sequel: *the value of $p \in \mathbb{P}_1$ on the edge that joins two vertices of the triangle depends only on the values of p on this two vertices.* In other words, the value of $p \in \mathbb{P}_1$ on an edge is uniquely determined by the degrees of freedom associated to the edge, namely, the values of p on the two vertices that lie on that edge.

2.2 Triangulations

So far we have functions on a single triangle. We now go for partitions of the domain into triangles. A triangulation of Ω is a subdivision of this domain into triangles. Triangles must cover all Ω but no more and must fulfill the following rule:

> *If two triangles have some intersection, it is either a common vertex or a common full edge. In particular, two different triangles do not overlap.*

Figure 1.3 shows two forbidden configurations. See Figure 1.5 to see how a triangulation looks like. There is another rule, related to the partition of Γ into Γ_D and Γ_N:

> *The triangulation must respect the partition of the boundary into Dirichlet and Neumann boundaries.*

This means that an edge of a triangle that lies on Γ cannot be part Dirichlet and part Neumann. Therefore if there is a transition from Dirichlet to Neumann boundaries, there must be a vertex of a triangle in that transition point. Note that this situation has to be taken into account only when there is a transition from Dirichlet to Neumann conditions inside a side of the polygon Ω.

The set of the triangles (that is, the list thereof) will be generally denoted $\mathcal{T}_h$. The subindex h makes reference to the diameter of the triangulation, defined as *the length of the longest edge of all triangles*, that is, the longest distance between vertices of the triangulation.

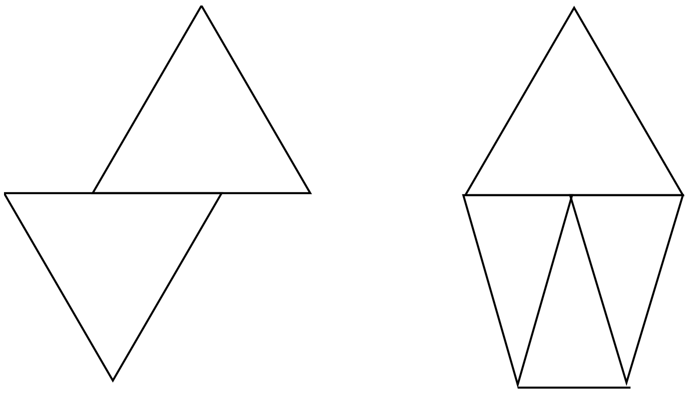


Figure 1.3: Situations not admitted in triangulations. In the second one we see the appearance of what is called a hanging node.

2.3 Piecewise linear functions on a triangulation

We now turn our attention to functions defined on the whole of the polygon Ω that has been triangulated as shown before.

Consider first two triangles sharing a common edge, say K and K' (see Figure 1.6). We take values at the four vertices of this figure and build a function that belongs to $\mathbb{P}_1$ on each of the triangles and has the required values on the vertices. Obviously we can define a unique function with this property. Moreover, since the value on the common edge depends only on the values on the two common vertices, the resulting function is continuous.

We can do this triangle by triangle. We end up with a function that is linear on each triangle and globally continuous. The space of such functions is

$$V_h = \big\{ u_h \in \mathcal{C}(\overline{\Omega}) \,:\, u_h|_K \in \mathbb{P}_1, \quad \forall K \in \mathcal{T}_h \big\}.$$

If we fix values on the set of vertices of the triangulation $\mathcal{T}_h$, there exists a unique $u_h \in V_h$ with those values on the vertices. Therefore an element of V_h is uniquely determined by its values on the set of vertices of the triangulation. The values on the vertices of the whole triangulation are the degrees of freedom that determine an element of V_h. In this context we will call **nodes** to the vertices in their role as points where we take values. (In forthcoming lessons there will be other nodes in addition to vertices.)

Elements of the space V_h are called linear finite element functions or simply $\mathbb{P}_1$ finite elements.

Let us take now a numbering of the set of nodes (that is, vertices) of the triangulation. At this moment any numbering goes[1]. In Figure 1.7 we have a numbering of the nodes of the triangulation of our model domain. The vertices will be generically denoted $\mathbf{p}_i$ with i varying from one to the number of vertices, say N.

[1] And in many instances this will be so to the end of the discretization process. Using one numbering or another has a great influence on the shape of the linear system we will obtain in Section 3, but this shape is relevant only for some choices of the method to solve the corresponding linear system.

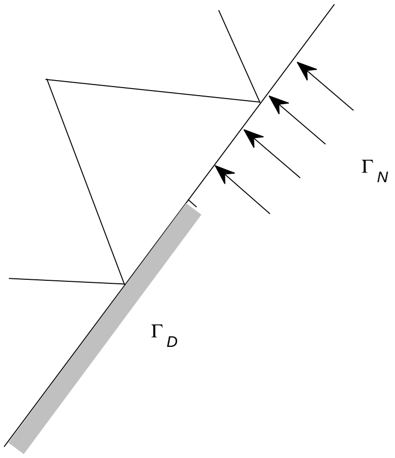


Figure 1.4: A forbidden transition of Dirichlet to Neumann boundary conditions happening inside an edge. Graphical notation for Dirichlet a Neumann boundaries as shown in many Mechanics books are given in the graph.

Because of what we have explained above, if we fix one node (vertex) and associate the value one to this node and zero to all others, there exists a unique function $\varphi_i \in V_h$ that has these values, that is,

$$\varphi_i(\mathbf{p}_j) = \delta_{ij} = \begin{cases} 1, & j = i, \\ 0, & j \neq i. \end{cases}$$

The aspect of one of these functions is shown in Figure 1.8.

Notice that if a triangle K has not $\mathbf{p}_i$ as one of its vertices, φ_i vanishes all over K, since the value of φ_i on the three vertices of K is zero. Therefore, the support of φ_i (the closure of the set of points where φ_i is not zero) is the same as the union of triangles that share $\mathbf{p}_i$ as vertex. In Figure 1.9 you can see the type of supports you can find.

There is even more. Take $u_h \in V_h$. It is simple to see that

$$u_h = \sum_{j=1}^{N} u_h(\mathbf{p}_j)\varphi_j.$$

Why? Let me explain. Take the function $\sum_{j=1}^{N} u_h(\mathbf{p}_j)\varphi_j$ and evaluate it in $\mathbf{p}_i$: you obtain

$$\sum_{j=1}^{N} u_h(\mathbf{p}_j)\varphi_j(\mathbf{p}_i) = \sum_{j=1}^{N} u_h(\mathbf{p}_j)\delta_{ji} = u_h(\mathbf{p}_i).$$

Therefore, this function has exactly the same nodal values as u_h and must be u_h. The fact that two functions of V_h with the same nodal values are the same function is the linear independence of the nodal functions $\{\varphi_i\}$. What we have proved is the fact that $\{\varphi_i \,:\, i = 1, \ldots, N\}$ is a basis of V_h and therefore

$$\dim V_h = N = \#\{\text{vertices}\}.$$

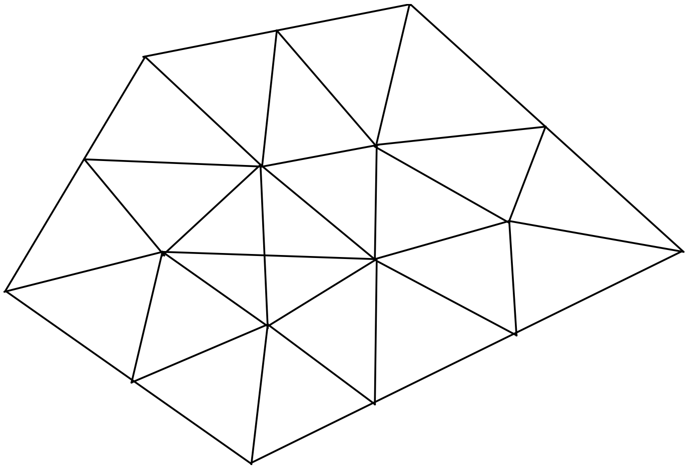


Figure 1.5: A triangulation of Ω.

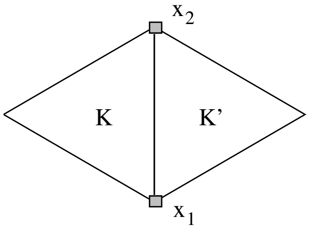


Figure 1.6: Two triangles with a common edge.

There is a particularly interesting aspect of this basis of V_h that makes it special. In general if you have a basis of V_h you know that you can decompose elements of V_h as a unique linear combination of the elements of the basis, that is,

$$u_h = \sum_{j=1}^{N} u_j \, \varphi_j$$

is a general element of V_h. With this basis, the coefficients are precisely the values of u_h on the nodes, that is, $u_j = u_h(\mathbf{p}_j)$. Hence, the coefficients of u_h in this basis are something more than coefficients: there are values of the function on points.

An important result. As you can see, when defining the space V_h we have just glued together $\mathbb{P}_1$ functions on triangles. Thanks to the way we have made the triangulation and to the way we chose the local degrees of freedom, what we obtained was a continuous function. One can think, is this so important? Could I take something discontinuous? At this level, the answer is a very loud and clear NO! The reason is the following result that allows us to know whether certain functions are in $H^1(\Omega)$ or not.

Theorem. *Let u_h be a function defined on a triangulation of Ω such that*

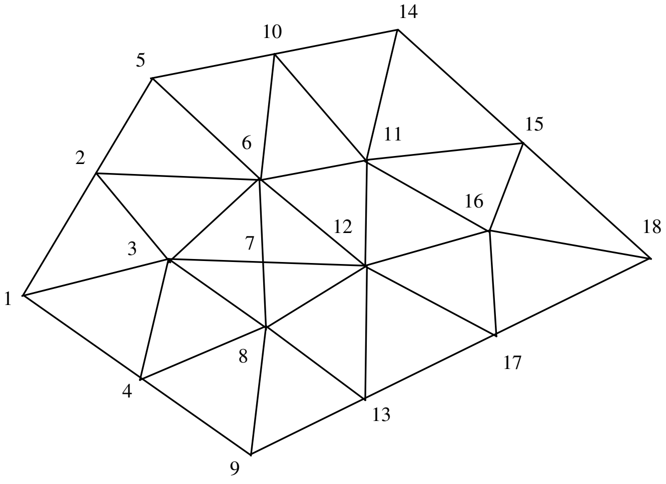


Figure 1.7: Global numbering of nodes.

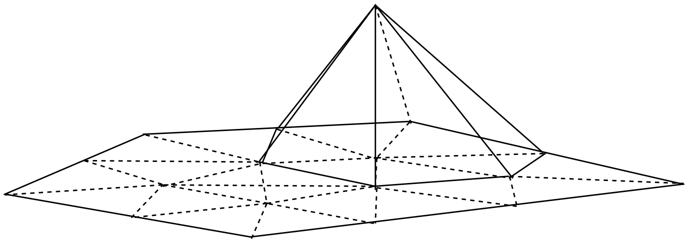


Figure 1.8: The graph of a nodal basis function: it looks like a camping tent.

restricted to each triangle it is a polynomial (or smooth) function. Then

$$u_h \in H^1(\Omega) \qquad \Longleftrightarrow \qquad u_h \textit{ is continuous.}$$

There is certain intuition to be had on why this result is true. If you take a derivative of a piecewise smooth function, you obtain Dirac distributions along the lines where there are discontinuities. Dirac distributions are not functions and it does not make sense to see if the are square-integrable or not. Therefore, if there are discontinuities, the function fails to have a square-integrable gradient. □

2.4 Dirichlet nodes

So far we have taken into account the discrete version of the domain Ω but not the partition of its boundary Γ into Dirichlet and Neumann sides. We first need some terminology. A Dirichlet edge is an edge of a triangle that lies on Γ_D. Similarly a **Neumann edge** is an edge of a triangle that is contained in Γ_N. The vertices of the Dirichlet edges are called **Dirichlet nodes**. The doubt may arise in transitions from the Dirichlet to the Neumann part of the boundary. If a node belongs to both Γ_N and Γ_D, it is a Dirichlet node.

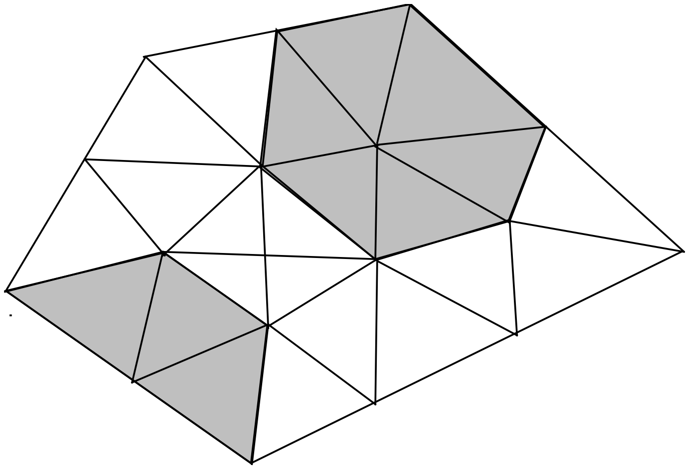


Figure 1.9: Supports of two nodal basis functions.

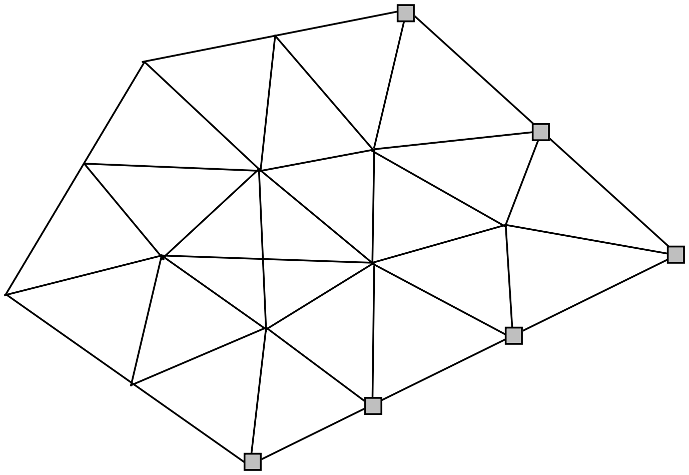


Figure 1.10: Dirichlet nodes corresponding to the domain as depicted in Figure 1.1

In truth, in parallel to what happens with how the Dirichlet and Neumann boundary conditions are treated in the weak formulation, we will inherit two different discrete entities:

- Dirichlet nodes, and
- Neumann edges.

Let us now recall the space

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

We might be interested in the space

$$V_h^{\Gamma_D} = V_h \cap H^1_{\Gamma_D}(\Omega) = \{v_h \in V_h \, : \, v_h = 0, \quad \text{on } \Gamma_D\}.$$

Recall now the demand we placed on the triangulation to respect the partition of Γ into Dirichlet and Neumann parts. Because of this, $v_h \in V_h$ vanishes on Γ_D if and only if it vanishes on the Dirichlet edges. Again, since values of piecewise linear functions on edges are determined by the values on the corresponding vertices, we have

> $v_h \in V_h$ *vanishes on* Γ_D *if and only if it vanishes on all Dirichlet nodes.*

The good news is the fact that we can easily construct a basis of $V_h^{\Gamma_D}$. We simply eliminate the elements of the nodal basis corresponding to Dirichlet nodes. To see that recall that when we write $v_h \in V_h$ as a linear combination of elements of the nodal basis, what we have is actually

$$v_h = \sum_{j=1}^{N} v_h(\mathbf{p}_j)\varphi_j.$$

Therefore $v_h = 0$ on Γ_D if and only if the coefficients corresponding to nodal functions of Dirichlet nodes vanish. To write this more efficiently we will employ two lists, Dir and Ind (as in *independent* or free nodes), to number separately Dirichlet and non-Dirichlet (independent/free) nodes. It is not necessary to number first one type of nodes and then the other, although sometimes it helps to visualize things to assume that we first numbered the free nodes and then the Dirichlet nodes.[2] With our model triangulation numbered as in Figure 1.7 and with the Dirichlet nodes marked in 1.10, the lists are

$$\begin{aligned}\text{Dir} &= \{9, 13, 14, 15, 17, 18\},\\ \text{Ind} &= \{1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 16\}.\end{aligned}$$

With these lists, an element of V_h can be written as

$$u_h = \sum_{j\in\text{Ind}} u_j\varphi_j + \sum_{j\in\text{Dir}} u_j\varphi_j, \qquad u_j = u_h(\mathbf{p}_j)$$

and an element of $V_h^{\Gamma_D}$ has the form

$$v_h = \sum_{j\in\text{Ind}} v_j\varphi_j.$$

Finally, this proves that

$$\text{dim}V_h^{\Gamma_D} = \#\text{Ind} = \#\{\text{nodes}\} - \#\{\text{Dirichlet nodes}\}.$$

[2] The reason for not doing this is merely practical. The triangulation is built without taking into account which parts of the boundary are Dirichlet and which are Neumann. As we will see in the next Lesson, the numbering of the nodes is inherent to the way the triangulation is given. In many practical problems we play with the boundary conditions for the same domain and it is not convenient to renumber the vertices each time.

3 The finite element method

3.1 The discrete variational problem

After almost fifteen pages of introducing concepts and formulas we can finally arrive at a numerical approximation of our initial problem. Recall that we wrote the problem in the following form

$$\left[\begin{array}{l} \text{find } u \in H^1(\Omega) \text{ such that} \\ u = g_0 \quad \text{on } \Gamma_D, \\ \displaystyle\int_\Omega \nabla u \cdot \nabla v + c \int_\Omega u\, v = \int_\Omega f\, v + \int_{\Gamma_N} g_1\, v \qquad \forall v \in H^1_{\Gamma_D}(\Omega). \end{array}\right.$$

The finite element method (with linear finite elements on triangles) consists of the following discrete version of the preceding weak formulation:

$$\left[\begin{array}{l} \text{find } u_h \in V_h \text{ such that} \\ u_h(\mathbf{p}) = g_0(\mathbf{p}) \quad \text{for every Dirichlet node } \mathbf{p}, \\ \displaystyle\int_\Omega \nabla u_h \cdot \nabla v_h + c \int_\Omega u_h\, v_h = \int_\Omega f\, v_h + \int_{\Gamma_N} g_1\, v_h \qquad \forall v_h \in V_h^{\Gamma_D}. \end{array}\right.$$

As you can easily see we have made three substitutions:

- We look for the unknown in the space V_h instead of on the whole Sobolev space. This means that we have reduced the problem to computing u_h in the vertices of the triangulation (in the nodes) and we are left with a finite number of unknowns.
- We have substituted the Dirichlet condition by fixing the values of the unknowns on Dirichlet nodes. This fact reduces the number of unknowns of the system to the number of free nodes.[3]
- Finally, we have reduced the testing space from $H^1_{\Gamma_D}(\Omega)$ to its discrete subspace $V_h^{\Gamma_D}$. We will show right now that this reduces the infinite number of tests of the weak formulation to a finite number of linear equations.

3.2 The associated system

We write again the discrete problem, specifying the numbering of Dirichlet nodes in the discrete Dirichlet condition:

$$\left[\begin{array}{l} \text{find } u_h \in V_h \text{ such that} \\ u_h(\mathbf{p}_i) = g_0(\mathbf{p}_i) \quad \forall i \in \mathrm{Dir}, \\ \displaystyle\int_\Omega \nabla u_h \cdot \nabla v_h + c \int_\Omega u_h\, v_h = \int_\Omega f\, v_h + \int_{\Gamma_N} g_1\, v_h \qquad \forall v_h \in V_h^{\Gamma_D}. \end{array}\right.$$

[3] This way of substituting the Dirichlet condition by a sort of interpolated Dirichlet condition is neither the only nor the best way of doing this approximation, but it is definitely the simplest, so we will keep it like this for the time being.

Our next claim is the following: the discrete equations

$$\int_\Omega \nabla u_h \cdot \nabla v_h + c \int_\Omega u_h\, v_h = \int_\Omega f\, v_h + \int_{\Gamma_N} g_1\, v_h \qquad \forall v_h \in V_h^{\Gamma_D}$$

are equivalent to the following set of equations

$$\int_\Omega \nabla u_h \cdot \nabla \varphi_i + c \int_\Omega u_h\, \varphi_i = \int_\Omega f\, \varphi_i + \int_{\Gamma_N} g_1\, \varphi_i \qquad \forall i \in \text{Ind}.$$

Obviously this second group of equations is a (small) part of the original one: it is enough to take $v_h = \varphi_i \in V_h^{\Gamma_D}$. However, because of the linearity of the first expression in v_h, if we have the second one for all φ_i, we have the equation for all possible linear combinations of these functions, that is for all $v_h \in V_h^{\Gamma_D}$. Summing up, the method is equivalent to this set of N equations to determine the function u_h:

$$\left[\begin{array}{l} \text{find } u_h \in V_h \text{ such that} \\ u_h(\mathbf{p}_i) = g_0(\mathbf{p}_i) \quad \forall i \in \text{Dir}, \\ \displaystyle\int_\Omega \nabla u_h \cdot \nabla \varphi_i + c \int_\Omega u_h\, \varphi_i = \int_\Omega f\, \varphi_i + \int_{\Gamma_N} g_1\, \varphi_i \qquad \forall i \in \text{Ind}. \end{array}\right.$$

In order to arrive at a linear system, we first have to write u_h in terms of the nodal basis functions

$$u_h = \sum_{j\in\text{Ind}} u_j \varphi_j + \sum_{j\in\text{Dir}} u_j \varphi_j.$$

We next substitute the discrete Dirichlet condition in this expression

$$u_h = \sum_{j\in\text{Ind}} u_j \varphi_j + \sum_{j\in\text{Dir}} g_0(\mathbf{p}_j) \varphi_j.$$

Finally we plug this expression into the discrete variational equation

$$\int_\Omega \nabla u_h \cdot \nabla \varphi_i + c \int_\Omega u_h\, \varphi_i = \int_\Omega f\, \varphi_i + \int_{\Gamma_N} g_1\, \varphi_i,$$

apply linearity, noticing that

$$\nabla u_h = \sum_{j\in\text{Ind}} u_j \nabla\varphi_j + \sum_{j\in\text{Dir}} g_0(\mathbf{p}_j) \nabla\varphi_j$$

and move to the right-hand side what we already know (the Dirichlet data)

$$\begin{aligned} \sum_{j\in\text{Ind}} \Big(\int_\Omega \nabla\varphi_j \cdot \nabla\varphi_i + c \int_\Omega \varphi_j \varphi_j \Big) u_j \;\; &= \;\; \int_\Omega f\, \varphi_i + \int_{\Gamma_N} g_1\, \varphi_i \\ &\quad - \sum_{j\in\text{Dir}} \Big(\int_\Omega \nabla\varphi_j \cdot \nabla\varphi_i + c \int_\Omega \varphi_j \varphi_j \Big) g_0(\mathbf{p}_j). \end{aligned}$$

This is a linear system with as many equations as unknowns, namely with $\#\text{Ind} = \dim V_h^{\Gamma_D}$ equations and unknowns. The unknowns are in fact the nodal values of u_h on the free (non-Dirichlet) vertices of the triangulation. After solving this linear system, the formula for u_h lets us recover the function everywhere, not only on nodes.

Remark Unlike the finite difference method, the finite element method gives as a result a function defined on the whole domain and not a set of point values. Reconstruction of the function from computed quantities is in the essence of the method and cannot be counted as a posprocessing of nodal values. □

3.3 Mass and stiffness

There are two matrices in the system above. Both of them participate in the final matrix and parts of them go to build the right-hand side. First we have the **stiffness matrix**

$$w_{ij} = \int_\Omega \nabla\varphi_j \cdot \nabla\varphi_i$$

and second the **mass matrix**

$$m_{ij} = \int_\Omega \varphi_j\,\varphi_i.$$

Both matrices are defined for $i, j = 1, \ldots, N$ (although parts of these matrices won't be used). Both matrices are symmetric. The mass matrix $\mathbf{M}$ is positive definite. The stiffness matrix is positive semidefinite and in fact almost positive definite: if we take an index i and erase the $i-$th row and the $i-$th column of $\mathbf{W}$, the resulting matrix is positive definite.

The system can be easily written in terms of these matrices, using the vector

$$b_i = \int_\Omega f\,\varphi_i + \int_{\Gamma_N} g_1\,\varphi_i, \qquad i \in \mathrm{Ind},$$

to obtain

$$\sum_{j\in\mathrm{Ind}} \big(w_{ij} + c\,m_{ij}\big)u_j = b_i - \sum_{j\in\mathrm{Dir}} \big(w_{ij} + c\,m_{ij}\big)g_0(\mathbf{p}_j), \qquad i \in \mathrm{Ind}.$$

This is clearly a square symmetric linear system. If $c = 0$ (then the original equation is the Poisson equation $-\Delta u = f$ and no reaction term appears), only the stiffness matrix appears. Therefore, stiffness comes from diffusion. Likewise mass proceeds from reaction.

The matrix is positive definite except in one special situation: when $c = 0$ and there are no Dirichlet conditions (i.e., $\Gamma_D = \varnothing$, i.e., $\mathrm{Ind} = \{1, \ldots, N\}$ and $V_h^{\Gamma_D} = V_h$). For the pure Neumann problem for the Laplace operator there are some minor solvability issues similar to the occurrence of rigid motions in mechanical problems. Let us ignore this minor complication for now.

Now look again at the figure showing the supports of nodal basis functions (we copy it right here for convenience) and look at the mass matrix

$$m_{ij} = \int_\Omega \varphi_j\,\varphi_i.$$

If the supports of φ_i and φ_j have no intersecting area, the integral defining m_{ij} vanishes. In fact, since the product of φ_i and φ_j is a non-negative function, $m_{ij} = 0$ if and only if

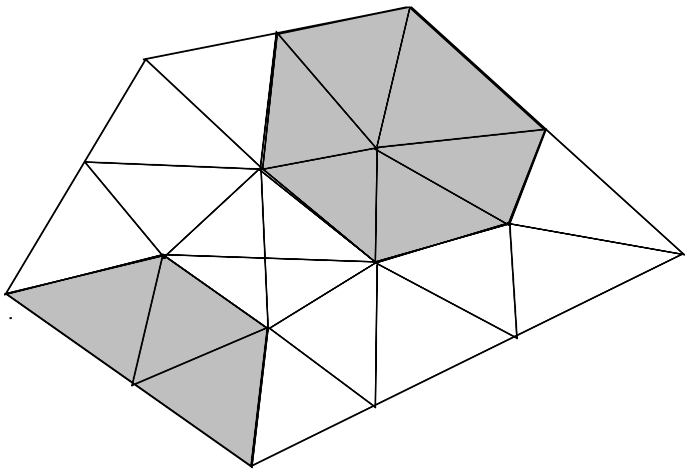


Figure 1.11: Supports of two nodal basis functions

the area of the intersection of the supports is zero[4]. This happens whenever $\mathbf{p}_i$ and $\mathbf{p}_j$ are not vertices of the same triangle.

We say that two nodes are **adjacent** *if they belong to the same triangle.*

In the case of the stiffness matrix we have a similar (maybe weaker result): if the nodes i and j are not adjacent, then $w_{ij} = 0$.

This fact makes the mass and stiffness matrices display a great sparsity character. Given a row i, there are only non-zero entries on positions related to nodes that are adjacent to the $i-$th node.

Going back to the system

$$\sum_{j\in\text{Ind}} \big(w_{ij} + c\, m_{ij}\big) u_j = b_i - \sum_{j\in\text{Dir}} \big(w_{ij} + c\, m_{ij}\big) g_0(\mathbf{p}_j), \qquad i \in \text{Ind},$$

let us remark some simple facts:

- As written now, all data appear in the right-hand side of the system (Neumann data and source terms are in the vector $\mathbf{b}$, Dirichlet data appear multiplying columns of the stiffness-plus-mass matrix).

- Of the full matrices $\mathbf{W}$ and $\mathbf{M}$ we discard rows corresponding to Dirichlet nodes (Dir indices), since no testing is done with the corresponding basis functions. The columns corresponding to these indices are not eliminated though: they are sent to the right-hand side multiplied by the values of the unknown in the Dirichlet nodes, which are known.

[4] By definition the support of a function includes the boundary of the set where the function is non-zero. Therefore, it is possible that the intersection is one edge. The integral is still zero.

4 Exercises

1. **Third type of boundary condition.** Let us consider our usual polygon Ω and the boundary value problem

$$\left[\begin{array}{ll} -\Delta u + u = f & \text{in } \Omega, \\ \partial_n u + k\, u = g & \text{on } \Gamma. \end{array}\right.$$

Here k is a positive parameter. This type of boundary condition is usually called a boundary condition of the third kind (first being Dirichlet and second Neumann) or a Robin (or Fourier) boundary condition.

   (a) Write down the weak formulation for this problem. Note that the condition is natural and there will not be essential boundary condition in the resulting formulation.

   (b) Write down in detail (as in Sections 3.2/ 3.3) the linear system that has to be solved when we apply the finite element method to this problem. Check that there is a new matrix that can be seen as a boundary-mass matrix. How many non-zero entries does each row of this new matrix have?

If we take ε very small and the following slightly modified version of the boundary condition

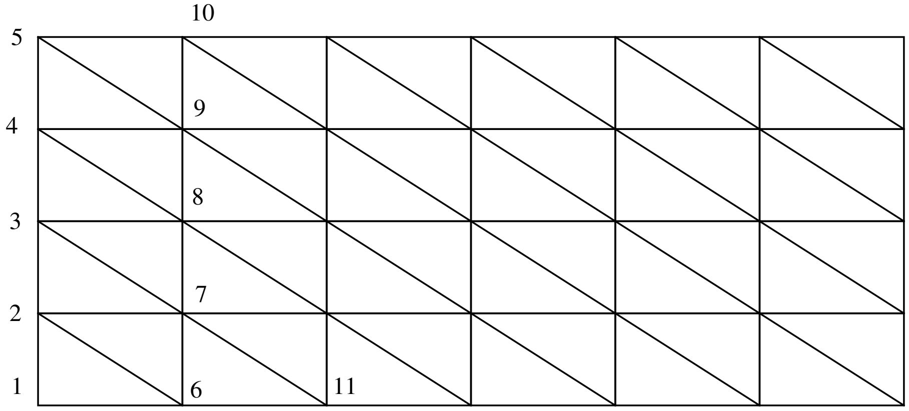
$$\varepsilon \partial_n u + u = g_0, \qquad \text{on } \Gamma$$

(take $k = \varepsilon^{-1}$ and $g = \varepsilon^{-1} g_0$), we are enforcing the Dirichlet condition in an approximate way. This is done in some commercial and open-source packages.

2. **A particular domain.** Consider the boundary problem of Section 1 on the domain given in the next figure and the following specification for Γ_N and Γ_N

the left and upper sides have Dirichlet conditions

and where numbering is done as shown. Let $\mathbf{A} = \mathbf{W} + \mathbf{M}$ be the matrix associated to the system obtained by discretizing with the $\mathbb{P}_1$ finite element method



(a) Write the index sets Dir and Ind.

(b) Write which elements of the 12th row of $\mathbf{A}$ are non-zero.

(c) Identify on the figure the support of the nodal basis function φ_{13}.

(d) What's the size of the system that has to be solved?

(e) We call the profile of the matrix $\mathbf{A}$ to the following vector:

$$m(i) = \inf\{j \,:\, a_{ij} \neq 0\}, \qquad i = 1, \ldots, \#\{\text{nodos}\}$$

that is, $m(i)$ indicates the column number where the first non-zero entry of the ith row is. Compute the profile of $\mathbf{W} + \mathbf{M}$ (without eliminating Dirichlet rows and columns). Draw the form of the matrix using the profile.

(f) In the preceding graph mark which rows and columns will be modified by introduction of Dirichlet conditions. Compute the profile of the reduced matrix (without Dirichlet rows and columns).

(g) What happens if we number nodes horizontally?

Lesson 2

Theoretical and practical notions

1 Assembly

The first lesson left us with a linear system to solve in order to approximate the boundary value problem with the finite element method. There is however the trick question on how to compute all the integrals that appear in the matrix and right-hand side of the system. This is done by a clever process called **assembly** of the system, another of the many good deeds of the finite element method that has made it so extremely popular (as in popular among scientists and engineers, of course) in the last decades.

At this moment we need the polygonal domain Ω and:

- a triangulation $\mathcal{T}_h$,
- a numbering of the nodes $\{\mathbf{p}_i\}$ (nodes are the vertices of the triangles),
- the set of the nodal basis functions $\{\varphi_i\}$.

In this section, nNod will be the global number of nodes.

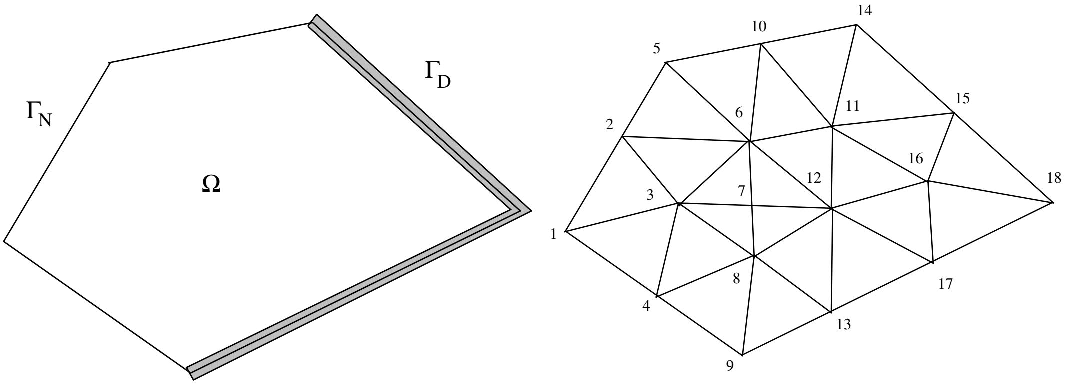


Figure 2.1: Geometry of the problem and triangulation

1.1 The mass and stiffness matrices

We are going to center our attention in the efficient construction of the stiffness matrix

$$w_{ij} = \int_\Omega \nabla\varphi_j \cdot \nabla\varphi_i$$

and of the mass matrix

$$m_{ij} = \int_\Omega \varphi_j \, \varphi_i.$$

Integrals over Ω can be decomposed as the sum of integrals over the different triangles

$$w_{ij} = \int_\Omega \nabla\varphi_j \cdot \nabla\varphi_i = \sum_K \int_K \nabla\varphi_j \cdot \nabla\varphi_i = \sum_K w_{ij}^K.$$

On each triangle we are going to define three local nodal basis functions. First assign a number to each of the three vertices of a triangle K:

$$\mathbf{p}_1^K, \quad \mathbf{p}_2^K, \quad \mathbf{p}_3^K.$$

Then consider the functions

$$N_1^K, \quad N_2^K, \quad N_3^K \quad \in \mathbb{P}_1$$

that satisfy

$$N_\alpha^K(\mathbf{p}_\beta^K) = \delta_{\alpha\beta}, \qquad \alpha, \beta = 1, 2, 3.$$

It is simple to see that the nodal basis function φ_i restricted to the triangle K is either zero (this happens when $\mathbf{p}_i$ is not one of the three vertices of K) or one of the N_α^K functions. More precisely, let n_α be the global number of the local node with number α in the triangle K. This means that

$$N_\alpha^K = \varphi_{n_\alpha}, \qquad \text{on the triangle } K.$$

We can now compute the 3×3 matrix $\mathbf{K}^K$

$$k_{\alpha\beta}^K = \int_K \nabla N_\beta^K \cdot \nabla N_\alpha^K, \qquad \alpha, \beta = 1, 2, 3.$$

This is due to be simple, since the functions N_α^K are polynomials of degree one (unlike the functions φ_i that are only piecewise polynomials). Later on, we will see strategies to do this computation. Note at this moment that computation of this matrix depends only on the triangle K and does not take into account any other element of the triangulation.

Therefore

$$k_{\alpha\beta}^K = w_{n_\alpha n_\beta}^K$$

All other elements of the matrix $\mathbf{W}^K$ are zero. Recall again that $\mathbf{W}^K$ is a nNon × nNod matrix and that

$$\mathbf{W} = \sum_K \mathbf{W}^K.$$

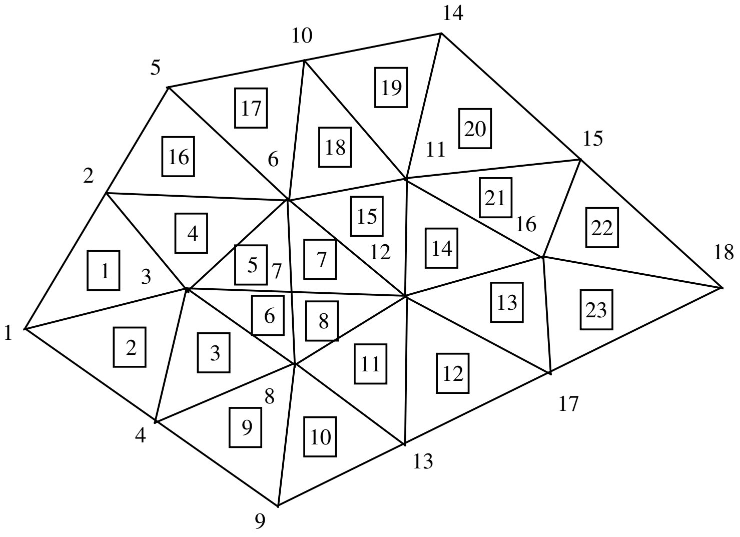
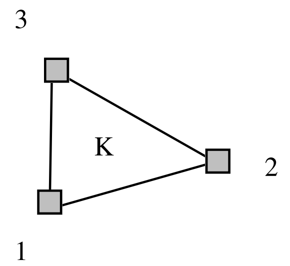


Figure 2.2: A numbering of the triangles.



local		global
1	↔	12
2	↔	16
3	↔	11

Figure 2.3: The 14th triangle and their vertex numberings.

The assembly process requires then a given numbering of triangles as shown in Figure 2.2. The order of this numbering is only used to do the computations but does not modify the shape of the final result.

The process to assemble the mass matrix is the same. Effective assembly of the mass and stiffness matrices can be done at the same time. Instead of computing separately the matrix $\mathbf{K}^K$ and a similar one for the mass matrix, we can directly try to compute the 3×3 matrix with elements

$$\int_K \nabla N_\beta^K \cdot \nabla N_\alpha^K + c \int_K N_\beta^K \, N_\alpha^K, \qquad \alpha, \beta = 1, 2, 3.$$

1.2 The reference element

To compute the elements

$$\int_K \nabla N^K_\beta \cdot \nabla N^K_\alpha \qquad \text{and} \qquad \int_K N^K_\beta \, N^K_\alpha$$

we need: (a) either and effective way of evaluating the functions N^K_α and their gradients; (b) or a closed form for the resulting integrals. Both possibilities are done usually by moving to the so-called reference element.

For triangles, the reference element is the triangle with vertices

$$\widehat{\mathbf{p}}_1 = (0,0), \qquad \widehat{\mathbf{p}}_2 = (1,0), \qquad \widehat{\mathbf{p}}_3 = (0,1).$$

To distinguish variables in the reference element and in a general triangle (in this context we say a physical element) it is customary to use the variables (ξ, η) in the reference element and (x, y) in the physical element. In the mathematical literature for FEM it is also usual to put a hat on the name of the variables in the reference element, so that $(\hat{x}, \hat{y})$ would be used to denote coordinates in the reference configuration.

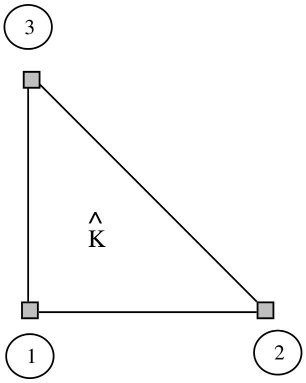


Figure 2.4: The reference element

An unimportant detail. Some people prefer to use a different reference triangle, with the same shape but with vertices on $(-1,-1)$, $(1,-1)$ and $(-1,1)$. Some details of the forthcoming computations have to be adapted if this choice is taken. □

The local nodal functions in the reference triangles are three $\mathbb{P}_1$ functions satisfying

$$\widehat{N}_\alpha(\widehat{\mathbf{p}}_\beta) = \delta_{\alpha\beta}, \qquad \alpha, \beta = 1, 2, 3.$$

These functions are precisely

$$\widehat{N}_1 = 1 - \xi - \eta, \qquad \widehat{N}_2 = \xi, \qquad \widehat{N}_3 = \eta$$

or, if you prefer hatting variables (this is the last time we will write both expressions)

$$\widehat{N}_1 = 1 - \widehat{x} - \widehat{y}, \qquad \widehat{N}_2 = \widehat{x}, \qquad \widehat{N}_3 = \widehat{y}$$

Let us now take the three vertices of a triangle K

$$\mathbf{p}_1^K = (x_1, y_1), \qquad \mathbf{p}_2^K = (x_2, y_2), \qquad \mathbf{p}_3^K = (x_3, y_3).$$

The following affine transformation[1]

$$\begin{aligned} \begin{bmatrix} x \\ y \end{bmatrix} &= \underbrace{\begin{bmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{bmatrix}}_{\mathbf{B}_K} \begin{bmatrix} \xi \\ \eta \end{bmatrix} + \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} \\ &= \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} (1 - \xi - \eta) + \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} \xi + \begin{bmatrix} x_3 \\ y_3 \end{bmatrix} \eta \end{aligned}$$

maps the triangle $\widehat{K}$ bijectively into K. In fact, if we call this transformation F_K, then

$$F_K(\widehat{\mathbf{p}}_\alpha) = \mathbf{p}_\alpha^K, \qquad \alpha = 1, 2, 3.$$

Notice that the second expression we have written for the transformation gives it in terms of the nodal basis functions in the reference domain. You can think of it as a coincidence. In a way it is: the coincidence stems from the fact that the type of functions we are using for finite elements is the same as the functions needed to transform linearly triangles in the plane.

It is simple now to prove that

$$\widehat{N}_\alpha = N_\alpha^K \circ F_K, \qquad \alpha = 1, 2, 3,$$

or, what is the same

$$N_\alpha^K = \widehat{N}_\alpha \circ F_K^{-1}, \qquad \alpha = 1, 2, 3.$$

The $\circ$ symbol is used for composition. In the last expression, what we have is

$$N_\alpha^K(x, y) = \widehat{N}_\alpha(F_K^{-1}(x, y)).$$

Since computing F_K^{-1} is straightforward from the explicit expression for F_K, this formula gives a simple way of evaluating the functions N_α^K. The fact of representing the local basis for the physical in terms of the basis in the reference configuration, $N_\alpha^K = \widehat{N}_\alpha \circ F_K^{-1}$, is referred to as *pushing forward the basis* on the reference element[2].

To evaluate the gradient of N_α^K we have to be more careful, since we have to apply the chain rule. Let us denote briefly gradients as

$$\nabla = \begin{bmatrix} \partial_x \\ \partial_y \end{bmatrix}, \qquad \widehat{\nabla} = \begin{bmatrix} \partial_\xi \\ \partial_\eta \end{bmatrix}.$$

(Note that we are writing gradients as column vectors.) The following formula is the result of applying the chain rule

$$\mathbf{B}_K^\top \big(\nabla\phi \circ F_K\big) = \widehat{\nabla}(\phi \circ F_K).$$

[1] Many mesh generators prepare number triangle locally by ordering nodes counterclockwise. This makes $\det \mathbf{B}_K > 0$.

[2] The opposite process, bringing something from the physical element to the reference one, is called *pull-back*.

$\mathbf{B}_K^\top$ is the transpose of the matrix of the linear transformation F_K. Taking $\phi = N_\alpha^K$ in this expression and moving things a little, we obtain a formula for the gradient of the local basis functions

$$\nabla N_\alpha^K = \mathbf{B}_K^{-\top}\big((\widehat{\nabla}\widehat{N}_\alpha) \circ F_K^{-1}\big).$$

The expression may look complicated but it is very simple to use. If we want to compute the value of the gradient of N_α^K at a point $(x, y) \in K$, we first compute the transformed point $(\xi, \eta) = F_K^{-1}(x, y)$ in the reference triangle, evaluate the gradient of $\widehat{N}_\alpha$ at this point and then multiply it by the matrix $\mathbf{B}_K^{-\top}$, which is the transpose of the inverse of $\mathbf{B}_K$, i.e.,

$$\mathbf{B}_K^{-\top} = \frac{1}{\det \mathbf{B}_K} \begin{bmatrix} y_3 - y_1 & -(y_2 - y_1) \\ -(x_2 - x_1) & x_2 - x_1 \end{bmatrix}$$

with

$$\det \mathbf{B}_K = (x_2 - x_1)(y_3 - y_1) - (y_2 - y_1)(x_3 - x_1)$$

(remember that $|\det \mathbf{B}_K| = 2 \operatorname{area} K$). In fact, for this very elementary method, the gradients of the three basis functions on the reference element are constant vectors

$$\widehat{\nabla}\widehat{N}_1 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \qquad \widehat{\nabla}\widehat{N}_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \widehat{\nabla}\widehat{N}_3 = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

so computation of the constant vectors ∇N_α^K is very simple, and we don't even have to use the inverse transformation F_K^{-1} for the gradients. We do, however, to evaluate N_α^K.

1.3 Computing with quadrature rules

Depending on the complications of the problem (we are dealing with a very simple model problem), all the computations can be carried out to the reference element or we can try to do things directly on the physical triangle K. Let us mention here two popular quadrature rules for triangles: the three point rule with the vertices

$$\int_K \phi \approx \frac{\operatorname{area} K}{3}\Big(\phi(\mathbf{p}_1^K) + \phi(\mathbf{p}_2^K) + \phi(\mathbf{p}_3^K)\Big)$$

and the midpoints approximation

$$\int_K \phi \approx \frac{\operatorname{area} K}{3}\Big(\phi(\mathbf{m}_1^K) + \phi(\mathbf{m}_2^K) + \phi(\mathbf{m}_3^K)\Big),$$

where $\mathbf{m}_\alpha^K$ are the midpoints of the edges of K. If ϕ is a polynomial of degree one, the first formula gives the exact value. The second formula is even better: if ϕ is a polynomial of degree two, the edge-midpoints formula is exact.

In the very simple case of $\mathbb{P}_1$ elements, we have ∇N_α^K constant and therefore

$$\int_K \nabla N_\beta^K \cdot \nabla N_\alpha^K = (\operatorname{area} K)\, \nabla N_\beta^K \cdot \nabla N_\alpha^K,$$

and this computation is very simple. For the mass matrix, we note that $N_\beta^K\, N_\alpha^K$ is a polynomial of degree two and therefore, the edge-midpoints formula gives the exact value of the integrals

$$\int_K N_\beta^K\, N_\alpha^K$$

with just three evaluations of the functions.

1.4 Doing everything on the reference element

This section gives another idea on how to compute the local mass and stiffness matrices. You can skip it without losing continuity and go to Section 1.5. The change of variables applied to the integral of the local mass matrix gives

$$\int_K N_\beta^K\, N_\alpha^K = |\det \mathbf{B}_K| \int_{\widehat{K}} \widehat{N}_\beta \widehat{N}_\alpha.$$

Therefore everything is done once we have the 3×3 matrix

$$\widehat{\mathbf{K}}_0 = \left[\int_{\widehat{K}} \widehat{N}_\beta \widehat{N}_\alpha \right]_{\alpha,\beta} = \tfrac{1}{24} \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$$

For derivatives, we have to be more careful

$$\begin{aligned} \int_K \nabla N_\beta^K \cdot \nabla N_\alpha^K &= |\det \mathbf{B}_K| \int_{\widehat{K}} \big(\nabla N_\beta^K \circ F_K\big) \cdot \big(\nabla N_\alpha^K \circ F_K\big) = \\ &= |\det \mathbf{B}_K| \int_{\widehat{K}} \Big(\mathbf{B}_K^{-\top} \widehat{\nabla} \widehat{N}_\beta\Big) \cdot \Big(\mathbf{B}_K^{-\top} \widehat{\nabla} \widehat{N}_\alpha\Big) = \\ &= |\det \mathbf{B}_K| \int_{\widehat{K}} \mathbf{C}_K \widehat{\nabla} \widehat{N}_\beta \cdot \widehat{\nabla} \widehat{N}_\alpha \end{aligned}$$

where

$$\mathbf{C}_K = \mathbf{B}_K^{-1} \mathbf{B}_K^{-\top} = \begin{bmatrix} c_{11}^K & c_{12}^K \\ c_{12}^K & c_{22}^K \end{bmatrix}$$

is a symmetric 2×2 matrix that depends only on the triangle. If we compute the following 3×3 matrices in the reference element

$$\begin{aligned} \widehat{\mathbf{K}}_{\xi\xi} &= \left[\int_{\widehat{K}} \partial_\xi \widehat{N}_\beta \, \partial_\xi \widehat{N}_\alpha \right]_{\alpha,\beta} = \tfrac{1}{2} \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\ \widehat{\mathbf{K}}_{\eta\eta} &= \left[\int_{\widehat{K}} \partial_\eta \widehat{N}_\beta \, \partial_\eta \widehat{N}_\alpha \right]_{\alpha,\beta} = \tfrac{1}{2} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix} \\ \widehat{\mathbf{K}}_{\xi\eta} &= \left[\int_{\widehat{K}} \partial_\xi \widehat{N}_\beta \, \partial_\eta \widehat{N}_\alpha \right]_{\alpha,\beta} = \tfrac{1}{2} \begin{bmatrix} 1 & 0 & -1 \\ -1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \end{aligned}$$

we have

$$\left[\int_K \nabla N_\beta^K \cdot \nabla N_\alpha^K \right]_{\alpha,\beta} = |\det \mathbf{B}_K| \left(c_{11}^K \widehat{\mathbf{K}}_{\xi\xi} + c_{22}^K \widehat{\mathbf{K}}_{\eta\eta} + c_{12}^K (\widehat{\mathbf{K}}_{\xi\eta} + \widehat{\mathbf{K}}_{\xi\eta}^\top) \right).$$

1.5 Right-hand sides

Construction of the right-hand side of the linear system requires the computation of two vectors:

$$\int_\Omega f\,\varphi_i, \qquad \int_{\Gamma_N} g_1\,\varphi_i.$$

In principle, this has to be done for indices of free nodes ($i \in \mathrm{Ind}$), but in practice what is done is to compute them for all i and then discard the elements corresponding to Dirichlet nodes.

The surface forces (source terms) can be treated in a similar way to the stiffness and mass matrices:

$$\int_\Omega f\,\varphi_i = \sum_K \int_K f\,\varphi_i.$$

For each triangle K we compute the vector

$$\int_K f\,N_\alpha^K, \qquad \alpha = 1, 2, 3$$

and then add these elements in the positions (n_1, n_2, n_3) of the full vector. This process can be done at the same time as the matrix assembly, since it goes triangle by triangle. For the $\mathbb{P}_1$ element, the following extremely simple approximation is enough:

$$\begin{aligned}\int_K f\,N_\alpha^K &\approx \tfrac{1}{3}\sum_{\beta=1}^3 f(\mathbf{p}_\beta^K)\int_K N_\alpha^K = \frac{|\det \mathbf{B}_K|}{3}\sum_{\beta=1}^3 f(\mathbf{p}_\beta^K)\int_{\widehat K}\widehat N_\alpha \\ &= \frac{|\det \mathbf{B}_K|}{18}\sum_{\beta=1}^3 f(\mathbf{p}_\beta^K).\end{aligned}$$

A simpler options is

$$\int_K f\,N_\alpha^K \approx f(\mathbf{b}^K)\int_K N_\alpha^K = f(\mathbf{b}^K)\frac{|\det \mathbf{B}_K|}{6},$$

where

$$\mathbf{b}^k = \frac{1}{3}(\mathbf{p}_1^K + \mathbf{p}_2^K + \mathbf{p}_3^K)$$

is the barycenter of K. (This second option is wiser when f has discontinuities that are captured by the triangulation, that is, when f is allowed to have jumps across element interfaces.)

The three integrals related to the element K are approximated by the same number. We have actually approximated f by a function that is constant over each triangle: the constant value on the triangle is the average of the values on its vertices (or its value at the barycenter). Otherwise, we can try a quadrature rule to approximate the integrals. It is important at this stage to note that the choice of an adequate quadrature rule has to take into account two facts:

- it has to be precise enough not to lose the good properties of the finite element method, but

- it has to be simple enough not to be wasting efforts in computing with high precision a quantity that is only needed with some precision.

In principle, we could think of using a very precise rule to compute the integrals as exactly as possible. This is overdoing it and forgetting one of the most important principles of well-understood scientific computing: errors from different sources have to be balanced. It doesn't make much sense to spend time in computing exactly a quantity when that number is to be used in the middle of many approximate computations.

The presence of Neumann boundary conditions imposes the computation of the following integrals

$$\int_{\Gamma_N} g_1\, \varphi_i.$$

This process is made separately to the ones of computing domain integrals for the matrices and the source terms. First of all we have to decompose the Neumann boundary in the set of edges that lie on it (for that we will need a numbering of the Neumann edges):

$$\int_{\Gamma_N} g_1\, \varphi_i = \sum_L \int_L g_1\, \varphi_i.$$

Note first that unless $\mathbf{p}_i$ is on the Neumann boundary, this integral vanishes.

Next, for each edge consider the two vertices that delimit it: $\mathbf{p}_1^L$ and $\mathbf{p}_2^L$. As we had with triangular elements, we will need the relation between the extremal points of each Neumann edge and the global numbering. If

$$\mathbf{p}_1^L = (x_1, y_1), \qquad \mathbf{p}_2^L = (x_2, y_2),$$

the function

$$[0,1] \ni t \longmapsto \phi_L(t) = (1-t)\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} + t\begin{bmatrix} x_2 \\ y_2 \end{bmatrix}$$

is a parameterization of the segment L. We now consider the following two functions

$$\psi_1 = 1 - t, \qquad \psi_2 = t.$$

They are just the nodal basis functions on the reference element $[0,1]$ for the space of linear polynomials in one dimension. It is simple to see that

$$(\varphi_i \circ \phi_L)(t) = \begin{cases} \psi_1(t), & \text{if } \mathbf{p}_i = \mathbf{p}_1^L, \\ \psi_2(t), & \text{if } \mathbf{p}_i = \mathbf{p}_2^L, \\ 0, & \text{otherwise.} \end{cases}$$

The integrals to be computed are

$$\int_L g_1 \varphi_{n_\alpha} = \operatorname{length} L \int_0^1 (g_1 \circ \phi_L)(t) \psi_\alpha(t) \mathrm{d}t, \qquad \alpha = 1, 2$$

(as before n_α denotes the global index for the local node α). We can the use numerical quadrature for this line integral. Alternatively we can approximate

$$\int_L g_1\, \varphi_{n_\alpha} \approx g_1(\mathbf{m}^L) \int_L \varphi_i = \frac{\operatorname{length} L}{2} g_1(\mathbf{m}^L), \qquad \alpha = 1, 2,$$

where $\mathbf{m}^L = \frac{1}{2}(\mathbf{p}_1^L + \mathbf{p}_2^L)$ is the midpoint of L.

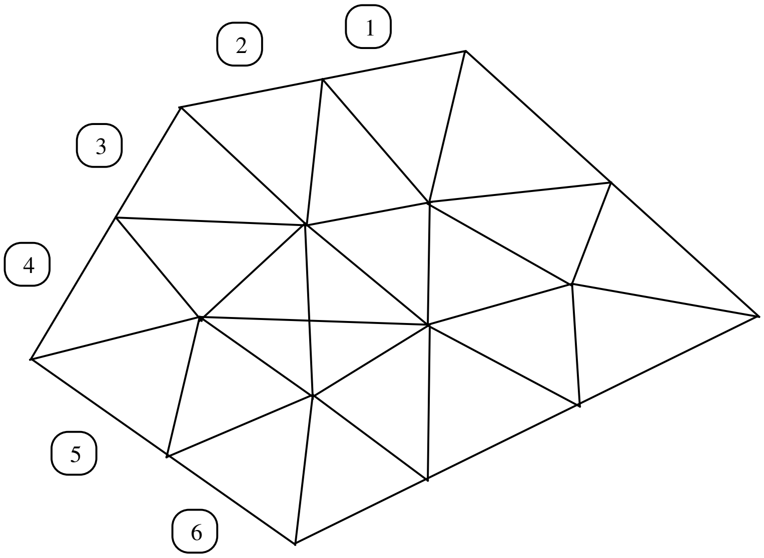
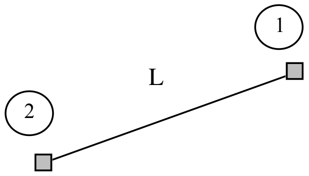


Figure 2.5: A numbering of Neumann edges/elements.



local		global
1	↔	10
2	↔	5

Figure 2.6: The 2nd Neumann edge and its numberings. For this edge, $n_1 = 10$ and $n_2 = 5$. It is common to number boundary edges positively from the point of view of the interior domain, that is, when going from the first node to the second, we leave the interior domain to the left.

2 A taste of the theory

2.1 Abstract frame

Because many of the ideas that we will develop on and on in this course are quite independent from the particular problem, let us rewrite everything in a slightly more abstract language. We have two spaces,

$$V = H^1(\Omega) \quad \text{and} \quad V_0 = H^1_{\Gamma_D}(\Omega),$$

a bilinear form (related only to the partial differential operator)

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v + c \int_\Omega u\, v$$

and a linear form where most of the data appear

$$\ell(v) = \int_\Omega f\, v + \int_{\Gamma_N} g_1\, v.$$

Finally there is a linear operator γ that serves us to describe the essential conditions: for us γu is the value of u on the boundary Γ_D. Notice that

$$V_0 = \{v \in V \, : \, \gamma v = 0\}.$$

The problem admits then this simple form

$$\left[\begin{array}{l} \text{find } u \in V \text{ such that} \\[1ex] \gamma u = g_0, \\[1ex] a(u,v) = \ell(v) \quad \forall v \in V_0 \end{array}\right. .$$

Only when $g_0 = 0$ (or when there's no Γ_D and the whole boundary is a Neumann boundary), the problem reduces to an even simpler one

$$\left[\begin{array}{l} \text{find } u \in V_0 \text{ such that} \\[1ex] a(u,v) = \ell(v) \quad \forall v \in V_0. \end{array}\right.$$

Therefore, when the essential condition is homogeneous (or when there is no essential condition), the set where we look for u and the test space are the same. In other cases, the restriction imposed to the tests v is the homogeneous version of the essential condition.

2.2 Well-posedness

Let us recall that the natural norm in our space $V = H^1(\Omega)$ was

$$\|u\| = \|u\|_{1,\Omega} = \left(\int_\Omega |\nabla u|^2 + \int_\Omega |u|^2 \right)^{1/2}.$$

There are several conditions that ensure the well-posedness of the problem

$$\left[\begin{array}{l} \text{find } u \in V \text{ such that} \\[1ex] \gamma u = g_0, \\[1ex] a(u,v) = \ell(v) \quad \forall v \in V_0, \end{array}\right.$$

or of its homogeneous version ($g_0 = 0$)

$$\left[\begin{array}{l} \text{find } u \in V_0 \text{ such that} \\[1ex] a(u,v) = \ell(v) \quad \forall v \in V_0. \end{array}\right.$$

Well-posedness means existence and uniqueness of solution and continuity of the solution with respect to the data.

Let us first list the properties that are satisfied in all the situations we are addressing in this course:

- V is a Hilbert space (a vector space, with an inner product so that the space is complete with respect to the associate norm)[3],
- V_0 is a closed subspace of V,
- the bilinear form a is continuous in V, that is, there exists $M > 0$ such that

$$|a(u,v)| \le M\|u\|\,\|v\|, \qquad \forall u, v \in V,$$

- the linear form ℓ is continuous

$$|\ell(v)| \le C_\ell \|v\|, \qquad \forall v \in V.$$

As already mentioned, all of these properties are satisfied in our case. In fact

$$C_\ell^2 \le \int_\Omega |f|^2 + C_\Omega \int_{\Gamma_N} |g_1|^2.$$

There is a last property, called **ellipticity** or **coercivity**, which reads: there exists $\alpha > 0$ such that

$$a(v,v) \ge \alpha \|v\|^2, \qquad \forall v \in V_0.$$

Note that the property is only demanded on the set V_0. In our case it is not satisfied in all situations. In fact, it is satisfied in all but one case:

- if $c > 0$ the property is satisfied with α depending only on c,
- if $c = 0$ and length $\Gamma_D > 0$, the property is satisfied with α depending on Ω and on the partition of the boundary in Dirichlet and Neumann parts.

If all the properties mentioned above hold, then the problem

$$\left[\begin{array}{l} \text{find } u \in V_0 \text{ such that} \\[2ex] a(u,v) = \ell(v) \quad \forall v \in V_0, \end{array}\right.$$

has a unique solution and

$$\|u\| \le C_\ell/\alpha.$$

If $g_0 \neq 0$ then the problem

$$\left[\begin{array}{l} \text{find } u \in V \text{ such that} \\[2ex] \gamma u = g_0, \\[2ex] a(u,v) = \ell(v) \quad \forall v \in V_0, \end{array}\right.$$

has a unique solution if there exists a $u_0 \in V$ such that $\gamma u_0 = g_0$. In that case, the continuity of the solution with respect to the data has a more complicated expression

$$\|u\| \le C_\ell/\alpha + (M/\alpha + 1) \inf \big\{ \|u_0\| \,:\, \gamma u_0 = g \big\}.$$

[3] Maybe this sentence looks too hard. You should know what a vector space and also what an inner (or scalar) product is. When you have an inner product, you have an associated norm and with it a concept of convergence of sequences of elements of V. Completeness is a property that ensures that all Cauchy sequences have a limit. In essence, it means that convergence has to happen inside the space. We cannot have a sequence of elements of V converging to something that is not in V.

Remark. For the pure Neumann problem with $c = 0$

$$\left[\begin{array}{ll} -\Delta u = f & \text{in } \Omega, \\ \partial_n u = g_1 & \text{on } \Gamma, \end{array}\right.$$

we cannot verify the conditions to prove existence and uniqueness. In fact, existence is not guaranteed and we never have uniqueness. First of all, because of the divergence theorem we must have

$$\int_\Omega \Delta u = \int_\Omega \operatorname{div}(\nabla u) = \int_\Gamma \partial_n u$$

and therefore the data have to satisfy the compatibility condition

$$\int_\Omega f + \int_\Gamma g_1 = 0.$$

If this condition is satisfied, there is more that one solution, since constant functions satisfy the problem

$$\left[\begin{array}{ll} -\Delta u = 0 & \text{in } \Omega, \\ \partial_n u = 0 & \text{on } \Gamma. \end{array}\right.$$

□

2.3 Galerkin methods

A Galerkin method for the problem

$$\left[\begin{array}{l} \text{find } u \in V_0 \text{ such that} \\ a(u, v) = \ell(v) \quad \forall v \in V_0, \end{array}\right.$$

consists of the choice of a finite dimensional space

$$V_h^0 \subset V_0$$

and on the consideration of the discrete problem

$$\left[\begin{array}{l} \text{find } u_h \in V_h^0 \text{ such that} \\ a(u_h, v_h) = \ell(v_h) \quad \forall v_h \in V_h^0. \end{array}\right.$$

The $\mathbb{P}_1$ finite element method for the reaction-diffusion problem with homogeneous Dirichlet conditions is therefore an example of Galerkin method[4].

The Galerkin equations are equivalent to a linear system. Let us do here the detailed argument, although you will see that we already did exactly this in Section 3 of the previous lesson.

[4] Galerkin comes from Boris Galerkin. A good pronunciation of the word would be something more like *Galyorkin*, with emphasis on the *lyor* syllable. Most English speakers pronounce it however as if it were an English word.

First we need a basis of V_h^0: $\{\varphi_i \,:\, i \in \mathrm{Ind}\}$. The index set Ind is now anything you want to use in order to number the finite basis of the set. In general we would number form one to the dimension of the space, but in our model problem the numbering proceeds from eliminating some indices from a wider numbering. Then we notice that the abstract set of equations

$$a(u_h, v_h) = \ell(v_h) \quad \forall v_h \in V_h^0$$

is equivalent to

$$a(u_h, \varphi_i) = \ell(\varphi_i) \quad \forall i \in \mathrm{Ind}.$$

Finally, we decompose

$$u_h = \sum_{j \in \mathrm{Ind}} u_j \varphi_j$$

and substitute this expression above to obtain the linear system

$$\sum_{j \in \mathrm{Ind}} a(\varphi_j, \varphi_i) u_j = \ell(\varphi_i), \qquad i \in \mathrm{Ind}.$$

There are as many unknowns as there are equations here. In this abstract setting, the values u_j are not nodal values, since an arbitrary basis of a linear space has nothing to do with nodes or evaluations of functions.

If the hypotheses of Section 2.2 hold, this system has a unique solution. Furthermore we have the following result, which is popularly referred to as Céa's Lemma[5]:

$$\|u - u_h\| \le \frac{M}{\alpha} \inf \big\{ \|u - v_h\| \,:\, v_h \in V_h^0 \big\}.$$

The result might not seem to say much at first sight. There are however some aspects that have to be remarked here:

- The result gives an upper bound of the error between the exact solution u and the approximate solution u_h (the finite element solution) and this error bound is measured in the *energy* norm and not in any other one.

- The term
$$\inf \big\{ \|u - v_h\| \,:\, v_h \in V_h^0 \big\}$$
is just an approximation error, completely unrelated to the original problem. It measures how well the (unknown) exact solution can be approximated by elements of the space where we are looking for the solution. Because of how this term is estimated in particular situations (in FEM, for instance) many people call this an interpolation error. We will see a bit of this in the following section. This approximation error is measured also in the energy norm, of course[6].

[5] Céa, as in Jean Céa. French. Do you best with the pronunciation of the name.

[6] There's a well-established tradition to keep the infimium in the right-hand side of Céa's estimate. The infimum is actually a minimum, as guaranteed by elementary functional analysis arguments. Céa's estimate is also called the *quasioptimality* of the Galerkin method.

- The only other constants in the inequality depend on the problem, but not on data. Note however that complicated solutions (solutions that vary a lot, or that have large gradients, or anything you can think of as difficult to grasp with a simple approximation) will not necessarily be approximated as well as simple smooth solutions. Since we do not know the solution (by definition, it is the unknown), how can we have an idea of this error? The answer is the lifetime work of numerical analysts and computational scientists. Just three ideas:
  - for simple smooth solutions, numerical analysis shows usually how error behaves quite precisely, which gives us a hint of the best possible behavior of our method;
  - PDE theory sometimes helps in understanding where things can go wrong and we can do some effort in concentrating approximation in that area;
  - finally, there is a whole new (new as in only thirty years old or so) branch of computational knowledge related to error estimation and adaptivity, allowing you to improve your computations with information you harvest from the already performed computations.

The theoretical frame for the case with non-homogeneous Dirichlet conditions is somewhat more delicate, because we have to go someplace more abstract to write correctly the approximation of the condition

$$u = g_0 \quad \text{on } \Gamma_D$$

by

$$u_h(\mathbf{p}) = g_0(\mathbf{p}) \quad \forall \mathbf{p} \text{ Dirichlet node},$$

without making any use of the particularities of the finite element space $\mathbb{P}_1$. This can be done in several ways, and we are not going to detail them. Particularized to FEM the result will look like this

$$\|u - u_h\| \le (1 + \frac{M}{\alpha}) \inf \big\{ \|u - v_h\| \,:\, v_h \in V_h, \quad v_h(\mathbf{p}) = g_0(\mathbf{p}) \quad \forall \mathbf{p} \text{ Dirichlet node} \big\}.$$

Note that the approximation error in the right-hand side includes the imposition of the discrete essential boundary condition.

2.4 Convergence of the $\mathbb{P}_1$ finite element method

How does all of this work for the $\mathbb{P}_1$ finite element? Let us go back to the case with homogeneous boundary conditions. As mentioned, the error can be bounded as

$$\|u - u_h\|_{1,\Omega} \le \frac{M}{\alpha} \inf \big\{ \|u - v_h\|_{1,\Omega} \,:\, v_h \in V_h^0 \big\}.$$

Let us emphasize again that the norm for measuring the error is imposed by the problem (see Section 2.1). Assume now that u is a well-behaved function. For example, that it is continuous. Then we can construct a function $\pi_h u$ by taking nodal values of u on the vertices of the triangulation and creating with them an element of V_h. This is, obviously, interpolation in V_h, that is, interpolation with continuous piecewise linear functions. Because of the Dirichlet boundary condition u vanishes on Dirichlet nodes, and so does consequently $\pi_h u$. Therefore $\pi_h u \in V_h^0$ and we can use the bound

$$\|u - u_h\|_{1,\Omega} \le \frac{M}{\alpha} \|u - \pi_h u\|_{1,\Omega}.$$

We have therefore bounded the error of the finite element method by the error of interpolation of the exact solution in the finite element space. A nice thing about this interpolation process is the fact that it is done triangle-by-triangle, so actually, the global error for interpolation is the sum of the errors that we have done element-by-element.

In basic courses on numerical methods you will have seen that it is possible to estimate the error of interpolation without knowing the solution, but that this bound of the error is proportional to some quantity depending on a high order derivative of the function that is interpolated. You will have seen this in one space dimension. In several space dimensions, it is a bit more difficult but not so much. The result is the following: there exists a constant C that depends on the minimum angle of the triangulation such that

$$\|u - \pi_h u\|_{1,\Omega} \le C h \Big(\int_\Omega |\partial_{xx} u|^2 + |\partial_{xy} u|^2 + |\partial_{yy} u|^2 \Big)^{1/2},$$

where h is the size of the longest edge of the triangulation. The expression on the right-hand side is an example of a Sobolev seminorm. It is denoted usually as

$$|u|_{2,\Omega} = \Big(\int_\Omega |\partial_{xx} u|^2 + |\partial_{xy} u|^2 + |\partial_{yy} u|^2 \Big)^{1/2}.$$

The whole bound is

$$\|u - u_h\|_{1,\Omega} \le C' h |u|_{2,\Omega}$$

with the constant C' depending on the coefficients of the problem, on the geometry of the physical setting and on the smallest angle. If the triangles are very flat (the ratio between the longest edge and the inradius[7] is very small), the constant gets to be very large.

First of all, let us remark that the error bound requires the second derivatives of the solution to be square-integrable, which is not always the case. Second, note that if u is a polynomial of degree one, this error bound is zero and u_h is exactly u. You can use this as a way of constructing exact solutions to validate your own coding of the method. Third, the fact that the bound is proportional to h makes the method a **method of order one**. This means that if you make the longest edge half its size, you should only expect the error to be divided by two. Be aware that the argument on error decrease is done on the bound, since the error itself is unknown. In fact the error could decrease much faster, but in principle you should not expect this to happen.

3 Quadratic elements

Its very low order makes the $\mathbb{P}_1$ method not very attractive. Just to expect having an additional digit in precision you should have edges ten times shorter, which amounts to

[7] Inradius is the geometric term for the radius of the inscribed circumference.

increasing dramatically the number of unknowns. Instead, it is often recommended to use a higher order method, which is exactly what we are going to do right now.

3.1 Local and global descriptions

Let us consider the space of polynomials in two variables with degree at most two

$$\mathbb{P}_2 = \big\{ a_0 + a_1\, x + a_2\, y + a_2\, x^2 + a_4\, y^2 + a_5\, x\, y \ : \ a_0, \ldots, a_5 \in \mathbb{R} \big\}.$$

An element of $\mathbb{P}_2$ is determined by six independent parameters (the quantities a_i), that is, the space $\mathbb{P}_2$ has dimension equal to six. Let us take a triangle K and let us mark six points as **nodes**:

- the three vertices of the triangle,
- the midpoints of the three edges.

The following result is easy to prove: *a function in $\mathbb{P}_2$ is uniquely determined by its values on the six nodes of the triangle.* Take now two points $\mathbf{p}_1$ and $\mathbf{p}_2$. The function

$$[0,1] \ni t \longmapsto (1-t)\,\mathbf{p}_1 + t\,\mathbf{p}_2$$

parameterizes linearly the segment between these two points. If $p \in \mathbb{P}_2$, then a simple computation shows that

$$p((1-t)\mathbf{p}_1 + t\,\mathbf{p}_2) \in \mathbb{P}_2(t) = \big\{ b_0 + b_1\, t + b_2\, t^2 \ : \ b_0, b_1, b_2 \in \mathbb{R} \big\},$$

that is, seen on any segment (on any straight line actually), an element of $\mathbb{P}_2$ is a parabolic function, which, as everyone knows, is determined by three different points. Therefore *the value of a function in $\mathbb{P}_2$ on an edge of the triangle is uniquely determined by its three values on the nodes that lie on that edge* (two vertices and one midpoint).

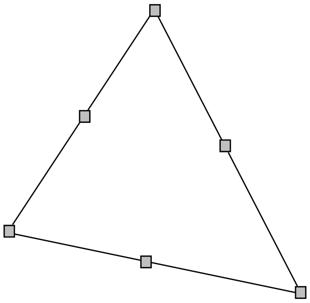


Figure 2.7: The nodes (local degrees of freedom) of a $\mathbb{P}_2$ triangle.

Because of this last property, we can glue together two $\mathbb{P}_2$ triangles as we did in the $\mathbb{P}_1$ case. Take a triangulation in the usual conditions, fix values of a function in all the nodes (vertices and midpoints) and on each triangle construct the only function in $\mathbb{P}_2$

that matches the given values. The resulting function is continuous. In fact it is a general element of the space

$$V_h = \big\{ u_h \in \mathcal{C}(\overline{\Omega}) \,:\, u_h|_K \in \mathbb{P}_2, \quad \forall K \in \mathcal{T}_h \big\}.$$

All the arguments presented in Lesson 1 hold also here. The dimension of this space is

$$\dim V_h = \#\{\text{vertices}\} + \#\{\text{edges}\},$$

since there is one midpoint per edge.

As before, we give a global numbering to the set of nodes (vertices and midpoints of edges): $\{\mathbf{p}_1, \ldots, \mathbf{p}_N\}$ and construct the functions $\varphi_i \in V_h$ satisfying

$$\varphi_i(\mathbf{p}_j) = \delta_{ij} = \left\{ \begin{array}{ll} 1, & j = i, \\ 0, & j \neq i. \end{array} \right.$$

For the same reasons as in the $\mathbb{P}_1$ case, these functions constitute a basis of V_h and any function of this space can be expressed as

$$u_h = \sum_{j=1}^N u_h(\mathbf{p}_j)\varphi_j.$$

There are two types of basis functions now:

- those associated to vertices, whose support is the set of triangles surrounding the vertex,
- those associated to midpoints, whose support is the set of two triangles (only one if the edge is on the boundary) that share the edge.

Take the usual triangulation and make yourself some drawing of the form of the supports of the nodal basis functions.

The concept of a Dirichlet node is the same: it is any node on a Dirichlet edge, Dirichlet edges being edges on the Dirichlet boundary Γ_D. The following result is then a straightforward consequence of the fact that value on edges is determined by degrees of freedom on edges:

$v_h \in V_h$ vanishes on Γ_D if and only if it vanishes on all Dirichlet nodes.

Therefore, it is very simple to construct a basis of

$$V_h^{\Gamma_D} = V_h \cap H^1_{\Gamma_D}(\Omega) = \{ v_h \in V_h \,:\, v_h = 0 \quad \text{on } \Gamma_D \}$$

by simply ignoring nodal basis functions φ_i associated to Dirichlet nodes. Can you notice that I am copy-pasting formulas from Lesson 1?

Very important. The whole of Section 3 in Lesson 1 can be read with these adapted concepts. There's nothing new at all, but the different concepts of local spaces and nodes. *You should have a detailed look again at that section* to convince yourself that this is so. In particular pay attention to mass and stiffness matrices and note that the number of adjacent nodes for each node is increased with respect to $\mathbb{P}_1$ triangles (we will explore this in an exercise). □

Bookkeeping for quadratic elements. Counting local and global degrees of freedom on quadratic elements gets us into a new world of (minor) difficulties. So far we had the lists of *vertices* of the triangulation, and the lists of *elements*. For quadratic elements, we need to number the edges of the triangulation. This is a list of what vertices of the triangulation are the vertices surrounding each of the edges. This list gives an automatic numbering of the midpoints of the edges, which are nodes in the $\mathbb{P}_2$ elements. We can then consider that the list of all nodes is built as follows:

- first all the vertices,
- then all the (midpoints of) the edges.

With this numbering, the degrees of freedom corresponding to the midpoints of the edges come at the end. We also have to relate the local and the global lists. This can be easily done with yet another list: we now produce the list of the three edges (global numbering) for each of the elements. (We can typically think that the first edge is the opposed to the first vertex, etc.) At the time of the assembly, the indices referred to vertices are taken from the list of elements, and the indices referred to midpoints are taken from the list of edges, adding the number of vertices, so that it is correlative.

3.2 The reference element

If we want to implement the $\mathbb{P}_2$ we need to compute the usual integrals for the mass and stiffness matrices (an also, of course, the right-hand sides, that include the influence of data). For that, we need a way of evaluating the nodal basis functions on each triangle.

Since the argument is, again, exactly the same as for the $\mathbb{P}_1$ element, let us work now in the opposite sense. In the reference triangle we mark the six nodes as shown in Figure 2.8. As usual (ξ,η) are the coordinates in the reference configuration.

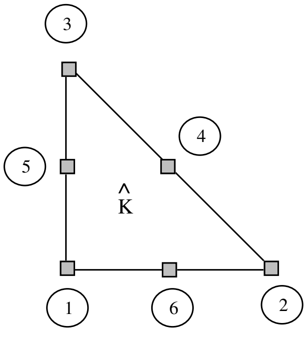


Figure 2.8: The $\mathbb{P}_2$ reference triangle.

Each of the functions

$$\widehat{N}_1 = (1-\xi-\eta)(1-2\xi-2\eta), \qquad \widehat{N}_2 = \xi(2\xi-1), \qquad \widehat{N}_3 = \eta(2\eta-1)$$

$$\widehat{N}_4 = 4\xi\eta, \qquad \widehat{N}_5 = 4\eta(1-\xi-\eta), \qquad \widehat{N}_6 = 4\xi(1-\xi-\eta)$$

takes the unit value on the corresponding node (the one numbered with the subindex) and vanishes in all other five nodes.

Let's do again some copy–pasting. The functions

$$N_\alpha^K = \widehat{N}_\alpha \circ F_K^{-1}, \qquad \alpha = 1, \ldots, 6$$

have the same property as the $\widehat{N}_\alpha$ functions, only on the triangle K, that is mapped from $\widehat{K}$ via the linear transformation F_K. These functions are polynomials of degree two (do you see why?) and therefore

$$N_\alpha^K = \varphi_{n_\alpha}, \qquad \text{in } K$$

where n_α is the global index corresponding to the local node α in K. Here is again the formula for the gradient

$$\nabla N_\alpha^K = \mathbf{B}_K^{-\top}\big((\widehat{\nabla}\widehat{N}_\alpha) \circ F_K^{-1}\big).$$

Note that now $\widehat{\nabla}\widehat{N}_\alpha$ is not constant, so the inverse transformation F_K^{-1} is needed also to evaluate the gradient.

We then compute the local matrices, which are 6×6 matrices,

$$\int_K \nabla N_\beta^K \cdot \nabla N_\alpha^K \qquad \text{and} \qquad \int_K N_\beta^K \, N_\alpha^K \,,$$

put the elements in the global positions

$$\int_K \nabla\varphi_{n_\beta} \cdot \nabla\varphi_{n_\beta} \qquad \text{and} \qquad \int_K \varphi_{n_\beta}\, \varphi_{n_\alpha}$$

and add the contributions of all triangles to assemble the full stiffness and mass matrices.

3.3 Convergence

The general error bound

$$\|u - u_h\|_{1,\Omega} \le (1 + \frac{M}{\alpha}) \inf \big\{ \|u - v_h\|_{1,\Omega} \,:\, v_h \in V_h,\ v_h(\mathbf{p}) = g_0(\mathbf{p})\ \forall \mathbf{p} \text{ Dirichlet node} \big\}.$$

still holds here. In the case of homogeneous Dirichlet conditions, we can use the same arguments as in the preceding section to obtain a full bound like

$$\|u - u_h\|_{1,\Omega} \le C h^2 |u|_{3,\Omega},$$

where:

- the constant C depends on the PDE operator, on the geometry and on the smallest angle (becoming worse as the triangles become flatter)

- the new Sobolev seminorm $|u|_{3,\Omega}$ uses the third order partial derivatives of u.

The result is valid only when this last seminorm is finite, which is much more to require than what we had at the beginning. Note that the **order two** in energy norm ($H^1(\Omega)$ norm) is good news, since using smaller triangles really pays off and the gain of precision is due to be much faster than in the $\mathbb{P}_1$ case. In the final exercise of this section we will explore what's the price to be paid (there's no free lunch, you know).

4 Cubic elements and static condensation

4.1 The $\mathbb{P}_3$ element

Can we do better than order two? The answer is yes, and besides, it is easy to do better. We will just give some hints on the order three case, because something new appears and we really want to deal with new ideas instead of doing the same thing over and over. Look first at Figure 2.9. There are ten nodes in the triangle:

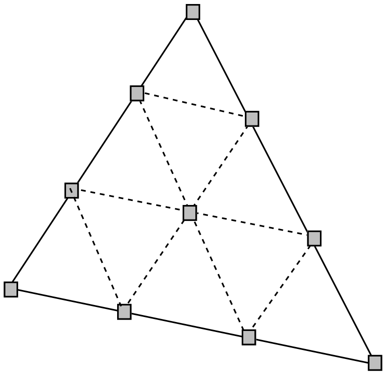


Figure 2.9: The $\mathbb{P}_3$ triangle

- the three vertices,
- two points per side, at relative distances 1/3 and 2/3 from the vertices,
- the barycenter, which is computed by averaging the coordinates of the three vertices

$$\tfrac{1}{3}\mathbf{v}_1^K + \tfrac{1}{3}\mathbf{v}_2^K + \tfrac{1}{3}\mathbf{v}_3^K.$$

Note also that each edge has four nodes on it. The local space is that of polynomials of degree up to three $\mathbb{P}_3$. Instead of writing a general element of this space, let us list the monomials that are used:

$$\begin{array}{llll} 1 & & & \\ x & y & & \\ x^2 & xy & y^2 & \\ x^3 & x^2y & xy^2 & y^3 \end{array}$$

Count them. Ten monomials (i.e., ten coefficients) and ten nodes. Well, that's a surprise! Two other surprises:

- *a function in $\mathbb{P}_3$ is uniquely determined by its values on the ten nodes of the triangle,*
- *the value of a function in $\mathbb{P}_3$ on an edge of the triangle is uniquely determined by its four values on the nodes that lie on that edge.*

Note that a $\mathbb{P}_3$ function restricted to a segment (straight line) is a cubic function of one variable.

We are almost done here. We can construct the spaces V_h, the nodal basis functions φ_i, the subspace $V_h^{\Gamma_D}$ by eliminating Dirichlet nodes, etc. The dimension of V_h is

$$\#\{\text{vertices}\} + 2\,\#\{\text{edges}\} + \#\{\text{triangles}\}.$$

4.2 Static condensation

There is however a new entity here, and that's the very isolated interior node. I say isolated because that node is only adjacent to the other nodes of the same triangle. This has some consequences at the practical level that we are going to explore right now.

Let φ_i be the nodal basis function associated to a node that is the barycenter of the triangle K. Then $\operatorname{supp}\varphi_i = K$. Therefore

$$a(\varphi_j, \varphi_i) = \int_K \nabla\varphi_j \cdot \nabla\varphi_i + c \int_K \varphi_j\,\varphi_i \qquad \forall j,$$

and

$$\ell(\varphi_i) = \int_K f\,\varphi_i$$

(do you see why there is no Neumann term here?) which means that once we have gone through the element K in the assembly process, we will have finished the i-th row of the system, with no contributions from other elements. The idea of **static condensation** is simple: get rid of that equation and unknown in the same process of assembly.

Let us consider that the 0-th node is the barycenter of K. Let $\mathbf{K}^K$ and $\mathbf{b}^K$ be the local matrix and right–hand side contributions from the triangle K

$$k^K_{\alpha\beta} = \int_K \nabla N^K_\beta \cdot \nabla N^K_\alpha + c \int_K N^K_\beta\, N^K_\alpha, \qquad b^K_\alpha = \int_K f\, N^K_\alpha, \qquad \alpha,\beta = 0,\ldots,9.$$

Now we decompose the matrix and the vector into blocks, separating the contribution from the interior node from all others:

$$\begin{bmatrix} \mathbf{K}^K_{00} & \mathbf{K}^K_{01} \\ \mathbf{K}^K_{10} & \mathbf{K}^K_{11} \end{bmatrix}, \qquad \begin{bmatrix} \mathbf{b}^K_0 \\ \mathbf{b}^K_1 \end{bmatrix},$$

with

$$\mathbf{K}^K_{00} = \begin{bmatrix} k^K_{0,0} \end{bmatrix}, \qquad \mathbf{K}^K_{01} = \begin{bmatrix} k^K_{0,1} & \ldots & k^K_{0,9} \end{bmatrix}, \qquad \mathbf{b}^K_0 = \begin{bmatrix} b^K_0 \end{bmatrix}$$

$$\mathbf{K}^K_{10} = \begin{bmatrix} k^K_{1,0} \\ \vdots \\ k^K_{9,0} \end{bmatrix}, \qquad \mathbf{K}^K_{11} = \begin{bmatrix} k^K_{1,1} & \ldots & k^K_{1,9} \\ \vdots & & \vdots \\ k^K_{9,1} & \ldots & k^K_{9,9} \end{bmatrix}, \qquad \mathbf{b}^K_1 = \begin{bmatrix} b^K_1 \\ \vdots \\ b^K_9 \end{bmatrix}.$$

You will be wondering why are we calling matrices to blocks 1×1 (scalars) and 1×9 or 9×1 (row or column vectors). The reason is twofold: first, the role these scalars and vectors are playing are the ones of blocks in a matrix so we'd better use block notation, independent of their shape; second, we will thus be able to recycle all that comes right now for more complicated situations.

The (ten) equations related to the nodes of the element K are

$$\begin{aligned} &\mathbf{K}_{00}^K \mathbf{u}_0^K + \mathbf{K}_{01}^K \mathbf{u}_1^K = \mathbf{b}_0^K, \\ &\mathbf{K}_{10}^K \mathbf{u}_0^K + (\mathbf{K}_{11}^K + \mathbf{A})\mathbf{u}_1^K + \mathbf{B}\mathbf{u}_{\text{other}} = \mathbf{b}_1^K + \mathbf{b}. \end{aligned}$$

The unknowns are separated in the same blocks (1 plus 9) and are denoted with local numbering, that is $\mathbf{u}_0^K$ is the unknown associate to the barycenter of K and $\mathbf{u}_1^K$ is the column vector of the nine unknowns associated to all the other nodes on K.

- The matrix $\mathbf{A}$ includes all contributions from other elements to nodes of K. It will be added in the assembly process when we go through these elements.
- The block $\mathbf{B}$ includes all contributions from other triangles to other unknowns (generically written as $\mathbf{u}_{\text{other}}$), that is, unknowns on nodes that are not on K but are adjacent to those on K.
- Finally, $\mathbf{b}$ includes all contributions from other triangles and possibly also from Neumann edges, to the right-hand side.

Now we can write $\mathbf{u}_0^K$ (which, in this case, is just the unknown corresponding to the barycenter of K) as

$$\mathbf{u}_0^K = \left(\mathbf{K}_{00}^K\right)^{-1} \mathbf{b}_0^K - \left(\mathbf{K}_{00}^K\right)^{-1} \mathbf{K}_{01}^K \mathbf{u}_1^K$$

and substitute this expression in the block of the remaining equations for the triangle K (the non–interior unknowns), obtaining

$$\left(\mathbf{K}_{11}^K - \mathbf{K}_{10}^K \left(\mathbf{K}_{00}^K\right)^{-1} \mathbf{K}_{01}^K + \mathbf{A}\right) \mathbf{u}_1^K + \mathbf{B}\mathbf{u}_{\text{other}} = \mathbf{b}_1^K - \mathbf{K}_{10}^K \left(\mathbf{K}_{00}^K\right)^{-1} \mathbf{b}_0^K + \mathbf{b}$$

This means that instead of assembling the full (10×10) block from K and its corresponding right–hand side, we can forget about the interior nodes (just one) on condition of assembling

$$\mathbf{K}_{\text{cond}}^K = \mathbf{K}_{11}^K - \mathbf{K}_{10}^K \left(\mathbf{K}_{00}^K\right)^{-1} \mathbf{K}_{01}^K, \qquad \mathbf{b}_{\text{cond}}^K = \mathbf{b}_1^K - \mathbf{K}_{10}^K \left(\mathbf{K}_{00}^K\right)^{-1} \mathbf{b}_0^K$$

instead of the original matrix. Once we have solved the system, the interior variables are solved using the local equations

$$\mathbf{K}_{00}^K \mathbf{u}_0^K + \mathbf{K}_{01}^K \mathbf{u}_1^K = \mathbf{b}_0^K,$$

that work element–by–element.

Remark. This is a method for implementing the $\mathbb{P}_3$ FEM in a way that the information of the interior nodes is incorporated to the assembly process directly without having to use the corresponding unknown. This doesn't mean that the node is not there. We only compute it separately after having added its contribution to assembly directly. So don't confuse this, which is nothing else than an implementation trick, with some finite elements (in the class of the so-called exotic or serendipity elements) that avoid interior nodes. □

Maybe I've left you wondering about that strange Algebra in the assembly process and it somehow rings a bell. It should. Write the extended matrix

$$\left[\begin{array}{cc|c} \mathbf{K}_{00}^K & \mathbf{K}_{01}^K & \mathbf{b}_0^K \\ \mathbf{K}_{10}^K & \mathbf{K}_{11}^K & \mathbf{b}_1^K \end{array}\right]$$

and apply Gaussian block elimination (the $\mathbf{K}_{00}^K$ block is just 1×1, so this is just Gauss elimination) you obtain

$$\left[\begin{array}{cc|c} \mathbf{K}_{00}^K & \mathbf{K}_{01}^K & \mathbf{b}_0^K \\ \mathbf{0} & \mathbf{K}_{11}^K - \mathbf{K}_{10}^K \left(\mathbf{K}_{00}^K\right)^{-1} \mathbf{K}_{01}^K & \mathbf{b}_1^K - \mathbf{K}_{10}^K \left(\mathbf{K}_{00}^K\right)^{-1} \mathbf{b}_0^K \end{array}\right].$$

Ta daaaa! There they are. The blocks you wanted. Again, our diagonal block was a scalar, so this was easy. What would have happened if it was a matrix? Do you have to compute that inverse and apply all that Algebra? No, you don't. Gauss block elimination is a nice way of writing the result of Gauss elimination. The point is you apply row elimination to create all those zeros, with no row changes and without trying to create any other zeros. Blocks of the form

$$\mathbf{K}_{11}^K - \mathbf{K}_{10}^K \left(\mathbf{K}_{00}^K\right)^{-1} \mathbf{K}_{01}^K$$

are called Schur complements. If the original matrix is symmetric and positive definite, they are still symmetric and positive definite.

4.3 Convergence, $\mathbb{P}_4$ and higher

We haven't mentioned convergence of the $\mathbb{P}_3$ method yet. In the best possible conditions, this is a method of order three in the $H^1(\Omega)$ Sobolev norm:

$$\|u - u_h\|_{1,\Omega} \le C h^3 |u|_{4,\Omega}$$

(can you guess what's in $|u|_{4,\Omega}$?). These best possible conditions include the fact that triangles do not become too flat, since the constant C becomes worse and worse as triangles get flatter and flatter. Note that if you apply static condensation to the $\mathbb{P}_3$ you complicate the assembly process but you end up with a system of order

$$\#\{\text{vertices}\} + 2\,\#\{\text{edges}\}$$

(minus the number of Dirichlet nodes), which is smaller than the one you obtain without condensation. There is an additional advantage of applying condensation. With the usual information of a grid generator (you will have to read the Appendix for that) you can easily construct a coherent numbering including vertices and edges, which works for $\mathbb{P}_2$ elements. Going from $\mathbb{P}_2$ to $\mathbb{P}_3$ means that you have to double the number of unknowns per edge (which is easy) and add the triangles. The numbering of triangles becomes then relevant. It is not, I insist, for the assembly process. If you apply static condensation, you avoid the unknowns related to barycenter and the numbering of vertices-and-edges is enough for the $\mathbb{P}_3$ element.

The $\mathbb{P}_4$ element is constructed easily following these lines:

- You divide each edge into five equally sized pieces. Then you join these new points on different sides with lines that run parallel to the edges. With that you have created a grid of 15 nodes: three vertices, three points per edge, three interior points, placed on the intersections of the interior lines.
- The space is $\mathbb{P}_4$, which has dimension 15. Everything goes on as usual.
- The three interior nodes can be treated with static condensation: the $\mathbf{K}^K_{00}$ blocks are now 3×3 blocks. With this you reduce in three times the number of triangles the size of the global system to be solved without affecting convergence.
- Order of the method is.... four! (That was easy)

It is possible to create $\mathbb{P}_k$ methods for arbitrary k. You will find people around that will assert that these methods are useless or just of theoretical interest. Be warned: maybe they find them useless, but some other people work with really high order methods and find many advantages in them[8]. However, if you go from $\mathbb{P}_4$ upwards, you implement the method in a very different way. Nodal bases are not the best choice in that case and there is a different way of constructing node-free bases. We will deal with this in Lesson 7.

5 Exercises

1. **Basis functions for the $\mathbb{P}_2$ element.** Try to sketch the form of the nodal basis functions for a $\mathbb{P}_2$ finite element space (similar as Figure 1.8). Note that there are two different types of functions, those associated to vertices and those associated to midpoints of edges.

2. **The plane elasticity system.** The problem of plane deformations in linear elasticity can be reduced to the variational problem:[9]

$$\left[\begin{array}{l} \text{find } u_1, u_2 \in H^1(\Omega) \text{ such that} \\ u_1 = g_x, \quad u_2 = g_y \qquad \text{on } \Gamma_D, \\ \displaystyle\int_\Omega \Big((\lambda+2\mu)\frac{\partial u_1}{\partial x} + \lambda\frac{\partial u_2}{\partial y}\Big)\frac{\partial v}{\partial x} + \mu\Big(\frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x}\Big)\frac{\partial v}{\partial y} = \int_\Omega v\, f_x + \int_{\Gamma_N} v\, t_x \quad \forall v \in H^1_{\Gamma_D}(\Omega), \\ \displaystyle\int_\Omega \mu\Big(\frac{\partial u_1}{\partial y} + \frac{\partial u_2}{\partial x}\Big)\frac{\partial v}{\partial x} + \Big(\lambda\frac{\partial u_1}{\partial x} + (\lambda+2\mu)\frac{\partial u_2}{\partial y}\Big)\frac{\partial v}{\partial y} = \int_\Omega v\, f_y + \int_{\Gamma_N} v\, t_y \quad \forall v \in H^1_{\Gamma_D}(\Omega), \end{array}\right.$$

where:

[8] Be always prepared to find opinionated people in the scientific computing community. Sometimes they are right, sometimes they are partially right, sometimes they are plain wrong.

[9] **Warning.** For reasons that are not so easy to explain as many people think, $\mathbb{P}_1$ elements are never used in elasticity problems because their performance is rather bad. Note that in what you have done here $\mathbb{P}_1$ or $\mathbb{P}_k$ is all the same, so you can be applying this to $\mathbb{P}_2$ elements, which work well for this problem.

- Ω is the plane section of the cylindrical solid
- Γ_D is the part of the boundary of Ω where we know displacements $g_0 = (g_x, g_y)$
- Γ_N is the part of the boundary where we know normal stresses $t = (t_x, t_y)$
- $f = (f_x, f_y)$ are the volume forces
- λ and $\mu = G$ are the Lamé parameters

$$\lambda = \frac{\nu E}{(1+\nu)(1-2\nu)}, \qquad \mu = \frac{E}{2(1+\nu)}$$

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

We are given a triangulation $\mathcal{T}_h$, the associated $\mathbb{P}_1$ nodal basis functions (φ_i), etc. We call Ind and Dir to the usual index sets. We approximate the pair (u_1, u_2) by the discrete functions

$$u_h^1 = \sum_j u_j^1 \varphi_j, \qquad u_h^2 = \sum_j u_j^2 \varphi_j$$

Alternating tests with the two variational equations, and grouping both unknowns on the same node (u_j^1, u_j^2) prove that the resulting finite element system can be written in the form

$$\sum_{j\in\mathrm{Ind}} A_{ij} \begin{bmatrix} u_j^1 \\ u_j^2 \end{bmatrix} = F_i + T_i - \sum_{j\in\mathrm{Dir}} A_{ij} \begin{bmatrix} g_j^1 \\ g_j^2 \end{bmatrix}, \qquad i \in \mathrm{Ind}\,.$$

where A_{ij} are 2×2 matrices. What's the dimension of the system? Prove that $A_{ij}^\top = A_{ji}$ and deduce that the system is symmetric.

3. **Comparison of $\mathbb{P}_1$ and $\mathbb{P}_2$.** Consider the simple triangulation depicted in Figure 2.10

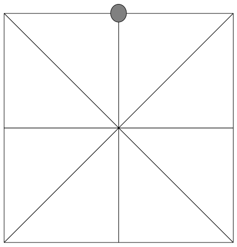


Figure 2.10: A simple triangulation with a marked node

(a) If you consider the Neumann problem there (no Dirichlet nodes), how many unknowns are there in the system corresponding to the $\mathbb{P}_2$ method?

(b) What are the adjacent nodes to the node that is marked on the figure?

(c) A red refinement of a triangle consists of taking the midpoints of the edges and joining them to create four triangles per triangle (see Figure 2.11). If you apply a red refinement to all the elements of the triangulation above and the apply the $\mathbb{P}_1$ element, how many unknowns do you have in the system? Which nodes are adjacent to the same marked nodes in this new triangulation for the $\mathbb{P}_1$ method?

(d) **Discussion.** The error of the $\mathbb{P}_2$ method is bounded by something times h^2. The error of the $\mathbb{P}_1$ method on the uniform red refinement is something else times $h/2$. The constant (the unspecified something) for each case is different. In principle, when the triangulation is fine enough h^2 wins over $h/2$ (it is smaller). With the same number of unknowns one method is better than the other. Where's the difference?

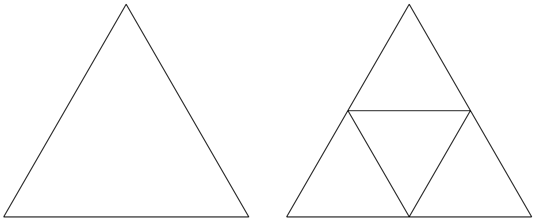


Figure 2.11: A red refinement of a triangle

4. **Bookkeeping for $\mathbb{P}_2$ elements.** Consider the triangulation given in Figure 2.10.

(a) Number vertices, edges, and elements. Construct all the lists that you would need to use $\mathbb{P}_2$ FEM:

- the list of vertices for each element,
- the list of vertices for each edge,
- the list of edges for each element.

For the list of vertices for each edge, *choose always a positive orientation for boundary edges.* This means that if you go from the first edge to the second, you are leaving the exterior domain to the right.

(b) Additionally, build a matrix (list) with the same shape as the one of edges-counted-by-element where you specify if the orientation of the edge: positive or negative. This means the following. If the first edge of an element connects the nodes n_1 and n_2 in the element (counting counterclockwise) and the edge is listed as n_1 going to n_2 you assign a + sign. If the edge is listed as n_2 going to n_1, you assign a minus sign.

(c) Using the above information, choose one element, and say where you would assemble a local mass matrix in the global matrix.

Lesson 3

New classes of elements

1 The lowest order element on parallelograms

Sometimes dividing a domain into triangles is not the best idea. Some domains, such as rectangles, are much better subdivided into smaller rectangles. Also sometimes triangulations become really messy. For instance Figure 3.1 shows a typical triangular grid of a rectangle as produced by the PDE Toolbox of Matlab. Because of the way these triangulations are produced, working from the boundary to the interior and avoiding very acute angles, they display a very disorganized and non-symmetric pattern. If your problem favors directions, maybe this is not the best way to begin your discretization.

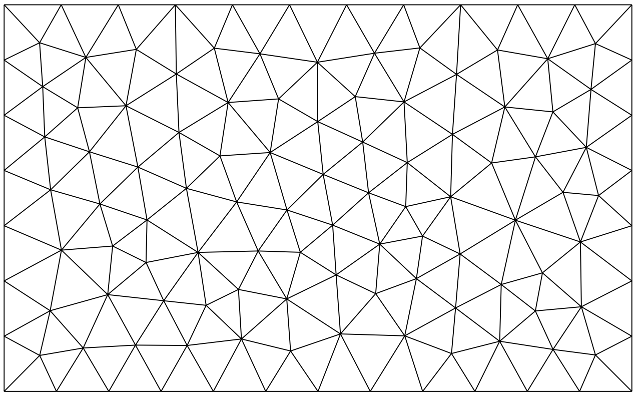


Figure 3.1: A typical triangular grid of a rectangle.

We are going to introduce here finite elements on rectangles and parallelograms. These elements share many features with finite elements on triangles, but there are plenty of novelties. To learn about finite elements on arbitrary quadrilaterals (trapezes and trapezoids) you will have to wait to Lesson 4. They constitute a different species and have to be studied later on to grasp their difficulties.

1.1 The reference space

First of all we need a new polynomial space, which we are going to introduce in reference variables,

$$\mathbb{Q}_1 = \big\{ a_0 + a_1\xi + a_2\eta + a_3\xi\,\eta \;:\; a_0, a_1, a_2, a_3 \in \mathbb{R} \big\}.$$

These are polynomials in two variables that are of degree at most one in each variable separately. Note that this space contains $\mathbb{P}_1$. The reference square $\widehat{K}$ is going to be the one with vertices on

$$\widehat{\mathbf{p}}_1 = (-1,-1), \qquad \widehat{\mathbf{p}}_2 = (1,-1), \qquad \widehat{\mathbf{p}}_3 = (1,1), \qquad \widehat{\mathbf{p}}_4 = (-1,1),$$

that is $\widehat{K} = [-1,1] \times [-1,1]$. Note that many books prefer to take the unit square $[0,1] \times [0,1]$ as reference element. Some details change if this choice is made[1]. This is not so important. Note that I have chosen to number the vertices in rotating order. Whether we do this in this way (rotating clockwise or counter-clockwise is immaterial) or in a different way is relevant and we have to be very careful with this. Unlike what happens with triangles, here we really have to know what points are vertices of each edge and we need to fix an order to say that.

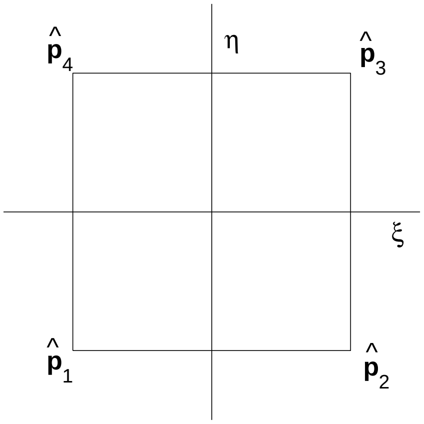


Figure 3.2: The reference square.

Restricted to a horizontal line (η constant) or to a vertical line (ξ constant), functions of $\mathbb{Q}_1$ are polynomials of degree one, that is, seen on horizontal or vertical lines, functions of $\mathbb{Q}_1$ are linear functions. They are however not linear functions (flat plane functions), because of the crossed product $\xi\,\eta$.

Two simple observations, in the line of what we have been doing for triangles:

- the value of an element of $\mathbb{Q}_1$ is uniquely determined by its values on the four vertices of $\widehat{K}$,
- the value of an element of $\mathbb{Q}_1$ on each of the four sides of $\widehat{K}$ is uniquely determined by the value on the extreme points of that side.

[1] In a way, I'm mixing choices in this course, because I chose the unit reference triangle in a form and the reference square in another form.

As usual, we can construct functions $\widehat{N}_\alpha \in \mathbb{Q}_1$ such that

$$\widehat{N}_\alpha(\widehat{\mathbf{p}}_\beta) = \delta_{\alpha\beta}, \qquad \alpha, \beta = 1, \ldots, 4.$$

These functions are

$$\widehat{N}_1 = \tfrac{1}{4}(1-\xi)(1-\eta), \qquad \widehat{N}_2 = \tfrac{1}{4}(1+\xi)(1-\eta),$$

$$\widehat{N}_3 = \tfrac{1}{4}(1+\xi)(1+\eta), \qquad \widehat{N}_4 = \tfrac{1}{4}(1-\xi)(1+\eta).$$

The nice joint formula $\frac{1}{4}(1\pm\xi)(1\pm\eta)$ for the whole set justifies the choice of this reference square over $[0,1] \times [0,1]$.

1.2 The local spaces

Take now a parallelogram K and write its four vertices in rotating order (clockwise or counter-clockwise, it doesn't matter)

$$\mathbf{p}_\alpha^K = (x_\alpha, y_\alpha), \qquad \alpha = 1, \ldots, 4.$$

Consider now a linear map that transforms $\widehat{K}$ into K. For instance, this one does the job:

$$\left[\begin{array}{c} x \\ y \end{array}\right] = -\frac{1}{2}(\xi+\eta)\left[\begin{array}{c} x_1 \\ y_1 \end{array}\right] + \frac{1}{2}(1+\xi)\left[\begin{array}{c} x_2 \\ y_2 \end{array}\right] + \frac{1}{2}(1+\eta)\left[\begin{array}{c} x_4 \\ y_4 \end{array}\right].$$

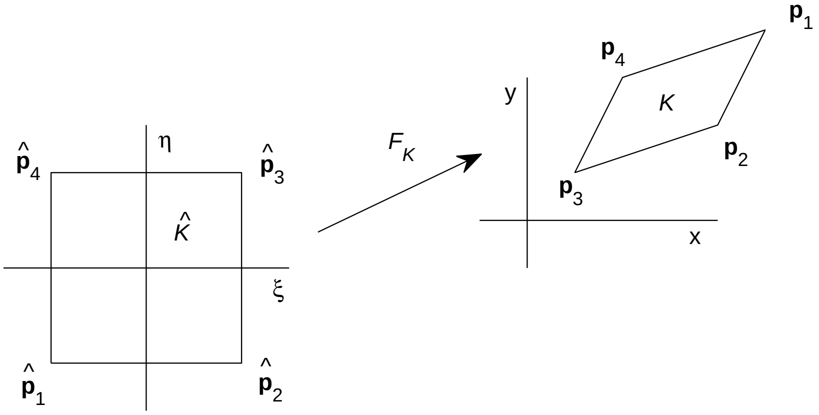


Figure 3.3: The reference square $\widehat{K}$ is mapped to the physical domain K. Note that vertices are notated in rotating order, even if the sense is different.

If we had taken before the reference triangle with vertices on $\widehat{\mathbf{p}}_1$, $\widehat{\mathbf{p}}_2$ and $\widehat{\mathbf{p}}_4$, the $\mathbb{P}_1$ basis functions we would had found would have been

$$-\tfrac{1}{2}(\xi+\eta), \qquad \tfrac{1}{2}(1+\xi), \qquad \tfrac{1}{2}(1+\eta).$$

What we are doing is mapping this triangle into the triangle with vertices $\mathbf{p}_1$, $\mathbf{p}_2$ and $\mathbf{p}_4$. The additional point $\widehat{\mathbf{p}}_3$ is mapped automatically to $\mathbf{p}_3$, because K is a parallelogram and linear transformations preserve parallelism. Let's not worry about the explicit formula for the transformation. We'll call it $F_K : \widehat{K} \to K$ and write simply

$$\begin{bmatrix} x \\ y \end{bmatrix} = \mathbf{B}_K \begin{bmatrix} \xi \\ \eta \end{bmatrix} + \mathbf{b}_K.$$

or $(x, y) = F_K(\xi, \eta)$. We finally get to the local polynomial space

$$\begin{aligned} \mathbb{Q}_1(K) &= \{q : K \to \mathbb{R} \, : \, q \circ F_K \in \mathbb{Q}_1\} \\ &= \{\widehat{q} \circ F_K^{-1} \, : \, \widehat{q} \in \mathbb{Q}_1\}. \end{aligned}$$

Note that the space is defined by transforming (pushing forward) the space $\mathbb{Q}_1$ on the reference element to the physical element K. In a way, that happened also with the $\mathbb{P}_k$, only with the simplicity that in that case

$$\mathbb{P}_k(K) = \mathbb{P}_k$$

and the space in physical and reference variables was the same.

Before giving properties of $\mathbb{Q}_1(K)$ (we need concepts like local degrees of freedom, the possibility of gluing together different elements, et cetera), let's have a look at the functions in this space. A function in $\mathbb{Q}_1$ is of the form

$$\widehat{q} = a_0 + a_1 \, \xi + a_2 \, \eta + a_3 \, \xi \, \eta.$$

The reference variables can be written in terms of the physical variables by inverting the transformation F_K. We obtain something of the form:

$$\begin{bmatrix} \xi \\ \eta \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} e \\ f \end{bmatrix} = \begin{bmatrix} a\,x + b\,y + e \\ c\,x + d\,y + f \end{bmatrix}$$

(the actual numbers $a, \ldots, f$ are not important). Therefore

$$\begin{aligned} \widehat{q} \circ F_K^{-1} &= a_0 + a_1(a\,x + b\,y + e) + a_2(c\,x + d\,y + f) + a_3(a\,x + b\,y + e)(c\,x + d\,y + f) \\ &= b_0 + b_1\,x + b_2\,y + a_3(a\,c\,x^2 + b\,d\,y^2 + (a\,d + b\,c)\,x\,y), \end{aligned}$$

which means that functions in $\mathbb{Q}_1(K)$ have a linear part plus a term of degree two that depends on the element K (see how the coefficients of F_K^{-1} are there). Actually, it looks like the space depends on the transformation chosen to map K from $\widehat{K}$, but that's not so. The following list of facts is of easy verification:

- The space $\mathbb{Q}_1(K)$ depends only on K, not on the concrete transformation F_K we have chosen. This is an important property, that means that we have not to worry about the way in which we ordered the vertices of the parallelogram K. We only need a list in rotating order and nothing else.

- If K is a rectangle with sides parallel to the cartesian axes (and in fact, only in this case), the space $\mathbb{Q}_1(K)$ is simply $\mathbb{Q}_1$.

- In all cases
$$\mathbb{P}_1 \subset \mathbb{Q}_1(K) \subset \mathbb{P}_2,$$
so $\mathbb{Q}_1(K)$ contains all polynomials of degree at most one and is a space of polynomials of degree at most two. The first part of this property is what will give order of convergence to the finite element method using this space.

- The space $\mathbb{Q}_1(K)$ has dimension four. The functions
$$N_\alpha^K = \widehat{N}_\alpha \circ F_K^{-1}, \qquad \alpha = 1, \ldots, 4$$
form a basis of this space. In fact
$$N_\alpha^K(\mathbf{p}_\beta^K) = \delta_{\alpha\beta}, \qquad \alpha, \beta = 1, \ldots, 4.$$

- Restricted to any of the sides of K, a function of $\mathbb{Q}_1(K)$ is a linear function of one variable, that is, if $\mathbf{p}_i$ and $\mathbf{p}_{i+1}$ are two consecutive vertices of K and $q \in \mathbb{Q}_1(K)$, then
$$t \ni [0,1] \longmapsto q((1-t)\mathbf{p}_i + t\,\mathbf{p}_{i+1}) \in \mathbb{P}_1(t) = \{a + b\,t \,:\, a, b \in \mathbb{R}\}.$$

From the last two bullet points of this list we easily recover the needed properties to construct finite element spaces. First

> *a function of $\mathbb{Q}_1(K)$ is uniquely determined by its values on the four vertices of K*

and

> *the form a function of $\mathbb{Q}_1(K)$ restricted to an edge is independent of the shape of K and is uniquely determined by its values on the two vertices of this side.*

You might have noticed that the second property looks longer than usual. It has to be like that. What we assert there is not only that the function on an edge (side) depends only on the values on the two vertices that lie on that edge, but also that the type of function itself does not depend on where the other two vertices are. Restricted to one of the sides we always have a linear function.

1.3 The $\mathbb{Q}_1$ finite element method

We are done locally. Now we have to divide the domain into parallelograms and glue the local spaces together. Note here the fact that not all domains can be decomposed into parallelograms, but that we are speaking of something else than rectangles and similar domains.

A partition of a domain in parallelograms (we will call **elements** to the parallelograms) has also to respect the rules we gave to partitions with triangles:

- two different elements can meet only on a common vertex or a full edge of both, and
- the partition has to respect the division of the boundary in Dirichlet and Neumann parts.

Recall that the last property is used only when there is a transition point from Γ_D to Γ_N somewhere inside a side of Ω. The properties are exactly the same as those demanded to triangulations. In fact, there is a tradition to calling simply elements to the constitutive figures (triangles or parallelograms) and triangulation to the partition, even if it is a 'parallelogramization' (that's an ugly word!), a tradition we are going to honor.

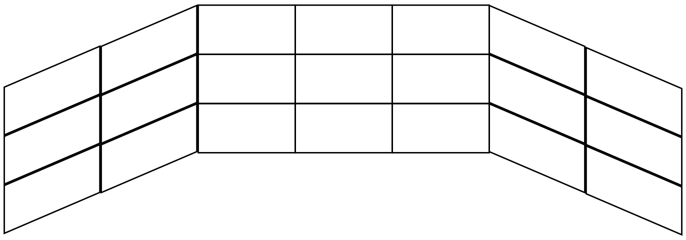


Figure 3.4: A 'triangulation' made of parallelograms.

If we now fix values on the nodes (and now **nodes** are vertices of the elements again), we can construct a unique function on each K such that it belongs to the local space $\mathbb{Q}_1(K)$ and matches the values on the vertices. Because of the second local property, the functions on the edges do not depend really on the particular space $\mathbb{Q}_1(K)$ (i.e., on the shape of the element). They are always linear and fixed by the values on the corresponding two vertices. Therefore, what we obtain is a globally continuous function, an element of

$$V_h = \big\{u_h \in \mathcal{C}(\overline{\Omega}) \,:\, u_h|_K \in \mathbb{Q}_1(K), \quad \forall K \in \mathcal{T}_h\big\}.$$

We have done this enough times so that you already now what would come here if we just bothered to rewrite all the details:

- First, we note that the space

  $$V_h^{\Gamma_D} = V_h \cap H^1_{\Gamma_D}(\Omega) = \{v_h \in V_h \,:\, v_h = 0 \quad \text{on } \Gamma_D\}$$

  is the same as the space of elements of V_h that are zero on Dirichlet nodes.

- We then number nodes and define the nodal basis functions, to obtain a basis of V_h with functions that have small support (four elements at most in a triangulation like the one of Figure 3.4, although there could be more with some special displays of parallelograms). Ignoring functions related to Dirichlet nodes we obtain a basis of $V_h^{\Gamma_D}$.

- We go on and copy the whole of Section 3 in Lesson 1. We have a finite element method, a linear system, mass and stiffness matrices,...

- In the assembly process we notice that, restricted to an element K, a nodal basis function φ_i is either the zero function or one of the four N_α^K. Computing local 4×4 matrices and assembling them in the usual fashion, we can construct effectively the matrix of the system. The same thing applies to the right-hand side. Whenever we want to evaluate N_α^K or its gradient, we have the usual formulas. Note that in this case, the gradients are not constant.

Are we there yet? Almost. We were forgetting about the order. The process is the same one. For the homogeneous Dirichlet problem we obtain

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

The constant depends on the domain Ω (as well as on the division of Γ into Dirichlet and Neumann parts) and also on a parameter that measures the maximum flatness of elements.

Unlike in the case of triangles, flatness of elements is not given by extreme acuteness of angles, but can happen with elongated rectangles. Now, the measurement of flatness has to be the ratio between the maximum distance between points of an element and the radius of the largest circumference we can inscribe in K. This measurement of flatness (some people call it chunkiness) is also valid for triangles and is the one that is given usually in textbooks. As a general rule, in this type of error analysis, elements cannot become too flat.

You will notice that, at least in theory, performance of $\mathbb{P}_1$ and $\mathbb{Q}_1$ elements seems to be very similar. Linear elements on triangles can be adapted to more complicated geometries and make assembly a bit simpler. In some cases (in particular in many important test cases in mechanics) using rectangles as elements reflects better the inherent geometrical features of the problem and it is advised to do so. In a forthcoming exercise we will observe that $\mathbb{Q}_1$ elements are just a bit stiffer (more rigid) than $\mathbb{P}_1$ elements.

1.4 Combination of $\mathbb{P}_1$ and $\mathbb{Q}_1$ elements

You might be thinking... if I cannot divide (triangulate) my polygon Ω with parallelograms, am I completely done with the whole $\mathbb{Q}_1$ stuff? Is that it? First of all, let me mention that you will have to wait to the next lesson to see how to construct elements on general quadrilaterals, elements that are, by the way, much more complicated to use. Anyhow, there's another possibility, which I personally find one of the simplest proofs of the great versatility of finite element methods and of the great idea that assembly is. Wait for it.

Let me recall something we just saw. In the global finite element space for the $\mathbb{Q}_1$ method

$$V_h = \big\{ u_h \in \mathcal{C}(\overline{\Omega}) \,:\, u_h|_K \in \mathbb{Q}_1(K), \quad \forall K \in \mathcal{T}_h \big\},$$

the local space depends on the particular element. You could think that this makes the method complicated. What is complicated is the explanation of the method. The assembly process does not see this difference of local space since it sends evaluations of the local basis functions to the reference domain.

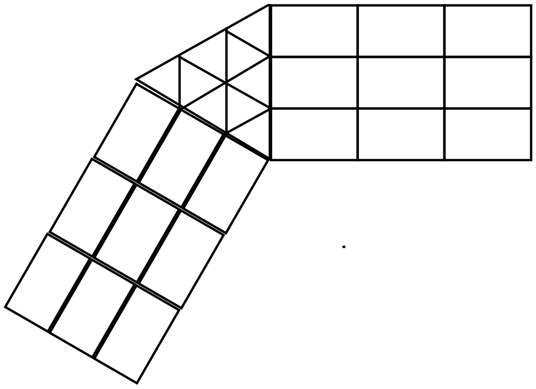


Figure 3.5: A triangulation made of triangles and rectangles.

We can think of domains that admit a triangulation made of triangles and rectangles, such as the one of Figure 3.5. The rectangular pieces of the domain are perfect for a division in rectangles, but the connecting piece seems to demand the use of triangles, so we decide to use both.

An element of this new type of triangulation can be either a triangle or a parallelogram. The triangulation has to fulfill the usual requirements: intersections can only happen in common vertices or common edges and Dirichlet-Neumann partition of the boundary has to be respected. The local space will depend on the type of element:

$$\mathbb{P}(K) = \left[\begin{array}{ll} \mathbb{P}_1, & \text{if } K \text{ is a triangle,} \\ \mathbb{Q}_1(K), & \text{if } K \text{ is a parallelogram.} \end{array} \right.$$

Nodes are vertices, as usual, and the global space

$$V_h = \big\{ u_h \in \mathcal{C}(\overline{\Omega}) \,:\, u_h|_K \in \mathbb{P}(K), \quad \forall K \in \mathcal{T}_h \big\}$$

is easily defined by gluing elements because of the following facts:

> *a function of $\mathbb{P}(K)$ is uniquely determined by its values on the nodes (three or four) of K*

and

> *the form a function of $\mathbb{P}(K)$ restricted to an edge is independent of the type of K and is uniquely determined by its values on the two vertices of this side.*

Seen on edges, all these discrete functions are linear, so we can glue triangles with parallelograms of any shape, as we were able to glue different parallelograms together.

Other than this, life goes on as usual. In the process of assembly is where we use whether an element is a parallelogram or a rectangle: the reference domain is different depending on which and local matrices have different sizes (3 × 3 for triangles, 4 × 4 for parallelograms). This looks more complicated but you have to think in terms of the grid generator. If it gives you triangles and rectangles, either they are given in a different list

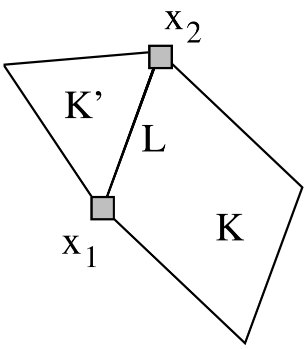


Figure 3.6: A triangle and a parallelogram sharing an edge.

(and you assemble first ones and then the other) or it gives you information about the type of geometry of each element, which you obviously learn by only looking at the list of vertices of the element.

What about error? Let's stick again to the case of homogeneous Dirichlet condition. Céa's lemma still applies

$$\|u-u_h\| \le \frac{M}{\alpha} \inf \left\{ \|u-v_h\| \,:\, v_h \in V_h^0 \right\}$$

and the right-hand side is an approximation error which can be bounded locally, element by element. Hence, the error of the method can be bounded by the error of approximating with linear functions on triangles and with $\mathbb{Q}_1(K)$ functions on parallelograms. In both cases, we have an h-type error. Order one.

2 Higher order methods on parallelograms

Once here, we should make a fast review of the novelties of introducing the $\mathbb{Q}_1$ method. Note that it took its time, compared to how simple it was to introduce the $\mathbb{P}_2$ elements once we had done everything on the $\mathbb{P}_1$ very clear. The novelties were: (a) there is a new reference element and therefore a new concept of triangulation, plus (b) the local space depends on the particular element, but (c) restricted to edges the local spaces do not depend on the element. That was basically all of it. Let us go for higher order.

2.1 The $\mathbb{Q}_2$ elements

The space $\mathbb{Q}_2$ uses all linear combinations of these monomials

$$1, \quad \xi, \quad \eta, \quad \xi^2, \quad \xi\eta, \quad \eta^2, \quad \xi^2\eta, \quad \xi\eta^2, \quad \xi^2\eta^2.$$

There is nine of them. (We will need nine nodes in the reference domain.) Looking carefully you'll see that $\mathbb{Q}_2$ is the space of polynomials in the variables ξ and η that have degree at most two in each variable separately. It includes therefore all polynomials of degree two but goes up to polynomials of degree four.

There's a nice table that will simplify your life in remembering these spaces. It serves to compare $\mathbb{P}_2$ (the order two space for triangles) with $\mathbb{Q}_2$ (the order two space for squares)

$$\begin{array}{lll} \eta^2 & & \\ \eta & \xi\eta & \\ 1 & \xi & \xi^2 \end{array} \qquad\qquad \begin{array}{lll} \eta^2 & \xi\eta^2 & \xi^2\eta^2 \\ \eta & \xi\eta & \xi^2\eta \\ 1 & \xi & \xi^2 \end{array}$$

(You will have to recognize that that's clever.) We now consider nine points (nodes) on the reference square:

- the four vertices,

- the midpoints of the four sides,

- the center (the origin).

The two relevant properties here are the usual ones: (a) a function of $\mathbb{Q}_2$ is uniquely determined by its values on the nine nodes; (b) restricted to one of the sides of $\widehat{K}$, a function of $\mathbb{Q}_2$ is a polynomial of degree two in one variable and is therefore determined by its values on the three nodes that lie on the edge.

Note that if you use a linear map F_K to transform $\widehat{K}$ into the parallelogram K, midpoints of edges are mapped onto midpoints of edges and the center is mapped onto the 'center' of the parallelogram (the point where both diagonals meet, or where the lines joining midpoints of opposite sides meet). See Figure 3.7 for a sketch of this.

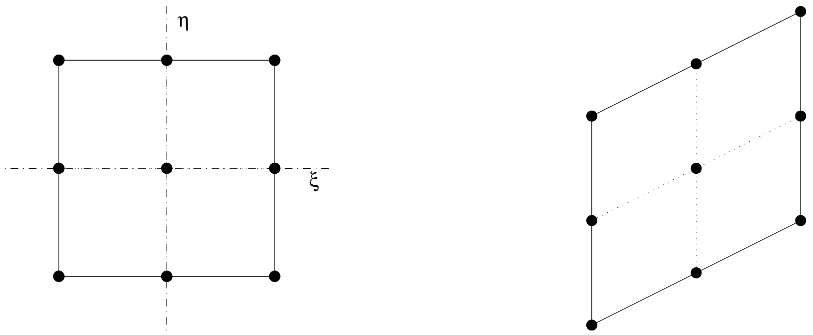


Figure 3.7: The $\mathbb{Q}_2$ reference element on the left. A parallelogram with the $\mathbb{Q}_2(K)$ nodes marked on the right.

We then create the local space

$$\mathbb{Q}_2(K) = \{q : K \to \mathbb{R} \,:\, q \circ F_K \in \mathbb{Q}_2\} = \{\widehat{q} \circ F_K^{-1} \,:\, \widehat{q} \in \mathbb{Q}_2\},$$

and observe again that:

- this 9-dimensional space depends on K but not on the particular transformation,

- $\mathbb{Q}_2(K) = \mathbb{Q}_2$ when K is a rectangle in the horizontal-vertical direction, but in general
$$\mathbb{P}_2 \subset \mathbb{Q}_2(K) \subset \mathbb{P}_4,$$
- we can construct nine nodal basis functions on $\widehat{K}$, $\{\widehat{N}_\alpha : \alpha = 1, \ldots, 9\}$ and transform them to
$$N_\alpha^K = \widehat{N}_\alpha \circ F_K^{-1}$$
to obtain a nodal basis of $\mathbb{Q}_1(K)$,
- the two nodal properties that hold on the reference square still hold on K; in particular, restriction of an element of $\mathbb{Q}_1(K)$ to one of the four sides is a polynomial of degree at most two, and is independent of the shape of K.

From here on, everything is just a copy of the $\mathbb{Q}_1$ case: global spaces, global nodal basis functions, restriction of those to elements giving the local nodal functions, Dirichlet nodes, etc. Note that the interior node is just that: interior. Therefore you can apply static condensation to this node. In $\mathbb{P}_k$ we had to wait to $k = 3$ to obtain an interior node.

The fact that the polynomial degree increases is something you cannot take too lightly. For instance, when computing the local mass matrices
$$\int_K N_\beta^K \, N_\alpha^K,$$
you have to compute an integral of a polynomial of degree eight.

Order of the method (in the best possible conditions) is two in the $H^1(\Omega)$ Sobolev norm. The method can be simultaneously used with $\mathbb{P}_2$ elements over triangles for triangulations that combine parallelograms and triangles. Can you see? That was really fast.

3 Three dimensional domains

3.1 Elements on tetrahedra

What we did at the beginning of Lesson 1 about formulating a boundary value problem in a weak form can be easily done for three dimensional domains Ω. Integrals over Ω become volume integrals. Integrals over Γ are now surface integrals. If Ω is a polyhedral domain, it is possible (although not easy) to divide it into tetrahedra. Triangulations with tetrahedra[2] have to follow the usual rules: (a) two different elements can intersect only on a common vertex, a common edge or a common face; (b) the Dirichlet/Neumann division of the boundary has to be respected by the discretized geometry. This last point is much trickier than before. If each face of the boundary of Ω is entirely included either on the Dirichlet boundary or on the Neumann boundary, everything is simple and condition (b)

[2] We are in three dimensions, but we keep on calling these things triangulations. For these ones some people prefer the very specific neololgism tetrahedrizations.

reduces to nothing. When there are transitions inside a face, these transitions have to be along straight lines or polygonal lines. Otherwise, the method introduces another kind of error, as the discrete geometry is not able to describe precisely the exact geometry. This is similar to what happens in curved boundaries, a problem that we will explore briefly in the following lesson.

An element of

$$\mathbb{P}_1 = \big\{ a_0 + a_1\, x + a_2\, y + a_3\, z \; : \; a_0, a_1, a_2, a_3 \in \mathbb{R} \big\}$$

is uniquely determined by its values on the four vertices of a general non-degenerate tetrahedron. See Figure 3.8 for a sketch of the tetrahedral finite element. More over, seen on each of the faces of the tetrahedron such a function is a linear function of two variables and seen on each of the six edges it is a linear function of a single variable. Therefore: the value of the function on each face is determined by the three degrees of freedom (nodal values) that lie on that face and the value on each edge is determined by its values on the two associated nodes.

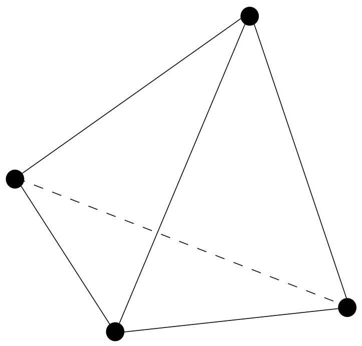


Figure 3.8: A tetrahedron and its four $\mathbb{P}_1$ degrees of freedom

With this in hand we can do our business as usual. Nothing is really changed by going to the three dimensional case. The reference element is usually taken as the tetrahedron with vertices on $(0, 0, 0)$, $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$. Fortunately, the order of the local numbering of vertices is irrelevant, since all permutations give valid numberings.

The price to pay for this simplicity is the fact that tetrahedra are much more strange animals than they look at first sight. In particular it is not that simple to fathom how to divide a given tetrahedron into pieces that are not too deformed. Look at what happens (Figure 3.10) when you cut out the four corners of a regular tetrahedron. Inside you obtain a regular octahedron that can be easily divided into two pyramids with square basis, each of which can be divided into two similar tetrahedra. The resulting interior four tetrahedra are not regular anymore. There are more ways of doing this kind of things. My point here is that tetrahedra are easy but not so easy.

Local dimension of the space is four. When you glue the corresponding $\mathbb{P}_1$ elements to create a finite element space the full dimension is the number of vertices. Dirichlet nodes are defined as usual (nodes on the Dirichlet boundary, or vertices of faces that are on the Dirichlet boundary). Order is one.

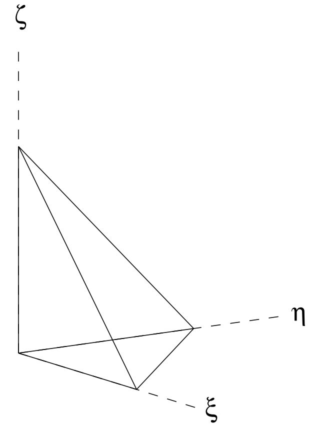


Figure 3.9: The reference tetrahedron, as seen from behind (sort of). The reference variables are ξ, η and ζ (some authors prefer z for the third one)

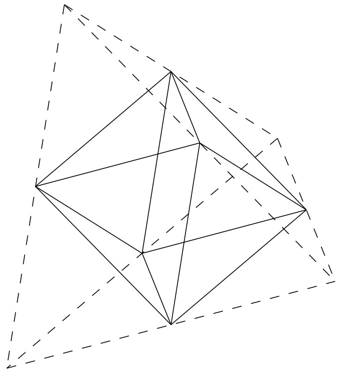


Figure 3.10: When you cut the four corners of a regular tetrahedron you end up with a regular octahedron

It is not difficult to define $\mathbb{P}_k$ elements on the tetrahedron for any k. Note that the local dimensions of the $\mathbb{P}_k$ spaces (as well as the number of necessary nodes) increase now much faster than in the two dimensional cases, because there are many more new monomials. The space $\mathbb{P}_k$ uses all monomials of the form

$$x^{i_1} y^{i_2} z^{i_3}, \qquad i_1, i_2, i_3 \geq 0, \quad i_1 + i_2 + i_3 \leq k.$$

For instance

$$\dim \mathbb{P}_2 = 9, \qquad \dim \mathbb{P}_3 = 19.$$

There's a formula for this but we will not give it here.

It is simple to give the nodes in the reference domain. For the $\mathbb{P}_k$ element, they are just the points with coordinates

$$\left(\tfrac{i_1}{k}, \tfrac{i_2}{k}, \tfrac{i_3}{k}\right), \qquad i_1, i_2, i_3 \geq 0, \quad i_1 + i_2 + i_3 \leq k.$$

We have to wait to $k = 4$ to obtain an interior node that we can condense statically.

3.2 Elements on parallelepipeds

The $\mathbb{Q}_k(K)$ elements are very naturally defined on parallelepipeds. The reference element is the cube $[-1,1] \times [-1,1] \times [-1,1]$ or also $[0,1] \times [0,1] \times [0,1]$, depending on personal preferences. The reference space $\mathbb{Q}_k$ is the one of all linear combinations of monomials

$$\xi^{i_1}\eta^{i_2}\zeta^{i_3}, \qquad 0 \le i_1, i_2, i_3 \le k$$

and has therefore dimension $(k+1)^3$. Nodes are easily found by subdividing uniformly the cube into equally sized smaller cubes. Interior nodes appear already with $k = 2$. The local spaces on parallelepipeds (the image of a cube under a linear transformation) are the new spaces $\mathbb{Q}_k(K)$ defined as usual.

One has to be extra careful here in giving always vertices in a coherent order, so that we don't try to map the figure incorrectly from the reference element. That is the price to pay for the geometrical simplicity. The increase of the polynomial degree is also a non-minor issue: for $\mathbb{Q}_1(K)$ elements we have polynomials of degree three!

4 Exercises

1. **Comparison of $\mathbb{P}_1$ and $\mathbb{Q}_1$.** Consider a square domain Ω and two triangulations of it as the ones given in Figure 3.11. In the first triangulation we consider a $\mathbb{P}_1$ method for the usual equation, only with Neumann conditions. In the second partition we consider a $\mathbb{Q}_1$ method.

   Check that we have the same number of unknowns in both cases. Draw the form of the mass-plus-stiffness matrices in both cases. Check that the $\mathbb{Q}_1$ has in principle more non-zero elements, since there are pairs of adjacent nodes that are not so in the triangular

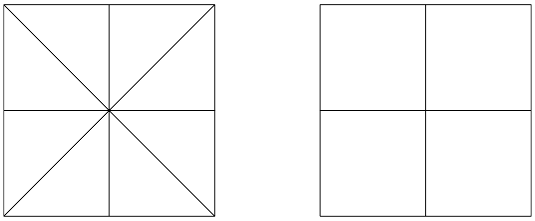


Figure 3.11: A triangle mesh and a parallelogram mesh of a square.

2. **The $\mathbb{Q}_3$ element in the plane.** What would be the nodes and the polynomial space for a generalization of the $\mathbb{Q}_k$ type elements to $k = 3$? How many interior nodes do you obtain?

3. **Elements on prisms.** The reference prism with triangular basis can be taken for instance as the set of points (ξ, η, ζ) with

$$0 \le \xi, \eta, \qquad \xi + \eta \le 1, \qquad 0 \le \zeta \le 1.$$

In the two plane variables it works like a triangle. In the vertical variables it works like a parallelogram. Propose a correct polynomial space in this reference configuration so that the six vertices are valid nodes for a finite element using prisms.

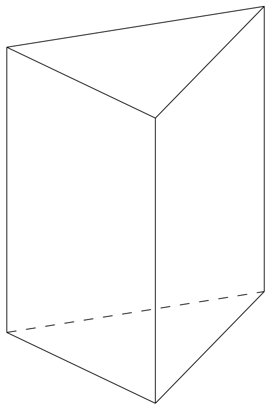


Figure 3.12: The reference prism.

Lesson 4

More advanced questions

In this lesson we are going to have a fast look at several different questions related to how the Finite Element Method is used (or adapted) in different situations. The section on eigenvalues is of particular importance, since we will be using it for the stability analysis of evolution problems.

1 Isoparametric elements

So far we have only dealt with polygonal domains. You will agree that in many instances boundaries are due to be curved, so we will have to take into account that fact.

First of all when creating a triangulation, you are substituting your real curved domain by a polygonal approximation. Your grid generator is going to take care of the following detail: *all boundary nodes of the triangulation have to be placed on the real boundary.* This means in particular that if you need smaller triangles, you cannot obtain them by simply subdividing your existing grid and you definitely have to call back your grid generator to give you new vertices that are on the boundary.

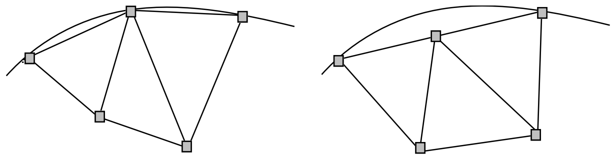


Figure 4.1: A correct and an incorrect triangulation of a curved domain.

You might think, well that's it then, isn't it? You have your triangles and you apply your triangular finite element scheme. The answer is yes if you are going to apply the $\mathbb{P}_1$ method.

Note for a moment that functions on $H^1(\Omega)$ are defined on Ω and functions of V_h on the approximated polygon. Therefore the discrete space V_h is not a subspace of $H^1(\Omega)$. However, an error analysis is still possible. What this error shows is that the error produced by the geometry approximation beats the error of the method if we try to do $\mathbb{P}_2$ elements or higher, so it doesn't pay off to use high order elements if the approximation to the boundary is so rough.

Let us see how to mend this for the $\mathbb{P}_2$ approximation. Note that this is not a purely theoretical question and that part of what we are going to learn here will be used to define finite elements on general quadrilaterals.

1.1 Deformed triangles

As usual, take $\widehat{K}$ to be the reference triangle and K^0 a triangle with vertices

$$\mathbf{p}_\alpha^K = (x_\alpha, y_\alpha), \qquad \alpha = 1,2,3.$$

With these points we construct the linear map $F_K^0 : \widehat{K} \to K^0$

$$\begin{aligned} \begin{bmatrix} x \\ y \end{bmatrix} &= \begin{bmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{bmatrix} \begin{bmatrix} \xi \\ \eta \end{bmatrix} + \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} \\ &= (1 - \xi - \eta) \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} + \xi \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} + \eta \begin{bmatrix} x_3 \\ y_3 \end{bmatrix}. \end{aligned}$$

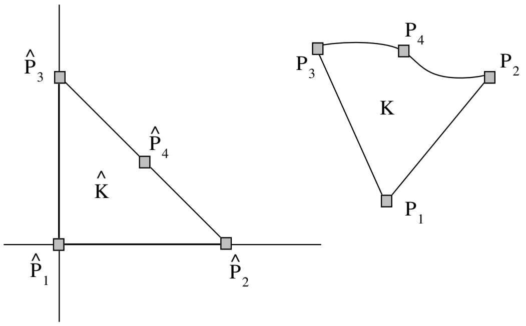


Figure 4.2: The reference triangle and a deformation of the image triangle.

Let us now call $\mathbf{p}_4^K$ to the midpoint of the segment that joins $\widehat{\mathbf{p}_2}$ and $\widehat{\mathbf{p}}_3$, that is,

$$\widehat{\mathbf{p}}_4 = (\tfrac{1}{2}, \tfrac{1}{2}).$$

Take a fourth point in the physical space, $\mathbf{p}_4^K = (x_4, y_4)$, and compute its deviation from the midpoint of $\mathbf{p}_2^K$ and $\mathbf{p}_3^K$

$$\begin{bmatrix} \delta_x \\ \delta_y \end{bmatrix} = \begin{bmatrix} x_4 \\ y_4 \end{bmatrix} - \begin{bmatrix} \dfrac{x_2 + x_3}{2} \\ \dfrac{y_2 + y_3}{2} \end{bmatrix}.$$

Finally take the transformation $F_K : \widehat{K} \to \mathbb{R}^2$ given by

$$F_K(\xi, \eta) = F_K^0(\xi, \eta) + 4\,\xi\,\eta \begin{bmatrix} \delta_x \\ \delta_y \end{bmatrix}.$$

Note that this is a linear transformation plus a correction term. The transformation F_K satisfies the following properties, all of them of easy verification:

- It sends the chosen points in the reference domain to the ones in the physical space

$$F_K(\widehat{\mathbf{p}}_\alpha) = \mathbf{p}_\alpha^K, \qquad \alpha = 1, \ldots, 4.$$

- If $\xi = 0$, then

  $$F_K(0,t) = F_K^0(0,t).$$

  This means that the image of the vertical edge in reference coordinates is the segment joining $\mathbf{p}_1^K$ and $\mathbf{p}_3^K$, covered at constant velocity, as if we were using the linear transformation. The same thing happens to the horizontal side of $\widehat{K}$.

- If $\mathbf{p}_4^K$ is aligned with $\mathbf{p}_2^K$ and $\mathbf{p}_3^K$, then the image of the edge that joins $\widehat{\mathbf{p}}_2$ and $\widehat{\mathbf{p}}_3$ is the segment that joins $\mathbf{p}_2^K$ and $\mathbf{p}_3^K$. However, this segment is parameterized at constant velocity only when $\mathbf{p}_4^K$ is the midpoint of $\mathbf{p}_2^K$ and $\mathbf{p}_3^K$ (in that case $\delta_x = \delta_y = 0$ and we have only the linear term in the transformation F_K).

- The Jacobian matrix of F_K is not constant:

$$\mathbf{B}_K = \mathrm{D}F(\xi,\eta) = \mathbf{B}_K^0 + 4 \begin{bmatrix} \eta \\ \xi \end{bmatrix} \begin{bmatrix} \delta_x & \delta_y \end{bmatrix} = \begin{bmatrix} x_2 - x_1 + 4\eta\delta_x & x_3 - x_1 + 4\eta\delta_y \\ y_2 - y_1 + 4\xi\delta_x & y_3 - y_1 + 4\xi\delta_y \end{bmatrix}$$

When $\mathbf{p}_4^K$ is not too far from the midpoint of $\mathbf{p}_2^K$ and $\mathbf{p}_3^K$, that is, when the deviation (δ_x, δ_y) is not too large, it is possible to prove that the image of $\widehat{K}$ under this transformation $K = F_K(\widehat{K})$ is mapped bijectively from the reference element and therefore we can construct an inverse to

$$F_K : \widehat{K} \to K.$$

1.2 Local spaces

Now we have the physical element, K, which is defined as the image of $\widehat{K}$ by the transformation F_K, so we have gone one step further from the beginning, as now the physical element is only defined from the reference element. With this element in hand we define the local space by transforming $\mathbb{P}_2$ on reference variables (as we did with all $\mathbb{Q}_k(K)$ spaces):

$$\mathbb{P}_2(K) = \{p : K \to \mathbb{R} \,:\, p \circ F_K \in \mathbb{P}_2\} = \{\widehat{p} \circ F_K^{-1} \,:\, \widehat{p} \in \mathbb{P}_2\}.$$

The degrees of freedom are placed in the following six nodes:

- the three vertices,

- the midpoints of the two straight sides,

- the point $\mathbf{p}_4^K$.

We do not have an explicit expression of how elements of $\mathbb{P}_2(K)$ are, but we know that if $\widehat{N}_\alpha$ are the six nodal basis functions of the $\mathbb{P}_2$ reference element, then the functions

$$N_\alpha^K = \widehat{N}_\alpha \circ F_K^{-1}$$

form a basis of $\mathbb{P}_2(K)$. The following properties are simple to prove:

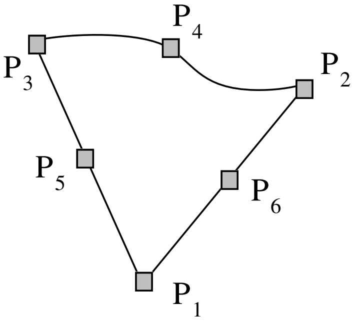


Figure 4.3: The local nodes in an isoparametric $\mathbb{P}_2$ triangle

- A function in $\mathbb{P}_2(K)$ is uniquely determined by the values on the six nodes on K.
- Restricted to any of the two straight sides of K, a function in $\mathbb{P}_2(K)$ is a polynomial of degree two in one variable (that is, the form of the function does not depend on the geometry of the element) and is therefore uniquely determined by its values on the three nodes that lie on that side.
- The value of a function in $\mathbb{P}_2(K)$ on the curved edge of K is uniquely determined by its value on the three nodes that lie on that edge.

The first property allows us to use the six nodes as local degrees of freedom. The second one allows as to glue $\mathbb{P}_2(K)$ on curved triangles with $\mathbb{P}_2$ elements on straight triangles, since the values on the straight sides are just polynomials.

If K is a usual straight triangle and we take $\mathbf{p}_4^K$ to be the midpoint of the corresponding edge, then F_K is a linear map and $\mathbb{P}_2(K) = \mathbb{P}_2$.

1.3 Finite element spaces with isoparametric triangles

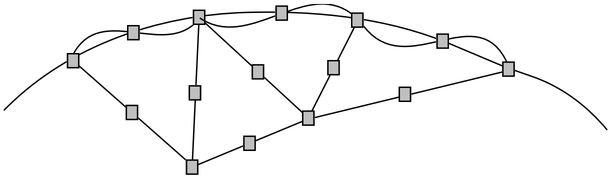


Figure 4.4: A triangulation using isoparametric elements

Let us then begin with an approximate triangulation of a curved domain following the rules:

- Intersection of two different triangles can only happen in a common vertex or edge.
- There must be a vertex placed on each transition point from Dirichlet to Neumann boundaries.

- Triangles with an edge on the of the approximating polygon can have only one edge on this boundary and both vertices have to be on the exact boundary Γ.

Look again at Figure 4.1 to see what we mean. We not only want boundary triangles to hang from the real boundary, but we want to avoid a triangle to have two edges on the boundary[1].

The second part of the triangulation process consists of choosing a point on the exact boundary for each boundary edge. This point should be close to the midpoint of the straight edge that approximates the real curved boundary. We use this new point to construct an isoparametric triangle with the same vertices for each boundary triangle.

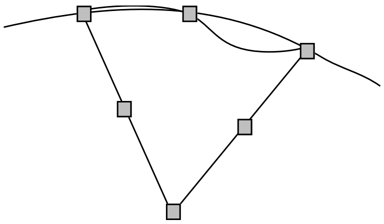


Figure 4.5: Substituting a straight triangle on the boundary by a isoparametric triangle.

When we write the equations of the finite element method using these local spaces, we must have in mind that the union of all triangles (curved on the boundary, straight otherwise) is not the original domain Ω, but an approximation of it, which we will call Ω_h. We will still call **Dirichlet nodes** to nodes on the Dirichlet boundary, remarking that these nodes are in the correct boundary Γ, so we will be able to read data on them when needed. The full finite element space is

$$V_h = \{u_h \in \mathcal{C}(\overline{\Omega}) \,:\, u_h|_K \in \mathbb{P}_2(K), \quad \forall K \in \mathcal{T}_h\},$$

and the subspace with homogeneous Dirichlet boundary conditions is

$$V_h^{\Gamma_D} = \{v_h \in V_h \,:\, v_h(\mathbf{p}) = 0, \forall \mathbf{p}\,\text{Dirichlet node}\}.$$

Note that functions of $V_h^{\Gamma_D}$ are not really zero on the Dirichlet boundary Γ_D but on the curved approximation of that boundary, an approximation hanging from the vertices of the initial triangulation and from the additional point per edge that was used to create the isoparametric elements. It will not come out as a surprise, since the process of gluing spaces is the same as what we did with $\mathbb{P}_2$ elements, that the dimension of V_h is the number of nodes (that is the number of vertices plus the number of edges) and the dimension of $V_h^{\Gamma_D}$ is the number of non-Dirichlet edges. A nodal basis can be constructed as usual. The restriction to elements of nodal basis functions will be again the local basis functions, themselves defined as the transformed local nodal functions on the reference element.

[1] Many grid generators, even for polygonal domains, avoid putting two edges of the same triangle on the boundary. There is a simple reason for that: if two edges of a triangle are in a homogeneous Dirichlet boundary and we are using $\mathbb{P}_1$ elements, the function vanishes in the whole triangle, which is a poor result.

The discrete bilinear form is $a_h : V_h \times V_h \to \mathbb{R}$

$$a_h(u_h, v_h) = \int_{\Omega_h} \nabla u_h \cdot \nabla v_h + \int_{\Omega_h} u_h \, v_h,$$

and the linear form is $\ell_h : V_h \to \mathbb{R}$

$$\ell_h(v_h) = \int_{\Omega_h} f \, v_h + \int_{\Gamma_N^h} g \, v_h.$$

With them we obtain the numerical method

$$\left[\begin{array}{l} \text{find } u_h \in V_h \text{ such that} \\ u_h(\mathbf{p}_i) = g_0(\mathbf{p}_i), \qquad \forall i \in \text{Dir}, \\ a_h(u_h, \varphi_i) = \ell_h(\varphi_i), \qquad \forall i \in \text{Ind}\,. \end{array}\right.$$

Note that the bilinear form poses no problem whatsoever. The fact that we are working on the approximate domain is sort of invisible to the assembly process: we will go element by element transforming from the reference configuration and after having added all terms we will have computed an integral over Ω_h instead of Ω. More on this at the end of this section.

The issue of the data functions is a little more delicate. When we want to compute

$$\int_K f \, \varphi_i \qquad \text{or, in fact,} \qquad \int_K f \, N_\alpha^K$$

for one of the curved domains K, it is perfectly possible that the source function is not defined in parts of K. Look at Figure 4.5 and see how there is a small piece of the discrete geometry that lies outside the domain. Several theoretically sound possibilities can be proposed to mend this. In practice, and since you are due to use quadrature formulas for this integrals, just avoid using quadrature points in those areas.

The situation for the Neumann conditions (given normal derivative) is even more complicated and I have been deliberately vague in writing

$$\int_{\Gamma_N^h} g \, \varphi_i$$

without specifying what I mean by Γ_N^h. The fact is g_1 is defined in the exact Neumann boundary and φ_i in its approximation, so the integral is just a way of speaking. Assembly of this term will be done edge-by-edge. For each edge integral we could just try to use a quadrature formula that evaluates only on the three common points between the exact and the discrete geometry or think of something more clever. Let's not do this right now. I just wanted you to see that complications arise very easily.

Even when computing the local integrals

$$\int_K \nabla N_\beta^K \cdot \nabla N_\alpha^K \qquad \int_K N_\beta^K N_\alpha^K$$

for curved K, we still have to be careful. Let us begin with the easy one, the mass matrix. We have a formula for the local basis functions

$$N_\alpha^K = \widehat{N}_\alpha \circ F_K^{-1}.$$

If we want to evaluate N_α^K in a point of K, say (x, y), we need to compute $F_K^{-1}(x, y)$. This is the same as solving the non-linear system

$$\begin{aligned} x &= (x_2 - x_1)\,\xi + (x_3 - x_1)\,\eta + x_1 + 4\xi\eta\delta_x \\ y &= (y_2 - y_1)\,\xi + (y_3 - y_1)\,\eta + y_1 + 4\xi\eta\delta_y. \end{aligned}$$

It is only a 2×2 system and equations are quadratic, but it is still a non-linear system and you will need Newton's method or something similar to get an approximate solution. Of course we know the exact solution for six points (the six nodes), since they are mapped back to the six nodes of the reference domain, so using these points is for free. It looks like we are done, but you have still to notice that the integral is happening over a very strange domain for which we don't have quadrature formulas. What is the wise thing to do? Move everything back to the reference domain:

$$\int_K N_\beta^K N_\alpha^K = \int_{\widehat{K}} |\det \mathbf{B}_K| \widehat{N}_\beta \, \widehat{N}_\alpha^K.$$

With this strategy, the integral is defined on a plain triangle and we just need to compute the non-constant determinant of

$$\mathbf{B}_K = \begin{bmatrix} x_2 - x_1 + 4\eta\delta_x & x_3 - x_1 + 4\eta\delta_y \\ y_2 - y_1 + 4\xi\delta_x & y_3 - y_1 + 4\xi\delta_y \end{bmatrix}$$

on the chosen quadrature points. The stiffness matrix is more challenging. Instead of trying to work the integral on the curved domains (with the complication of having to invert F_K every time we need an evaluation), what we do is go backwards to the reference domain and write

$$\int_{\widehat{K}} |\det \mathbf{B}_K| \, (\mathbf{C}_K \nabla \widehat{N}_\alpha \cdot \nabla \widehat{N}_\beta),$$

where

$$\mathbf{C}_K = \mathbf{B}_K^{-1} \mathbf{B}_K^{-\top}$$

(we did this in Lesson 2) is a non-constant matrix that requires inversion of $\mathbf{B}_K$ every time an evaluation is needed.

The whole thing looks more complicated than it is, because there are many aspects to take care of at the same time. The lesson you have to learn here is that evaluating anything (a basis function, its gradient, etc) has a price so you should try to balance a sufficiently precise approximation (exact computation is not possible any longer) of the integrals with taking as few quadrature points as possible.

2 Elements on quadrilaterals

Going back to the case of polygonal domains, we might be still more interested in using grids of quadrilaterals type than triangular grids. It can be a question of your geometry being described in a simpler form by using quadrilaterals, a preference for $\mathbb{Q}_k$ elements or a physical motivation to prioritize directions in the discrete level[2]. Whatever your reasons are, here is a way of defining finite elements of quadrilaterals that are not parallelograms. By the way, many people say **quads**, which is a nice shortening. I'm not going to write it again.

The construction is reminiscent of that of isoparametric elements. We begin with the reference square $[-1,1] \times [-1,1]$ and recall the four $\mathbb{Q}_1$ basis functions

$$\frac{1}{4}(1 \pm \xi)(1 \pm \eta)$$

(it is easy to check which one corresponds to each vertex). Now we take a general convex quadrilateral[3] and take its four vertices in rotating order: $\mathbf{p}_1^K, \ldots, \mathbf{p}_4^K$.

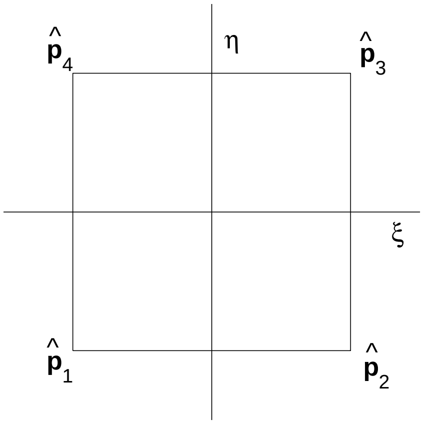


Figure 4.6: The reference square again.

Since the functions $\widehat{N}_\alpha$ satisfy

$$\widehat{N}_\alpha(\widehat{\mathbf{p}}_\beta) = \delta_{\alpha\beta}, \qquad \alpha, \beta = 1, \ldots, 4,$$

it is clear that the map $F_K : \widehat{K} \to \mathbb{R}^2$

$$\begin{bmatrix} x \\ y \end{bmatrix} = \widehat{N}_1(\xi,\eta) \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} + \widehat{N}_2(\xi,\eta) \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} + \widehat{N}_3(\xi,\eta) \begin{bmatrix} x_3 \\ y_3 \end{bmatrix} + \widehat{N}_4(\xi,\eta) \begin{bmatrix} x_4 \\ y_4 \end{bmatrix},$$

sends vertices to vertices, that is

$$F_K(\widehat{\mathbf{p}}_\alpha) = \mathbf{p}_\alpha^K, \qquad \alpha = 1, \ldots, 4.$$

[2] In any case, remember that from a quadrilateral grid you can always obtain a triangular one doubling the number of elements and with a little less stiffness (remember the exercise in Lesson 2).

[3] Do not waste your time with non-convex quadrilaterals. For that it's better to use pairs of triangles

Moreover, the restriction of F_K to one of the four sides of $\widehat{K}$ is mapped at constant velocity to the corresponding edge of K. Obviously, by continuity, the interior of $\widehat{K}$ is mapped to the interior of K.

The map F_K in fact transforms a uniform Cartesian into something very similar in K, as shown in Figure 4.7. Computation of F_K^{-1} is therefore very simple on points of this special grid in the quadrilateral.

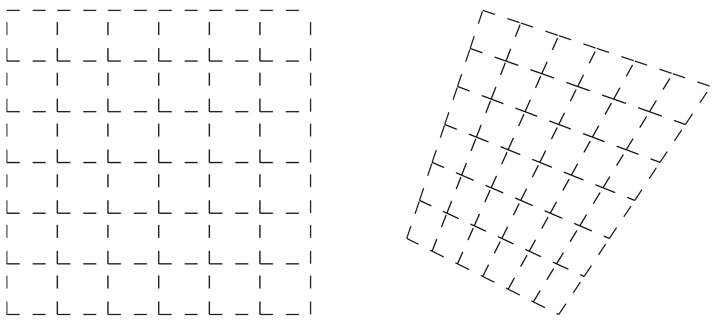


Figure 4.7: The image by the bilinear map of a Cartesian grid in the reference square.

With this transformation in hand we can define the local spaces

$$\mathbb{Q}_k(K) = \{q : K \to \mathbb{R} \,:\, q \circ F_K \in \mathbb{Q}_k\}.$$

I think you can already imagine what the local nodes are, how we can glue elements on different quadrilaterals and so on. If K is a parallelogram, F_K is a linear map and we obtain the usual $\mathbb{Q}_k$ spaces, which are composed of polynomial functions. In other cases, elements of the space are functions of the form $\widehat{q} \circ F_K^{-1}$ with $\widehat{q}$ a polynomial. The inverse map F_K^{-1} is not linear anymore. Now it is rational function. Therefore, elements of $\mathbb{Q}_k(K)$ are not polynomials any longer, which is not really relevant, since we are only going to use the basis functions, which we obtain by transforming from the reference element.

Just a fast list of facts:

- The spaces $\mathbb{Q}_k(K)$ depend on the quadrilateral and not on the order we have given the vertices to construct the transformation.
- The image of the $\mathbb{Q}_k$ nodes by F_K are valid degrees of freedom in $\mathbb{Q}_k(K)$.
- Restricted to the four sides, functions of $Q_k(K)$ are just polynomials of degree up to k in one variable. Therefore, the type of functions is independent of the shape of the quadrilateral and the values on sides is determined by the values on nodes that are on the side.

Thanks to these properties we can easily construct finite element spaces on quadrilateral grids (composed of convex quadrilaterals). Parallelograms are a particular case of these elements, so we can use all types of quadrilaterals together . Therefore, we can combine these elements with triangular elements of the same degree: for instance $\mathbb{Q}_1(K)$ elements on quadrilaterals with $\mathbb{P}_1$ elements on triangles.

3 Mass lumping

Let us just here add some comment about the mass matrix in the finite element method. For the usual $\mathbb{P}_k$ and $\mathbb{Q}_k$, we should expect the mass matrix

$$\int_\Omega \varphi_j \, \varphi_i$$

to be well-conditioned. Recall that the mass matrix is symmetric and positive definite. The spectral condition number of this type of matrices is the ratio between its largest and its smallest eigenvalue (all of them are real and positive), and good conditioning means that this ratio is not large. In particular it means that if

$$u_h = \sum_j u_j \varphi_j$$

then the constants in the inequality

$$C_1 \sum_j |u_j|^2 \le \int_\Omega |u_h|^2 \le C_2 \sum_j |u_j|^2$$

are of the same size and thus and the Euclidean norm of the vector of coefficients represents faithfully the $L^2(\Omega)$-norm of the function up to a scaling factor. In its turn, good conditioning means that the use of the most common iterative methods for systems with symmetric positive definite matrices (such as Conjugate Gradient) is going to converge quickly.

However, sometimes it seems convenient to substitute the mass matrix by an even simpler matrix. In the next lesson we will see a situation where this seems justified. Substitution of the mass matrix by a diagonal matrix is called mass lumping, since it lumps mass on the nodes instead of distributing it along pairs of nodes. We are going to explain this process for the $\mathbb{P}_1$ case.

Recall briefly the three-vertex quadrature rule on triangles (we mentioned it in the lesson on assembly)

$$\int_K \phi \approx \frac{\operatorname{area} K}{3} \sum_{\alpha=1}^{3} \phi(\mathbf{p}_\alpha^K),$$

where $\mathbf{p}_\alpha^K$ are the three vertices of K. This formula integrates exactly all polynomials of degree one. However, it introduces error when applied to a polynomial of degree two. In particular, the approximation

$$\int_K N_\beta^K N_\alpha^K \approx \frac{\operatorname{area}}{K} \sum_{\gamma=1}^{3} N_\alpha^K(\mathbf{p}_\gamma^K) N_\beta^K(\mathbf{p}_\gamma^K) = \frac{\operatorname{area} K}{3} \delta_{\alpha\beta}$$

is not exact. (We have accumulated as many as three indices in the last expression. Do you see why the result holds? Note that local basis functions are one on a single vertex and zero on the other two.) If we apply this approximation at the assembly process for the mass matrix, we are substituting the 3 × 3 local mass matrices by a 3 × 3 diagonal matrix. Adding up all contributions we are approximating

$$\int_\Omega \varphi_j \varphi_i \approx 0, \qquad i \neq j$$

and

$$\begin{aligned} \int_\Omega |\varphi_i|^2 &= \sum_K \int_K |\varphi_i|^2 \approx \sum \Big\{ \frac{\mathrm{area}\, K}{3} \,:\, K \,\text{such that}\; \mathbf{p}_i \in K \Big\} \\ &= \tfrac{1}{3}\,\mathrm{area}\,\big(\mathrm{supp}\,\varphi_i\big). \end{aligned}$$

Once again, the support of φ_i is the set of triangles that surround the node $\mathbf{p}_i$.

In an exercise at the end of this lesson we will see a very, very simple way of computing the lumped mass matrix once the exact mass matrix has been computed.

4 The discrete eigenvalues

While the mass matrix is well conditioned, the stiffness matrix is not. And it has to be so, because it is trying to approximate an intrisically ill-conditioned problem. We are going to have a look at this. Note that this section is really important to understand the stability analysis of the application of FEM methods for evolution problems, so take your time to understand what's being told here. Note that the results here are considerably deeper than what we have been using so far.

4.1 The Dirichlet eigenvalues of the Laplace operator

For simplicity, let us concentrate our efforts in problems only with Dirichlet conditions. In fact, with homogeneous Dirichlet conditions. Instead of trying to solve a boundary value problem, we are going to study an eigenvalue problem: find numbers λ such that there exists non-zero u satisfying

$$\left[\begin{array}{ll} -\Delta u = \lambda u, & \quad \text{in } \Omega, \\[1ex] u = 0, & \quad \text{on } \Gamma. \end{array}\right.$$

Note two things. First of all, $u = 0$ is not an interesting solution since it always satisfies the conditions, no matter what λ is. Second, boundary conditions in eigenvalue problems have to be zero. If you have two different solutions u for the same λ, any linear combination of them is another solution. The set of **eigenfunctions** (that's u) for a given **eigenvalue** (that's λ) is a subspace of ... (wait for it).

In this problem $\Gamma_D = \Gamma$. The space $H^1_\Gamma(\Omega)$ is given a different name. This one

$$H^1_0(\Omega) = \{u \in H^1(\Omega) \,:\, u = 0, \quad \text{on } \Gamma\}.$$

The set of eigenfunctions for a given eigenvalue is a subspace of $H_0^1(\Omega)$. Therefore, also of $H^1(\Omega)$ and of $L^2(\Omega)$, which are bigger and bigger spaces. Substituting the definition of eigenfunction inside Green's formula

$$\int_\Omega \Delta u\, v + \int_\Omega \nabla u \cdot \nabla v = \int_\Gamma (\partial_n u)\, v$$

and proceeding as usual, we obtain

$$-\lambda \int_\Omega u\, v + \int_\Omega \nabla u \cdot \nabla v = \int_\Gamma (\partial_n u)\, v = 0, \qquad \text{if } v = 0 \text{ on } \Gamma.$$

So we arrived easily to the weak formulation of the eigenvalue problem:

$$\left[\begin{array}{l} \text{find } \lambda \text{ such that there exists } 0 \neq u \in H_0^1(\Omega) \text{ satisfying} \\ \displaystyle\int_\Omega \nabla u \cdot \nabla v = \lambda \int_\Omega u\, v, \qquad \forall v \in H_0^1(\Omega). \end{array}\right.$$

Do we know how many eigenvalues are going to appear? Yes, we do. Infinitely many. But among those infinitely many, not so many, since we will be able to count them. I am going to try and break up the theoretical result in many pieces so that you really grasp what's in here:

- All eigenvalues are real and positive.

- They can be numbered and they diverge to infinity. There is therefore no accumulation point of eigenvalues. In other words, if you choose a finite interval, there is only a finite number of eigenvalues in it.

- Two eigenfunctions corresponding to two different eigenvalues are $L^2(\Omega)$-orthogonal. In more detail, assume that

$$\left[\begin{array}{ll} -\Delta u = \lambda u, & \text{in } \Omega, \\ u = 0, & \text{on } \Gamma, \end{array}\right. \qquad \text{and} \qquad \left[\begin{array}{ll} -\Delta v = \mu v, & \text{in } \Omega, \\ v = 0, & \text{on } \Gamma, \end{array}\right.$$

  with $\lambda \neq \mu$. Then

$$\int_\Omega u\, v = 0.$$

- The multiplicity of each eigenvalue is finite, that is, if λ is an eigenvalue, there is only a finite number of linearly independent eigenfunctions associated to it.

Let's try to put all these properties together. For each eigenvalue we take a set of linearly independent eigenfunctions. Using the Gram-Schmidt orthogonalization method, we can make them mutually $L^2(\Omega)$-orthogonal and with unit square integral. Instead of numbering the different eigenvalues, we take k copies of each eigenvalue with multiplicity k. Then we can order all the eigenvalues in increasing order and we obtain a list

$$0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n \leq \ldots, \qquad \lambda_n \to \infty$$

and associate to each eigenvalue an eigenfunction

$$\left[\begin{array}{l} -\Delta\phi_n = \lambda_n \phi_n, \qquad \text{in } \Omega, \\ \phi_n = 0, \qquad \text{on } \Gamma, \end{array}\right.$$

so that

$$\int_\Omega \phi_n\, \phi_m = \delta_{nm}$$

and we have taken all possible eigenvalues and (linearly independent) eigenfunctions. Note again that eigenfunctions for different eigenvalues are per se orthogonal and that we enforce orthogonality of eigenfunctions of the same eigenvalue by an orthogonalization process.

There is an additional property that needed this kind of numbering of eigenvalues and eigenfunctions to be properly introduced:

- The sequence of eigenfunctions we have just obtained is a complete orthogonal set in $L^2(\Omega)$, which means that if $u \in L^2(\Omega)$, then

  $$u = \sum_{j=1}^\infty u_j\, \phi_j, \qquad u_j = \int_\Omega u\, \phi_j,$$

  with convergence of the series in the norm of $L^2(\Omega)$, i.e.,

  $$\int_\Omega \Big| u - \sum_{j=1}^n u_j \phi_j \Big|^2 \overset{n\to\infty}{\longmapsto} 0.$$

4.2 The discrete Dirichlet eigenvalues

Assume now that Ω is a polygon and take $V_h \subset H^1(\Omega)$, any of our choices of finite element spaces. By eliminating the Dirichlet nodes we obtain a basis of the space

$$V_h^0 = V_h \cap H_0^1(\Omega) = \{u_h \in V_h \,:\, u_h = 0, \quad \text{on } \Gamma\}.$$

We now substitute the problem

$$\left[\begin{array}{l} \text{find } \lambda \text{ such that there exists } 0 \neq u \in H_0^1(\Omega) \text{ satisfying} \\ \displaystyle\int_\Omega \nabla u \cdot \nabla v = \lambda \int_\Omega u\, v, \qquad \forall v \in H_0^1(\Omega), \end{array}\right.$$

by its finite element approximation

$$\left[\begin{array}{l} \text{find } \lambda_h \text{ such that there exists } 0 \neq u_h \in V_h^0 \text{ satisfying} \\ \displaystyle\int_\Omega \nabla u_h \cdot \nabla v_h = \lambda_h \int_\Omega u_h\, v_h \qquad \forall v_h \in V_h^0. \end{array}\right.$$

Consider the matrices $\mathbf{W}$ and $\mathbf{M}$

$$w_{ij} = \int_\Omega \nabla\varphi_j \cdot \nabla\varphi_i, \qquad m_{ij} = \int_\Omega \varphi_j \varphi_i, \qquad i, j \in \text{Ind}.$$

Note that these are just the parts of the stiffness and mass matrices related to non-Dirichlet nodes. Let $N = \#\mathrm{Ind}$ be the number of non-Dirichlet nodes. The discrete eigenvalue problem is equivalent to this other problem

$$\left[\begin{array}{l} \text{find } \lambda_h \text{ such that there exists } \mathbf{0} \neq \mathbf{u} \in \mathbb{R}^N \text{ satisfying} \\ \mathbf{W}\mathbf{u} = \lambda_h \mathbf{M}\mathbf{u}. \end{array}\right.$$

This last problem is a generalized eigenvalue problem for matrices. Let me condense the main properties of this problem for you. Recall that $N = N(h)$ is the dimension of the problem.

- (Generalized) eigenvalues are real and positive.
- Eigenvectors corresponding to different eigenvalues are orthogonal with respect to $\mathbf{M}$: if
$$\mathbf{W}\mathbf{u} = \lambda_h \mathbf{M}\mathbf{u}, \qquad \mathbf{W}\mathbf{v} = \mu_h \mathbf{M}\mathbf{v}$$
with $\lambda_h \neq \mu_h$, then
$$\mathbf{u} \cdot (\mathbf{M}\mathbf{v}) = 0$$
- There are N linearly independent eigenvectors.

Note that unlike in the original problems, there is no question about having more than N eigenvalues, since we are dealing with an $N \times N$ matrix. Counting eigenvalues as many times as their multiplicity we have N of them that we can arrange in increasing order

$$0 < \lambda_{h,1} \leq \lambda_{h,2} \leq \ldots \leq \lambda_{h,N}.$$

Choosing linearly independent eigenvectors in case we have multiplicity higher than one, we can choose vectors $\boldsymbol{\phi}_n$ such that

$$\mathbf{W}\boldsymbol{\phi}_n = \lambda_{h,n}\mathbf{M}\boldsymbol{\phi}_n$$

and

$$\boldsymbol{\phi}_n \cdot (\mathbf{M}\boldsymbol{\phi}_m) = \delta_{nm}.$$

The vectors $\boldsymbol{\phi}_n$ give a basis of $\mathbb{R}^N$. The corresponding finite element functions

$$\phi_{h,n} = \sum_{j=1}^N \phi_{n,j}\varphi_j, \qquad \boldsymbol{\phi}_n = (\phi_{n1}, \ldots, \phi_{nN})^\top$$

form a basis for V_h^0. Note that the $\mathbf{M}$-orthogonality of the eigenvectors is just the matrix form of the orthogonality condition

$$\int_\Omega u_{h,n}\, u_{h,m} = \delta_{nm}.$$

4.3 Convergence

So far we have two different problems. The continuous problem (the Dirichlet eigenvalues of the Laplace operator) has infinitely many solutions. The discrete problem (approximation by finite elements of the weak formulation of the eigenvalue problem) has a finite number of solutions. We will not deal in full detail with convergence of the discrete solutions to the exact solutions. We will however mention here two important properties. The first one is, let's say so, peculiar:

> *with the increasing order of continuous and discrete eigenvalues that takes into account their multiplicity, discrete eigenvalues always overestimate continuous eigenvalues*
> $$\lambda_n \le \lambda_{h,n}, \qquad n = 1, \ldots, N.$$

The second one is what we would expect from a discretization method:

> *discrete eigenvalues converge to continuous eigenvalues; for fixed (but arbitrary n)*
> $$\lambda_{h,n} \overset{h\to 0}{\longmapsto} \lambda_n,$$
> *if the triangulations become finer.*

You can think of you matrix eigenvalue problem as having infinetily many solutions: the N eigenvalues and then $\lambda_{h,N+1} = \lambda_{h,N+2} = \ldots = +\infty$. These non-finite eigenvalues obviously overestimate the corresponding continuous ones. When you increase the dimension of the space (you refine the mesh) you bring some newer eigenvalues from infinity. They begin to approximate the corresponding higher eigenvalues of the exact problems, which are larger as the dimension of the discrete space increases. Note that

$$\frac{\lambda_N}{\lambda_1} \approx \frac{\lambda_N}{\lambda_{h,1}} \le \frac{\lambda_{h,N}}{\lambda_{h,1}}.$$

This means that the ratio between the largest and smallest generalized eigenvalue of $\mathbf{W}$ diverges. Because the mass matrix is in principle well-conditioned, we can prove with this that the stiffness matrix is ill-conditioned. How bad the conditioning is depends on how fast the Dirichlet eigenvalues diverge. Anyway, you have to expect bad behavior of the stiffness matrix in anything that depends on conditioning.

5 Exercises

1. **Bad quadrilaterals.** Figure 4.7 will help you to solve both questions in this exercise.

   (a) Take four points that define a convex quadrilateral, given in rotating order: $\mathbf{p}_1^K$, $\mathbf{p}_2^K$, $\mathbf{p}_3^K$ and $\mathbf{p}_4^K$. (They could be, for instance, the vertices of the reference square). If we give the third and the fourth vertices in the wrong order to the transformation, what is the transformed figure we obtain?

(b) Take now the four vertices of a non-convex quadrilateral given in rotating order and consider the usual bilinear transformation from the reference square. Using the fact that vertical and horizontal lines in the reference square are mapped to straight lines in the physical element, what kind of figure are we mapping?

2. **Computation of the $\mathbb{P}_1$ lumped mass matrix.** We will go step by step. Using the three vertex formula compute exactly the following integral
$$\int_K N_\alpha^K.$$
Adding the previous results, prove that
$$\int_\Omega \varphi_i = \tfrac{1}{3}\,\mathrm{area}\,\big(\mathrm{supp}\,\varphi_i\big).$$
Prove now that the sum of all nodal basis functions is the unit function
$$\sum_j \varphi_j \equiv 1.$$
(To do this, compare nodal values of both functions and note that constant functions belong to V_h.) Finally use the following trick based on the preceding identity
$$\int_\Omega \varphi_i = \sum_j \int_\Omega \varphi_j \varphi_i$$
to prove that
$$\tfrac{1}{3}\,\mathrm{area}\,\big(\mathrm{supp}\,\varphi_i\big) = \sum_j m_{ij}$$
and therefore the i-th diagonal element of the lumped mass matrix can be computed by adding all the elements of the i-th row of the mass matrix.

3. **Generalized eigenvalues from FE approximation.** Assume that we are given a discrete problem
$$\left[\begin{array}{l} \text{find } \lambda_h \text{ such that there exists } 0 \neq u_h \in V_h^0 \text{ satisfying} \\ \displaystyle\int_\Omega \nabla u_h \cdot \nabla v_h = \lambda_h \int_\Omega u_h\, v_h \qquad \forall v_h \in V_h^0, \end{array}\right.$$
where V_h^0 is any subspace of $H_0^1(\Omega)$ for which we have a basis $\{\varphi_i \,:\, i = 1, \ldots, N\}$. Following the ideas of Lesson 1 (when we converted the Galerkin equations to a system of linear equations), prove that this problem is equivalent to the generalized eigenvalue problem
$$\left[\begin{array}{l} \text{find } \lambda_h \text{ such that there exists } \mathbf{0} \neq \mathbf{u} \in \mathbb{R}^N \text{ satisfying} \\ \mathbf{W}\mathbf{u} = \lambda_h \mathbf{M}\mathbf{u}. \end{array}\right.$$
for the matrices
$$w_{ij} = \int_\Omega \nabla\varphi_j \cdot \nabla\varphi_i, \qquad m_{ij} = \int_\Omega \varphi_j\, \varphi_i.$$

Lesson 5

Evolution problems

There are many different approaches in the application of finite element techniques to evolution problems. In fact, there are also many different types of evolution problems. In this lesson we are going to concentrate on two evolution equations:

- the heat equation, a good example of parabolic behavior (transient diffusion),
- the wave equation, the simplest model of hyperbolic equation of the second order.

We can group the FEM-based approaches for these evolution equations:

- methods that discretize space and time simultaneously, and
- methods that discretize one of these variables and then the other.

We are going to do as follows. We'll first take the heat equation and do time discretization with finite differences and then space discretization with finite elements. Afterwards we will see how discretization only of the space variable with FEM leads to a system of ordinary differential equations, for which you can use a great variety of methods. (This approach is often called the *method of lines.*) If we use a finite element type method for the time variable we end up with something very similar to applying a FEM discretization at the same time to space-and-time. Finally, we'll go for the wave equation and show some basic ideas.

1 Forward Euler FEM for the heat equation

First of all, let us state the problem. We are given a polygon Ω in the (x, y)-variable space. We are going to consider only Dirichlet boundary conditions, since they are usually more complicated. The extension to mixed boundary conditions is not very difficult, provided that the Dirichlet and Neumann boundaries remain fixed in time. The origin of times will be $t_0 = 0$. We will state the problem for all positive time values, from 0 to ∞, although we will be mainly thinking of solving the problem in a finite time interval $(0, T)$. The problem is the following: find $u(\mathbf{x},t) = u(x,y,t)$ such that

$$\left[\begin{array}{ll} u_t = \Delta_{\mathbf{x}} u + f, & \text{in } \Omega \times (0,\infty), \\ u(\,\cdot\,,0) = \omega, & \text{in } \Omega, \\ u(\,\cdot\,,t) = g, & \text{on } \Gamma \text{ for all } t > 0. \end{array}\right.$$

Many new things again, so let's go step by step:

- u_t is the partial derivative with respect to t and $\Delta_{\mathbf{x}} u$ is the Laplacian in the (x,y)-variables.
- $f : \Omega \times (0,\infty) \to \mathbb{R}$ is a given function of time and space.
- $g : \Gamma \times (0,\infty) \to \mathbb{R}$ is a given function of time and of the space variable on the boundary. It gives the enforced Dirichlet boundary condition for all times.
- When both f and g are independent of time, there is still evolution, but we will see that it is just a transient state converging to a steady-state solution. We'll talk about this on the section about stability.
- $\omega : \Omega \to \mathbb{R}$ is a function of the space variable and represents the initial condition.

In principle, we are going to make ourselves our life simple by assuming that u_0 is a continuous function (and we can evaluate it without difficulty) and that f and g_0 are continuous in the time variable.

1.1 Time semidiscretization with the forward Euler method

Let us first take a partition in time, which can be non-uniform (variable time-step)

$$0 = t_0 < t_1 < t_2 < \ldots < t_n < \ldots$$

If our time interval is finite (as it is for any practical problem), the partition finishes in a certain point t_M. This is not important right now. The local time-step is

$$\delta_n = t_{n+1} - t_n.$$

For us, doing a time step will be moving from an already computed approximation if time t_n to time t_{n+1}. The time-steps δ_n are given as if they were known from the beginning. Unless you are taking it to be uniform (which is not really a very good idea in most practical situations), time-steps are computed with information about the behavior of the numerical solution solution and about the performance of the method as we proceed in discrete time. For the sake of exposition, we do as if we already knew all time-steps in advance.

We freeze the source term and boundary data at each time t_n by simple evaluation

$$f^n = f(\,\cdot\,,t_n) : \Omega \to \mathbb{R}, \qquad g^n = g(\,\cdot\,,t_n) : \Gamma \to \mathbb{R}.$$

When the data functions are not continuous we should be willing to average in a time interval around t_n instead. Time semidiscretization strives to obtain approximations

$$u(\,\cdot\,, t_n) \approx u^n : \Omega \to \mathbb{R}.$$

The first attempt we will do is the forward (or explicit) Euler method. It consists of looking at the equation in discrete time n and approximating a time derivative in this time by the forward quotient

$$\phi'(t_n) \approx \frac{\phi(t_{n+1}) - \phi(t_n)}{t_{n+1} - t_n} = \frac{\phi(t_{n+1}) - \phi(t_n)}{\delta_n}.$$

If we take the heat equation and use this forward Euler approximation, we obtain the recurrence

$$\frac{u^{n+1} - u^n}{\delta_n} = \Delta u^n + f^n$$

or in explicit form

$$u^{n+1} = u^n + \delta_n \Delta u^n + \delta_n f^n.$$

This recurrence is started at $n = 0$ with the initial data function $u^0 = \omega$. Bullet points again:

- Note that all functions are functions of the space variable, so the Laplace operator in space variables is just the Laplace operator. Time is now discrete time and appears as the n superscript everywhere.

- In principle this formula gives you u^{n+1} from u^n and the source function. Where is g? We've lost it in the way! There seems to be no way of imposing the boundary condition without entering in open conflict with the recurrence.

- There's more. If you begin with u^0, you take two derivatives to compute u^1. Then another two to compute u^2 and so on and so on. You had better have many space derivatives available! How could we possibly think of approximating u^n by a finite element function, which only has the first derivatives?

The answer to the last two questions comes from the use of a weak formulation for the recurrence. The price will be losing this explicit recurrence character that made the forward Euler approximation really explicit.

Consider Green's Theorem applied to u^n. Yes, I know we want to compute u^{n+1} (we already know u^n). Follow me anyway. We have

$$\int_\Omega (\Delta u^n)\, v + \int_\Omega \nabla u^n \cdot \nabla v = \int_\Gamma (\partial_n u^n)\, v.$$

Multiply this by δ_n and for lack of knowledge of the Neumann boundary data function, impose v to be zero on the boundary. We have therefore

$$\delta_n \int_\Omega (\Delta u^n)\, v + \delta_n \int_\Omega \nabla u^n \cdot \nabla v = 0, \qquad \text{for all } v \text{ such that } v = 0 \text{ on } \Gamma.$$

Substitute now the Laplacian of u^n, that is

$$\delta_n \Delta u^n = u^{n+1} - u^n - \delta_n f^n$$

and move what you know (data and functions at time n) to the right-hand side to obtain

$$\int_\Omega u^{n+1} v = \int_\Omega u^n\, v - \delta_n \int_\Omega \nabla u^n \cdot \nabla v + \delta_n \int_\Omega f^n\, v, \qquad v = 0 \text{ on } \Gamma.$$

Now there seems to be room for imposing the missing Dirichlet boundary condition, implicitly at time $n+1$, since the test is satisfying the homogeneous Dirichlet boundary condition. The sequence of problems would be consequently: begin with u_0 and then for each n,

$$\left[\begin{array}{l} \text{find } u^{n+1} \in H^1(\Omega) \text{ such that} \\ u^{n+1} = g^{n+1}, \qquad \text{on } \Gamma, \\ \displaystyle\int_\Omega u^{n+1} v = \int_\Omega u^n\, v - \delta_n \int_\Omega \nabla u^n \cdot \nabla v + \delta_n \int_\Omega f^n\, v, \qquad \forall v \in H_0^1(\Omega). \end{array}\right.$$

Recall (last section of the previous Lesson) that

$$H_0^1(\Omega) = \{ v \in H^1(\Omega) \, : \, v = 0, \quad \text{on } \Gamma\}.$$

The problem looks more like what we have been solving so far[1]. Only there is no stiffness-term for the unknown, which is a problem (there is no way we will obtain ellipticity of the bilinear form in energy norm), and at the same time there is a stiffness term in the right–hand side, which is more complicated than usual. Don't worry, we are getting near something reasonable.

1.2 Full discretization

We are there. Take a finite element method. Any of the methods exposed in Lessons 1, 2 and 3 will do the job. We have the space

$$V_h \subset H^1(\Omega)$$

associated to a triangulation of the domain, a nodal basis, the concept of Dirichlet nodes (all nodes on the boundary) and the subspace

$$V_h^0 = V_h \cap H_0^1(\Omega) = \{ v_h \in V_h \, : \, v_h = 0, \quad \text{on } \Gamma\}.$$

Nodes are numbered as usual and we take two lists: Dir, the one of indices of Dirichlet nodes, and Ind, the remaining nodes. The Dirichlet nodes are then

$$\mathbf{p}_i, \qquad i \in \text{Dir}.$$

[1] For knowledgeable mathematicians, I know, this sequence of problems is giving you the creeps. It is so ill-posed! You will have to wait to the fully discretized problem to get some satisfaction.

The main point now is to substitute all the infinite–dimensional elements of the problem

$$\left[\begin{array}{l} \text{find } u^{n+1} \in H^1(\Omega) \text{ such that} \\ u^{n+1} = g^{n+1}, \qquad \text{on } \Gamma, \\ \displaystyle\int_\Omega u^{n+1} v = \int_\Omega u^n\, v - \delta_n \int_\Omega \nabla u^n \cdot \nabla v + \delta_n \int_\Omega f^n\, v, \qquad \forall v \in H_0^1(\Omega), \end{array}\right.$$

by the their discrete counterparts, which is easy: for each n we have to

$$\left[\begin{array}{l} \text{find } u_h^{n+1} \in V_h \text{ such that} \\ u_h^{n+1}(\mathbf{p}_i) = g^{n+1}(\mathbf{p}_i), \qquad \forall i \in \text{Dir}, \\ \displaystyle\int_\Omega u_h^{n+1} v_h = \int_\Omega u_h^n\, v_h - \delta_n \int_\Omega \nabla u_h^n \cdot \nabla v_h + \delta_n \int_\Omega f^n\, v_h, \qquad \forall v_h \in V_h^0. \end{array}\right.$$

This looks more like something we can do. Before going for matrices, we have to give a starting point for this recurrence: $u_h^0 \in V_h$ can be computed by interpolating in the nodes of the grid the initial data function ω. This is not the best option, but it is definitely the simplest one.

We need to reintroduce matrices and vectors to give a simpler idea of what we are doing here in each time step. The nodal values of u_h^n are given in the vector $\mathbf{u}^n$. They are divided into values on free/interior nodes $\mathbf{u}^n_{\text{Ind}}$ and values on the Dirichlet nodes $\mathbf{u}^n_{\text{Dir}}$. Actually, the Dirichlet condition states that

$$\mathbf{u}^{n+1}_{\text{Dir}} = \mathbf{g}^{n+1},$$

where $\mathbf{g}^{n+1}$ is the vector of values of $g^{n+1} = g(\,\cdot\,, t_{n+1})$ on Dirichlet nodes.

We are going to pick up two pieces of the mass matrix

$$\mathbf{M}_{\text{Ind}} = \Big[\int_\Omega \varphi_j \varphi_i \Big]_{i,j\in\text{Ind}}, \qquad \mathbf{M}_{\text{Dir}} = \Big[\int_\Omega \varphi_j \varphi_i \Big]_{i\in\text{Dir}, j\in\text{Ind}}.$$

The matrix $\mathbf{M}_{\text{Ind}}$ is square shaped, with as many rows as there are interior nodes. On the other hand $\mathbf{M}_{\text{Dir}}$ is rectangular, with as many rows as there are interior nodes and one column per Dirichlet node. We will glue them together in the rectangular matrix

$$\mathbf{M}_{\text{all}} = \left[\ \mathbf{M}_{\text{Ind}} \,\middle|\, \mathbf{M}_{\text{Dir}}\ \right].$$

This division is made so that we can write products

$$\mathbf{M}_{\text{all}} \mathbf{u}^{n+1} = \mathbf{M}_{\text{Ind}} \mathbf{u}^{n+1}_{\text{Ind}} + \mathbf{M}_{\text{Dir}} \mathbf{u}^{n+1}_{\text{Dir}}.$$

Its rectangular shape reflects the fact that testing with nodal basis function is ignored. We similarly construct the matrices $\mathbf{W}_{\text{Dir}}$, $\mathbf{W}_{\text{Ind}}$ and $\mathbf{W}_{\text{all}}$.

At this stage of the course we have seen this kind of arguments enough times so that you will easily recognize that the step in variational form

$$\left[\begin{array}{l} \text{find } u_h^{n+1} \in V_h \text{ such that} \\ u_h^{n+1}(\mathbf{p}_i) = g^{n+1}(\mathbf{p}_i), \qquad \forall i \in \text{Dir}, \\ \displaystyle\int_\Omega u_h^{n+1} v_h = \int_\Omega u_h^n\, v_h - \delta_n \int_\Omega \nabla u_h^n \cdot \nabla v_h + \delta_n \int_\Omega f^n\, v_h, \qquad \forall v_h \in V_h^0, \end{array}\right.$$

is the same as the system

$$\left[\begin{array}{l}\mathbf{u}_{\mathrm{Dir}}^{n+1} = \mathbf{g}^{n+1},\\ \mathbf{M}_{\mathrm{all}}\mathbf{u}^{n+1} = \mathbf{M}_{\mathrm{all}}\mathbf{u}^{n} - \delta_n \mathbf{W}_{\mathrm{all}}\mathbf{u}^{n} - \mathbf{f}^{n},\end{array}\right.$$

where $\mathbf{f}^n$ is the vector with elements

$$\int_\Omega f^n\,\varphi_i, \qquad i \in \mathrm{Ind}.$$

We can also write each step as the solution of the system

$$\mathbf{M}_{\mathrm{Ind}}\mathbf{u}_{\mathrm{Ind}}^{n+1} = \mathbf{M}_{\mathrm{all}}\mathbf{u}^{n} - \delta_n \mathbf{W}_{\mathrm{all}}\mathbf{u}^{n} - \mathbf{f}^{n} - \mathbf{M}_{\mathrm{Dir}}\mathbf{g}^{n+1}$$

to compute only values on free nodes. Values on Dirichlet nodes are incorporated to this formulation but we have also to remind ourselves to keep them in the full vector $\mathbf{u}^{n+1}$, that will be used in the next time–step.

1.3 Some features of the method

Making the method really explicit. It may come to you as a surprise to see that working the explicit equations of the forward Euler method with Finite Elements you end up with a system to be solved in each time–step, so the explicit method is not so explicit after all. Note however that:

- The matrix is always the same and it is always the mass matrix, so you have good conditioning of the system together with symmetry and positive definiteness.

- Therefore if you do some preprocess (a factorization of the matrix) or you build a good preconditioner, it's going to be useful for all time steps. Moreover, for each time step you have a linear system that you can try to solve with an iterative method (Conjugate Gradient looks like the best option), but you have a guess of the starting point for the iterations: why not begin with the value in the previous time?

- If you are not happy yet with the implicit character of these equations, you can substitute the mass matrix (at least the one that appears on the left hand side) by the **lumped mass matrix**, which is diagonal. A diagonal system is immediate to solve.

Diffusion or propagation of heat? There are some good reasons to make the method completely explicit: you compute the time steps faster, since you don't have to solve any linear system, no matter how well conditioned this is. There are reasons not to make it fully explicit. In fact the argument I'm going to give to you here is somewhat tricky and you'll have to take it with a grain of salt. The real reason for going implicit is given in the stability analysis.

Let us consider just the first time step in the case where $f \equiv 0$ and $g \equiv 0$. We only have to compute the free nodes in all steps, because the boundary condition is homogeneous.

Let us consider the $\mathbb{P}_1$ method and let us take a free node that is completely surrounded by free nodes. As initial condition we take an impulse in that node, that is, if the node is given the index i, we are starting with

$$u_h^0 = \varphi_i.$$

In matrix form we are beginning with the vector $\mathbf{e}_i$, that has all components zero but the i-th that is one. This is the system we solve:

$$\mathbf{M}_{\mathrm{Ind}}\mathbf{u}^1_{\mathrm{Ind}} = \mathbf{M}_{\mathrm{Ind}}\mathbf{e}_i - \delta_0 \mathbf{W}_{\mathrm{Ind}}\mathbf{e}_i.$$

Note that the i-th row of $\mathbf{M}_{\mathrm{Ind}}$ and $\mathbf{W}_{\mathrm{Ind}}$ is the only one used in the right-hand side. It contains non–zero elements only on the positions of adjacent (neighboring) nodes. The vector $\mathbf{M}_{\mathrm{Ind}}\mathbf{e}_i - \delta_0 \mathbf{W}_{\mathrm{Ind}}\mathbf{e}_i$ propagates the unit value on the i-th node to its neighboring nodes. All other elements of this vector are still zero.

If you do mass lumping, that's all that is going to be non–zero in $\mathbf{u}^1$. In the next step, we will reach the following set of neighbors (neighbors of neighbors of the i-th node). What we are doing here is propagating heat at finite speed: the physics are all wrong! Heat diffusion is done at infinite speed. A unit impulse in time zero heats all the domain at any positive time. In truth, the values far from the heating source are very, very small at small times, but they are non zero. If we keep the mass matrix without lumping, at least it looks like we can reach all nodes in the first time step. The reason is the fact that $\mathbf{M}_{\mathrm{Ind}}^{-1}$ has most (if not all) elements non–zero. The process is much more similar to diffusion, although what we call diffusion, that's done by $\mathbf{W}_{\mathrm{Ind}}^{-1}$. But I cannot explain why right now.

Changing spaces with time. In some cases, with highly varying source terms and boundary conditions it could be wiser to change the finite element space from time to time, maybe even at all time–steps[2]. Think in the step $n \mapsto (n+1)$. Assume that u_h^n we computed with a $\mathbb{P}_1$ finite element on a given triangulation. The space is denoted $V_{h,n}$ and $V_{h,n}^0$ is the subspace obtained by eliminating the Dirichlet nodes. For whichever the reason, we are going to change grid and compute u_{n+1}^h in a new space $V_{h,n+1}$. In principle these are the discrete variational equations:

$$\left[\begin{array}{l} \text{find } u_h^{n+1} \in V_{h,n+1} \text{ such that} \\ u_h^{n+1}(\mathbf{p}_i) = g^{n+1}(\mathbf{p}_i), \qquad \forall i \in \mathrm{Dir}(n+1), \\ \displaystyle\int_\Omega u_h^{n+1} v_h = \int_\Omega u_h^n\, v_h - \delta_n \int_\Omega \nabla u_h^n \cdot \nabla v_h + \delta_n \int_\Omega f^n\, v_h, \qquad \forall v_h \in V_{h,n+1}^0. \end{array}\right.$$

It looks the same but it isn't exactly the same. If you add a superindex with the discrete time to the nodal bases, you will see that in the left–hand side, you have a usual mass matrix for the current space

$$\int_\Omega \varphi_j^{n+1}\varphi_i^{n+1}.$$

[2] This change of space with time is the daily bread in finite element methods for waves, but since we are taking the heat equation as the first model problem, it's okay if we have a look at this here.

However, since u_n^h was computed on the old grid, the two matrices that appear on the right–hand side are

$$\int_\Omega \varphi_j^n \varphi_i^{n+1} \qquad \text{and} \qquad \int_\Omega \nabla\varphi_j^n \cdot \nabla\varphi_i^{n+1}.$$

These matrices do not need even to be square. But there's more. The very nice idea of assembly is much more complicated if the triangulations are not related and what we did in Lesson 2 is definitely not valid here anymore. With this naïve approach things really get messy.

What can be done in practice is taking a very different approach, consisting of pre-processing the solution in time n to move it to the grid of time $n+1$. In essence it is like interpolating u_h^n to the new space $V_{h,n+1}$. This can be a somewhat complicated process but has the advantage that the u_h^n we input in the right–hand side is now in the same space as the u_h^{n+1} we want to compute and the assembly process can be used again.

1.4 Stability analysis

Let's simplify again the problem to have source term and boundary conditions that do not depend on time. The problem is therefore

$$\left[\begin{array}{ll} u_t = \Delta_{\mathbf{x}} u + f, & \text{in } \Omega \times (0,\infty), \\[1ex] u(\,\cdot\,,0) = \omega, & \text{in } \Omega, \\[1ex] u(\,\cdot\,,t) = g, & \text{on } \Gamma \text{ for all } t > 0, \end{array}\right.$$

with f and g independent of time. If we ignore the initial condition we can look for the only steady–state solution to the problem

$$\left[\begin{array}{ll} -\Delta u_{\text{lim}} = f, & \text{in } \Omega, \\[1ex] u_{\text{lim}} = g, & \text{on } \Gamma. \end{array}\right.$$

Assume now that we know all Dirichlet eigenvalues and eigenfunctions of the Laplace operator in Ω:

$$\left[\begin{array}{ll} -\Delta \phi_k = \lambda_k \phi_k, & \text{in } \Omega, \\[1ex] \phi_k = 0. & \text{on } \Gamma. \end{array}\right.$$

The solution to the heat diffusion problem is

$$u(\mathbf{x},t) = u_{\text{lim}}(\mathbf{x}) + \sum_{k=1}^\infty c_k\, e^{-\lambda_k t} \phi_k(\mathbf{x}), \qquad c_k = \int_\Omega (\omega - u_{\text{lim}})\, \phi_k.$$

This formula[3] shows that the solution goes exponentially fast to the steady–state solution. The occurrence of negative exponentials at increasing velocities (λ_k diverges as k goes to infinity) makes the initial times very hard to compute with precision.

[3] You might (should) recognize it from your course(s) on differential equations. It is the solution obtained by separation of variables

In case we are dealing with zero data

$$f \equiv 0, \qquad g \equiv 0,$$

the formula for the solution is really simple: it's just diffusion of the initial condition towards the zero solution

$$u(\mathbf{x}, t) = \sum_{k=1}^{\infty} c_k \, e^{-\lambda_k \, t} \phi_k(\mathbf{x}), \qquad c_k = \int_\Omega \omega \, \phi_k.$$

Let us see what the numerical method does. Since boundary conditions vanish we don't have to take into account Dirichlet nodes. In the n-th time-step we solve

$$\mathbf{M}_{\mathrm{Ind}} \mathbf{u}_{\mathrm{Ind}}^{n+1} = \mathbf{M}_{\mathrm{Ind}} \mathbf{u}_{\mathrm{Ind}}^{n} - \delta_n \mathbf{W}_{\mathrm{Ind}} \mathbf{u}_{\mathrm{Ind}}^{n}.$$

Let us drop the Ind subindex and keep in mind that we are only computing in the interior nodes. Also for simplicity assume that $\delta_n = \delta$ for all n, that is, we are using a fixed time step. This is the very simple n-th time step:

$$\mathbf{M}\mathbf{u}^{n+1} = \mathbf{M}\mathbf{u}^{n} - \delta \mathbf{W}\mathbf{u}^{n}.$$

There is only a finite number of linearly independent eigenvectors (that are nodal values of the discrete eigenvectors):

$$\mathbf{W}\boldsymbol{\phi}_k = \lambda_{h,k} \mathbf{M}\boldsymbol{\phi}_k.$$

Maybe you should go back to Section 4 of Lesson 4 to review this. Recall that $\lambda_{h,k} \geq \lambda_k$ approximates this k-th exact eigenvalue for h sufficiently small. Take $\mathbf{u}^0 = \boldsymbol{\phi}_k$ as initial condition in the recurrence that determines the discrete time steps. Then the equation for the first time step is

$$\mathbf{M}\mathbf{u}^{1} = \mathbf{M}\boldsymbol{\phi}_k - \delta \mathbf{W}\boldsymbol{\phi}_k = (1 - \lambda_{h,k}\delta) \, \mathbf{M}\boldsymbol{\phi}_k.$$

Therefore, using the fact that $\mathbf{M}$ is invertible, we have $\mathbf{u}^1 = (1 - \delta\lambda_{h,k})\boldsymbol{\phi}_k$. The following time steps are similar and we obtain the following formula for all the time steps

$$\mathbf{u}^n = (1 - \lambda_{h,k}\delta)^n \boldsymbol{\phi}_k.$$

Note that $\lambda_{h,k}$ is trying to approximate λ_k and $\boldsymbol{\phi}_k$ is trying to approximate the nodal values of ϕ_k. The formula for the recurrence is trying to approximate the diffusive solution

$$e^{-\lambda_k \delta n} \phi_k = e^{-\lambda_k t_n} \phi_k.$$

Is it doing a good job? Independently of whether this approximation is good or not, let us just look at the asymptotic behavior. The exact solution goes to zero as n goes to infinity. What about the discrete solution? Well, not always. It will do the right thing if

$$|1 - \lambda_{h,k}\delta| < 1,$$

which is equivalent (note that δ and $\lambda_{h,k}$ are positive) to

$$\lambda_{h,k}\delta < 2.$$

This should be satisfied for all discrete eigenvalues. Since we have ordered them from smallest to largest, it has to be satisfied by the largest of them

$$\lambda_{h,N(h)}\delta < 2.$$

Why do I say that it has? The fact is that any initial condition can be decomposed as

$$\mathbf{u}^0 = \sum_{k=1}^{N(h)} c_k \boldsymbol{\phi}_k$$

and the corresponding discrete evolution is therefore

$$\mathbf{u}^n = \sum_{k=1}^{N(h)} c_k (1 - \lambda_{h,k}\delta)^n \boldsymbol{\phi}_k.$$

The orthogonality condition of the discrete eigenvector proves that $\mathbf{u}^n$ goes to zero as $n \to \infty$ (that is, it has the correct asymptotic value) if and only if all conditions $\lambda_{h,k}\delta < 2$ hold.

Let's discuss the condition

$$\lambda_{h,N(h)}\delta < 2.$$

If we take the fixed time-step to begin with, the condition is of the form

$$\lambda_{h,N(h)} < 2/\delta.$$

Note that $\lambda_{h,N(h)} \geq \lambda_{N(h)}$. If we take a very fine grid (a very precise finite element method) it is very likely that you are getting to capture a very large eigenvalue and the stability condition does not hold any longer. This **conditional stability** says that given the time-step you can only try to do *this good* with finite elements, but if you try to be too precise you lose stability. This may be a shock to you. One would think that each part of the discretization process can be done as precisely as possible without taking care of the others. The conditional stability denies that.

If you fix the finite element grid, the inequality can be read as

$$\delta < 2/\lambda_{h,N(h)}$$

which says that you have to take time-steps that are short enough in order not to lose stability[4]. It is difficult to make oneself an idea of how the largest discrete eigenvalue grows with finer grids. For the one dimensional problem the precise formula is known. Given the variety of polygonal domains you can think of, the question is less clear in two dimensions.

[4] People in the ODE discretization community call this problem stiff and say that explicit methods are not linearly stable and should not be applied (or applied with great care) to stiff problems. More on this in Section 3.

Remark. In fact, the term $(1-\lambda_{h,k}\delta)^n$ can be oscillating even when going to zero, so we even might like it to be positive in addition to convergent to zero. The condition is then $\lambda_{h,k}\delta < 1$. □

What else? Convergence of course. Well, let's not do this here. The issue becomes really difficult. Note only that: (a) use of forward Euler in time means you should expect no more that error proportional to time step (order one); (b) the effort made in the space discretization should agree with the low order in time; (c) imposition of non-homogeneous Dirichlet conditions becomes openly critical here. Doing the simplest thing here makes you lose convergence order. You have to look at the theory (and we are so not going to to that now) to understand why. Anyway, never use high order in space with low order in time. You are wasting your efforts. Second, be careful with stability. You don't have it for free! If you have fixed your time-step you cannot try to be too precise in space.

2 Backward Euler FEM for the heat equation

It has taken time, but we have had a close look at the very simplest discretization method for the heat equation. If you have your FEM code for the steady state problem, it is easy to create a FEM code for the forward Euler and FEM discretization of the heat equation. We move now to improve our method.

2.1 Time semidiscretization with the backward Euler method

First thing we have to improve is conditional stability. That condition is definitely not the best thing to have, in particular since you really don't know precisely whether it holds or not unless you compute the largest generalized eigenvalue of $\mathbf{W}$.

We begin from scratch. Almost. The backward Euler discretization uses the same quotient as the forward Euler method but to approximate the value of the derivative in discrete time $(n+1)$

$$\phi'(t_{n+1}) \approx \frac{\phi(t_{n+1})-\phi(t_n)}{t_{n+1}-t_n} = \frac{\phi(t_{n+1})-\phi(t_n)}{\delta_n}.$$

Correspondingly, we look at the heat equation in time t_{n+1} and impose the backward Euler approximation

$$\frac{u^{n+1}-u^n}{\delta_n} = \Delta u^{n+1} + f^{n+1},$$

or equivalently

$$-\delta_n \Delta u^{n+1} + u^{n+1} = u^n + \delta_n f^{n+1}.$$

Let's not forget the boundary condition, which now enters the game in a more standard way

$$u^{n+1} = g^{n+1}, \qquad \text{on } \Gamma.$$

Equation and boundary condition constitute a boundary value problem like those we have been studying all along this course. Note that the diffusion parameter is the time step (it is very small) but that this parameter is also multiplying the source term. If you formally take it to zero, what you obtain is a constant solution, which is what happens with evolution when you stop the clock counting times.

The boundary value problem to obtain u_{n+1} has nothing special. Its weak formulation is done in the usual way, as if there was no time in the equation

$$\left[\begin{array}{l} \text{find } u^{n+1} \in H^1(\Omega) \text{ such that} \\ u^{n+1} = g^{n+1}, \qquad \text{on } \Gamma, \\ \displaystyle \delta_n \int_\Omega \nabla u^{n+1} \cdot \nabla v + \int_\Omega u^{n+1} v = \int_\Omega u^n\, v + \delta_n \int_\Omega f^{n+1}\, v, \qquad \forall v \in H^1_0(\Omega). \end{array}\right.$$

2.2 Full discretization

Taking the finite element space instead of the exact Sobolev space, we obtain a sequence of problems

$$\left[\begin{array}{l} \text{find } u_h^{n+1} \in V_h \text{ such that} \\ u_h^{n+1}(\mathbf{p}_i) = g^{n+1}(\mathbf{p}_i), \qquad \forall i \in \text{Dir}, \\ \displaystyle \delta_n \int_\Omega \nabla u_h^{n+1} \cdot \nabla v_h + \int_\Omega u_h^{n+1} v_h = \int_\Omega u_h^n\, v_h + \delta_n \int_\Omega f^{n+1}\, v_h, \qquad \forall v_h \in V_h^0. \end{array}\right.$$

The recurrence (the time-steps) has to be started with an initial condition of u_h^0 given, as we had in the explicit method. You can go back to the previous section and you will notice that the only serious change is the stiffness term changing sides. It is implicit now.

Using the same notations for the vectors of unknowns and for the pieces of the matrices, we have a fully implicit method now

$$\left[\begin{array}{l} \mathbf{u}_{\text{Dir}}^{n+1} = \mathbf{g}^{n+1}, \\ \Big(\delta_n \mathbf{W}_{\text{all}} + \mathbf{M}_{\text{all}}\Big)\mathbf{u}^{n+1} = \mathbf{M}_{\text{all}}\mathbf{u}^n + \mathbf{f}^{n+1}, \end{array}\right.$$

Note again that the stiffness matrix has changed sides in the system. The system to be solved in each time step is actually

$$\Big(\delta_n \mathbf{W}_{\text{Ind}} + \mathbf{M}_{\text{Ind}}\Big)\mathbf{u}_{\text{Ind}}^{n+1} = \mathbf{M}_{\text{all}}\mathbf{u}^n + \mathbf{f}^{n+1} - \Big(\delta_n \mathbf{W}_{\text{Dir}} + \mathbf{M}_{\text{Dir}}\Big)\mathbf{g}^{n+1}.$$

You can take from here a first idea: the cost of programming the forward and the backward Euler is exactly the same. The main difference is that in the implicit method you have to solve a linear system in each time step and there is not diagonal approximation for the corresponding matrix. The matrix itself varies with time-step, but if you have to look for a preconditioner, you just have to take care of the stiffness matrix, which is the bad guy here (mass=good, stiffness=bad) in terms of conditioning. For fixed time stepping, the matrix is always the same, by the way.

If you put a point source in time zero, it diffuses instantaneously to the whole domain thanks to the inverse of the matrix of the system.

2.3 Stability analysis

With vanishing boundary conditions and zero sources as well as with fixed time-step we solve the recurrence

$$(\delta\mathbf{W}+\mathbf{M})\mathbf{u}^{n+1}=\mathbf{M}\mathbf{u}^n$$

to follow the free evolution of the system with initial condition $\mathbf{u}^0$. If $\mathbf{u}^0=\boldsymbol{\phi}_k$ (we use the same notation as in the corresponding subsection for the forward method), then the eigenvectors satisfy

$$(\delta\mathbf{W}+\mathbf{M})\boldsymbol{\phi}_k=(1+\lambda_{h,k}\delta)\mathbf{M}\boldsymbol{\phi}_k.$$

Therefore, it is simple to check that

$$\mathbf{u}^n=(1+\lambda_{h,k}\delta)^{-n}\boldsymbol{\phi}_k.$$

This discrete evolution is always correct, since $0<(1+\lambda_{h,k}\delta)^{-1}<1$. The method is therefore **unconditionally stable**. Expected convergence is similar to the one of the forward Euler approximation, since both time discretizations have the same order. What changes here is stability.

3 Doing first space and then time

In one of the exercises I've proposed to have a look at the scheme developed by John Crank and Phyllis Nicolson using the same quotient to approximate the average of the derivatives in both points. It leads to a sort of average of the forward and backward Euler methods[5]. This is an easy way of increasing order of convergence in time: formally it goes up to order two. Doing better with finite differences in time requires using more time points for each steps. We could also forget about finite differences in time and do Galerkin (finite elements) also in that variable.

Instead we are going to try something else. The following approach is the origin of many ideas but definitely requires that your space triangulation remains fixed, so forget about it if things are changing really fast and you want to remesh from time to time.

We are back to the heat diffusion problem. Here it is again

$$\left[\begin{array}{ll} u_t=\Delta_{\mathbf{x}}u+f, & \text{in } \Omega\times(0,\infty),\\[6pt] u(\,\cdot\,,0)=\omega, & \text{in } \Omega,\\[6pt] u(\,\cdot\,,t)=g, & \text{on } \Gamma \text{ for all } t>0.\end{array}\right.$$

For the moment, let us think of time as an additional parameter, forget the initial condition and deal with this as an elliptic problem. For each t, the space function $u=u(\,\cdot\,,t)$ (mathematicians, forgive me for not changing the name) satisfies:

$$\left[\begin{array}{ll} -\Delta u+u_t=f, & \text{in } \Omega,\\[6pt] u=g, & \text{on } \Gamma.\end{array}\right.$$

[5] Note that properly speaking the Crank-Nicolson scheme uses also finite differences for the space variables.

Using Green's Theorem we obtain a weak formulation

$$\left[\begin{array}{l} u = g, \qquad \text{on } \Gamma, \\ \displaystyle\int_\Omega \nabla u \cdot \nabla v + \int_\Omega u_t\, v = \int_\Omega f\, v, \qquad \forall v \in H_0^1(\Omega). \end{array}\right.$$

Hey, teacher! Your forgot to write the space for u! No, I didn't. We can try to think of u as a function that for each t, gives an element of $H^1(\Omega)$, but I really prefer not to write the correct spaces. First of all, because they are complicated. Second,... because they are complicated, if we want to have the right spaces where we are certain to have a solution and not some safe spaces where everything looks nice but we will never be able to show that there is a solution.

Instead, let us go to discretization. The idea is the same: for each time we associate a function in the finite element space (it will be the same space for all times). So, fix V_h and V_h^0 as usual. A time-dependent element of V_h is something of the form

$$u_h(t, \mathbf{x}) = \sum_{j=1}^N u_j(t)\, \varphi_j(\mathbf{x}).$$

The coefficients vary with time, but the global basis is always the same since the triangulation is fixed. In fact, when we are dealing with the nodal basis functions $u_j(t) = u_h(t, \mathbf{p}_j)$, so we are following the nodal values of the discrete function. The partial derivative of this function with respect to time is

$$\sum_{j=1}^N \dot{u}_j\, \varphi_j.$$

Then, the semidiscrete in space problem looks for u_h such that for all t

$$\left[\begin{array}{l} u_h(\,\cdot\,, t) \in V_h, \\ u_h(\mathbf{p}, t) = g(\mathbf{p}, t), \qquad \text{for every Dirichlet node } \mathbf{p}\ , \\ \displaystyle\int_\Omega \nabla_{\mathbf{x}} u_h \cdot \nabla v_h + \int_\Omega u_{h,t}\, v_h = \int_\Omega f\, v_h, \qquad \forall v \in V_h^0. \end{array}\right.$$

We also need an initial condition

$$u_h(\,\cdot\,, 0) = \sum_j u_j(0) \varphi_h = u_h^0,$$

where $u_h^0 \in V_h$ approximates the initial condition ω. If we decide ourselves for interpolating data, this means that we are giving an initial condition to the coefficients

$$u_j(0) = \omega(\mathbf{p}_j), \qquad \forall j.$$

The problem can be easily written using these coefficients

$$\left[\begin{array}{l} u_i(t) = g(\mathbf{p}_i, t), \qquad \forall i \in \text{Dir}, \\ \displaystyle\sum_{j=1}^N \Big(\int_\Omega \nabla\varphi_j \cdot \nabla\varphi_i\Big) u_j(t) + \sum_{j=1}^N \Big(\int_\Omega \varphi_j\, \varphi_i\Big) \dot{u}_j(t) = \int_\Omega f\, \varphi_i, \qquad \forall i \in \text{Ind}. \end{array}\right.$$

This system holds for all t. This is a somewhat non-standard but simple differential system. We can get rid of the algebraic (non-standard) part by simply substituting the Dirichlet conditions inside the formulation to obtain

$$\sum_{j\in\mathrm{Ind}} w_{ij}u_j(t) + \sum_{j\in\mathrm{Ind}} m_{ij}\dot{u}_j(t) = \int_\Omega f\varphi_i - \sum_{j\in\mathrm{Dir}}\Big(w_{ij}g(\mathbf{p}_j,t) + m_{ij}g_t(\mathbf{p}_j,t)\Big), \qquad \forall i \in \mathrm{Ind}.$$

This looks much more like a system of linear differential equations. Let us simplify the expression by improving notation. We consider the following functions of time:

$$f_i(t) = \int_\Omega f(\,\cdot\,,t)\varphi_i, \qquad i \in \mathrm{Ind},$$

$$g_j(t) = g(\mathbf{p}_j,t), \qquad j \in \mathrm{Dir}.$$

The system is therefore

$$\sum_{j\in\mathrm{Ind}} w_{ij}u_j(t) + \sum_{j\in\mathrm{Ind}} m_{ij}\dot{u}_j(t) = f_i(t) - \sum_{j\in\mathrm{Dir}}\Big(w_{ij}g_j(t) + m_{ij}\dot{g}_j(t)\Big), \qquad \forall i \in \mathrm{Ind}.$$

You will have noticed that this way of discretizing the problem, imposes the need to compute the time derivative of the Dirichlet data. It's because they are essential (Neumann data would appear like source terms, happily placed inside integral signs). If you want to avoid this derivative of data, you have to deal with the algebraic-differential system as was first obtained.

Using the matrix notation introduced in the first section of this lesson, we can write

$$\mathbf{W}_{\mathrm{Ind}}\mathbf{u}_{\mathrm{Ind}} + \mathbf{M}_{\mathrm{Ind}}\dot{\mathbf{u}}_{\mathrm{Ind}} = \mathbf{f} - \mathbf{W}_{\mathrm{Dir}}\mathbf{g} - \mathbf{M}_{\mathrm{Dir}}\dot{\mathbf{g}}.$$

Now, we write everything together, more in the style of how we write differential systems:

$$\left[\begin{array}{l} \mathbf{u}_{\mathrm{Ind}}(0) = \mathbf{u}_0, \\ \mathbf{M}_{\mathrm{Ind}}\dot{\mathbf{u}}_{\mathrm{Ind}} = -\mathbf{W}_{\mathrm{Ind}}\mathbf{u}_{\mathrm{Ind}} + \mathbf{f} - \mathbf{W}_{\mathrm{Dir}}\mathbf{g} - \mathbf{M}_{\mathrm{Dir}}\dot{\mathbf{g}}. \end{array}\right.$$

This is a linear system of differential equations (with initial values) given in implicit form. To make it explicit you would have to premultiply by $\mathbf{M}_{\mathrm{Ind}}^{-1}$. In principle you don't have to compute the inverse of the mass matrix to know how to multiply by it. The reason is the fact that

the vector $\mathbf{M}_{\mathrm{Ind}}^{-1}\mathbf{v}$ *is the solution to the system* $\mathbf{M}_{\mathrm{Ind}}\mathbf{x} = \mathbf{v}$.

Therefore, you just need to know how to solve linear systems with $\mathbf{M}_{\mathrm{Ind}}$ as matrix. You don't even need that much. Most packages that solve numerically systems of differential equations (with Runge-Kutta methods for instance) already consider the implicit situation, where the derivative is premultiplied by an invertible matrix.

This approach allows you to use high order in space and high order in time very easily, because the processes are separated. In fact, many people in the numerical ODE community use the heat equation after space discretization as a benchmark for their methods, since the resulting system is stiff. Remember all those fastly decaying exponentials in the separation of variable solutions? In the differential system they become large negative eigenvalues, which are difficult to handle. For stiff problems, the safe bet is the use implicit methods. Anything explicit will be understandably conditionally convergent, requiring short time steps or a very rigid step control strategy.

Remark. If you apply the forward or backward Euler method to this differential system you obtain the methods you had in Sections 1 and 2 if:

- g is independent of time
- g depends on time but you substitute the occurrence of $\dot{\mathbf{g}}$ in the n-th time step by the quotient $(\mathbf{g}_{n+1} - \mathbf{g}_n)/\delta_n$.

This coincidence of lower order methods in the simplest cases is something you find over and over in numerical analysis. □

4 Some ideas about the wave equation

There is a long stretch since the beginning of this course, ninety-something pages ago. We need to put an end to it, but it would be wrong (for me) to end a lesson of evolution problems with nothing on the wave equation[6]. You'll see how this is very simple to introduce. To make it simpler we will use homogeneous Dirichlet conditions in the entire boundary of the domain.

The wave propagation problem is then

$$\left[\begin{array}{ll} u_{tt} = \Delta_{\mathbf{x}} u + f, & \text{in } \Omega \times (0,\infty), \\ u(\,\cdot\,,0) = u_0, & \text{in } \Omega, \\ u_t(\,\cdot\,,0) = v_0, & \text{in } \Omega, \\ u(\,\cdot\,,t) = 0, & \text{on } \Gamma \text{ for all } t > 0. \end{array}\right.$$

If we try the finite difference in time approach, the simplest thing to do is to apply the central difference approximation (some people call this Newmark's method[7]) to the second derivative. If we take a fixed time step, this means approximating

$$\phi''(t_n) \approx \frac{\phi(t_{n+1}) - 2\phi(t_n) + \phi(t_{n-1})}{\delta^2}.$$

When applied to the time-variable in the wave equation we obtain the explicit time step

$$\frac{u_{n+1} - 2u_n + u_{n-1}}{\delta^2} = \Delta u_n + f_n.$$

After doing the weak formulation and introducing finite element spaces and bases, we end up with

$$\mathbf{M}\mathbf{u}_{n+1} = 2\mathbf{M}\mathbf{u}_n - \mathbf{M}\mathbf{u}_{n-1} - \delta^2 \mathbf{W}\mathbf{u}_n + \delta^2 \mathbf{f}_n.$$

[6] You can easily claim that I'm not dealing with conservation laws either. True. You are right. That's not my turf.

[7] As far as I know about this, the method proposed by Nathan Newmark is something more general destined to approximate second order equations. There is however a developed habit of calling this central difference approximation for the time derivative in the wave equation, Newmark's method.

(Only free nodes appear in all the expressions, since we have taken homogeneous Dirichlet boundary conditions). The initial value for $\mathbf{u}_0$ is easy. You have data. You still need $\mathbf{u}_1$ (the nodal values of u_1^h. For that, you can do very easy (and not very well) by taking a Taylor approximation

$$u_1 = u_0 + \delta v_0,$$

or take a false discrete time -1 and use the equation

$$\frac{u_1 - 2u_0 + u_{-1}}{\delta^2} = \Delta u_0 + f_0$$

together with the central difference approximation

$$\frac{u_1 - u_{-1}}{2\delta} = v_0$$

to obtain the equation

$$u_1 = \tfrac{1}{2}\delta^2 \Delta u_0 + u_0 + \delta v_0 + \tfrac{1}{2}\delta^2 f_0.$$

Then you need to give a weak formulation of this too. And do all the finite element stuff. Nothing you don't know how to do. Some really fast last strokes:

- Space discretization has made the equations implicit but it's only with the mass matrix. To obtain the good physics (finite velocity of propagation), the use of the lumped mass matrix is highly recommended. Wait for a couple of points to know more about this.

- The method is explicit so it is going to be **conditionally stable**. The stability condition is a bit harder to derive in this situation. It reads like

  $$\delta^2 \lambda_{h,N} < 4$$

  and it is called a Courant-Friedrichs-Lewy condition[8] and always refered by the initials CFL condition.

- Things with the wave equation happen quite fast so most people are willing to accept the short time-step imposed by a CFL condition, since they want to observe the propagation anyway.

- Implicit methods have the advantage of unconditional stability but get the physics wrong. When you are trying to follow the propagation of wave-fronts you sort of dislike the diffusion that would be caused by the presence of the inverse of the stiffness matrix.

- Wave propagation is however a delicate matter. If you take the explicit method, made fully explicit by the use of mass lumping, you move (in the $\mathbb{P}_1$ method) from node to node in each time step. That is, the speed of numerical propagation is

[8] We have already met Richard Courant, moral father of the $\mathbb{P}_1$ element. Now, meet Kurt Friedrichs and Hans Lewy. All three of them were German (Lewy's birthplace counts as Poland nowadays) and moved to America.

controlled by the time step. If you take a very, very short time-step to be sure that you are satisfying the CFL condition, you may be going too fast, so you have to play it safe but not too safe. This balance between stability and correct speed of propagation makes the discretization of wave phenomena a difficult but extremely interesting problem.

5 Exercises

1. **Crank-Nicolson and FEM for the heat equation.** The Crank-Nicolson[9] scheme consists of using the quotient to approximate the average of the derivative in t_n and t_{n+1}:

$$\frac{1}{2}\phi'(t_{n+1}) + \frac{1}{2}\phi'(t_n) \approx \frac{\phi(t_{n+1}) - \phi(t_n)}{t_{n+1} - t_n} = \frac{\phi(t_{n+1}) - \phi(t_n)}{\delta_n}.$$

We can apply this to the heat equation and propose this problem as n-th time step:

$$\left[\begin{array}{l} \dfrac{u_{n+1} - u_n}{\delta_n} = \dfrac{1}{2}\Big(\Delta u_n + \Delta u_{n+1}\Big) + \dfrac{1}{2}(f_n + f_{n+1}), \qquad \text{in } \Omega \\ u_{n+1} = g_{n+1}, \qquad \text{on } \Gamma. \end{array}\right.$$

- Write the preceding time-step as a reaction-diffusion problem to compute u_{n+1}.
- Write a weak formulation taking care of not having the Laplacian of u_n in the right-hand side but a stiffness term (you will have to use Green's formula twice, once in t_n and once in t_{n+1}).
- Write the discrete equations obtained from the FEM discretization of the weak formulation.
- Show that the method is unconditionally stable (use the same particular case: fixed time-step, $f \equiv 0$ and $g \equiv 0$).

2. **Full discretization of the wave equation.** We have already said that from the three terms recurrence

$$\frac{u_{n+1} - 2u_n + u_{n-1}}{\delta^2} = \Delta u_n + f_n,$$

a finite element method gives you this other full discrete three-term recurrence

$$\mathbf{M}\mathbf{u}_{n+1} = 2\mathbf{M}\mathbf{u}_n - \mathbf{M}\mathbf{u}_{n-1} - \delta^2\mathbf{W}\mathbf{u}_n + \delta^2\mathbf{f}_n.$$

Prove it. (You just have to follow step by step what we did for the heat equation and the forward Euler discretization. Note again the we have dropped the subscript Ind everywhere.)

[9] This method is named after John Crank and Phyllis Nicolson in the aftermath of WWII. The Mathematics building at Brunel University in West London is called after John Crank, a fact that, apparently, many students there find hilarious.

3. **Space semidiscretization of the wave equation.** We begin again

$$\left[\begin{array}{ll} u_{tt} = \Delta_{\mathbf{x}} u + f, & \text{in } \Omega \times (0,\infty), \\ u(\,\cdot\,,0) = u_0, & \text{in } \Omega, \\ u_t(\,\cdot\,,0) = v_0, & \text{in } \Omega, \\ u(\,\cdot\,,t) = 0, & \text{on } \Gamma \text{ for all } t > 0. \end{array}\right.$$

(Note that we have homogeneous Dirichlet boundary conditions). Taking the approach of space-first, prove that we arrive at a system of differential equations of the second order:

$$\mathbf{M}_{\mathrm{Ind}}\ddot{\mathbf{u}}_{\mathrm{Ind}} + \mathbf{W}_{\mathrm{Ind}}\mathbf{u}_{\mathrm{Ind}} = \mathbf{f}.$$

You just have to follow carefully the same process for the heat equation, with the additional simplification of having zero boundary conditions. To finish, note that we have two initial conditions that we can incorporate to the differential system.

Lesson 6

A posteriori error estimation and adaptivity

The general goal of **a posteriori error estimation** is the search for a computable quantity which is cheap to calculate from the numerical solution, and which gives an upper bound of the error committed in the simulation without being too distant from this. In this lesson we are going to review some basic methods for a posteriori error estimation and we will see what can be expected from them. We will also learn something about **adaptivity**, that is, mesh-refinement led by an a posteriori error estimate that leads to an automatic sequence of solutions that approaches the exact solution at the desired level of precision. I want to emphasize that this area is hot at the time of writing these notes and that there are still many difficult open questions in the understanding of adaptivity.

1 Goals and terminology

1.1 Error estimators

If u_h is the Finite Element solution of a problem whose exact solution is u and the natural norm for the problem (the one in which we have obtained a priori estimates) is $\|\cdot\|$, then an a posteriori error estimate is any quantity

$$\mathrm{Est}_h := \mathrm{Est}(u_h, \mathrm{data})$$

such that there exist two positive constants C_1, C_2, independent of the solution and of the data, and such that

$$C_1 \,\mathrm{Est}_h \le \|u - u_h\| \le C_2 \,\mathrm{Est}_h.$$

The second inequality is probably the most important one: it is called **reliability** of the estimator and it says that the estimator gives an upper bound of the error. The first inequality (**efficiency**) shows the impossibility that the estimator rejects a well-computed solution. One desirable property of a good a posteriori estimator is the closeness of C_1 and C_2 to the value one. In fact, there is an entire school of thought emphasizing estimators for which the upper bound constant C_2 is explictly known.

Sometimes **reliability** is not reachable for a particular estimator and we accpet inequalities of the form

$$\|u - u_h\| \le C_2 \operatorname{Est}_h + \operatorname{Osc}(h),$$

where the term $\operatorname{Osc}(h)$ (oscillation error) must converge to zero and can be estimated using the problem data, and not its solution. This term is typically related to the variation (oscillation) of the data function. It is also common to include data oscillation in the error estimators themselves.

A practical demand on the estimator (which, I insist, can only depend on the numerical solution and on the data) is the fact that it has to be fast to compute, that is, the time needed to compute it has to be much lower than the time devoted to solving the problem. Additionally, the reliability bound cannot use any kind of smoothness assumption on the data (except those that hold true for every solution). The reason for this is that we want to estimate the error for solutions with minimal regularity, especially when we do not know if the solution is smooth or not.

A final demand that is applied to most estimators is their **local efficiency.** This will happen in different forms and shapes, but the idea can be condensed in a couple of lines. We expect the estimator to be built from a list of nodes, elements or edges as

$$\operatorname{Est}_h^2 = \sum_j \operatorname{Est}_{j,h}^2,$$

where $\operatorname{Est}_{j,h}$ only uses information (about the data and the numerical solution) close to the element, node or edge tagged with the index j. The estimator is local if we can bound

$$C_3 \operatorname{Est}_{j,h} \le \|u - u_h\|_j,$$

where $\|u - u_h\|_j$ is the error in a region that surrounds the element, node or edge j. If these local domains allow for an inequality of the form

$$\sum_j \| \cdot \|_j^2 \le C_4 \| \cdot \|^2,$$

(this means that the global norm can be estimated with local pieces), then the local efficiency of the estimator implies its global efficiency. Thanks to the local character of the estimator we can know (as opposed to guess) that if some terms in the error estimator concentrate most of its value, then the error proceeds from that part of the geometry and not from some place else. This is less obvious than it looks! It might happen that the estimator is computed with quantities localized in an area but the error is provoked by an effect taking place in a different region.

1.2 Adaptivity

Let us now briefly observe a simple adaptive scheme to compute Finite Element solutions, provided we have a locally efficient and reliable error estimator. We start with an initial grid and compute the numerical solution u_h. The method is iterated in the following way.

- We compute the estimator Est_h and compare it with a prescribed tolerance

  $$\mathrm{Est}_h \le \mathrm{Tol}.$$

  If this inequality holds we consider the solution is good enough and stop the computation. The tolerance needs to be relative to a guess on the size of the solution. If the inequality does not hold, we continue with the next step.

- We now reorder the local estimators by size,

  $$\mathrm{Est}_{1,h} \le \mathrm{Est}_{2,h} \le \ldots \le \mathrm{Est}_{N-1,h} \le \mathrm{Est}_{N,h},$$

  and we then decide how many of them we want to consider to lead our refinement strategy. For instance we can fix a parameter $\eta < 1$ and find the maximum number k such that

  $$\eta\,\mathrm{Est}_h^2 \le \sum_{j=k}^{N} \mathrm{Est}_{j,h}^2,$$

  that is, we take the minimum number of local estimators that contribute to a η-significant portion of the global error. Since the local estimators are associated to geometric elements (triangles, nodes, edges), we then **mark** these elements. This way of deciding which local estimators are relevent (ordering and choosing the largest ones up to a percentage of the error) is commonly referred to as Dörfler marking[1].

- We then apply a **mesh refinement** strategy. This strategy looks for a finer grid where all marked elements have been refined. This has to be done carfully since we need the triangulation to satisfy the initial requirements (no hanging nodes) and we do not want the angles of the triangulation to progressively degenerate. Note that we might be forced to refine elements that were not in the marked list (in order to avoid hanging nodes), but we want to avoid the refinement to be extended to too many elements.

As already mentioned, these three steps are repeated until the desired tolerance is reached. Some adaptive procedures use **mesh coarsening** in order to avoid an excessive increase in the number of global degrees of freedom. Mesh coarsening is logically applied in areas where there clearly very little error. While coarsening is a useful and interesting technique it requires rethinking the data structures for the triangulations, since we need to keep some memory of previous refinements in order to undo them if needed.

1.3 A model problem

In order to be able to focus on the novelties, the remainder of this lesson will be restricted to a very particular problem. We have a polygon $\Omega \subset \mathbb{R}^2$, whose boundary is subdivided

[1] Named after Willy Dörfler, a German mathematician who started the standarization of the theory of adaptive FEM.

into Dirichlet and Neumann parts. The problem is

$$\left[\begin{array}{ll} -\Delta u = f & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_D, \\ \partial_n u = g & \text{on } \Gamma_N, \end{array}\right.$$

for given data functions $f : \Omega \to \mathbb{R}$ and $g : \Gamma_N \to \mathbb{R}$. We are going to assume homogeneous Dirichlet boundary conditions. Including non-homogeneous conditions complicates what comes next in non-trivial ways. The essential difficulties are very much the same but arguments are harder to state and some additional notation is required.

The weak formulation works in the space

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

and can be written as

$$\left[\begin{array}{l} u \in V, \\ \displaystyle\int_\Omega \nabla u \cdot \nabla v = \int_\Omega f\, v + \int_{\Gamma_N} g\, v \qquad \forall v \in V. \end{array}\right.$$

We are given a triangulation $\mathcal{T}_h$ in the usual conditions. The effective finite element space is

$$V_h = \{u_h : \Omega \to \mathbb{R} \, : \, u_h|_K \in \mathbb{P}_1 \quad \forall K \in \mathcal{T}_h, \quad u_h|_{\Gamma_D} = 0\}$$

and the finite element approximation is the solution to the problem

$$\left[\begin{array}{l} u_h \in V_h, \\ \displaystyle\int_\Omega \nabla u_h \cdot \nabla v_h = \int_\Omega f\, v_h + \int_{\Gamma_N} g\, v_h \qquad \forall v_h \in V_h. \end{array}\right.$$

Some notation related to the triangulation will be useful. Locally, given $K \in \mathcal{T}_h$ we consider the sets

- $\mathcal{E}(K)$, the edges of K,
- $\mathcal{V}(K)$, the vertices of K.

We also consider the global sets

- $\mathcal{E}_h$, all edges of the triangulation, separated into

  $$\mathcal{E}_h^\circ \cup \mathcal{E}_h^D \cup \mathcal{E}_h^N$$

  (interior, Dirichlet, Neumann), and

- $\mathcal{V}_h$, all the vertices (nodes) of the triangulation, separated into

  $$\mathcal{V}_h^\circ \cup \mathcal{V}_h^D \cup \mathcal{V}_h^N,$$

  where, as usual, Dirichlet vertices are all vertices on the Dirichlet boundary (including the transition nodes).

Given an edge $E \in \mathcal{E}_h$, we consider a unit normal vector $\mathbf{n}_E$, obtained by rotating $\pi/2$ counter-clockwise from the given orientation of the edge (going from the first vertex to the second vertex of the edge). We also consider the set:

- $\mathcal{V}(E)$, the vertices (endpoints) of E.

Finally, when a function u is defined on both sides of an edge, we define its jump

$$[\![u]\!]_E = u^+|_E - u^-|_E$$

where $u^\pm|_E$ are the limits obtained from both sides considering the normal vector to be pointing from $-$ to $+$. In other words if $E = K^+ \cup K^-$ for $K^\pm \in \mathcal{T}_h$ and $\mathbf{n}_E$ points from K^- to K^+, then

$$[\![u]\!]_E = u^+|_E - u^-|_E, \qquad u^\pm = u|_{K^\pm}.$$

2 A posteriori estimators

2.1 Residual estimators

Given a function $w \in V = H^1_{\Gamma_D}(\Omega)$ the residual associated to the weak formulation

$$\left[\begin{array}{l} u \in V, \\ \displaystyle\int_\Omega \nabla u \cdot \nabla v = \int_\Omega f\, v + \int_{\Gamma_N} g\, v \qquad \forall v \in V, \end{array}\right.$$

is the expression

$$R(v; w) := \int_\Omega f\, v + \int_{\Gamma_N} g\, v - \int_\Omega \nabla w \cdot \nabla v.$$

Fixing w, we can understand the residual as a functional $R(\cdot; w) : V \to \mathbb{R}$. Moreover

$$R(v; w) = 0 \quad \forall v \in V \qquad \Longleftrightarrow \qquad w = u,$$

that is, the only function that cancels the residual is the solution of the problem. Note also that the discrete solution

$$\left[\begin{array}{l} u_h \in V_h, \\ \displaystyle\int_\Omega \nabla u_h \cdot \nabla v_h = \int_\Omega f\, v_h + \int_{\Gamma_N} g\, v_h \qquad \forall v_h \in V_h \end{array}\right.$$

satisfies

$$R(v_h; u_h) = 0 \quad \forall v_h \in V_h.$$

The discrete residual is the quantity

$$\mathcal{R}_h := \sup_{0 \neq v \in V} \frac{R(v; u_h)}{\|v\|_{1,\Omega}}.$$

Intuitively speaking, the residual measures how the numerical solution does not satisfy the exact equation when tested by functions that are not discrete. Because of the fact that

$$\begin{aligned} R(v;u_h) &= \int_\Omega f\,v + \int_{\Gamma_N} g\,v - \int_\Omega \nabla v\cdot\nabla u_h \\ &= \int_\Omega \nabla u\cdot\nabla v - \int_\Omega \nabla v\cdot\nabla u_h = \int_\Omega \nabla(u-u_h)\cdot\nabla v, \end{aligned}$$

it is possible to prove that

$$C_1\|u-u_h\|_{1,\Omega} \le \mathcal{R}_h \le C_2\|\nabla u-\nabla u_h\|_\Omega \le C_2\|u-u_h\|_{1,\Omega}.$$

This shows that a good way to estimate the error is to estimate $\mathcal{R}_h$. Before we go on, let us do a simple computation. It only involves integration by parts and reorganizing the information in edges and elements:

$$\begin{aligned} R(v,;u_h) &= \sum_{K\in\mathcal{T}_h}\int_K f v - \sum_{K\in\mathcal{T}_h}\int_K \nabla u_h\cdot\nabla v + \sum_{E\in\mathcal{E}_h^N}\int_E g\,v \\ &= \sum_{K\in\mathcal{T}_h}\int_K (f+\Delta u_h)\,v - \sum_{K\in\mathcal{T}_h}\sum_{E\in\mathcal{E}(K)}\int_E \partial_n u_h\,v + \sum_{E\in\mathcal{E}_h^N}\int_E g\,\partial_n u_h \\ &= \sum_{K\in\mathcal{E}_h}\int_K (f+\Delta u_h)\,v + \sum_{E\in\mathcal{E}_h^N}\int_E (g-\partial_n u_h)\,v - \sum_{E\in\mathcal{E}_h^\circ}\int_E [\![\partial_n u_h]\!]\,v. \end{aligned}$$

While we won't work on the theory of the estimators, this formula can give a good idea on why the next estimators look like they do.

A global estimator. With some integration by parts and Cauchy-Schwarz inequalities, it is possible to show that

$$\mathcal{R}_h \le \mathrm{Est}_h,$$

where

$$\begin{aligned} \mathrm{Est}_h^2 &:= \sum_{K\in\mathcal{T}_h} h_K^2\int_K |f|^2 + \sum_{E\in\mathcal{E}_h^N} h_E\int_E |g-\partial_n u_h|^2 + \sum_{E\in\mathcal{E}_h^\circ} h_E\int_E [\![\partial_n u_h]\!]^2 \\ &= \sum_{K\in\mathcal{T}_h} h_K^2\int_K |f+\Delta u_h|^2 + \sum_{E\in\mathcal{E}_h^N} h_E\int_E |g-\partial_n u_h|^2 + \sum_{E\in\mathcal{E}_h^\circ} h_E\int_E [\![\partial_n u_h]\!]^2. \end{aligned}$$

Let us have a look at the estimator:

- The first term has to be examined in the second line (note that we are using linear elements and, therefore, $\Delta u_h = 0$ on each element). It measures how the PDE $(f+\Delta u = 0)$ is not being satisfied inside the elements. We assume that the triangles do not degenerate and therefore

  $$C_1 h_K^2 \le |K| \le C_2 h_K^2,$$

  where $|K|$ is the area of K.

- The second term measures (on Neumann edges) how the Neumann condition is not being satisfied on the edges. Here h_E is the length of E.
- Finally, the third term looks at how discontinuous the gradient is across element interfaces (interior edges).

Note that all quantities in this estimator are computable: they come from data and from the numerical solution. It is also interesting to observe how the weight in front of each term scales like the measure (area or length) of the set over which we are integrating.

Local estimators. We can also compute a slightly different version of the estimator, using averages of the data functions

$$f_K := \frac{1}{|K|}\int_K f, \qquad g_E := \frac{1}{h_E}\int_E g_E.$$

On the triangle $K \in \mathcal{T}_h$ we define the very easy to compute quantity

$$\mathrm{Est}_R(K)^2 := h_K^2 \int_K |f_K + \Delta u_h|^2 + \sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_h^N} h_E \int_E |g_E - \partial_n u_h|^2 + \sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_h^\circ} h_E \int_E [\![\partial_n u_h]\!]^2.$$

The new global estimator is the sum of the local estimators

$$\mathrm{Est}_{R,h}^2 := \sum_{K\in\mathcal{T}_h} \mathrm{Est}_R(K)^2.$$

This estimator is reliable up to oscillation terms, namely,

$$\mathcal{R}_h^2 \le C\mathrm{Est}_{R,h}^2 + \sum_{K\in\mathcal{T}_h} \mathrm{Osc}_K(f)^2 + \sum_{E\in\mathcal{E}_h^N} \mathrm{Osc}_E(g)^2,$$

where

$$\begin{aligned}\mathrm{Osc}_K(f) &:= h_K\|f - f_K\|_K = \sqrt{h_K^2\int_K |f-f_K|^2},\\ \mathrm{Osc}_E(g) &:= \sqrt{h_E}\|g - g_E\|_E = \sqrt{h_E\int_E |g-g_E|^2}.\end{aligned}$$

Local efficiency. The measure of local efficiency requires the introduction of a new set:

$$\omega_K := \cup\{K' \in \mathcal{T}_h \,:\, \mathcal{E}(K)\cap\mathcal{E}(K') \neq \varnothing\}.$$

This so-called **macroelement** is formed by K and the (at most three) triangles that share an edge with K. The local efficiency can be stated in the following form

$$\mathrm{Est}_{R,h}(K)^2 \le C\|u-u_h\|_{1,\omega_K}^2 + \sum_{K'\subset\omega_K} \mathrm{Osc}_{K'}(f)^2 + \sum_{E\in\mathcal{E}(K)\cap\mathcal{E}_h^N} \mathrm{Osc}_E(g)^2.$$

We finish this section noting that the oscillation terms can be considered to be of higher order: if f and g are piecewise smooth with respect to the triangulation (which does not imply that u is smooth), the

$$\mathrm{Osc}_K(f) \le C h_K^2, \qquad \mathrm{Osc}_E(g) \le C h_E^{3/2}.$$

Provable adaptive techniques include oscillation terms in the error estimators. This is just logical (not even taking into account all the above inequalities): we cannot expect to reduce the error without having the mesh take care of details of the data functions.

2.2 Bubbles and hierarchical estimators

Edge bubble functions. Let $E \in \mathcal{E}_h$ be any edge of the triangulation. We can then consider the continuous function $b_E : \Omega \to \mathbb{R}$ such that $b_E|_K \in \mathbb{P}_2(K)$ for all K, $b_E(\mathbf{m}_E) = 1$ (where $\mathbf{m}_E$ is the midpoint of E), and $b_E \equiv 0$ on all other edges of the triangulation. This is one of the global basis functions of the $\mathbb{P}_2$ Finite Element space on the same triangulation. It satisfies:

- The support of b_E is contained in the macroelement

$$\omega_E = \cup\{K \in \mathcal{T}_h \,:\, E \in \mathcal{E}(K)\}.$$

  This macroelement is made of the two elements sharing E, unless E is on the boundary.

- It takes values in the unit inverval:

$$0 \le b_E \le 1.$$

- The edge integral of b_E is of the order of the length of the edge.

$$C h_E \le \int_E b_E \le h_E.$$

- The gradient scales in the following form

$$\int_{\omega_E} |\nabla b_E|^2 \le C_1 h_E^{-2} \int_{\omega_E} |b_E|^2 \le C_2.$$

Note also that if we take the nodal basis $\{\varphi_1, \ldots, \varphi_N\}$ for V_h and we consider all edge bubble functions not associated to Dirichlet edges $\{b_E \,:\, E \in \mathcal{E}_h^\circ \cup \mathcal{E}_h^N\}$, then we have a basis for the space of $\mathbb{P}_2$ finite elements on the same triangulation (see exercises). This basis is not the nodal basis of that space, but it is a basis nonetheless. It has the advantage of being **hierachical**: we start with the basis for the $\mathbb{P}_1$ F.E. space, and we then add bubble functions to obtain a basis for the richer $\mathbb{P}_2$ space.

Discrete edge bubble functions. Given the triangulation $\mathcal{T}_h$, we can easily build a red refinement $\mathcal{T}_{h/2}$, where every triangle has been substituted by four triangles of the same shape (see Figure 6.1) by joining the midpoints of the edges. We then consider the function $b_E : \Omega \to \mathbb{R}$ in $V_{h/2}$ (the corresponding Finite Element space in the refined triangulation $\mathcal{T}_{h/2}$) such that $b_E(\mathbf{m}_E) = 1$ and b_E vanishes on all other nodes (vertices) of $\mathcal{T}_{h/2}$. Note that this new bubble function has the same four properties of the edge bubble function (the bulletpoints above) although the support of the function is slightly reduced.

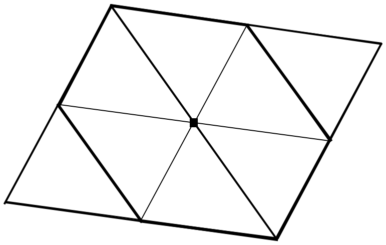


Figure 6.1: A red (uniform) refinement of two triangles and the support of the discrete bubble function b_E associated to the node added to the interior edge. The thicker line delineates the support of b_E.

Element bubbles. Let $K \in \mathcal{T}_h$ and let $\mathbf{b}_K$ be its barycenter. We can then define $b_K \in \mathbb{P}_3(K)$ such that $b_K(\mathbf{b}_K) = 1$ and $b_K \equiv 0$ on ∂K. The following properties are simple to show:

- The support of b_K is contained in K.
- The function takes values in the unit interval:
$$0 \le b_K \le 1.$$
- The area integral of b_K is of the order of the area of the element
$$C|K| \le \int_K b_K \le |K|.$$
- The gradient scales in the following form:
$$\int_K |\nabla b_K|^2 \le C_1 h_K^{-2} \int_K |b_K|^2 \le C_2.$$

If we consider the nodal basis $\{\varphi_i\}$, the $\mathbb{P}_2$ edge bubbles $\{b_E \,:\, E \in \mathcal{E}_h\}$ and the element bubbles $\{b_K \,:\, K \in \mathcal{T}_h\}$ we get a linearly independent set of functions. Its span is a proper subset of the $\mathbb{P}_3$ finite element space defined on $\mathcal{T}_h$. Namely, if

$$W_h = \{u_h \in \mathcal{C}(\overline{\Omega}) \,:\, u_h|_K \in \mathbb{P}_3, \quad \forall K \in\},$$

then the span of all the functions above is the set

$$\{u_h \in W_h \,:\, u_h|_E \in \mathbb{P}_2(E) \quad \forall E \in \mathcal{E}_h\}.$$

Discrete element bubbles. Let now $\mathcal{T}_{h/4}$ be the result of a double red refinement of a triangulation $\mathcal{T}_h$ (see Figure 6.2) and let $V_{h/4}$ be the $\mathbb{P}_1$ finite element space on this triangulation. We then consider $b_K \in V_{h/4}$ such that $b_K = 1$ on the three interior nodes and $b_K = 0$ on the other nodes. This discrete bubble shares the same properties as the $\mathbb{P}_3$ bubble defined above. If we now consider the set formed by the $\mathbb{P}_1$ basis functions in $\mathcal{T}_h$, the $\mathbb{P}_1$ discrete bubbles b_E (elements of the refined space $V_{h/2}$, and the $\mathbb{P}_1$ discrete bubbles b_K (elements of the doubly refined space $V_{h/4}$) we obtain a linearly independent set spanning a proper subset of $V_{h/4}$.

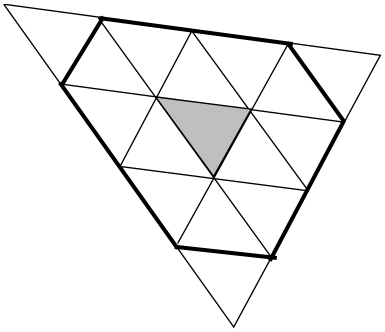


Figure 6.2: Quadruple uniform refinement of a triangle. The thick line shows the boundary of the support for the discrete bubble b_K. The grey area shows where $b_K \equiv 1$.

An edge-based estimator. Let b_E be any of the two edge bubble functions defined above. We then consider the quantity

$$\begin{aligned}
r_E \quad &:= \quad R(b_E; u_h) = \int_{\omega_E} \nabla(u - u_h) \cdot \nabla b_E \\
&= \quad \int_\Omega f\, b_E + \int_{\Gamma_N} g\, b_E - \int_\Omega \nabla u_h \cdot \nabla b_E \\
&= \quad \left\{ \begin{array}{ll} \displaystyle\int_{\omega_E} f\, b_E - \int_E [\![\partial_n u_h]\!]\, b_E & \text{if } E \in \mathcal{E}_h^\circ, \\ \displaystyle\int_K f\, b_E + \int_E (g - \partial_n u_h)\, b_E & \text{if } E \in \mathcal{E}_h^N, \quad E \in \mathcal{E}(K). \end{array} \right.
\end{aligned}$$

The global estimator is

$$\mathrm{Est}^2_{H,h} := \sum_{E \in \mathcal{E}_h^\circ \cup \mathcal{E}_h^N} r_E^2.$$

It is locally efficient

$$|r_E| \le C \|u - u_h\|_{1,\omega_E}$$

and globally reliable up to oscillation terms, which are of higher order for smooth enough data. Note that the local efficiency is derived from the scaling properties of the gradient of b_E.

Edge- and element-based estimator. Let b_K be the element bubble functions (or their discrete counterparts) defined above. We then define the local residual

$$\begin{aligned} r_K &:= R(b_K; u_h) = \int_K \nabla(u - u_h) \cdot \nabla b_K \\ &= \int_\Omega f\, b_K + \int_{\Gamma_N} b_K - \int_\Omega \nabla u_h \cdot \nabla b_K = \int_K f\, b_K. \end{aligned}$$

The new global estimator adds the element contributions to the edge contributions

$$\mathrm{Est}^2_{H,h} := \sum_{E \in \mathcal{E}_h^\circ \cup \mathcal{E}_h^N} r_E^2 + \sum_{K \in \mathcal{T}_h} r_K^2.$$

Its local efficiency follows from the scaling properties of the gradients of the bubble functions. The reliability of this estimator follows from the reliability of the edge-based estimator.

2.3 Gradient averaging

A commonly employed strategy for a posteriori error estimation arises from the construction of an improved approximation of the gradient using local averaging techniques and then comparing the averaged gradient with the original discrete gradient as an error estimator. Let $\{\mathbf{v}_i \,:\, i = 1, \ldots, N\}$ be a numbering of all the vertices of the triangulation. For each node $\mathbf{v}_i$ we consider the macroelement

$$\omega_i := \cup\{K \in \mathcal{T}_h \,:\, \mathbf{v}_i \in \mathcal{V}(K)\}.$$

We then consider the weighted average of all the gradients that we obtain at this point (one on each element, given the fact that we expect ∇u_h to be discontinuous)

$$\mathbf{g}_i := \sum_{K \subset \omega_i} \frac{|K|}{|\omega_i|} \nabla u_h|_K \quad \in \mathbb{R}^2.$$

Finally, we attach these nodal values to the global basis functions

$$\mathbf{G}_h u_h = \sum_{i=1}^N \varphi_i \mathbf{g}_i : \Omega \to \mathbb{R}^2, \qquad \mathbf{G}_h u_h \in \mathbf{V}_h,$$

where

$$\mathbf{V}_h := V_h^2 := \{\mathbf{v}_h = (v_h^x, v_h^y) : \Omega \to \mathbb{R}^2 \,:\, v_h^x, v_h^y \in V_h\}.$$

The a posteriori error estimator (sometimes called the Zienkiewicz-Zhu[2] estimator) is the quantity

$$\mathrm{Est}_{ZZ,h} := \|\nabla u_h - \mathbf{G}_h u_h\|_\Omega.$$

[2] after Olgierd Zienckiewicz and Jian Zhong Zhu

Another point of view. The construction of the 'improved' gradient by local weighted averaging of gradients in the elements surrounding a node can be given several possible interpretations. Here's one. Recall that the three-vertex formula

$$\int_K \phi \approx \frac{|K|}{3} \sum_{\mathbf{p}\in\mathcal{V}(K)} \phi(\mathbf{p}),$$

which was used in mass lumping, is exact for all polynomials of degree one. We can then define the discrete product

$$(\mathbf{v},\mathbf{w})_h = \sum_{K\in\mathcal{T}_h} \frac{|K|}{3} \left(\sum_{\mathbf{p}\in\mathcal{V}(K)} \mathbf{v}(\mathbf{p})\cdot\mathbf{w}(\mathbf{p}) \right),$$

for every pair of functions that are well defined on triangles (up to the boundary). Note that if $\mathbf{v}_h \in \mathbf{V}_h$ and $u_h \in V_h$, then

$$(\nabla u_h, \mathbf{v}_h)_h = \int_\Omega \nabla u_h \cdot \mathbf{v}_h.$$

(The fact that $\mathbf{v}_h$ is continuous does not play any role in this argument. Only the fact that $\mathbf{v}_h$ is linear elementwise matters.) It is quite simple to show that the averaged gradient is the only solution to the reconstruction problem

$$\left[\begin{array}{l} \mathbf{G}_h u_h \in \mathbf{V}_h, \\ (\mathbf{G}_h u_h, \mathbf{v}_h)_h = (\nabla u_h, \mathbf{v}_h)_h \quad \forall \mathbf{v}_h \in \mathbf{V}_h. \end{array}\right.$$

This shows that we pick the piecewise constant (and therefore discontinuous) gradient and project it on the space $\mathbf{V}_h$, using a discrete inner product (the same one we used for mass lumping) to obtain the averaged gradient. If we had used the continuous inner product, we would be forced to solve a global linear system and we would be losing the localization effect.

A word on efficiency and reliability. While residual estimators have had a very well understood theory since the very beginning, the performance of ZZ-style estimators are based on some kind of super-approximation properties. For instance, in most theoretical expositions of this kind of estimators, it is assumed that

$$\int_\Omega |\nabla u - \mathbf{G}_h u_h|^2 \le \beta \int_\Omega |\nabla u - \nabla u_h|^2 \qquad \text{with } 0 < \beta < 1.$$

There are several considerations and hypotheses to be made to show that this approximation holds. Whether these hypotheses hold or not is a different story. However, there has been much theoretical development in recent years, justifying what was observed in practice: these estimators are actually performing their task. Note that these estimators are extremely easy to code, which is a point in their favor. Note, at the same time, that they completely ignore the data.

3 Refinement of triangular meshes

At this point we have defined several strategies to decide which elements (or which edges) are producing the most error in our computation. Imagine that we have marked some elements and/or some edges. The goal of the next algorithm is to find a finer triangulation where:

- all marked elements are refined (at least by being divided into two subelements),
- all marked edges are refined (at least divided by two),
- not too many unmarked elements/edges are refined,
- the triangulation does not generate elements with too acute angles.

We note that the fact that triangulations cannot have hanging nodes forces the refinement of triangles that might not be in the list of marked triangles. The **Newest Vertex Bisection** algorithm suceeds in producing a refined triangulation with the above criteria. The edges will be subdivided into two equally sized subedges. The key idea is to think in terms of edges of triangles to set up a refinement condition:

> If an edge of a triangle is to be subdivided, then the longest edge of that triangle should be subdivided as well.

Let us try to see a way of deciding how to mark edges. Assume that we have a marking strategy for triangles and edges. We first look at the triangles. If a triangle is marked, we mark its three edges. With this rule, we have added marked edges to the possible list of marked edges. We now go ahead and apply the rule about the longest edge[3]. If there are triangles that have marked edges but their longest edges is not marked, we mark their longest edges as well. We repeat this process untill we have a marking of edges that satisfies the above rule.

At this stage, we go back to looking at triangles. Triangles fall into four categories:

- Those with no marked edges.
- Those with the longest edge marked.
- Those with two marked edges, one of which is the longest edge.
- Those with the three edges marked.

Even in the case where two or three edges are marked, the longest edge plays a role. The refinement strategy is better shown with a picture. See Figure 6.3 for the possible refinements.

[3] It might happen that one triangle has two equally sized longest edges. In this case, one of them is considered to be the longest edge and the other one is ignored in this capacity.

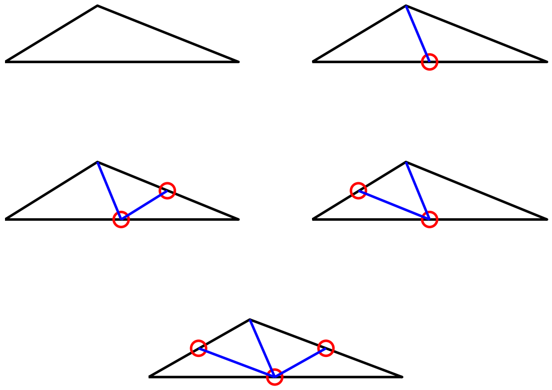


Figure 6.3: The refinement strategy of the Newest Vertex Bisection method. Triangles are refined differently depending on whether one edge (the longest one) is marked, or two edges are marked, or all edges are marked. If only one edge is marked, its midpoint connects to the opposed vertex. Otherwise, midpoints of the edges connect to the midpoint of the longest edge.

4 Exercises

1. Let b_E be the $\mathbb{P}_2$ edge bubble function associated to the edge $E \in \mathcal{T}_h$. Let $\{\varphi_i \,:\, i = 1, \ldots, N\}$ be the $\mathbb{P}_1$ nodal basis for the same triangulation. Show that

$$\{\varphi_i \,:\, i = 1, \ldots, N\} \cup \{b_E \,:\, E \in \mathcal{E}_h\}$$

is a basis for the $\mathbb{P}_2$ finite element space on $\mathcal{T}_h$. (Hint. Count dimensions and show linear independence. Note that the edge bubbles vanish on all the vertices.)

2. Let $\mathcal{T}_{h/2}$ be the red refinement of a triangulation $\mathcal{T}_h$ and let V_h and $V_{h/2}$ be the $\mathbb{P}_1$ finite element spaces on $\mathcal{T}_h$ and $\mathcal{T}_{h/2}$ respectively. Let $\{\varphi_i \,:\, i = 1, \ldots, N\}$ be the nodal basis for $\mathcal{T}_h$. For each edge $E \in \mathcal{E}_h$ we consider the function $b_E \in \mathcal{V}_{h/2}$ that takes the unit value on the midpoint of the edge and zero on all other nodes of $\mathcal{E}_{h/2}$. Show that

$$\{\varphi_i \,:\, i = 1, \ldots, N\} \cup \{b_E \,:\, E \in \mathcal{E}_h\}$$

is a basis for $V_{h/2}$. (Hint. See the previous exercise.)

3. Let $N^k_\alpha = \lambda^K_\alpha$ be the local $\mathbb{P}_1$ basis functions for a triangle K. Use them to write explicit formulas for the edge and element bubble functions.

4. Consider the simple triangulation of Figure 6.4 and assume that we have marked one element (the one with a circle in the middle). Draw the Newest Vertex Bisection refinement of the mesh.

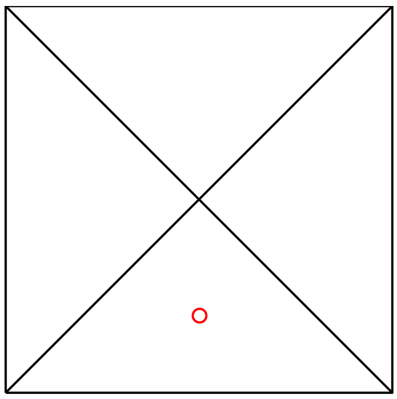


Figure 6.4: A simple triangulation with one marked triangle.

Lesson 7

Polynomial bases and bookkeeping

In the first lessons we have introduced the basic finite element methods based on triangular partitions of a polygon. In this lesson we are going to learn about how to organize the bookkeeping for finite elements of moderately high order. We will also discuss polynomial bases that allow for better implementation than the traditional Lagrange basis.

1 Counting degrees of freedom locally

1.1 Barycentric coordinates

The continuous reference triangle. We have seen the reference element in the plane

$$\widehat{K} := \{(\xi,\eta) \,:\, \xi,\eta,1-\xi-\eta \geq 0\}$$

several times in this course. Let us now introduce the barycentric coordinates associated to this element

$$\widehat{\lambda}_1 := 1-\xi-\eta, \qquad \widehat{\lambda}_2 := \xi, \qquad \widehat{\lambda}_3 := \eta.$$

Note that these coordinates can be seen as functions of the reference variables (ξ,η). As such, they are just the $\mathbb{P}_1$ basis functions in the reference element. Recall that the $\mathbb{P}_1$ basis functions have also been used to build the affine map from the reference element to a general physical element. Every point of the plane can be represented with three barycentric coordinates $(\widehat{\lambda}_1,\widehat{\lambda}_2,\widehat{\lambda}_3)$ satisfying

$$\widehat{\lambda}_1+\widehat{\lambda}_2+\widehat{\lambda}_3 = 1.$$

Coordinate sets satisfying this kind of properties are called homogeneous coordinates. Note that on the reference triangle, the second and the third barycentric coordinates are the reference coordinates (ξ,η). A point $(\widehat{\lambda}_1,\widehat{\lambda}_2,\widehat{\lambda}_3)$ (with, let me repeat, $\widehat{\lambda}_1+\widehat{\lambda}_2+\widehat{\lambda}_3 = 1$, since otherwise these cannot be considered coordinates) is in the closed reference triangle $\widehat{K}$ if and only if the three coordinates are non-negative. Some special points are easy to recognize:

- The vertices are

$$(1,0,0), \qquad (0,1,0), \qquad (0,0,1).$$

- The edges of the triangle are

$$\begin{aligned} \widehat{e}_1 &:= \{(1-t,t,0) \,:\, 0 \le t \le 1\}, \\ \widehat{e}_2 &:= \{(0,1-t,t) \,:\, 0 \le t \le 1\}, \\ \widehat{e}_3 &:= \{(t,0,1-t) \,:\, 0 \le t \le 1\}. \end{aligned}$$

  With this numbering the first edge goes from the first vertex to the second, the second edges goes from the second vertes to the third, and the third edge goes down from $(0,1)$ to the origin. This nice rotation of the edges of the reference triangle is elegantly represented by the rotation of coordinates in the above parametrizations. If we keep $0 < t < 1$, we get the edges without the endpoints.

- Once again, the entire triangle can be represented as

$$\{(1-s-t,s,t) \,:\, 0 \le s \le 1, \quad 0 \le t \le 1-s\}.$$

  If we make the inequalities strict, this corresponds the points in the interior of the triangle.

Barycentric coordinates on a general triangle. Before we move on to the discrete case, which will be our way of counting degrees of freedom, let us briefly mention the barycentric coordinates of a triangle. Pick three unaligned points in the plane

$$\mathbf{v}_i = \left[\begin{array}{c} x_i \\ y_i \end{array}\right] \qquad i = 1,2,3$$

and order them in a way that

$$\det \left[\begin{array}{cc} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{array}\right] > 0.$$

Does this condition sound familiar? (It will not be relevant for what comes next, but it's always useful to have positive orientation.) Associated to the triangle K with the given vertices, we can consider the functions

$$\lambda_i^K \in \mathbb{P}_1(K) \quad \lambda_i^K(\mathbf{v}_j) = \delta_{ij} \quad i,j = 1,2,3.$$

To avoid being too wordy, let me give you some properties here. You should be able to prove all these statements quite easily. (Try it!)

- If $F_K : \widehat{K} \to K$ is the affine transformation mapping the vertices of $\widehat{K}$ to the vertices of K in the preset order, then $\lambda_i^K = \widehat{\lambda} \circ F_K^{-1}$. (Note that we wrote this as a transformation of nodal bases for the $\mathbb{P}_1$ element when we dealt with the reference element for the first time.)

- The sum of the barycentric coordinates of a point is one

$$\lambda_1^K + \lambda_2^K + \lambda_3^K = 1.$$

- The barycentric coordinates of the vertices are the canonical vectors

$$\mathbf{e}_1 = (1,0,0), \quad \mathbf{e}_2 = (0,1,0), \quad \mathbf{e}_3 = (0,0,1).$$

- The barycentric coordinates of a physical point (x,y) can be found as the solution of the system

$$\lambda_1^K \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} + \lambda_2^K \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} + \lambda_3^K \begin{bmatrix} x_3 \\ y_3 \end{bmatrix} = \begin{bmatrix} x \\ y \end{bmatrix}$$

  satisfying the restriction of homogeneity

$$\lambda_1^K + \lambda_2^K + \lambda_3^K = 1.$$

  Written in a different way, we solve the system

$$\begin{bmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \lambda_1^K \\ \lambda_2^K \\ \lambda_3^K \end{bmatrix} = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}.$$

- Points interior to K are characterized by having positive barycentric coordinates.

- Points on the boundary of K (except the vertices) are characterized by having one vanishing barycentric coordinate while the other two are positive. Moreover, the sets

$$\{(1-t,t,0) \,:\, t \in \mathbb{R}\}, \quad \{(0,1-t,t) \,:\, t \in \mathbb{R}\}, \quad \{(t,0,1-t) \,:\, t \in \mathbb{R}\}$$

  are the lines containing the edges of the triangle. In each case, by looking at what coordinate vanishes we can see which vertex is not included in the line.

1.2 The principal lattice

The discrete reference element. At the time of discretizing, we can think of the reference element as heving been reduced to the points

$$\tfrac{1}{k}(i,j), \qquad i,j,k-i-j \geq 0.$$

Note that all points $(\frac{i}{k},\frac{j}{k})$ are in the reference element $\widehat{K}$ as long as the parameters (i,j) satisfy the above restrictions. Instead of counting with the indices (i,j), we can count with $(k-i-j,i,j)$ or equivalently with

$$\boldsymbol{\alpha} = (\alpha_1,\alpha_2,\alpha_3), \qquad |\boldsymbol{\alpha}| = \alpha_1+\alpha_2+\alpha_3 = k.$$

Note that $\frac{1}{k}\boldsymbol{\alpha}$ are the barycentric coordinates (with respect to the reference element) of a point in $\widehat{K}$.

A general element. The $\mathbb{P}_k$ finite element is based on placing nodes on the principal lattice of the triangle K. The points have barycentric coordinates

$$\tfrac{1}{k}(\alpha_1, \alpha_2, \alpha_3), \qquad \alpha_1 + \alpha_2 + \alpha_3 = k, \qquad 0 \le \alpha_j \in \mathbb{Z}$$

For instance, for $k = 4$ we have

$$1 + 2 + 3 + 4 + 5 = 15 = 1 + \ldots + k + (k+1) = \binom{k+2}{2}$$

points, which can be separated as:

- the three vertices,
- 3 points $(k-1)$ per edge,
- and 3 points $((k-1)(k-2)/2 = 1 + \ldots + k - 2)$ inside the triangle.

A counting strategy. We are going to count the points on the principal $k-$lattice in a geometric fashion.

- We first count the three vertices in rotating order

$$k\mathbf{e}_1 = (k, 0, 0), \qquad k\mathbf{e}_2 = (0, k, 0), \qquad k\mathbf{e}_3 = (0, 0, k).$$

- We then pick the points on the edges, starting in the first vertex and moving counterclockwise from there:

$$\begin{array}{ll} (k-i, i, 0) & i = 1, \ldots, k-1, \\ (0, k-i, i) & i = 1, \ldots, k-1, \\ (i, 0, k-i) & i = 1, \ldots, k-1. \end{array}$$

  (This edge count is only done for $k \ge 2$, since the lowest order case involves only the three vertices.)

- We finally choose an order for the points inside the triangle, for instance

$$(k-i-j, i, j) \qquad i = 1, \ldots, k-2, \quad j = 1, \ldots, k-1-i.$$

  Interior nodes appear only when $k \ge 3$.

The total count for points is:

$$\begin{aligned} \binom{k+2}{k} = \dim \mathbb{P}_k = & 3 & & \text{(vertices)} \\ & + 3(k-1) & & \text{(edges)} \\ & + \frac{(k-2)(k-1)}{2} & & \text{(interior)} \end{aligned}$$

1.3 Geometric properties of the Lagrange basis

The Lagrange basis. Let $k \geq 1$ and consider the points on the principal k-lattice, numbered as

$$\mathbf{p}_{\boldsymbol{\alpha}}, \qquad \boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3), \qquad \alpha_1 + \alpha_2 + \alpha_3 = k.$$

(We will not make explicit the requirement of the indices to be non-negative integers.) The nodal (Lagrange) basis associated to these points is the set of polynomials $L_{\boldsymbol{\alpha}} \in \mathbb{P}_k$ such that

$$L_{\boldsymbol{\alpha}}(\mathbf{p}_{\boldsymbol{\beta}}) = \delta_{\boldsymbol{\alpha}\boldsymbol{\beta}} = \begin{cases} 1 & \text{if } \boldsymbol{\alpha} = \boldsymbol{\beta}, \\ 0 & \text{otherwise.} \end{cases}$$

Note that the count of indices equals the count of points on the lattice and, therefore, the dimension of the space of polynomials $\mathbb{P}_k$. Note also that if

$$p = \sum_{\boldsymbol{\alpha}\,:\,|\boldsymbol{\alpha}|=k} c_{\boldsymbol{\alpha}} L_{\boldsymbol{\alpha}}$$

then $p(\mathbf{p}_{\boldsymbol{\beta}}) = c_{\boldsymbol{\beta}}$, which implies the linear independence of these polynomials. (Proving that the values on points of the principal lattice determine uniquely the polynomial takes some Linear Algebra, which I'll leave for a better occasion.)

Geometric properties. Here are some properties and/or observations:

(G.1) The only element of the basis that does not vanish at the vertex $\mathbf{v}_i$ is $L_{k\mathbf{e}_i}$. Therefore, barring the three vertex functions $L_{k\mathbf{e}_i}$, all other elements of the basis vanish on the vertices.

(G.2) The basis functions associated to points inside an edge can be considered as edge bubbles. They vanish on the two other edges of the element. Moreover, the $k+1$ basis functions associated to an edge (the interior of the edge and the vertices) fully determine the value of the polynomial on the edge.

(G.3) The basis functions associated to the interior of the triangle vanish on the boundary of the triangle. (They can be considered as element bubbles.)

1.4 The Bernstein-Bézier basis

The Bernstein-Bézier basis on the element K is defined in terms of the barycentric coordinates $(\lambda_1, \lambda_2, \lambda_3)$ —we omit the dependence on K— as

$$B_{\boldsymbol{\alpha}} := \binom{k}{\alpha_1\, \alpha_2\, \alpha_3} \lambda_1^{\alpha_1} \lambda_2^{\alpha_2} \lambda_3^{\alpha_3}, \qquad |\boldsymbol{\alpha}| = k$$

where

$$\binom{k}{\alpha_1\ \alpha_2\ \alpha_3} = \frac{k!}{\alpha_1!\alpha_2!\alpha_3!}$$

are the trinomial coefficients. Recall that the barycentric coordinates are linear functions of (x,y) (elements of $\mathbb{P}_1$), from where it is clear that $B_{\boldsymbol\alpha} \in \mathbb{P}_k$. The expression in the reference element is sometimes useful to see

$$\widehat{B}_{\boldsymbol\alpha} := \binom{k}{\alpha_1\ \alpha_2\ \alpha_3}(1-\xi-\eta)^{\alpha_1}\xi^{\alpha_2}\eta^{\alpha_3}, \qquad |\boldsymbol\alpha| = k.$$

Note that

$$\sum_{\boldsymbol\alpha\,:\,|\boldsymbol\alpha|=k} B_{\boldsymbol\alpha} \equiv 1.$$

One advantage of this basis is simple to see: the elements of this basis are always nonnegative inside the element K. This is opposed to the Lagrange basis, which is forced to oscillate quite a lot due to its need to cancel at many points in the principal lattice. An apparent disadvantage with respect to the Lagrange basis is the fact that the coefficients $c_{\boldsymbol\alpha}$ in the expression

$$\sum_{\boldsymbol\alpha\,:\,|\boldsymbol\alpha|=k} c_{\boldsymbol\alpha} B_{\boldsymbol\alpha}$$

are not point values (actually, the ones corresponding to the vertices are the point values there). In other words, except for the vertex coefficients, the coefficients $c_{\boldsymbol\alpha}$ do not provide the value of the polynomial at a nodal point; they are just coefficients of a linear combination. This is compensated by the existence of very efficient and stable evaluation algorithms for linear combinations of the Bernstein-Bézier polynomials. Other than that, let us emphasize the most important fact about this basis: the geometric properties (G.1), (G.2), and (G.3) of the Lagrange basis hold for the Bernstein-Bézier basis as well.

2 Hierarchical bases on triangles

In this section we are going to describe a third basis of polynomials on a general triangle (it will be defined in terms of barycentric coordinates) with the geometric properties (G.1)-(G.3) and one additional advantage: the basis will be hierarchical. This means that the basis for $\mathbb{P}_{k+1}(K)$ will be an extension of the basis for $\mathbb{P}_k(K)$, including some new polynomials. This is not true for the Lagrange and Bernstein-Bézier bases.

We will need to parametrize edges from time to time, and we will be using the interval $(-1,1)$ to define some functions. To improve readability we will always take

$$t \in (0,1) \qquad x \in (-1,1).$$

2.1 Some univariate sequences

The Legendre polynomials. The Legendre polynomials can be defined with the following three-term recurrence:

$$\begin{aligned} L_0(x) &:= 1, \\ L_1(x) &:= x, \\ L_j(x) &:= \frac{2j-1}{j} x L_{j-1}(x) - \frac{j-1}{j} L_{j-2}(x), \qquad j \ge 2. \end{aligned}$$

It is clear that the degree of L_j is exactly j and, therefore, for all k

$$\{L_0, L_1, \ldots, L_k\} \text{ is a basis for } \mathbb{P}_k(x).$$

Using the recurrence, it is also simple to see that $L_j(1) = 1$ for all j. The Legendre polynomials have also the following parity property: even-indexed polynomials are even (they only contain even powers of x), whie odd-indexed polynomials are odd. Furthermore, they are orthogonal in $(-1, 1)$,

$$\int_{-1}^{1} L_i(x) L_j(x) \mathrm{d}x = 0 \qquad i \neq j.$$

A particular instance of this property (think that $L_0(x) = 1$) is

$$\int_{-1}^{1} L_j(x) \mathrm{d}x = 0 \qquad \forall j \geq 1.$$

The Lobatto functions. We now describe a different basis for the space of univariate polynomials. The first two polynomials are taken from the nodal (Lagrange, not Legendre) basis in $(-1, 1)$,

$$\Psi_0(x) := \tfrac{1}{2}(1 - x), \qquad \Psi_1(x) := \tfrac{1}{2}(1 + x).$$

Note thus that

$$\Psi_0(-1) = 1, \quad \Psi_0(1) = 0, \qquad \Psi_1(-1) = 0, \quad \Psi_1(1) = 1.$$

The rest of the basis is built by integrating Legendre basis

$$\Psi_j(x) := c_j \int_{-1}^{x} L_{j-1}(y) \mathrm{d}y,$$

where the normalization factor c_j is given by

$$c_j^{-2} = \int_{-1}^{1} |L_{j-1}(x)|^2 \mathrm{d}x.$$

Now, let us just observe that $\{\Psi_0, \Psi_1\}$ are a basis for $\mathbb{P}_1(x)$, while the degree of Ψ_j is exactly j for all j. This proves that for all k

$$\{\Psi_0, \Psi_1, \ldots, \Psi_k\} \text{ is a basis for } \mathbb{P}_k(x).$$

Some further properties of this basis are explored in the exercise list. What we care about here is the following

$$\Psi_j(-1) = 0, \qquad \Psi_j(1) = \int_{-1}^{1} L_{j-1}(x) \mathrm{d}x = 0, \qquad j \geq 2.$$

The kernel functions. There is a third sequence of univariate polynomials that is derived from the previous one. Note that $\Psi_j(\pm 1) = 0$ for all $j \ge 2$ and that Ψ_j has degree exactly j. We can therefore factor $1 - x^2 = (1-x)(1+x)$ from all of them. We thus define

$$\Phi_k(x) := \frac{4}{1-x^2}\Psi_{k+2}(x) = \frac{\Psi_{k+2}(x)}{\Psi_0(x)\Psi_1(x)} \in \mathbb{P}_k(x).$$

2.2 Hierarchical bases

Some preliminary computations. Everything we are going to say from now on can be written simultaneously in a general physical element or in the reference element. Readers who find barycentric coordinates hard to grasp should repeat all the calculations using the substitutions

$$\widehat{\lambda}_1 = 1 - \xi - \eta, \qquad \widehat{\lambda}_2 = \xi, \qquad \widehat{\lambda}_3 = \eta.$$

The differences between pairs of barycentric coordinates are linear functions characterized by the values

	$\mathbf{v}_1$	$\mathbf{v}_2$	$\mathbf{v}_3$
$\lambda_2 - \lambda_1$	-1	1	0
$\lambda_3 - \lambda_2$	0	-1	1
$\lambda_1 - \lambda_3$	1	0	-1

We will also consider the maps to the three edges, running in the order we have assigned from the beginning of this Lesson:

$$\begin{aligned}
[0,1] \ni t &\quad\longmapsto\quad \boldsymbol{\phi}_{12}(t) = (1-t)\mathbf{v}_1 + t\,\mathbf{v}_2,\\
[0,1] \ni t &\quad\longmapsto\quad \boldsymbol{\phi}_{23}(t) = (1-t)\mathbf{v}_2 + t\,\mathbf{v}_3,\\
[0,1] \ni t &\quad\longmapsto\quad \boldsymbol{\phi}_{31}(t) = (1-t)\mathbf{v}_3 + t\,\mathbf{v}_1.
\end{aligned}$$

In the particular case of the reference element, the maps are just

$$t \mapsto (t, 0), \qquad t \mapsto (1-t, t), \qquad t \mapsto (0, 1-t).$$

If we follow parametrization after parametrization, we run along the boundary of the element in counterclockwise orientation. Note that if, in what follows, we do all the computations in the reference element and then push the basis forward to the physical elements we get exactly the same basis.

The vertex functions. For any $k \ge 1$, the hierarchical basis starts with the functions

$$\lambda_1, \qquad \lambda_2, \qquad \lambda_3.$$

This means that the $\mathbb{P}_1$ basis will always be a part of the local basis, no matter the degree.

The edge functions. For $k \ge 2$ we also include the functions

$$\begin{aligned} H_{12,j} &:= \lambda_2\lambda_1\Phi_j \circ (\lambda_2-\lambda_1), \qquad j=0,\ldots,k-2,\\ H_{23,j} &:= \lambda_3\lambda_2\Phi_j \circ (\lambda_3-\lambda_2), \qquad j=0,\ldots,k-2,\\ H_{31,j} &:= \lambda_1\lambda_3\Phi_j \circ (\lambda_1-\lambda_3), \qquad j=0,\ldots,k-2. \end{aligned}$$

Let us pay attention to the first edge. The factor $\lambda_2\lambda_1$ creates the edge bubble: it vanishes on the second and third edge. Note also that

$$\lambda_1 \circ \boldsymbol{\phi}_{12}(t) = 1-t, \qquad \lambda_2 \circ \boldsymbol{\phi}_{12}(t) = t, \qquad (\lambda_2-\lambda_1)\circ \boldsymbol{\phi}_{12}(t) = 2t-1.$$

This is quite easy to verify noticing that $\lambda_1 \circ \boldsymbol{\phi}_{12} : [0,1] \to \mathbb{R}$ and $\lambda_2 \circ \boldsymbol{\phi}_{12} : [0,1] \to \mathbb{R}$ are linear functions with very particular values at $t=0$ and $t=1$. Moreover, it's not complicated to show that

$$\Psi_0(2t-1) = 1-t, \qquad \Psi_1(2t-1) = t.$$

Therefore

$$\begin{aligned} H_{12,j} \circ \boldsymbol{\phi}_{12}(t) &= (1-t)t\Phi_j(2t-1)\\ &= \Psi_0(2t-1)\Psi_1(2t-1)\Phi_j(2t-1)\\ &= \Psi_{j+2}(2t-1), \qquad\qquad j=0,\ldots,k-2. \end{aligned}$$

This means that, restricted to the first edge, the edge-counted basis functions are just the Lobatto functions with indices $j=2,\ldots,k$. However, we have already observed that the vertex-counted basis functions are just Ψ_0 and Ψ_1 when restricted to the edge. This implies that the functions associated to the edge produce the Lobatto basis when restricted to the edge. This one-dimensional basis is independent of the rest of the triangle and will therefore allow us to glue the local basis to the next triangle. The same arguments can be repeated for all other edges by rotating the indices $(1,2) \to (2,3) \to (3,1)$.

The element bubbles. We have, so far, $3+3(k-1)$ functions for the basis. We finish by adding the internal bubbles (for $k \ge 3$), that will vanish on the boundary of the element:

$$\lambda_1\lambda_2\lambda_3\big(\Phi_{j-1}\circ(\lambda_2-\lambda_1)\big)\big(\Phi_{l-1}\circ(\lambda_1-\lambda_3)\big), \qquad j,l \ge 1, \quad j+l \le k-1.$$

The hierarchy. We can renumber the above basis in different ways. A key property is the fact that the basis for $\mathbb{P}_k$ is a subset of the basis for $\mathbb{P}_{k+1}$. This was not true for the Lagrange or Bernstein-Bézier bases.

We can now renumber the basis 'geometrically':

$$H_{\boldsymbol{\alpha}}, \qquad |\boldsymbol{\alpha}| = k.$$

The elements $H_{k\mathbf{e}_j}$ are just the vertex functions λ_j. We add the edge functions by counting along the boundary of ∂K as if we were counting the nodes for the Lagrange basis and adding polynomial degrees. Similarly, we can count the interior bubbles as if they were associated to the interior nodes of the principal lattice. It is important to emphasize that the principal lattice is used for numbering purposes only and that this basis has no relation to that set of points. The following properties are exact copies of the ones satisfied by the Lagrange and BB bases:

(G.1) The only element of the basis that does not vanish at the vertex $\mathbf{v}_i$ is $H_{k\mathbf{e}_i}$. Therefore, barring the three vertex functions $H_{k\mathbf{e}_i}$, all other elements of the basis vanish on the vertices.

(G.2) The basis functions associated to points inside an edge can be considered as edge bubbles. They vanish on the two other edges of the element. Moreover, the $k+1$ basis functions associated to an edge (the interior of the edge and the vertices) fully determine the value of the polynomial on the edge.

(G.3) The basis functions associated to the interior of the triangle vanish on the boundary of the triangle.

2.3 Transition elements

3 Assembly

In this short section we are going to explore a way of organizing the assembly process for high order $\mathbb{P}_k$ finite elements on triangulations. We have already given a local numbering to the basis. This local way of numbering is compatible with assembling elements (gluing elements). This is an important issue that is invisible with Lagrange bases and comes to the foreground when we use other kind of bases. We can condense the issue in a couple of sentences:

- only one local basis functions is non-zero at each one of the vertices;
- only $k+1$ basis functions are non-zero on each one of the vertices;
- the restriction of the basis functions to an edge depends only on the edge, and not on the shape of the triangle.

Here's now some basic information needed from the mesh generator (part of it can be postprocessed):

- An ordered list of the vertices $\mathbf{v}_i$, $i = 1, \ldots, N_{\text{vert}}$.
- An ordered list of the edges e_i, $i = 1, \ldots, N_{\text{edg}}$. The edge information is a pair of indices for vertices. The order of the vertices gives the intrinsic orientation of the edge. It is convenient to assume that boundary edges are positively oriented, that is, when we move from the first vertex to the second, we leave the domain on the left.

- An ordered list of the elements K_i, $i = 1, \ldots, N_{\rm elt}$. The element information is an ordered triple of vertex indices. We assume that for every element, the orientation is positive.

- A cross-referenced list of edges counted by elements. This would be a list $N_{\rm elt} \times 3$ containing numbers from 1 to $N_{\rm edg}$. The part of the list corresponding to an element contains the global edge numbers (with the numbering given in the edge list) for the three edges of the triangle. This list has to be given in the same order that we decided for the basis (or, in other words, in a pre-established order given in the reference element).

- Finally, a list of orientations for the edges counted by elements. The edges have a local orientation (positive orientation from the point of view of the element), and a global orientation (implicit to the edge description as the segment joining two vertices). This table tells us if they match.

A **global count of degrees of freedom** for the $\mathbb{P}_k$ finite element can be easily achieved by:

- counting first the $N_{\rm vert}$ vertices

- counting the all $(k-1)$ degrees of freedom for each edge, in order of edge, thus counting the d.o.f.
$$N_{\rm vert} + 1, \ldots, N_{\rm vert} + (k-1)N_{\rm edg},$$

- counting finally the internal degrees of freedom
$$N_{\rm vert} + (k-1)N_{\rm edg} + 1, \ldots, N_{\rm vert} + (k-1)N_{\rm edg} + \frac{(k-1)(k-2)}{2} N_{\rm elt}.$$

If static condensation is used to eliminate the interior degrees of freedom, the last group of indices are never used in the assembly process. We can think of a function d.o.f. acting on vertices, edges, and elements, delivering the sets of indices associated to the corresponding geometric element:

$$\begin{aligned}
\text{d.o.f.}(n_i) &= n_i, \\
\text{d.o.f.}(e_i) &= N_{\rm vert} + (k-1)(e_i - 1) + [1, 2, \ldots, k-1], \\
\text{d.o.f.}(K) &= N_{\rm vert} + (k-1)N_{\rm edg} + [1, \ldots, \tfrac{(k-1)(k-2)}{2}].
\end{aligned}$$

From local to global. Assume that the element K is decribed by the vertices

$$\left[\begin{array}{ccc} n_1^K & n_2^K & n_3^K \end{array}\right], \qquad n_i^K \in \{1, \ldots, N_{\rm vert}\}$$

the edges

$$\left[\begin{array}{ccc} e_1^K & e_2^K & e_3^K \end{array}\right], \qquad e_i^K \in \{1, \ldots, N_{\rm edg}\},$$

with orientations

$$\left[\begin{array}{ccc} s_1^K & s_2^K & s_3^K \end{array}\right], \qquad s_i^K \in \{-1, 1\}.$$

We have counted the basis locally as explained in the first section. The global d.o.f. associated to K are given by the list

$$\text{d.o.f.}(n_i^K) = n_i^K, \qquad i = 1, 2, 3,$$

followed by

$$\text{d.o.f.}(e_i^K) \quad \text{if } s_i^K = 1, \qquad \text{or} \qquad \text{flip}(\text{d.o.f.}(e_i^K)) \quad \text{if } s_i^K = -1, \quad i = 1, 2, 3$$

and finally by

$$\text{d.o.f.}(K).$$

4 Exercises

1. Consider the Lobatto functions of Section 2. Show that

$$\int_{-1}^{1} \Psi_i'(x)\Psi_j'(x)\mathrm{d}x = \delta_{ij}, \qquad i, j \geq 2.$$

(This explains the normalization factor in their definitions.) Show also that

$$\int_{-1}^{1} \Psi_0'(x)\Psi_j'(x)\mathrm{d}x = \int_{-1}^{1} \Psi_1'(x)\Psi_j'(x)\mathrm{d}x = 0 \qquad j \geq 2.$$

2. Show that the Lobatto functions $\{\Psi_2, \ldots, \Psi_k\}$ define a basis for $\{p \in \mathbb{P}_k \, : \, p(\pm 1) = 0\}$.

3. **The Bernstein basis.** The polynomials

$$B_i(t) := \binom{k}{i}(1-t)^{k-i}t^i, \qquad i = 0, \ldots, k,$$

form a basis for $\mathbb{P}_k(t)$. Show that

$$B_i(0) = 0 \qquad i \geq 1, \qquad B_i(1) = 0, \qquad i \leq k-1.$$

Show that the restriction of the Bernstein-Bézier basis to one edge (you can parametrize it as $(1-t)\mathbf{v} + t\mathbf{w}$, where $\mathbf{v}$ and $\mathbf{w}$ are the endpoints of the edge) is the Bernstein basis.

4. **Bernstein-Bézier and hierarchical bases on the reference square.** Show that

$$B_{i,j}(\xi, \eta) := B_i(\xi)B_j(\eta) \qquad \text{and} \qquad \Psi_{i,j}(\xi, \eta) := \Psi_i(\xi)\Psi_j(\eta), \qquad 0 \leq i, j \leq k,$$

where $\{B_i\}$ is the Bernstein basis (exercise 2) and $\{\Psi_i\}$ is the Lobatto basis, are bases for $\mathbb{Q}_k$. Describe, in terms of the pairs of indices (i, j) which of these functions are associated to vertices, edges, and the interior of the element.

Lesson 8

Scaling arguments and FEM analysis

One of the important features of the Finite Element Method applied to elliptic problems (extensions of the Laplace operator) is that it converges for any solution, as long as it is in the energy (Sobolev) space. An you can actually prove that it does. No additional regularity is needed to prove convergence of the discrete solutions to the exact solution. FEM analysis is an interesting combination of simple results stated in an abstract language with some carefully crafted analysis on Sobolev spaces where you really need to roll up your sleeves to get the best results. In this lesson we are going to explore the language of FEM analysis and build an intuition on how it proceeds. We will do this for simple polygonal or polyhedral domains, but we will keep the equation quite general. You, the reader, will be asked to believe that some key results hold. It takes some time to get to understand the nitty-gritty of the details of proofs of some results in Sobolev spaces, so we will leave this for a better occasion and try to construct here a global understanding on how the FEM is analyzed.

(We have already met some of the forthcoming abstract ideas and arguments, and even the names of some of the key results, in the initial lessons. We will repeat everything again, as if we had never heard about them.)

1 Moving towards Hilbert spaces

1.1 Energy norm analysis

The domain, the problem, and the spaces. We are already quite versed in how to move from boundary value problems to variational formulations, but let us repeat again the main ideas, applied to a slightly more general problem than before. In a polygonal domain $\Omega \subset \mathbb{R}^2$ or a polyhedral domain $\Omega \subset \mathbb{R}^3$ we consider the elliptic partial differential equation

$$-\nabla \cdot (\kappa \nabla u) + c\, u = f \qquad \text{in } \Omega.$$

We will allow κ to be matrix-valued. We thus assume that

$$\kappa : \Omega \to \mathbb{R}^{d\times d}_{\mathrm{sym}},$$

where $\mathbb{R}^{d\times d}_{\mathrm{sym}}$ is the set of all symmetric $d \times d$ matrices ($d = 2$ or 3). The radiation coefficient is also variable $c : \Omega \to \mathbb{R}$. The boundary $\Gamma = \partial\Omega$ is subdivided into two non-overlapping parts Γ_D and Γ_N. We will assume that each of them is composed of full edges (resp. faces) of the boundary of the domain. We will next impose boundary conditions on Γ:

$$u = 0 \quad \text{on } \Gamma_D, \qquad (\kappa\nabla u)\cdot\boldsymbol{\nu} = g \quad \text{on } \Gamma_N,$$

where $\boldsymbol{\nu}$ is the unit outward pointing normal vector on Γ_N. *Taking the Dirichlet conditions to be homogeneous is a needed simplification at this time.* The case of non-homogeneous Dirichlet conditions requires some additional ingredients that we will introduce at the end of this lesson. The space where we look for the solution is

$$V = H^1_{\Gamma_D}(\Omega) = \{u \in H^1(\Omega) \,:\, u = 0 \text{ in } \Gamma_D\},$$

with the norm

$$\|u\|^2_{1,\Omega} = \|u\|^2_\Omega + \|\nabla u\|^2_\Omega = \int_\Omega |u|^2 + \int_\Omega |\nabla u|^2.$$

As mentioned in the first lessons of these notes, it is not that easy to define $H^1(\Omega)$ in a completely precise way. You can think of the space $H^1(\Omega)$ as what happens when we take limits of sequences in $\mathcal{C}^1(\overline{\Omega})$, when the limit is taken in the above norm. In other words, a function $u : \Omega \to \mathbb{R}$ is in $H^1(\Omega)$ when there exists

$$(u_n) \subset \mathcal{C}^1(\overline{\Omega}) \quad \text{such that} \quad \lim_{n\to\infty} \|u_n - u\|_{1,\Omega} = 0.$$

Some easy analysis shows that it makes sense to define $\nabla u \in L^2(\Omega)^d$ whenever $u \in H^1(\Omega)$. It is more complicated to show that it also makes sense to define $u|_\Gamma \in L^2(\Gamma)$, when $u \in H^1(\Omega)$ and that

$$\|u\|_\Gamma = \left(\int_\Gamma |u|^2\right)^{1/2} \le C\|u\|_{1,\Omega} \quad \forall u \in H^1(\Omega). \tag{8.1}$$

The weak (or variational) form of our boundary value problem is:

$$\left[\begin{array}{l} u \in V, \\ \displaystyle\int_\Omega (\kappa\nabla u)\cdot\nabla v) + \int_\Omega c\,u\,v = \int_\Omega f\,v + \int_{\Gamma_N} g\,v \qquad \forall v \in V. \end{array}\right. \tag{8.2}$$

Our next goal is the study of what appears in this variational form.

A symmetric bilinear form. Consider the bilinear form:

$$\int_\Omega (\kappa\nabla u)\cdot\nabla v.$$

We are next going to look at general requirements on the matrix valued function κ ensuring well-posedness to our problem. We have already assumed that $\kappa^\top(\mathbf{x}) = \kappa(\mathbf{x})$ for all $\mathbf{x}$, that is, κ takes values on the space of symmetric matrices. We next assume that the entries of the matrix-valued function κ are bounded functions. To make it simpler to state, let us assume that there exists C_κ such that

$$|\kappa(\mathbf{x})\boldsymbol{\xi}| \le C_\kappa|\boldsymbol{\xi}| \qquad \forall\boldsymbol{\xi} \in \mathbb{R}^d \qquad \forall\mathbf{x} \in \Omega.$$

Therefore[1]

$$\left|\int_\Omega (\kappa\nabla u)\cdot\nabla v\right| \le \|\kappa\nabla u\|_\Omega\|\nabla v\|_\Omega \le C_\kappa\|\nabla u\|_\Omega\|\nabla v\|_\Omega \qquad \forall u,v\in H^1(\Omega). \tag{8.3}$$

The second hypothesis on κ is more technical. We can write it as follows: there exists $c_\kappa > 0$ such that

$$(\kappa(\mathbf{x})\boldsymbol{\xi})\cdot\boldsymbol{\xi} \ge c_\kappa|\boldsymbol{\xi}|^2 \qquad \forall\boldsymbol{\xi}\in\mathbb{R}^d \qquad \forall\mathbf{x}\in\Omega. \tag{8.4}$$

Let us first use it, and we will next discuss what it means. As a consequence of this inequality

$$\int_\Omega (\kappa\nabla u)\cdot\nabla u \ge c_\kappa\int_\Omega |\nabla u|^2 = c_\kappa\|\nabla u\|_\Omega^2 \qquad \forall u\in H^1(\Omega). \tag{8.5}$$

The inequality (8.4) implies that the matrix $\kappa(\mathbf{x})$ is positive definite for all $\mathbf{x}$. It goes beyond that in requiring some sort of uniform positivity of $\kappa(\mathbf{x})$. It can be stated equivalently by saying the the smallest eigenvalue of $\kappa(\mathbf{x})$ is larger than a fixed positive quantity c_κ.

We still have to pay attention to the reaction term in the bilinear form of (8.2). This one is much easier: at the beginning we just assume that

$$|c(\mathbf{x})| \le C_c \qquad \text{and} \qquad c(\mathbf{x})\ge 0 \qquad \forall\mathbf{x}\in\Omega.$$

These inequalities imply

$$\left|\int_\Omega c\,u\,v\right| \le C_c\|u\|_\Omega\|v\|_\Omega \qquad \forall u,v\in H^1(\Omega) \tag{8.6}$$

and

$$\int_\Omega c\,u\,u \ge 0 \qquad \forall u\in H^1(\Omega). \tag{8.7}$$

The energy norm. Let us now consider the following 'norm':

$$|\!|\!|u|\!|\!|^2 = \int_\Omega(\kappa\nabla u)\cdot\nabla u + \int_\Omega c\,|u|^2. \tag{8.8}$$

At this point, it is easy to see that most of the axioms needed to call this function a norm are satisfied, but that strict positivity might be missing. Combining (8.3) and (8.6) we can show that

$$\begin{aligned}\left|\int_\Omega(\kappa\nabla u)\cdot\nabla v + \int_\Omega c\,u\,v\right| &\le C_\kappa\|\nabla u\|_\Omega\|\nabla v\|_\Omega + C_c\|u\|_\Omega\|v\|_\Omega\\ &\le \max\{C_\kappa,C_c\}\left(\|\nabla u\|_\Omega\|\nabla v\|_\Omega + \|u\|_\Omega\|v\|_\Omega\right)\\ &\le \max\{C_\kappa,C_c\}\left(\|\nabla u\|_\Omega^2+\|u\|_\Omega^2\right)^{1/2}\left(\|\nabla v\|_\Omega^2+\|v\|_\Omega^2\right)^{1/2}\\ &= M\|u\|_{1,\Omega}\|v\|_{1,\Omega} \qquad \forall u,v\in H^1(\Omega).\end{aligned}$$

[1] In this chapter we are going to be using the Cauchy-Schwarz inequality very often. The reader who feels unsure about this should review some of the many forms we can state the CS inequality before proceeding with the rest of this section.

Therefore

$$|\!|\!| u |\!|\!| \le M^{1/2} \|u\|_{1,\Omega} \qquad \forall u \in H^1(\Omega). \tag{8.9}$$

Our next point in inquiry concerns the hypotheses under which we can write

$$|\!|\!| u |\!|\!| \ge \alpha^{1/2} \|u\|_{1,\Omega} \qquad \forall u \in V \tag{8.10}$$

(note that we are not requiring the property to hold in the entire $H^1(\Omega)$ but are happy with the space V). Another way of writing (8.10) is

$$\int_\Omega (\kappa \nabla u) \cdot \nabla u + \int_\Omega c\,|u|^2 \ge \alpha \|u\|_{1,\Omega}^2 \qquad \forall u \in V. \tag{8.11}$$

Here are situations when this inequality holds true:

- If
  $$c(\mathbf{x}) \ge c_0 > 0 \qquad \forall \mathbf{x} \in \Omega,$$
  then we can use (8.5) and prove that
  $$\int_\Omega (\kappa \nabla u) \cdot \nabla u + \int_\Omega c\,|u|^2 \ge c_\kappa \|\nabla u\|_\Omega^2 + c_0 \|u\|_\Omega^2 \ge \min\{c_\kappa, c_0\} \|u\|_{1,\Omega}^2 \qquad \forall u \in H^1(\Omega).$$
  (Note that we have $H^1(\Omega)$ again.)

- If we just request $c \ge 0$ (including the chance that $c = 0$ everywhere) we have
  $$\int_\Omega (\kappa \nabla u) \cdot \nabla u + \int_\Omega c\,|u|^2 \ge c_\kappa \|\nabla u\|_\Omega^2 \ge c_\kappa C_{\Gamma_D} \|u\|_{1,\Omega}^2 \qquad \forall u \in H^1(\Omega) \quad u|_{\Gamma_D} = 0,$$
  as long as Γ_D is not the empty set. The inequality
  $$\|u\|_\Omega \le C \|\nabla u\|_\Omega$$
  is needed for the previous argument. It obviously does not hold in $H^1(\Omega)$ as can be seen by taking $u \equiv 1$. When $\Gamma_D = \Gamma$ (no Neumann boundary) this inequality is called the Poincaré-Friedrichs inequality. When Γ_D is non-empty it is a generalization of the Poincaré-Friedrichs inequality that is often called the same.

- The hardest situation is when $\Gamma = \Gamma_N$ (no Dirichlet boundary) and we cannot resort to a Poincaré-Friedrichs inequality. In this case we need to include new hypotheses for c. The following very general hypothesis happens to be enough:
  $$c(\mathbf{x}) \ge 0 \quad \forall \mathbf{x} \in \Omega, \qquad \int_\Omega c > 0. \tag{8.12}$$
  The reason why this is true is considerably more complicated. It uses some general compactness arguments or ideas from something called generalized Poincaré

inequalities or the Deny-Lions lemma[2]. Let us at least have a look at a tiny part of the argument. If $|||u||| = 0$, then

$$0 = \int_\Omega (\kappa \nabla u)\cdot \nabla u \ge c_\kappa \|\nabla u\|_\Omega^2 \quad \text{and} \quad \int_\Omega c\,|u|^2 = 0. \tag{8.13}$$

Therefore $\nabla u = 0$ and we can infer[3] that u is constant, say $u = C$. The second part of (8.13) and the second hypothesis in (8.12) then prove that $u = 0$. This means that the additional hypothesis (8.12) at least gurantees that the energy norm is a norm in $V = H^1(\Omega)$. (Let me emphasize that this is not needed if there's some Dirichlet boundary in the geometry.)

In summary, we have given hypotheses on the coefficients (κ and c) and on the type of boundary conditions (basically assuming the existence of Dirichlet conditions in some cases) guaranteeing that

$$C_1\|u\|_{1,\Omega} \le |||u||| \le C_2 \|u\|_{1,\Omega} \qquad \forall u \in V. \tag{8.14}$$

The right hand side. The right hand side of (8.2) leads to some bounds. For instance, just assuming that $f \in L^2(\Omega)$ we can estimate

$$\left|\int_\Omega f\,v\right| \le \|f\|_\Omega \|v\|_\Omega \le \|f\|_\Omega \|v\|_{1,\Omega} \qquad \forall v \in H^1(\Omega). \tag{8.15}$$

The integral term arising from the Neumann boundary condition is estimated using (8.1):

$$\left|\int_{\Gamma_N} g\,v\right| \le \|g\|_{\Gamma_N}\|v\|_{\Gamma_N} \le C\|g\|_{\Gamma_N}\|v\|_{1,\Omega} \qquad \forall v \in H^1(\Omega). \tag{8.16}$$

We are very close to being able to wrap up our problem in a wide theory. Recall that V is our space, let us use $\|\cdot\|_V$ for the norm in V and let us consider the bilinear and linear forms:

$$\begin{aligned} a(u,v) &= \int_\Omega (\kappa\nabla u)\cdot\nabla v + \int_\Omega c\,u\,v \\ \ell(v) &= \int_\Omega f\,v + \int_{\Gamma_N} g\,v. \end{aligned}$$

With these definitions our problem can be written in a very abstract looking way:

$$\left[\begin{array}{l} u \in V, \\ a(u,v) = \ell(v) \quad \forall v \in V. \end{array}\right. \tag{8.17}$$

[2] In their simplest versions, many well known theorems and inequalities of Sobolev space theory imply each other, which is the reason why there's no general agreement on how to refer to some of them. The results might not be identical in the most general statements though.

[3] You would think this is easy. One of the hard parts of defining spaces by completion is being sure that we haven't introduced anything funny like a function that is not constant but has vanishing gradient. Luckily we haven't, although this is not easy to show.

Well-posedness. We restart the theory working on the more abstract problem (8.17). We need V to be a Hilbert space (an inner product space where all Cauchy sequences are convergent). We need the **bilinear form** $a : V \times V \to \mathbb{R}$ to be symmetric

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

and we need its associated energy norm to be well defined

$$|\!|\!| u |\!|\!| = a(u,u)^{1/2}$$

and to be equivalent to the norm of V:

$$C_1 \|u\|_V \le |\!|\!| u |\!|\!| \le C_2 \|u\|_V \qquad \forall u \in V. \tag{8.18}$$

We finally need the **linear form** $\ell : V \to \mathbb{R}$ to be bounded

$$|\ell(v)| \le C_\ell \|v\|_V \qquad \forall v \in V. \tag{8.19}$$

Problem (8.17) can be understood as a representation theorem problem. We are given a bounded functional ℓ on V and we look for u such that the functional $a(u,\cdot) : V \to \mathbb{R}$ equals ℓ. The positive answer to existence and uniqueness of solution to this problem is given by the Riesz-Fréchet representation theorem[4]. It says that in the given hypothesis, problem (8.17) has a unique solution and that

$$|\!|\!| u |\!|\!| = \sup_{0 \neq v \in V} \frac{\ell(v)}{|\!|\!| v |\!|\!|} \le \frac{C_\ell}{C_1}.$$

(The last inequality follows from (8.19) and (8.18).) We can therefore bound

$$\|u\|_V \le \frac{C_\ell}{C_1^2},$$

which is an estimate of the norm of the solution in terms of the norm of the data.

Galerkin orthogonality. The FEM is a particular instance of a Galerkin method. In a Galerkin method applied to approximate the solution of (8.17), we choose a finite dimensional space $V_h \subset V$ (here h does not mean anything geometric; it's just a subindex denoting discretization) and solve

$$\left[\begin{array}{l} u_h \in V_h, \\ a(u_h, v_h) = \ell(v_h) \quad \forall v_h \in V_h. \end{array}\right. \tag{8.20}$$

[4] Savy readers might be expecting the use of the Lax-Milgram lemma at this point (we'll get to it later). The Lax-Milgram lemma is only needed for non-symmetric problems, although everyone in the FEM community resorts to it for symmetric problems and using the Riesz-Fréchet theorem is typically seens as a crime of pedantry. You are warned! Since we are already in a footnote, let me mention the Maurice Fréchet discovered the representation theorem independently of Frigyes Riesz, the eldest of the Riesz brothers. The younder Riesz, Marcel, proved another very important representation theorem in functional analysis, which has made pre-Wikipedia people confuse them quite often.

This problem is uniquely solvable again. Two reasons! The difficult one is because we can use the same argument as in the continuous problem with V_h instead of V. The easy one is because we can write (8.20) as an equivalent system of linear equations whose associated matrix is symmetric and positive definite. We have seen the argument in the initial lessons of this course, so we will not repeat it again. Note however that

$$a(u_h, v_h) = \ell(v_h) = a(u, v_h) \qquad \forall v_h \in V_h,$$

since $V_h \subset V$ and therefore

$$a(u - u_h, v_h) = 0 \qquad v_h \in V_h. \tag{8.21}$$

This property is often called **Galerkin orthogonality** and is plain orthogonality between the error $e_h = u - u_h$ and the space V_h, using the inner product $a(\cdot, \cdot)$ (the one whose associated norm is the energy norm) instead of the inner product of V. The following argument is then very simple to understand: for any $v_h \in V_h$,

$$\begin{aligned} |||u - u_h|||^2 &= a(u - u_h, u - u_h) && \text{(definition of } |||\cdot|||) \\ &= a(u - u_h, u) && \text{(Galerkin orthogonality)} \\ &= a(u - u_h, u - v_h) && \text{(Galerkin orthogonality again)} \\ &\le |||u - u_h||| \, |||u - v_h|||. && \text{(Cauchy-Schwarz)} \end{aligned}$$

Therefore

$$|||u - u_h||| \le |||u - v_h||| \qquad \forall v_h \in V_h,$$

or equivalently

$$|||u - u_h||| = \inf_{v_h \in V_h} |||u - v_h|||. \tag{8.22}$$

This means that the Galerkin method (the FEM in our particular example) provides the best approximation to the solution in the energy norm. Using (8.18) we can show that

$$C_1 \|u - u_u\|_V \le |||u - u_h||| \le |||u - v_h||| \le C_2 \|u - v_h\|_V \qquad \forall v_h \in V_h,$$

that is,

$$\|u - u_h\|_V \le \frac{C_2}{C_1} \inf_{v_h \in V_h} \|u - v_h\|_V. \tag{8.23}$$

This inequality (often called Céa's lemma or estimate) says that, up to a multiplicative constant depending only on the bilinear form, the Galerkin solution provides the best approximation in the norm of V. For this reason, the estimate (8.23) is often called **quasioptimality** of the Galerkin method.

1.2 General non-symmetric problems

Lax-Milgram and Céa. Symmetry of the bilinear form is not needed for the FEM to work. Non-symmetric bilinear forms appear in boundary value problems for non-self-adjoint partial differential equations. Here's a convection-reaction-diffusion equation

$$-\nabla \cdot (\kappa \nabla u) + \mathbf{b} \cdot \nabla u + c\, u = f \qquad \text{in } \Omega.$$

The bilinear form that appears in the variational formulation (actually, in one possible variational formulation) is

$$a(u,v) = \int_\Omega (\kappa \nabla u)\cdot\nabla v + \int_\Omega (\mathbf{b}\cdot\nabla u)\, v + \int_\Omega c\, u\, v. \tag{8.24}$$

If $\mathbf{b} : \Omega \to \mathbb{R}^d$ is bounded, then

$$\left| \int_\Omega (\mathbf{b}\cdot\nabla u)\, v \right| \le C_b \|\nabla u\|_\Omega \|v\|_\Omega \qquad \forall u, v \in H^1(\Omega).$$

Therefore, with the previous hypotheses for κ and c (they have to be bounded), we have **continuity a.k.a. boundedness** of the bilinear form

$$|a(u,v)| \le M \|u\|_V \, \|v\|_V \qquad \forall u, v \in V. \tag{8.25}$$

The other property that is needed is **coercivity** of the bilinear form

$$a(u,u) \ge \alpha \|u\|_V^2 \qquad \forall u \in V. \tag{8.26}$$

There are different sets of hypotheses that make bilinear form (8.24) coercive in $H^1(\Omega)$ or in the subspace where the Dirichlet boundary conditions hold. For instance, with any of the possibilities for κ and c given above, the hypotheses

$$\nabla\cdot\mathbf{b} = 0 \quad \text{in } \Omega, \qquad \mathbf{b}\cdot\boldsymbol{\nu} = 0 \quad \text{on } \Gamma_N$$

are sufficient conditions for coercivity of (8.24). These hypotheses can be relaxed to hypotheses involving all coefficients of the equation simultaneously[5]. Assuming that V is a Hilbert space, that $a : V \times V \to \mathbb{R}$ is a bilinear form that is bounded (8.25) and coercive (8.26) the Lax-Milgram lemma[6] proves that for every bounded linear functional $\ell : V \to \mathbb{R}$

$$|\ell(v)| \le C_\ell \|v\|_V \qquad \forall v \in V,$$

the problem

$$\left[\begin{array}{l} u \in V, \\ a(u,v) = \ell(v) \quad \forall v \in V, \end{array} \right. \tag{8.27}$$

has a unique solution. It is then clear from the inequalities

$$\alpha \|u\|_V^2 \le a(u,u) = \ell(u) \le C_\ell \|u\|_V$$

that

$$\|u\|_V \le \frac{C_\ell}{\alpha},$$

[5] And this is not even the whole story. Some convection-diffusion problems are not coercive but they can still be treated with the FEM. The reasons behind are, however, much more complicated and we won't deal with them here.

[6] Peter Lax is one of the most renown numerical analysts (and many other things) of the past century. In the numerical analysis and PDE community Arthur Milgram seems to have been relegated to be the conamer of this 'famous' lemma.

which shows that the norm of the solution is bounded by the 'norm' of the data. For a Galerkin discretization of (8.27)

$$\left[\begin{array}{l} u_h \in V_h, \\ a(u_h, v_h) = \ell(v_h) \quad \forall v_h \in V_h, \end{array}\right.$$

the Lax-Milgram lemma applied in the space V_h ensures existence and uniqueness of solution, while the following argument

$$\begin{aligned} \alpha\|u - u_h\|_V^2 &\le a(u - u_h, u - u_h) && \text{(coercivity)} \\ &= a(u - u_h, u - v_h) && \text{(Galerkin orthogonality; } v_h \in V_h) \\ &\le M\|u - u_h\|_V \|u - v_h\|_V && \text{(boundedness)} \end{aligned}$$

shows that

$$\|u - u_h\|_V \le \frac{M}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|_V$$

(Céa's lemma or quasioptimality) provides a bound of the error for the Galerkin methods in terms of the best approximation of the solution in the discrete space V_h. Before we go on, let me emphasize two points:

- When the bilinear form is not symmetric, you don't think in terms of the energy norm any longer. The reason is that $a(\cdot, \cdot)$ does not define an inner product, so it is incorrect to think that the associated quadratic form $a(u, u)$ is going to define a norm.

- Céa's lemma moves the problem of analyzing the error of Galerkin methods (the Finite Element Method!) to a problem of approximation theory. How small is the quantity

  $$\inf_{v_h \in V_h} \|u - v_h\|_V = \min_{v_h \in V_h} \|u - v_h\|_V$$

  and how does this quantity depend on properties of u that we can know in advance? While approximation theory will let us bound the right-hand side of the Céa estimate, this will be done based on assumptions on the smoothness of the (unknown) exact solution u. Figuring out how smooth the solution of the PDE is, depending on how smooth the data and the coefficients are (and on geometric properties of the domain Ω) is the realm of regularity theory, which is way more advanced than what we want to deal with at this stage.

2 Scaling of Sobolev norms

2.1 The big picture

In this section we abandon the neat world of linear functional analysis in Hilbert spaces to plunge in some soft-core mathematical analysis. Let me give you a preview before we get into details. Let us assume that the solution u to weak PDE (8.27) is smooth enough so that it is continuous[7]. Let now

$$W_h = \{u_h \in \mathcal{C}(\overline{\Omega})\,:\, u_h \in \mathbb{P}_k \quad \forall K \in \mathcal{T}_h\},$$

where $\mathcal{T}_h$ is a triangulation (or tetrahedrization) of Ω. We can then define $\Pi_h u \in W_h$ to be the interpolant of u in the nodes of the triangulation: vertices of $\mathbb{P}_1$, vertices and midpoints of edges for $\mathbb{P}_2$, etc for higher order. Recall that if $u = 0$ on Γ_D, then $\Pi_h u = 0$ on Γ_D and therefore $\Pi_h u \in V_h$. We then use Céa's estimate and bound

$$\|u - u_h\|_{1,\Omega} \le \frac{M}{\alpha} \inf_{v_h \in V_h} \|u - v_h\|_{1,\Omega} \le \frac{M}{\alpha} \|u - \Pi_h u\|_{1,\Omega}.$$

We have suddenly reduced the problem to a problem of analyzing the error for interpolation by piecewise polynomial functions. However, one of the beauties of interpolation by Finite Elements (by piecewise polynomials) is the fact that it is a local operator: if we consider the local interpolant $\Pi_K u \in \mathbb{P}_k(K)$, then $\Pi_h u|_K = \Pi_K u$. Why is this important? Because we just need to study interpolation by polynomials on triangles or tetrahedra, which looks like a so much simpler task. The full error will be bounded by adding the local errors since

$$\begin{aligned}
\|u - \Pi_h u\|^2_{1,\Omega} &= \int_\Omega |\nabla u - \nabla \Pi_h u|^2 + \int_\Omega |u - \Pi_h u|^2 \\
&= \sum_{K \in \mathcal{T}_h} \left(\int_K |\nabla u - \nabla \Pi_h u|^2 + \int_K |u - \Pi_h u|^2 \right) \\
&= \sum_{K \in \mathcal{T}_h} \|u - \Pi_K u\|^2_{1,K}.
\end{aligned}$$

(Pay attention to how surreptitiously I've susbtituted $\Pi_h u$ by $\Pi_K u$ on each triangle.) We'll go even further. Recall that for each element K we had an affine map from the reference element $\widehat{K}$

$$\mathrm{F}_K : \widehat{K} \to K. \tag{8.28}$$

Let us use $\widehat{\Pi}$ to denote the nodal $\mathbb{P}_k$ interpolation operator in the reference element. If we denote

$$\widehat{u} = u \circ \mathrm{F}_K : \widehat{K} \to \mathbb{R} \tag{8.29}$$

(this is u being pulled back to the reference element) then

$$\widehat{\Pi}\widehat{u} = \widehat{\Pi_K u},$$

or, with more detail,

$$\widehat{\Pi}(u \circ \mathrm{F}_K) = (\Pi_K u) \circ \mathrm{F}_K.$$

With words: interpolating u and bringing it back to the reference element is the same as pulling u back to $\widehat{K}$ and interpolating there. I won't give a formal proof of this, although

[7] You might be surprised about this assumption, given the fact that we are dealing with a PDE. The issue with weak formulations is that the functions are only observed under integral sign and Sobolev spaces might contain functions that are not continuous. We'll talk about this later on.

the property is quite simple to understand. The space of polynomials is invariant by invertible affine maps and the interpolation points are always taken by fixing barycentric coordinates, which are invariant by affine transformations[8]. The main advantage of the equality (8.29) is that our problem is reduced to:

- Understanding how the interpolation operator behaves in the reference element (as opposed to in every possible physical element).
- Studying how Sobolev norms are modified when we apply linear transformations.

The second of these questions is solved using what people in the FEM community call **scaling arguments** (you'll see why, wait for it). The first question will take us back to a more abstract world, by using a simple very clean cut theorem.

2.2 Some geometric ideas

Understanding scaling. The transformation from the reference element to the physical element K is

$$\mathrm{F}_K(\mathbf{x}) = \mathrm{B}_K\mathbf{x} + \mathbf{b},$$

where B_K is an invertible $d \times d$ matrix. This matrix can be factorized using its Singular Value Decomposition

$$\mathrm{B}_K = \mathrm{P}\Sigma\mathrm{Q}^\top, \qquad \mathrm{P}^\top\mathrm{P} = \mathrm{I}, \quad \mathrm{Q}^\top\mathrm{Q} = \mathrm{I},$$

where

$$\Sigma = \left[\begin{array}{cc} \sigma_1^K & 0 \\ 0 & \sigma_2^K \end{array}\right] \qquad \text{or} \qquad \Sigma = \left[\begin{array}{ccc} \sigma_1^K & 0 & 0 \\ 0 & \sigma_2^K & 0 \\ 0 & 0 & \sigma_3^K \end{array}\right], \qquad \sigma_i^K > 0,$$

depending on whether we are working in two or three dimensions. For readers who are not much acquainted with the SVD, just believe that this factorization exists and continue with the argument. A key assumption is the following: there exist two constants $C_1, C_2 > 0$ such that

$$C_1 h_K \le \sigma_i^K \le C_2 h_K \qquad \forall i \quad \forall K \in \mathcal{T}_h. \tag{8.30}$$

Here h_K is a quantity that measures the size of K. Typically h_K is taken to be the **diameter** of K. that is, the diameter of the smallest ball that contains K. We can also make the arguments work with h_K equal to the length of the longest edge of K. The first magnitude is preferred in expositions of FEM theory because it is easier to extend to the cases of paralellograms/parallelepipeds[9]. The bound (8.30) can be understood as putting some limits to how deformed (how flat or elongated) the element K is allowed to be. Note that the upper bound in (8.30) is not really an assumption, since it follows from geometric arguments about the SVD and the definition of h_K. I've always found it very useful to keep a very simple example in mind:

$$\mathbf{B}_K = h_K\mathrm{P}, \qquad \mathrm{P}^\top\mathrm{P} = \mathrm{I}, \tag{8.31}$$

[8] Let me clarify that this interpolation operator is used for analysis of approximation error. It doesn't really matter what basis we used in our FEM implementation. This is all about interpolation error.

[9] The analysis for the $\mathbb{Q}_k$ FEM is surprisingly similar and it can be carried out simultaneously.

which says that K is the result of rotating (and/or symmetrizing) $\widehat{K}$ and then rescaling it. This h_K here is not exactly the same as before, but it is equivalent.

Some context. Different FE textbooks provide different ways of presenting the scaling arguments. The most popular one, which has been around since the early books on FEM, is the use of the diameter h_K and the inradius ρ_K (the radius of the largest ball that we can fit in K). The typical hypothesis (substituting (8.30)) is **shape-regularity**

$$\frac{h_K}{\rho_K} \le C \qquad \forall K \in \mathcal{T}_h. \tag{8.32}$$

Other expositions use h_K and a parameter called **chunkiness**, measuring the deformation of the elements. And one more issue before we go on. In order not to overload notation (which can become overwhelming quite easily) the theory is done for all elements of any triangulation of any polygonal/polyhedral domain, as long as this triangulation satisfies the shape-regularity bounds. The idea of dealing with every triangulation and every domain is frequently left unmentioned. Think that the analysis is done on any physical element K, so it doesn't matter that K is part of a triangulation; it just floats in space, nicely ignoring all other elements. You can then get your mindset to thinking of any triangle/tetrahedron as long as its not too deformed following (8.30) or whatever equivalent hypothesis you prefer to handle.

Some notation. Once we start with the bounds we will get many constants like the ones in the hypothesis (8.30). In order to avoid having to keep track of them, we will often write

$$a_K \lesssim b_K$$

when there exists a constant C, independent of K, such that

$$a_K \le C b_K.$$

This constant will typically depend on:

- the deformation parameters (8.30) or (8.32),
- the polynomial degree k (which will reappear shortly).

Keeping track of constants depending on deformation is very important for analysis of FEM on very anisotropic meshes (but then the scaling arguments have to be redone from scratch). Keeping track of the polynomial degree[10] is important for analysis of the p-FEM, that is, the FEM when the polynomial degree is increased instead of the mesh being refined (this is called h-method). Finally

$$a_K \approx b_K \qquad \text{means} \qquad a_K \lesssim b_K \lesssim a_K.$$

[10] The polynomial degree will be part of what we will call finite-dimensional-arguments, that will be coupled with scaling arguments in the final stages of the analysis.

Some inequalities. The hypothesis (8.30) on the singular values of $\mathrm{B}_K = \mathrm{DF}_K$ (once more, think of $\Sigma = h_K\mathrm{I}$ to make your life simpler) implies all the following:

$$|\det \mathrm{B}_K| = \frac{|K|}{|\widehat{K}|} = \prod_i \sigma_i^K \approx h_K^d. \tag{8.33}$$

(Note how we have written two inequalities and hidden the constants with the symbol $\approx$.) We also have

$$|\mathrm{B}_K^\top \boldsymbol{\xi}| \approx h_K|\boldsymbol{\xi}| \qquad \forall \boldsymbol{\xi} \in \mathbb{R}^d$$

or equivalently

$$|\mathrm{B}_K^{-\top} \boldsymbol{\xi}| \approx h_K^{-1}|\boldsymbol{\xi}| \qquad \forall \boldsymbol{\xi} \in \mathbb{R}^d. \tag{8.34}$$

(Note now how this is an equality when $\sigma_i^K = h_K$ for all i. In that case the orthogonal matrices P and Q do not change the size of the vector.)

2.3 Scaling inequalities

The L^2 norm. Using (8.33) and recalling the notation $\widehat{u} = u \circ \mathrm{F}_K$ it is clear that

$$\int_K |u|^2 = |\det \mathrm{B}_K| \int_{\widehat{K}} |u \circ \mathrm{F}_K|^2 \approx h_K^d \int_{\widehat{K}} |\widehat{u}|^2,$$

or, in short-hand

$$\|u\|_K \approx h_K^{d/2} \|\widehat{u}\|_{\widehat{K}}. \tag{8.35}$$

The gradient seminorm. We have seen the following computation at the time when we moved to the reference element for evaluation of the stiffness matrix:

$$(\nabla u) \circ \mathrm{F}_K = \mathrm{B}_K^{-\top} \nabla (u \circ \mathrm{F}_K).$$

Therefore, using (8.34) and (8.33) we can estimate

$$\begin{aligned} \int_K |\nabla u|^2 &= |\det \mathrm{B}_K| \int_{\widehat{K}} |(\nabla u) \circ \mathrm{F}_K|^2 \\ &= |\det \mathrm{B}_K| \int_{\widehat{K}} |\mathrm{B}_K^{-\top} \nabla (u \circ \mathrm{F}_K)|^2 \\ &\approx |\det \mathrm{B}_K| h_K^{-2} \int_{\widehat{K}} |\nabla (u \circ \mathrm{F}_K)|^2, \end{aligned}$$

that is,

$$h_K \|\nabla u\|_K \approx |\det \mathrm{B}_K|^{1/2} \|\nabla \widehat{u}\|_{\widehat{K}} \approx h_K^{\frac{d}{2}} \|\nabla \widehat{u}\|_{\widehat{K}}. \tag{8.36}$$

If you put (8.35) and (8.36) together, you will see that the scaling for these two terms is different:

$$\|u\|_{1,K}^2 = \|u\|_K^2 + \|\nabla u\|_K^2 \approx h_K^{d/2} (\|\widehat{u}\|_{\widehat{K}}^2 + h_K^{-2} \|\nabla \widehat{u}\|_{\widehat{K}}^2).$$

The lesson to learn in this double inequality is that the Sobolev H^1 norm doesn't scale well (it is not a scalable norm) but the L^2 norm and the H^1 seminorm do.

The higher order seminorms. The Sobolev seminorm of order m on the element K can be defined as

$$|u|^2_{m,K} = \sum_{|\alpha|=m} \int_K |\partial^\alpha u|^2,$$

where

$$\partial^\alpha u = \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2}}, \qquad \alpha = (\alpha_1, \alpha_2), \qquad |\alpha| = \alpha_1 + \alpha_2$$

or (when we are in three dimensions)

$$\partial^\alpha u = \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \partial x_3^{\alpha_3}} \qquad \alpha = (\alpha_1, \alpha_2, \alpha_3), \qquad |\alpha| = \alpha_1 + \alpha_2 + \alpha_3.$$

The scaling for the Sobolev seminorm is

$$h_K^m |u|_{m,K} \approx h_K^{d/2} |\widehat{u}|_{m,\widehat{K}}. \tag{8.37}$$

Note that (8.35) and (8.36) are particular cases with $m = 0$ and $m = 1$, respectively. We are not going to prove this inequality here, but I will give a couple of ideas:

- When $\mathrm{F}_K(\mathbf{x}) = h_K \mathbf{x} + \mathbf{b}$, the inequality is almost straightforward to prove. In this case the derivatives do not get mixed and each partial derivative of u corresponds to a partial derivative of $\widehat{u}$. The same thing happens with B_K is a diagonal matrix.
- The general case can be proved by induction. If we take $v = \partial^\alpha u$ with $|\alpha| = m - 1$ and apply (8.36), we can prove the general case, by being extremely careful with changes of variables.

3 Convergence estimates

3.1 Interpolation error

The Bramble-Hilbert lemma. The study of interpolation on the reference element is done using a quite general argument that we will only present here in the particular case we are interested in. Let us consider the interpolation operator in the reference element: given a continuous function $v : \widehat{K} \to \mathbb{R}$, we denote by $\widehat{\Pi} v$ the unique element of $\mathbb{P}_k = \mathbb{P}_k(\widehat{K})$ that interpolates v on the principal k-lattice points, that is,

$$\widehat{\Pi} v(\widehat{\mathbf{p}}_\beta) = v(\widehat{\mathbf{p}}_\beta) \qquad \forall \beta, \quad |\beta| \le k$$

where

$$\widehat{\mathbf{p}}_{(i,j)} = \tfrac{1}{k}(i,j) \qquad \text{or} \qquad \widehat{\mathbf{p}}_{(i_1,i_2,i_3)} = \tfrac{1}{k}(i_1, i_2, i_3).$$

Then, there exists C that depends only on the polynomial degree k such that

$$\|v - \widehat{\Pi} v\|_{1,\widehat{K}} \le C |v|_{k+1,\widehat{K}} \qquad \forall v \in H^{k+1}(\widehat{K}). \tag{8.38}$$

We will refer to this result as the Bramble-Hilbert lemma [11], although the Bramble-Hilbert lemma is more general, dealing with other 'interpolation' operators, reference elements, etc. In different communities, this result receives different names. It is common to refer to this kind of result as the Deny-Lions lemma. In any case, this result can be proved to be a consequence of a much more general theorem in Sobolev space theory, the Rellich-Kondrachov compactess theorem. Even without a proof (it requires too much handling of Sobolev spaces for what we are using in these notes), the reader can note how the result writes the error of interpolating with polynomials of degree k in terms of the derivatives of order $k+1$, which is exactly what one would expect.

Interpolation of the physical element. As we have already mentioned,

$$\widehat{\Pi_K u} = \widehat{\Pi}\widehat{u} \tag{8.39}$$

(interpolation and pullback to the reference element can be interchanged), and therefore

$$\begin{aligned}
h_K^2 \|u - \Pi_K u\|_{1,K}^2 &\lesssim \|u - \Pi_K u\|_K^2 + h_K^2 \|\nabla u - \nabla \Pi_K u\|_K^2 \\
&\lesssim h_K^d \|\widehat{u - \Pi_K u}\|_{\widehat{K}}^2 + h_K^d \|\nabla(\widehat{u - \Pi_K u})\|_{\widehat{K}}^2 && \text{(scaling)} \\
&= h_K^d \|\widehat{u} - \widehat{\Pi}\widehat{u}\|_{1,\widehat{K}}^2 && \text{(by (8.39))} \\
&\lesssim h_K^d |\widehat{u}|_{k+1,\widehat{K}}^2 && \text{(Bramble-Hilbert)} \\
&\lesssim h_K^{2(k+1)} |u|_{k+1,K}^2, && \text{(scaling)}
\end{aligned}$$

that is,

$$\|u - \Pi_K u\|_{1,K} \lesssim h_K^k |u|_{k+1,K}.$$

Global interpolation estimate. Assume, for a moment, that boundary conditions do not play any role. We start with a smooth enough function $u : \Omega \to \mathbb{R}$ and apply the interpolation operator Π_K on each of the elements $K \in \mathcal{T}_h$. The resulting function, $\Pi_h u$ is continuous and piecewise polynomial of degree k. Note now that the value of Π_K on a face/edge depends only on the values of u on that face/edge. Therefore, if $u = 0$ on Γ_D, then $\Pi_h u = 0$ on Γ_D. Finally, we aggregate the local interpolation estimates on the elements. Denoting

$$h := \max_{K \in \mathcal{T}_h} h_K, \tag{8.40}$$

we can bound

$$\begin{aligned}
\|u - \Pi_h u\|_{1,\Omega}^2 &= \sum_{K \in \mathcal{T}_h} \|u - \Pi_K u\|_{1,K}^2 \\
&\lesssim \sum_{K \in \mathcal{T}_h} h_K^{2k} |u|_{k+1,K}^2 \\
&\le h^{2k} \sum_{K \in \mathcal{T}_h} |u|_{k+1,K}^2,
\end{aligned}$$

[11] This Hilbert is not the famous David Hilbert, father of Hilbert spaces. This lemma is well-known in the Finite Element community and was proved by James Bramble and Stephen Hilbert.

and therefore

$$\|u - \Pi_h u\|_{1,\Omega} \lesssim h^k |u|_{k+1,\Omega} \qquad \forall u \in H^{k+1}(\Omega). \tag{8.41}$$

This is the typical estimate that you learn in basic Finite Element theory. Note that we have overestimated all the element sizes by the maximum of all of them ($h_K \le h$), but that we could think of keeping each element size with the contribution of K to the global H^{k+1} seminorm. The result is less neat, but it allows us to take bigger values of h_K in elements where the $H^{k+1}(K)$ seminorm is small.

3.2 Finite Element estimates

Estimate using full regularity. We are ready to obtain an estimate for the finite element solution of an elliptic linear PDE of the second order. Céa's lemma gives

$$\|u - u_h\|_{1,\Omega} \lesssim \inf_{v_h \in V_h} \|u - v_h\|_{1,\Omega}.$$

The constants hidden in the symbol $\lesssim$ are related to the coefficients of the equation and, in some case, to the geometric configuration of the boundary, namely, to how large the Dirichlet boundary Γ_D is in comparison to the full boundary. However, the interpolant $\Pi_h u$ is only well defined when u is continuous, which is not a requirement for a general solution of the problem. Let us further assume that $u \in H^{k+1}(\Omega)$. Then, we can concatenate two inequalities:

$$\inf_{v_h \in V_h} \|u - v_h\|_{1,\Omega} \le \|u - \Pi_h u\|_{1,\Omega} \lesssim h^k |u|_{k+1,\Omega}.$$

The last inequality hides many more constants: we have used scaling inequalities galore (hiding constants related to shape-regularity) and the Bramble-Hilbert lemma (hiding a constant depending on the polynomial degree). In any case, we have shown that the $\mathbb{P}_k$ FEM has order k for smooth enough solutions.

Estimate using reduced regularity. There are several ways of approaching the analysis of FEM when the solution is not in $H^{k+1}(\Omega)$. We can redo the local interpolation estimates (that requires a modification of the statement of the Bramble-Hilbert lemma), or we can proceed as follows. Let V_h^ℓ be the FE space of degree ℓ on the same triangulation, with $1 \le \ell < k$. Then

$$\|u - u_h\|_{1,\Omega} \lesssim \inf_{v_h \in V_h^k} \|u - v_h\|_{1,\Omega} \le \inf_{v_h \in V_h^\ell} \|u - v_h\|_{1,\Omega}$$

since $V_h^k \subset V_h^\ell$. Using the estimate for the interpolation operator on V_h^ℓ, it follows that

$$\|u - u_h\|_{1,\Omega} \lesssim h_K^\ell |u|_{\ell+1,\Omega} \qquad \text{if } u \in H^{\ell+1}(\Omega), \quad 1 \le \ell \le k.$$

This proves convergence for solutions that are, in the worst case, in $H^2(\Omega)$. Note that for solutions $u \in H^2(\Omega)$, this results predicts *convergence of order one, no matter what the polynomial degree is.* This would seem to indicate that there's no point in using high order FEM except for cases when we are confident that the solution is very smooth.

The situation is actually much more complicated. First of all, we have used many upper bounds that might be overestimating the error in significant ways. For instance, in areas where the solution is very smooth, the high order interpolation operator will definitely give better results, even if we cannot claim convergence of high order in the sense of having a higher power of h. Even in the worst case, we have

$$\inf_{v_h \in V_h^k} \|u - v_h\|_{1,\Omega} \le \inf_{v_h \in V_h^\ell} \|u - v_h\|_{1,\Omega},$$

which clearly hints at the fact that the higher order methods will give a better solution. This can be proved if we work in the energy norm for symmetric problems, since in that case

$$|\!|\!| u - u_h^k |\!|\!| = \inf_{v_h \in V_h^k} |\!|\!| u - v_h |\!|\!| \le \inf_{v_h \in V_h^\ell} |\!|\!| u - v_h |\!|\!| = |\!|\!| u - u_h^\ell |\!|\!| \qquad \ell \le k,$$

where u_h^m is the $\mathbb{P}_m$ Finite Element solution.

General convergence theorem. One of the great properties of the FEM is convergence of the method as $h \to 0$ (meshes are refined) no matter the regularity. The result follows from a density argument. It says something like the following: the space $H^2(\Omega)$ is dense in $H^1(\Omega)$ (which means that every function in $H^1(\Omega)$ can be approximated to arbitrary precision by functions on $H^2(\Omega)$) and the FEM converges in $H^1(\Omega)$ for $H^2(\Omega)$ solutions; therefore the FEM converges for any solution in $H^1(\Omega)$.

3.3 The Aubin-Nitsche trick

A neat trick. The Aubin-Nitsche trick was named after two mathematicians[12] that found this idea (sometimes referred to as a lemma) independently. It is also called an estimate by duality. It is a very elegant example of a global superconvergence result. It needs an additional regularity hypothesis that we will next discuss.

The adjoint problem. To make this as general as possible let us consider a non-symemtric PDE

$$-\nabla \cdot (\kappa \nabla u) + \mathbf{b} \cdot \nabla u + c\, u = f \qquad \text{in } \Omega,$$

with boundary conditions

$$u = 0 \quad \text{on } \Gamma_D, \qquad (\kappa \nabla u) \cdot \boldsymbol{\nu} = g \quad \text{on } \Gamma_N.$$

The variational formulation of this problem involves the space

$$V = \{ u \in H^1(\Omega) \, : \, u = 0 \text{ on } \Gamma_D \},$$

the bilinear form

$$a(u, v) = \int_\Omega (\kappa \nabla u) \cdot \nabla v + \int_\Omega (\mathbf{b} \cdot \nabla u) v + \int_\Omega c\, u\, v$$

[12] One in France, Jean-Pierre Aubin, and one in Germany, Joachim Nitsche

and the linear form

$$\ell(v) = \int_\Omega f\, v + \int_{\Gamma_N} g\, v.$$

The adjoint problem is the problem that arises from the variational formulation that uses the transposed bilinear form. We are going to feed this problem with a simple linear form in the right-hand side

$$\widetilde{\ell}(v) = \int_\Omega \theta\, v \qquad \theta \in L^2(\Omega),$$

and then look for the solution of

$$\left[\begin{array}{l} w \in V, \\ a(v,w) = \widetilde{\ell}(v) \qquad \forall v \in V. \end{array}\right.$$

You have to do some integration by parts to recognize that this problem is the variational formulation of the following PDE:

$$-\nabla \cdot (\kappa \nabla w) - \nabla \cdot (w\mathbf{b}) + c\, w = \theta \qquad \text{on } \Omega \tag{8.42}$$

with boundary conditions

$$w = 0 \quad \text{on } \Gamma_D, \qquad (\kappa \nabla w) \cdot \boldsymbol{\nu} + (\mathbf{b} \cdot \boldsymbol{\nu}) w = 0 \quad \text{on } \Gamma_N. \tag{8.43}$$

(This is not very difficult to check and you are suggested to try this as an exercise.) The Aubin-Nitsche trick requires the following hypothesis to hold: for every θ, the solution of (8.42) with boundary conditions (8.43) is in the space $H^2(\Omega)$ and can be bounded as follows

$$\|w\|_{2,\Omega} \le C \|\theta\|_\Omega. \tag{8.44}$$

It is not simple to find general conditions for this property (often called $H^2(\Omega)$-regularity of the adjoint problem) to hold. A particular case where (8.44) holds:

- The diffusion coefficients κ and the reaction coefficient c are smooth functions and there is no convection coefficient $\mathbf{b} \equiv 0$.

- The domain Ω is a convex polygon/polyhedron.

Superconvergence. Here is the trick. We solve the adjoint problem (8.42)-(8.43) with $\theta = u - u_h$. Let w be the adjoint solution

$$\left[\begin{array}{l} w \in V, \\ a(v,w) = (\theta, v)_\Omega \qquad \forall v \in V, \end{array}\right.$$

and let w_h be its associated Finite Element approximation

$$\left[\begin{array}{l} w_h \in V_h, \\ a(v_h,w_h) = (u - u_h, v_h)_\Omega \qquad \forall v_h \in V_h. \end{array}\right.$$

Then

$$\begin{aligned}
\|u-u_h\|_\Omega^2 &= (u-u_h,u-u_h)_\Omega \\
&= a(u-u_h,w) && \text{(def of } w) \\
&= a(u-u_h,w-w_h) && \text{(Galerkin orthogonality for } u-u_h) \\
&\le M\|u-u_h\|_{1,\Omega}\|w-w_h\|_{1,\Omega} && \text{(boundedness)} \\
&\lesssim \|u-u_h\|_{1,\Omega} h|w|_{2,\Omega} && \text{(estimate for } H^2(\Omega) \text{ solution)} \\
&\lesssim h\|u-u_h\|_{1,\Omega}\|u-u_h\|_\Omega. && (H^2(\Omega)\text{-regularity})
\end{aligned}$$

We have used the convergence estimate for the FEM applied to the adjoint problem (if the original problem is coercive, so is the adjoint; why?) and the $H^2(\Omega)$ regularity hypothesis (8.44) on the adjoint problem. We have thus proved that

$$\|u-u_h\|_\Omega \lesssim h\|u-u_h\|_{1,\Omega},$$

which means that, given the additional regularity hypothesis on the adjoint problem, the FEM converges with an additional order in the $L^2(\Omega)$ norm.

3.4 Dirichlet boundary conditions

The problem. In this section we will explore, just for a moment, what to do with Dirichlet boundary conditions. Let us thus consider our original problem now with non-zero Dirichlet boundary conditions:

$$\left[\begin{array}{l}
-\nabla\cdot(\kappa\nabla u) + c\,u = f \qquad \text{in } \Omega, \\
u = g_0 \qquad \text{on } \Gamma_D, \\
(\kappa\nabla u)\cdot\boldsymbol{\nu} = g \qquad \text{on } \Gamma_N.
\end{array}\right.$$

In the variational formulation we use the same notation as before, so there's no need to reintroduce the spaces, linear and bilinear form:

$$\left[\begin{array}{l}
u \in H^1(\Omega), \\
u|_{\Gamma_D} = g_0, \\
a(u,v) = \ell(v) \qquad \forall v \in V.
\end{array}\right.$$

Note that the space $H^1(\Omega)$ appears as a bigger space than V. To make notation a bit more general, let us write $W = H^1(\Omega)$ and keep V for the space where the boundary conditions are imposed[13]. How do you deal with this problem? It's actually quite easy: take any $u_{\mathrm{nh}} \in W$ such that $u_{\mathrm{nh}} = g_0$ on Γ_D and define $u_0 = u - u_{\mathrm{nh}}$, which obviously satisfies

$$\left[\begin{array}{l}
u_0 \in V, \\
a(u_0,v) = \ell(v) - a(u_{\mathrm{nh}},v) \qquad \forall v \in V.
\end{array}\right. \tag{8.45}$$

We can now see that the hypotheses for the Lax-Milgram lemma have to be modified (just a little) to deal with this problem, which as a more complicated right-hand-side:

[13] Typically this is done in the reverse order: V is the full space and V_0 is the subspace with zero boundary conditions. There's no point on changing notation now for just on small section though.

- We are fine with coercivity in V.
- Boundedness of ℓ in V is still enough.
- Boundedness of $a(\cdot,\cdot)$ has to be extended to the full space

  $$|a(u,v)| \leq M\|u\|_V\|v\|_V \qquad \forall u,v \in W.$$

  (Actually we can leave $v \in V$.) This is needed for the new linear map

  $$\xi(v) = \ell(v) - a(u_{\mathrm{nh}}, v)$$

  to still be bounded:

  $$|\xi(v)| \leq C_\ell\|v\|_V + M\|u_{\mathrm{nh}}\|_V\|v\|_V.$$

These hypotheses imply that the reduced problem (8.45) has a unique solution. Note that we recover the original solution by adding back u_{nh} and

$$\begin{aligned}
\|u\|_V &\leq \|u_0\|_V + \|u_{\mathrm{nh}}\|_V \\
&\leq \frac{C_\xi}{\alpha} + \|u_{\mathrm{nh}}\|_V && \text{(Lax-Milgram)} \\
&\leq \frac{1}{\alpha}(C_\ell + M\|u_{\mathrm{nh}}\|_V) + \|u_{\mathrm{nh}}\|_V && \text{(Bound for } C_\xi\text{)} \\
&\leq \frac{C_\ell}{\alpha} + \left(1 + \frac{M}{\alpha}\right)\|u_{\mathrm{nh}}\|_V.
\end{aligned}$$

The bound can be refined by noticing that u is unique even if we choose different u_{nh} as a **lifting** of the Dirichlet boundary condition:

$$\|u\|_V \leq \frac{C_\ell}{\alpha} + \left(1 + \frac{M}{\alpha}\right)\inf\{\|u_{\mathrm{nh}}\|_V \,:\, u_{\mathrm{nh}} \in W, u_{\mathrm{nh}} = g_0 \text{ on } \Gamma_D\}. \tag{8.46}$$

A fractional Sobolev norm. Let us recall the inequality (8.1)

$$\|u\|_\Gamma \leq C\|u\|_{1,\Omega} \qquad \forall u \in H^1(\Omega),$$

which gives an extension of the idea of restriction to the boundary for a function in $H^1(\Omega)$. In the set[14]

$$H^{1/2}(\Gamma) := \{g \in L^2(\Gamma) \,:\, g = u|_\Gamma \quad \text{for some } u \in H^1(\Omega)\},$$

we can define the so called **image norm**

$$\|g\|_{1/2,\Gamma} = \inf\{\|u\|_{1,\Omega} \,:\, u \in H^1(\Omega), u = g \text{ on } \Gamma\}.$$

This norm makes the space $H^{1/2}(\Gamma)$ a Hilbert space. Note that this is the expression that we get in the right-hand side of (8.46).

[14] There's no simple reason why we should call this set $H^{1/2}(\Gamma)$ apart from a vague notion that it can be proved to be in the 'middle' of $H^0(\Gamma) = L^2(\Gamma)$ and the space $H^1(\Gamma)$.

The discrete case. In the initial lessons of this course we have always imposed the non-homogeneous Dirichlet conditions by imposing the value of u_h to match the value of the Dirichlet data on the Dirichlet nodes. This was easy to do because we were thinking of the Lagrange basis, where coefficients are nodal values. The process is less clear when we use other types of bases for the FEM. As usual W_h is the full Finite Element space and $V_h \subset V$ is the space with zero Dirichlet boundary conditions. No matter what we do, we can think that there is a boundary space

$$\Phi_h := \{u_h|_{\Gamma_D} \,:\, u_h \in W_h\}.$$

This space can be understood as a space of $\mathbb{P}_k$ finite elements on the triangulation of Γ_D that is inherited from $\mathcal{T}_h$. We then choose $g_{0,h} \in \Phi_h$, for instance, by interpolating $g : \Gamma_D \to \mathbb{R}$ on all Dirichlet nodes and write the FEM approximation to our problem:

$$\left[\begin{array}{l} u_h \in W_h, \\ u_h|_{\Gamma_D} = g_{0,h}, \\ a(u_h, v_h) = \ell(v_h) \qquad \forall v_h \in V_h. \end{array}\right.$$

The error analysis is very similar to the one for the continuous case. We choose any

$$u_{\mathrm{nh},h} \in W_h \text{ such that } u_{\mathrm{nh},h}|_{\Gamma_D} = g_{0,h}$$

and think in terms of the function $u_{0,h} = u_h - u_{\mathrm{nh},h} \in V_h$. Then

$$\begin{aligned} \alpha\|u_h - u_{\mathrm{nh},h}\|_V^2 &\le a(u_h - u_{\mathrm{nh},h}, u_h - u_{\mathrm{nh},h}) && \text{(coercivity in } V_h \subset V) \\ &= a(u - u_{\mathrm{nh},h}, u_h - u_{\mathrm{nh},h}) && \text{(Galerkin orthogonality)} \\ &\le M\|u - u_{\mathrm{nh},h}\|_V \|u_h - u_{\mathrm{nh},h}\|_V && \text{(boundedness),} \end{aligned}$$

which implies that

$$\|u - u_h\|_V \le \|u - u_{\mathrm{nh},h}\|_V + \|u_{\mathrm{nh},h} - u_h\|_V \le \left(1 + \frac{M}{\alpha}\right)\|u - u_{\mathrm{nh},h}\|_V.$$

Since the **discrete lifting** $u_{\mathrm{nh},h}$ of the boundary condition $g_{0,h}$ is arbitrary, we have proved that

$$\|u - u_h\|_{1,\Omega} \le \left(1 + \frac{M}{\alpha}\right) \inf\{\|u - w_h\|_{1,\Omega} \,:\, w_h \in W_h, w_h|_{\Gamma_D} = g_{0,h}\}, \tag{8.47}$$

which is the traditional version of Céa's lemma for non-homogeneous boundary conditions.

Error bounds. If we assume that $u \in H^{k+1}(\Omega)$ and $g_{0,h}$ has been built by interpolation, that is, matching u_h and g_0 on all Dirichlet nodes, then we can bound

$$\|u - u_h\|_{1,\Omega} \le \left(1 + \frac{M}{\alpha}\right)\|u - \Pi_h u\|_{1,\Omega} \lesssim h^k |u|_{k+1,\Omega}.$$

If we have reduced regularity the argument that we used for the homogeneous problem does not work. However, the statement we gave for the Bramble-Hilbert lemma can be modified to say now that

$$\|v-\widehat{\Pi}v\|_{1,\widehat{K}}\le C|v|_{\ell+1,\widehat{K}}\qquad 1\le\ell\le k\quad\forall v\in H^{\ell+1}(\widehat{K}).$$

With this intermediate result, we can go ahead and proof convergence for solutions $u\in H^2(\Omega)$.

A final word. The correct imposition of Dirichlet boundary conditions and its analysis is a complicated issue, especially when the Dirichlet data and/or the solution are not very smooth functions. The way to proceed is to show an enhanced version of Céa's estimate (8.47)

$$\|u-u_h\|_{1,\Omega}\lesssim\inf_{w_h\in W_h}\|u-w_h\|_{1,\Omega}+\|g_0-g_{0,h}\|_{1/2,\Gamma_D},$$

which separates the effects of approximation of the solution in the Finite Element space from the initial approximation of the Dirichlet data. The proof of this result is not trivial though, requiring the introduction of some 'interpolation' operator for non-smooth functions, due to Ridgway Scott and Shangyou Zhang. It also requires handling approximation properties in fractional Sobolev norms on the boundary. We will keep away from these complications and move on to other problems.

4 Exercises

1. Show a scaling inequality for the rescaled Sobolev norm

$$\|u\|_{1,K,h}:=\sqrt{\|u\|_K^2+h_K^2\|\nabla u\|_K^2}.$$

2. Prove the scaling inequality (8.37) for the case when

$$\mathrm{B}_K=\left[\begin{array}{ccc}\sigma_1^K & & \\ & \sigma_2^K & \\ & & \sigma_3^K\end{array}\right]\qquad \sigma_i^K\approx h_K.$$

3. If Π_h is the interpolation operator by continuous piecewise polynomials of degree k, show that

$$\|u-\Pi_h u\|_\Omega\lesssim h^{k+1}|u|_{k+1,\Omega}.$$

4. Find the variational formulation of the problem

$$\left[\begin{array}{l}-\nabla\cdot(\kappa\nabla w+w\mathbf{b})+c\,w=\theta\qquad\text{in }\Omega,\\ w=0\quad\text{on }\Gamma_D,\\ (\kappa\nabla w+w\mathbf{b})\cdot\boldsymbol{\nu}=0\quad\text{on }\Gamma_N.\end{array}\right.$$

5. Assume that we can count on the following inequality

$$\|v - \widehat{\Pi} v\|_{1,\widehat{K}} \le C |v|_{\ell+1,\widehat{K}} \qquad 1 \le \ell \le k \quad \forall v \in H^{\ell+1}(\widehat{K})$$

for the $\mathbb{P}_k$ interpolation operator on the reference element. Repeat the scaling inequalities to show that

$$\|u - \Pi_h u\|_{1,\Omega} \lesssim h^{\ell} |u|_{\ell+1,\Omega}, \qquad 1 \le \ell \le k.$$

Lesson 9

An introduction to Raviart-Thomas elements

Mixed finite elements form a class of their own, although their name is slightly misleading, because they include problems with very different features, sharing a common structure. In this lesson we are going to explore the basic finite element discretizations for a **mixed formulation of an elliptic equation**. This formulation falls into a general category of **saddle point problems.** You might wonder why we should even try to formulate the problem in a different way and then try to overcome the difficulties of discretizing the new problem. Mixed formulations bring new features to the table: they compute an approximation of the flux (the gradient in the case of the Laplacian) with better quality than the FEM (it has some conservation properties the FEM solution hasn't) and they can be easily postprocessed element by element to yield higher order approximations. While we will not have a look at them in this course, there are problems that have mixed structure from the very beginning: the most popular one is the Stokes problem, modeling viscous fluid flow. Mixed finite elements are also a good excuse to present divergence-conforming finite elements, closely related to the curl-conforming finite elements that are used in electromagnetism.

1 The mixed Laplacian

1.1 Problems with mixed form

Being ambitious. I believe that at this time we can be ambitious and work in two and three dimensions at the same time. You will barely notice the difference at the time of formulating the problems. When we discretize, we will have to count depending on the dimension d (again, $d = 2$ or $d = 3$) and we will do some funny renaming to avoid notational duplications like referring to edges/faces depending on the dimension.

The model problem. We consider a symmetric-matrix-valued coefficient $\kappa : \Omega \to \mathbb{R}^{d\times d}_{\rm sym}$ that, as usual, is assumed to be bounded and uniformly positive definite

$$(\kappa(\mathbf{x})\boldsymbol{\xi}) \cdot \boldsymbol{\xi} \geq \kappa_0 |\boldsymbol{\xi}|^2 \qquad \forall \mathbf{x} \in \Omega, \quad \forall \boldsymbol{\xi} \in \mathbb{R}^d.$$

Note that this property is then inherited by the inverse matrix

$$a := \kappa^{-1} : \Omega \to \mathbb{R}^{d\times d}_{\mathrm{sym}}.$$

I am going to use the coefficient a to avoid having to write κ^{-1} way too many times. You'll see why soon. The problem is a generalized Laplacian with mixed boundary conditions

$$\left[\begin{array}{l} -\nabla\cdot(\kappa\nabla u) = f \qquad \text{in } \Omega, \\ u = g_0 \qquad \text{on } \Gamma_D, \\ (\kappa\nabla u)\cdot\boldsymbol{\nu} = g_1 \qquad \text{on } \Gamma_N. \end{array}\right.$$

We introduce a new variable, the flux[1]

$$\mathbf{q} := \kappa\nabla u.$$

The modified set of equations has a very particular structure:

$$\left[\begin{array}{llll} a\,\mathbf{q} - \nabla u & = 0 & \text{in } \Omega, & \text{(state equation)} \\ \nabla\cdot\mathbf{q} & = -f & \text{in } \Omega, & \text{(equilibrium)} \\ u = g_0 & & \text{on } \Gamma_D, & \\ \mathbf{q}\cdot\boldsymbol{\nu} = g_1 & & \text{on } \Gamma_N. & \end{array}\right.$$

This model is often called **Darcy flow**. It corresponds to a linearized version of the equations of fluid flow in saturated porous media.

Weak formulation. We are going to use integration by parts in the state equation, and leave the equilibrium equation as is (more or less). In that way $\mathbf{q}$ and its associated test function will be subject to the divergence operator, while all the derivatives will have disappeared from u. We choose a smooth vector field $\mathbf{p}$ such that $\mathbf{p}\cdot\boldsymbol{\nu} = 0$ on Γ_N (which is the homogeneous version of the boundary condition satisfied by $\mathbf{q}$). We multiply by $\mathbf{p}$ and integrate by parts (using the divergence theorem)

$$\begin{aligned} \int_\Omega (a\,\mathbf{q})\cdot\mathbf{p} &= \int_\Omega \nabla u\cdot\mathbf{p} \\ &= -\int_\Omega u\,(\nabla\cdot\mathbf{p}) + \int_\Gamma u\,(\mathbf{p}\cdot\boldsymbol{\nu}) \\ &= -\int_\Omega u\,(\nabla\cdot\mathbf{p}) + \int_{\Gamma_D} g_0\,(\mathbf{p}\cdot\boldsymbol{\nu}). \end{aligned}$$

In the last line we have substituted the Dirichlet boundary condition and eliminated the integral over Γ_N by imposing $\mathbf{p}\cdot\boldsymbol{\nu} = 0$ on Γ_N. The equilibrium equation (or conservation

[1] The actual flux, if we think of heat diffusion, would be $-\kappa\nabla u$. This leads to some sign changes in what follows. Both options are fine. I'll stick to the positive sign for no particular reason but to avoid a couple of minus signs in the formulation.

of mass, if this is fluid flow; or conservation of energy if this is heat transfer) is just written in the equivalent form

$$\int_\Omega (\nabla\cdot\mathbf{q})v = -\int_\Omega f\,v$$

for an arbitrary test function v. Pending the definition of the correct function spaces, here's the full variational formulation:

$$\left[\begin{array}{llll} \mathbf{q}\cdot\boldsymbol{\nu} = g_1 \text{ on } \Gamma_N, & & & \\ \displaystyle\int_\Omega (a\,\mathbf{q})\cdot\mathbf{p} + \int_\Omega u\,(\nabla\cdot\mathbf{p}) & = \displaystyle\int_{\Gamma_D} g_0\,(\mathbf{p}\cdot\boldsymbol{\nu}) & \forall\mathbf{p}, & \mathbf{p}\cdot\boldsymbol{\nu} = 0 \text{ on } \Gamma_N, \\ \displaystyle\int_\Omega (\nabla\cdot\mathbf{q})v & = -\displaystyle\int_\Omega f\,v & & \forall v. \end{array}\right.$$

This is probably a good moment to emphasize what has happened to the boundary conditions:

- the Dirichlet boundary condition has moved to the right-hand side of the first equation (it has become natural),
- the Neumann boundary condition is kept apart from the integral expressions (it has become essential).

This goes in line to a comment at the beginning of this course, saying that natural/essential for Neumann/Dirichlet was because of the role of these conditions in the weak formulation and nothing else.

A space for vector fields. Consider the set $\mathcal{C}^1(\overline{\Omega})^d$ of $\mathcal{C}^1$ vector fields up to the boundary of Ω and the norm

$$\|\mathbf{q}\|^2_{\mathrm{div},\Omega} := \|\mathbf{q}\|^2_\Omega + \|\nabla\cdot\mathbf{q}\|^2_\Omega = \int_\Omega \left(|\mathbf{q}|^2 + |\nabla\cdot\mathbf{q}|^2\right).$$

We can then close the space $\mathcal{C}^1(\overline{\Omega})^d$ with respect to this norm to make it a Hilbert space. The inner product is easy to figure out from the definition of the norm

$$(\mathbf{q},\mathbf{p})_{\mathrm{div},\Omega} = \int_\Omega \mathbf{q}\cdot\mathbf{p} + \int_\Omega (\nabla\cdot\mathbf{q})(\nabla\cdot\mathbf{p}).$$

The Hilbert space we obtain is

$$\mathbf{H}(\mathrm{div},\Omega) = \{\mathbf{q}:\Omega\to\mathbb{R}^d \,:\, \mathbf{q}\in L^2(\Omega)^d,\ \nabla\cdot\mathbf{q}\in L^2(\Omega)\},$$

where, as it was the case with the gradient in $H^1(\Omega)$, the divergence operator has to be understood in a generalized sense. For a vector field $\mathbf{q}\in\mathbf{H}(\mathrm{div},\Omega)$ it makes sense to define the normal trace $\mathbf{q}\cdot\boldsymbol{\nu}$. This normal component is defined in a very weak way: it is so weak that it cannot be identified to a function defined on Γ. We will not worry about this.

More and more notation. It is now the moment to define spaces:

$$\begin{aligned} \mathbf{V} &:= \mathbf{H}(\mathrm{div},\Omega), \\ \mathbf{V}_0 &:= \{\mathbf{q}\in\mathbf{H}(\mathrm{div},\Omega)\,:\,\mathbf{q}\cdot\boldsymbol{\nu}=0 \text{ on } \Gamma_N\}, \\ M &:= L^2(\Omega). \end{aligned}$$

(The letter M comes from *multiplier*. This will be clear when we talk about saddle point problems.) We have two bilinear forms

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

given by

$$\begin{aligned} a(\mathbf{q},\mathbf{p}) &:= \int_\Omega (a\,\mathbf{q})\cdot\mathbf{p} \\ b(\mathbf{q},v) &:= \int_\Omega (\nabla\cdot\mathbf{q})\,v \end{aligned}$$

and the linear maps

$$\ell(\mathbf{p}) := \int_{\Gamma_N} g_0(\mathbf{p}\cdot\boldsymbol{\nu}), \qquad \chi(v) = -\int_\Omega f\,v.$$

The weak formulation above can now be written in very precise terms:

$$\left[\begin{array}{lll} (\mathbf{q},u)\in\mathbf{V}\times M, & & \\ \mathbf{q}\cdot\boldsymbol{\nu}=g_1 \text{ on } \Gamma_N, & & \\ a(\mathbf{q},\mathbf{p})+b(\mathbf{p},u) & =\ell(\mathbf{p}) & \forall\mathbf{p}\in\mathbf{V}_0, \\ b(\mathbf{q},v) & =\chi(v) & \forall v\in M. \end{array}\right.$$

Note that

$$b(\mathbf{q},v)=\int_\Omega(\nabla\cdot\mathbf{q})\,v=0 \quad \forall v\in M, \qquad \Longleftrightarrow \qquad \nabla\cdot\mathbf{q}=0.$$

1.2 A taste of theory

What comes next. We are not going to enter into great details of the theory of mixed methods, mainly because is slightly harder than the nice Lax-Milgram-based theory of coercive bilinear forms. The functional analysis background is, however, not that complicated. It requires some familiarity with different versions of the closed graph theorem. At the continuous level, the main difficulty arises from having verify one of the hypothesis. The real McCoy in this theory is numerics. In the coercive world the *stability or well-posedness* of the discrete problem is inherited from the continuous problem. This is not the case with mixed methods, where some strong compatibility conditions between the spaces used to discretize $\mathbf{V}$ and M will need to be imposed.

The Babŭska-Brezzi theory. Let us go step by step in the requirements of the theory. (We will not giving it at its most general.) For simplicity, let us assume that $g_1 = 0$ and let us rename the space

$$V := \{\mathbf{q} \in \mathbf{H}(\mathrm{div}, \Omega) \, : \, \mathbf{q} \cdot \boldsymbol{\nu} = 0 \quad \text{on } \Gamma_N\},$$

so that our problem is

$$\left[\begin{array}{lll} (\mathbf{q}, u) \in V \times M, & & \\ a(\mathbf{q}, \mathbf{p}) + b(\mathbf{p}, u) & = \ell(\mathbf{p}) & \forall \mathbf{p} \in V, \\ b(\mathbf{q}, v) & = \chi(v) & \forall v \in M. \end{array}\right.$$

The first group of hypotheses are to be expected:

1. The spaces V and M are Hilbert spaces.
2. The bilinear form a is bounded
$$|a(\mathbf{q}, \mathbf{p})| \le C_a \|\mathbf{q}\|_V \|\mathbf{p}\|_V \qquad \forall \mathbf{q}, \mathbf{p} \in V.$$
3. The bilinear form b is bounded
$$|b(\mathbf{q}, v)| \le C_b \|\mathbf{q}\|_V \|v\|_M \qquad \forall \mathbf{q} \in V, \quad v \in M.$$
4. The linear form ℓ is bounded
$$|\ell(\mathbf{p})| \le C_\ell \|\mathbf{p}\|_V \qquad \forall \mathbf{p} \in \mathbf{V}.$$
5. The linear form χ is bounded
$$|\chi(v)| \le C_\chi \|v\|_M \quad \forall v \in M.$$

In our original case we had the bigger space $\mathbf{H}(\mathrm{div}, \Omega)$ and boundedness should be required there in every occurrence of $\mathbf{V}$ (hypotheses 1,2 3, and 4.) If we go back to our example we can see easily that the hypotheses 2, 3, and 5 are really easy to verify. Hypothesis 4 is a consequence of the following inequality:

$$\left| \int_{\Gamma_D} (\mathbf{p} \cdot \boldsymbol{\nu}) g_0 \right| \le C_{g_0} \|\mathbf{p}\|_{\mathrm{div},\Omega} \qquad \forall \mathbf{p} \in \mathbf{H}(\mathrm{div}, \Omega).$$

However, this is not a trivial inequality to prove unless we gather some familiarity with the space $H^{1/2}(\Gamma_D)$ mentioned in the last lesson, and its dual space. Before we add another hypothesis, we need to define a new space

$$Z = \{\mathbf{q} \in \mathbf{V} \, : \, b(\mathbf{q}, v) = 0 \quad \forall v \in M\}.$$

Here's the next hypothesis:

6. The bilinear form a is coercive in Z:

$$a(\mathbf{q},\mathbf{q}) \ge \alpha \|\mathbf{q}\|_V^2 \qquad \forall \mathbf{q} \in Z.$$

Going back to our particular case, we have already recognized

$$\mathbf{Z} = \{\mathbf{q} \in \mathbf{V}_0 \,:\, \nabla \cdot \mathbf{q}\}$$

and therefore

$$\begin{aligned} a(\mathbf{q},\mathbf{q}) &= \int_\Omega (a\mathbf{q}) \cdot \mathbf{q} \\ &\ge a_0 \int_\Omega |\mathbf{q}|^2 && \text{(positivity of } a\text{)} \\ &= a_0 \left(\int_\Omega |\mathbf{q}|^2 + \int_\Omega |\nabla \cdot \mathbf{q}|^2 \right). && (\mathbf{q} \in Z) \end{aligned}$$

What's missing is the difficult hypothesis, dealing with the compatibility of the spaces V and M through the bilinear form b. There are many ways of expressing this hypothesis. Let me write one here, and then we will move to a discussion and to some (surprisingly?) equivalent formulations:

7. There exists $\beta > 0$ such that

$$\sup_{0 \ne q \in V} \frac{b(\mathbf{q}, v)}{\|\mathbf{q}\|_V} \ge \beta \|v\|_M \qquad \forall v \in M.$$

The inf-sup condition. Let's keep it abstract for a moment. There are equivalent formulations for Hypothesis 7 above:

- There exists $\beta > 0$ such that

$$\inf_{0 \ne v \in M} \sup_{0 \ne q \in V} \frac{b(\mathbf{q}, v)}{\|v\|_M \|\mathbf{q}\|_V} \ge \beta.$$

  This is one traditional way of writing the hypothesis, which is often called the **infimum-supremum condition**, or, in short, the inf-sup condition. (More about names in a while.) A more natural way to write this condition is, obviously,

$$\inf_{0 \ne v \in M} \sup_{0 \ne q \in V} \frac{b(\mathbf{q}, v)}{\|v\|_M \|\mathbf{q}\|_V} > 0.$$

  (You just then define β to be the inf-sup in the left-hand side of the inequality.)

- For every $\chi : V \to \mathbb{R}$ linear and bounded, the problem

$$b(\mathbf{q}, v) = \chi(v) \qquad \forall v \in M$$

  admits at least one solution. (The solution will not be unique, because there's the set Z with all the functions $\mathbf{q}$ making the left-hand side vanish.)

- For every linear $\ell_0 : \mathbf{V} \to \mathbb{R}$ such that

  $$|\ell_0(\mathbf{p})| \le C_{\ell_0} \|\mathbf{p}\|_V, \qquad \ell_0(\mathbf{p}) = 0 \quad \forall \mathbf{p} \in Z,$$

  there exists a unique solution to the problem

  $$u \in M, \qquad b(\mathbf{p}, u) = \ell_0(\mathbf{p}) \quad \forall \mathbf{p} \in V,$$

  and we can bound

  $$\|u\|_M \le C \times C_{\ell_0}.$$

There's a long history of renaming and attributing this hypothesis (or the entire set of seven hypotheses above) to different authors. It is common to refer to it as the Babuška-Brezzi condition[2], or even the Ladyzhenskaya-Babuška-Brezzi condition[3], shortened to LBB. The hypothesis is actually implied by (or can be rephrased as a version of) the Banach Closed Graph Theorem. In our particular case, with $V = \mathbf{V}_0$ and $M = L^2(\Omega)$, this hypothesis is equivalent to the following property: for every $f \in L^2(\Omega)$ there exists at least one $\mathbf{q} \in \mathbf{V}_0$ such that

$$\nabla \cdot \mathbf{q} = -f.$$

(The minus sign is just because..., or not). How do we check this? Take $f \in L^2(\Omega)$ and solve the PDE

$$\left[\begin{array}{ll} -\Delta v = f & \text{in } \Omega, \\ v = 0 & \text{on } \Gamma_D, \\ \partial_\nu v = 0 & \text{on } \Gamma_N. \end{array}\right.$$

The answer to our problem is the vector field $\mathbf{q} = \nabla v$ (recall that $\Delta = \nabla \cdot \nabla$), although there are some minor details about Sobolev spaces to be taken care of.

Saddle point problems. Let us go back to the general problem

$$\left[\begin{array}{lll} (\mathbf{q}, u) \in V \times M, & & \\ a(\mathbf{q}, \mathbf{p}) + b(\mathbf{p}, u) & = \ell(\mathbf{p}) & \forall \mathbf{p} \in V, \\ b(\mathbf{q}, v) & = \chi(v) & \forall v \in M. \end{array}\right.$$

We now assume that

$$a(\mathbf{q}, \mathbf{p}) = a(\mathbf{p}, \mathbf{q})$$

(symmetry) and

$$a(\mathbf{q}, \mathbf{q}) \ge 0 \quad \forall \mathbf{q} \in V.$$

Then, the variational problem is equivalent to a minimization problem with restrictions:

$$\text{minimize } \tfrac{1}{2} a(\mathbf{q}, \mathbf{q}) - \ell(\mathbf{q}) \qquad \text{subject to } b(\mathbf{q}, v) = \chi(v) \quad \forall v \in M.$$

[2] Named after Ivo Babuška, a powerhouse in the theory and praxis of the Finite Element method, and Franco Brezzi, father of much of what is known in mixed Finite Element.

[3] Adding here the name of Olga Ladyzhenskaya, Russian PDE theorist extraordinaire.

Note that in this minimization problem the unknown $u \in M$ has disappeared. We can bring it back in the Lagrangian

$$\mathcal{L}(\mathbf{q}, u) := \tfrac{1}{2}a(\mathbf{q}, \mathbf{q}) - \ell(\mathbf{q}) + (b(\mathbf{q}, u) - \chi(u)) .$$

The constrained minimization problem above is equivalent to the saddle point problem

$$\mathcal{L}(\mathbf{q}, v) \le \mathcal{L}(\mathbf{q}, u) \le \mathcal{L}(\mathbf{p}, u) \qquad \forall (\mathbf{p}, v) \in V \times M.$$

This can be read as follows: if we move from the *equilibrium point* $(\mathbf{q}, u)$ in the direction of V we see the Lagrangian increase, while if we move in the direction of M, we see the Lagrangian decrease. This is why the problem in the Lagrangian is a **saddle point problem**, and why many people refer to the original variational formulation in $V \times M$ as a saddle point problem. The variable u plays the role of a Lagrange multiplier for the constrained minimization problem.

1.3 Galerkin approximation

Discretizations of $\mathbf{H}(\mathrm{div}, \Omega)$. We will need some time to define a nice approximation for the mixed Laplacian, so we will content ourselves with some general arguments. However, let us just first say what we should expect from a finite element style discretization of $\mathbf{H}(\mathrm{div}, \Omega)$. Let $\mathcal{T}_h$ be a triangulation/tetrahedrization of the polygonal/polyhedral domain Ω. For simplicity (to avoid separating two and three dimensions) we will use the following conventions:

- When in three dimensions ($d = 3$) we will call the faces of a tetrahedron *faces.* When in two dimensions, we'll call edges of a triangle *faces* as well.

- We will write $\mathcal{F}_h$ to denote the set of all faces of the triangulation. Each face will have a unit normal vector $\boldsymbol{\nu}_F$ assigned to it. When $F \in \mathcal{F}_h$ is in Γ, we will assume that $\boldsymbol{\nu}_F$ points outwards.

- We will write $\mathcal{F}(K)$ to denote the set of $d+1$ faces of $K \in \mathcal{T}_h$. When seen from the point of view of the element, the normals to the faces will point outwards. The local exterior orientation doesn't have to coincide with the intrinsic orientation assigned to the face.

Recall that the condition for a piecewise smooth function to be in $H^1(\Omega)$ was continuity. The requirements for vector fields in $\mathbf{H}(\mathrm{div}, \Omega)$ are less demanding: given $\mathbf{q}_h : \Omega \to \mathbb{R}^d$ such that $\mathbf{q}|_K$ is smooth for every $K \in \mathcal{T}_h$

$$\mathbf{q}_h \in \mathbf{H}(\mathrm{div}, \Omega) \quad \Longleftrightarrow \quad \mathbf{q}_h \cdot \boldsymbol{\nu}_F \text{ is continuous across } F \text{ for all } F \in \mathcal{F}_h.$$

In other words, if two elements K and K' meet in a face F with preassigned normal $\boldsymbol{\nu}_F$, we need

$$\mathbf{q}_h|_K \cdot \boldsymbol{\nu}_F = \mathbf{q}_h|_{K'} \cdot \boldsymbol{\nu}_F \quad \text{on } F.$$

A general discretization. Consider finite dimensional subspaces

$$V_h \subset V, \qquad M_h \subset M.$$

(Be warned, as of now, that many choices of pairs might not work from the point of view of even delivering a uniquely solvable system.) We then define the Galerkin discretization as the problem:

$$\left[\begin{array}{lll} (\mathbf{q}_h, u_h) \in V_h \times M_h, & & \\ a(\mathbf{q}_h, \mathbf{p}_h) + b(\mathbf{p}_h, u_h) & = \ell(\mathbf{p}_h) & \forall \mathbf{p}_h \in V_h, \\ b(\mathbf{q}_h, v_h) & = \chi(v_h) & \forall v_h \in M_h. \end{array}\right.$$

Take now two bases:

$$\{\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_N\} \quad \text{for } V_h, \qquad \text{and} \qquad \{\mu_1, \ldots, \mu_S\} \quad \text{for } M_h.$$

The first discrete equation is equivalent to

$$a(\mathbf{q}_h, \boldsymbol{\varphi}_i) + b(\boldsymbol{\varphi}_i, u_h) = \ell(\boldsymbol{\varphi}_i) \qquad i = 1, \ldots, N,$$

while the second one is equivalent to

$$b(\mathbf{q}_h, \mu_i) \qquad i = 1, \ldots, S.$$

We then decompose the unknowns in the given bases:

$$\mathbf{q}_h = \sum_{j=1}^{N} q_j \boldsymbol{\varphi}_j, \qquad u_h = \sum_{j=1}^{S} u_j \mu_j$$

and substitute in the above equations

$$\begin{aligned} \sum_{j=1}^{N} a(\boldsymbol{\varphi}_j, \boldsymbol{\varphi}_i) q_j + \sum_{j=1}^{S} b(\boldsymbol{\varphi}_i, \mu_j) u_j &= \ell(\boldsymbol{\varphi}_i) \qquad i = 1, \ldots, N, \\ \sum_{j=1}^{S} b(\boldsymbol{\varphi}_j, \mu_i) q_j &= \chi(\mu_i) \qquad i = 1, \ldots, S. \end{aligned}$$

We can organize everything with two matrices $\mathbf{A} \in \mathbb{R}^{N \times N}$ and $\mathbf{B} \in \mathbb{R}^{S \times N}$ with elements

$$\begin{aligned} a_{ij} &= a(\boldsymbol{\varphi}_j, \boldsymbol{\varphi}_i) \qquad i, j = 1, \ldots, N, \\ b_{ij} &= b(\boldsymbol{\varphi}_j, \mu_i) \qquad i = 1, \ldots, S, \qquad j = 1, \ldots, N, \end{aligned}$$

are two vectors for the right-hand sides $\boldsymbol{\ell} \in \mathbb{R}^N$ and $\boldsymbol{\chi} \in \mathbb{R}^S$ with elements

$$\ell_i = \ell(\boldsymbol{\varphi}_i), \qquad \chi_i = \chi(\mu_i).$$

The discrete problem is then equivalent to the system

$$\left[\begin{array}{cc} \mathbf{A} & \mathbf{B}^\top \\ \mathbf{B} & \mathbf{O} \end{array}\right] \left[\begin{array}{c} \mathbf{q} \\ \mathbf{u} \end{array}\right] = \left[\begin{array}{c} \boldsymbol{\ell} \\ \boldsymbol{\chi} \end{array}\right].$$

Note that I have used the letter $\mathbf{q}$ to denote the vector with the coefficients of the decomposition of $\mathbf{q}_h$ (which approximated $\mathbf{q}$) in the given bases. I hope this is not too confusing. Two pointers:

- If the bilinear form a is symmetric (case of saddle point problems), then the matrix

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^\top \\ \mathbf{B} & \mathbf{O} \end{bmatrix}$$

  is symmetric. It is highly indefinite though: note the $S \times S$ block of zeros in the diagonal.

- Note also that if $S > N$ (the dimension of M_h is larger than the dimension of V_h), then the system cannot be invertible, because we have a much too large block of zeros in the diagonal. You should try to think of the matrix $\mathbf{B}$ as rectangular with less rows than columns.

The hypothesis for discretization. Differently from what happened in coercive (Lax-Milgram based) problems, Galerkin methods for mixed problems do not have guaranteed quasioptimality. There's a simple reason why. You just need to take M_h to be of higher dimension than V_h and the matrix that you get cannot be invertible. We need to identify the following set:

$$Z_h := \{\mathbf{q}_h \in V_h \,:\, b(\mathbf{q}_h, v_h) = 0 \quad \forall v_h \in V_h\}.$$

The discrete hypotheses we will assume are:

- $Z_h \subset Z$, that is, if $\mathbf{q}_h \in V_h$ and

$$b(\mathbf{q}_h, v_h) = 0 \quad \forall v_h \in M_h \qquad \Longrightarrow \qquad b(\mathbf{q}_h, v) = 0 \quad \forall v \in M.$$

- A **uniform discrete inf-sup condition** is stated by assuming the existence of $\widetilde{\beta} > 0$, independent[4] of h, such that

$$\sup_{0 \neq \mathbf{q}_h \in V_h} \frac{b(\mathbf{q}_h, v_h)}{\|\mathbf{q}_h\|_V} \geq \widetilde{\beta} \|v_h\|_M \qquad \forall v_h \in M_h.$$

If these hypotheses (and the hypotheses for the continuous problem) are met, then the discrete problem has a unique solution and we can bound:

$$\|\mathbf{q} - \mathbf{q}_h\|_V + \|u - u_h\|_M \leq C \left(\inf_{\mathbf{p}_h \in V_h} \|\mathbf{q} - \mathbf{p}_h\|_V + \inf_{v_h \in M_h} \|u - v_h\|_M \right).$$

This is quasioptimality of the solution with respect to the best approximation in both discrete spaces. There is another way, probably more rigorous, of presenting the uniform discrete inf-sup conditions. We first define

$$\beta_h := \inf_{0 \neq v_h \in M_h} \sup_{0 \neq \mathbf{q}_h \in V_h} \frac{b(\mathbf{q}_h, v_h)}{\|v_h\|_M \|\mathbf{q}_h\|_V}$$

[4] This hypothesis might sound unclear at this stage. It moves you to the context of not having just two spaces V_h, M_h but a whole sequence parametrized in h.

and assume that $\beta_h > 0$. (If $\beta_h = 0$ it is possible to show that the system is not invertible and we would be done.) Then, assuming all other conditions,

$$\|\mathbf{q} - \mathbf{q}_h\|_V + \|u - u_h\|_M \le C_h \left(\inf_{\mathbf{p}_h \in V_h} \|\mathbf{q} - \mathbf{p}_h\|_V + \inf_{v_h \in M_h} \|u - v_h\|_M \right).$$

where C_h is an increasing function of the quantity $1/\beta_h$. Being very careful, it can be shown that $C_h \le C/\beta_h^3$, where C depends on other quantities for the continuous problem (the boundedness constants C_a and C_b and the coercivity constant α).

Back to our example. The discrete uniform inf-sup condition is never easy to verify. However, the condition $Z_h \subset Z$ is sometimes quite obvious. Imagine that we have subspaces

$$\mathbf{V}_h \subset \mathbf{H}(\mathrm{div}, \Omega) \qquad M_h \subset L^2(\Omega)$$

and that

$$\mathrm{div}\, \mathbf{p}_h \in M_h \qquad \forall \mathbf{p}_h \in \mathbf{V}_h.$$

Recall our bilinear form

$$b(\mathbf{q}, v) = \int_\Omega (\nabla \cdot \mathbf{q})\, v.$$

If $\mathbf{p}_h \in \mathbf{V}_h$ satisfies $b(\mathbf{p}_h, v_h) = 0$ for all $v_h \in M_h$, then

$$0 = b(\mathbf{p}_h, \mathrm{div}\, \mathbf{p}_h) = \int_\Omega |\nabla \cdot \mathbf{p}_h|^2$$

and therefore $\nabla \cdot \mathbf{p}_h = 0$, which implies $b(\mathbf{p}_h, v) = 0$ for all $v \in M$.

2 The Raviart-Thomas space

In this section we discuss the lowest order Raviart-Thomas space[5]. We will start with a local description of the space, move to a global description, and finally present how to change to the reference element, which is way less obvious than one might naively imagine.

2.1 The local space

The space. On an element (triangle or tetrahedron) K we define the space

$$\mathrm{RT}_0(K) := \{\mathbf{p}(\mathbf{x}) = \mathbf{a} + b\mathbf{x} \,:\, \mathbf{a} \in \mathbb{R}^d, \quad b \in \mathbb{R}\}.$$

Just to be clear, an element of $\mathrm{RT}_0(K)$ looks like this when $d = 2$

$$\left[\begin{array}{c} a_1 \\ a_2 \end{array}\right] + b \left[\begin{array}{c} x_1 \\ x_2 \end{array}\right]$$

[5] Named after Pierre-Arnaud Raviart and Jean-Marie Thomas, both of them French as is easy to guess from their names. On a personal note, Raviart was my academic grandfather, although I've never met him.

and like this

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} + b \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

when $d = 3$.

Its properties. All the following properties are easy to verify.

1. It is clear that the dimension of $\mathrm{RT}_0(K)$ is $d+1$ and

   $$\mathbb{P}_0(K)^d \subset \mathrm{RT}_0(K) \subset \mathbb{P}_1(K)^d,$$

   which places $\mathrm{RT}_0(K)$ between the d-dimensional space of constant vector fields and the $d(d+1)$-dimensional space of linear polynomials with vector-valued coefficients.

2. Noticing that $\nabla\cdot\mathbf{x} = d$, it is obvious that if $\mathbf{p}(\mathbf{x}) = \mathbf{a}+b\mathbf{x}$, then $\nabla\cdot\mathbf{p} = d\,b \in \mathbb{P}_0(K)$. Therefore, if $\mathbf{p} \in \mathrm{RT}_0(K)$ and $\nabla\cdot\mathbf{p} = 0$, then $\mathbf{p} \in \mathbb{P}_0(K)^d$.

3. If $\mathbf{p} \in \mathrm{RT}_0(K)$ and $\boldsymbol{\nu}_F$ is normal to $F \in \mathcal{F}(K)$, then

   $$\mathbf{p}\cdot\boldsymbol{\nu}_F|_F \in \mathcal{P}_0(F),$$

   that is, the normal component of $\mathbf{p}$ on ∂K is constant on each face. This is actually quite easy to verify. The plane containing F can be written with the equation $\boldsymbol{\nu}_F\cdot\mathbf{x} = c$. Then for $\mathbf{x} \in F$,

   $$\mathbf{p}(\mathbf{x})\cdot\boldsymbol{\nu}_F = \mathbf{a}\cdot\boldsymbol{\nu}_F + b(\mathbf{x}\cdot\boldsymbol{\nu}_F) = \mathbf{a}\cdot\boldsymbol{\nu}_F + b\,c.$$

   The integrated normal fluxes on the faces

   $$\int_F \mathbf{p}\cdot\boldsymbol{\nu}_F = |F|\mathbf{p}\cdot\boldsymbol{\nu}_F|_F \qquad F \in \mathcal{F}(K),$$

   determine uniquely an element $\mathbf{p} \in \mathrm{RT}_0(K)$. In other words, given numbers c_F for $F \in \mathcal{F}(K)$, there exists a unique $\mathbf{p} \in \mathrm{RT}_0(K)$ such that

   $$\int_F \mathbf{p}\cdot\boldsymbol{\nu}_F = c_F \quad \forall F \in \mathcal{F}(K).$$

   Let us prove uniqueness. Assume that

   $$\int_F \mathbf{p}\cdot\boldsymbol{\nu}_F = 0 \quad \forall F \in \mathcal{F}(K).$$

   Then

   $$0 = \sum_{F\in\mathcal{F}(K)} \mathbf{p}\cdot\boldsymbol{\nu}_F = \int_{\partial K} \mathbf{p}\cdot\boldsymbol{\nu} = \int_K \nabla\cdot\mathbf{p} = |K|\nabla\cdot\mathbf{p},$$

   by the divergence theorem and the fact that $\nabla\cdot\mathbf{p}$ is constant. This implies that $\mathbf{p}(\mathbf{x}) = \mathbf{a}$. Going back to the beginning of the argument, we have

   $$\mathbf{a}\cdot\boldsymbol{\nu}_F \quad \forall F \in \mathcal{F}(K),$$

but this implies that $\mathbf{a} = \mathbf{0}$. The argument for existence is quite standard in linear algebra. Consider the map

$$\begin{array}{rcl} \mathrm{RT}_0(K) & \longrightarrow & \mathbb{R}^{d+1} \\ \mathbf{p} & \longmapsto & \left(\int_F \mathbf{p} \cdot \boldsymbol{\nu}_F\right)_{F \in \mathcal{F}(K)} . \end{array}$$

This is a linear transformation between two spaces of the same dimension, and we have shown that it does not have a null-space. Therefore, it is invertible.

The local Raviart-Thomas interpolation operator. Let $\mathbf{q} \in \mathcal{C}(K)^d$ and consider

$$\mathbf{q}_K \in \mathrm{RT}_0(K) \qquad \text{such that} \qquad \int_F \mathbf{q}_K \cdot \boldsymbol{\nu}_F = \int_F \mathbf{q} \cdot \boldsymbol{\nu}_F \quad \forall F \in \mathcal{F}(K).$$

We will refer to $\mathbf{q}_K$ as the local Raviart-Thomas interpolant. Note that if $\mathbf{q} \in \mathcal{C}^1(K)^d$, then

$$\begin{aligned} |K| \nabla \cdot \mathbf{q}_K &= \int_K \nabla \cdot \mathbf{q}_K = \int_{\partial K} \mathbf{q}_K \cdot \boldsymbol{\nu} = \sum_{F \in \mathcal{F}(K)} \int_F \mathbf{q}_K \cdot \boldsymbol{\nu}_F \\ &= \sum_{F \in \mathcal{F}(K)} \int_F \mathbf{q} \cdot \boldsymbol{\nu}_F = \int_{\partial K} \mathbf{q} \cdot \boldsymbol{\nu} = \int_K \nabla \cdot \mathbf{q}, \end{aligned}$$

and therefore

$$\mathbb{P}_0(K) \ni \nabla \cdot \mathbf{q}_K = \frac{1}{|K|} \int_K \nabla \cdot \mathbf{q},$$

which can be written with words: the divergence of the RT interpolant is the best constant approximation (in the sense of $L^2(K)$) of the divergence of the field. A basis for $\mathrm{RT}_0(K)$ can be constructed by numbering the faces of K and then looking for

$$\mathbf{N}_\alpha^K \in \mathrm{RT}_0(K) \qquad \text{such that} \qquad \int_{F_\beta} \mathbf{N}_\alpha^K \cdot \boldsymbol{\nu}_{F_\beta} = \delta_{\alpha\beta}, \qquad \alpha, \beta = 1, \ldots, d+1.$$

2.2 The global space

Construction of the space. The full RT space is obtained by gluing local RT spaces making the normal fluxes coincide. The definition is the following:

$$\mathrm{RT}_h^0 = \{\mathbf{q}_h : \Omega \to \mathbb{R}^d \ : \ \mathbf{q}_h \in \mathbf{H}(\mathrm{div}, \Omega),\ \mathbf{q}_h|_K \in \mathrm{RT}_0(K) \quad \forall K \in \mathcal{T}_h\}.$$

An element of RT_h^0 is determined by the values of the integrals

$$\int_F \mathbf{q}_h \cdot \boldsymbol{\nu}_F, \qquad F \in \mathcal{F}_h.$$

Why? Choose these values to be whatever you want them to be. Then go to the elements. In each element you can reconstruct $\mathbf{q}_h$ based on its boundary fluxes (you might need to change the sign of the flux if the orientation of the normal vectors does not coincide). Since we have fixed the interelement fluxes, we know that the reconstructed element is an element of $\mathbf{H}(\mathrm{div}, \Omega)$.

Global basis functions. For each $F \in \mathcal{F}_h$, we can now define the function $\boldsymbol{\varphi}_F \in \mathrm{RT}_h^0$ given by the conditions

$$\int_{F'} \boldsymbol{\varphi}_F \cdot \boldsymbol{\nu}_{F'} = \delta_{F,F'}, \qquad \forall F' \in \mathcal{F}_h.$$

The support of $\boldsymbol{\varphi}_F$ is the set of two elements surrounding F (one element if $F \subset \Gamma$). It is important to note that $\boldsymbol{\varphi}_F$ is not zero on the other faces of the elements that share F as a face. It is only the normal component that vanishes on these faces. If we number the faces $\mathcal{F}_h = \{F_1, \ldots, F_N\}$, we end up with an indexed basis of RT_h^0. An element of this space is then represented as

$$\mathbf{q}_h = \sum_{F \in \mathcal{F}_h} q_F \boldsymbol{\varphi}_F = \sum_{F \in \mathcal{F}_h} \left(\int_F \mathbf{q}_h \cdot \boldsymbol{\nu}_F \right) \boldsymbol{\varphi}_F.$$

You can compare this with the Lagrange (nodal) basis of the space of $\mathbb{P}_1$ Finite Elements. In that case, coefficients of representation of a function in the basis were nodal values, while here we control fluxes on the faces. Note that the dimension of the RT space is the number of faces:

$$\dim \mathrm{RT}_h^0 = \#\mathcal{F}_h.$$

We will come to the relation between local and global bases when we study implementation issues for the RT space.

Raviart-Thomas interpolation. The RT interpolant of $\mathbf{q} \in \mathcal{C}(\overline{\Omega})^d$ is the only $\mathbf{q}_h \in \mathrm{RT}_h^0$ such that

$$\int_F \mathbf{q}_h \cdot \boldsymbol{\nu}_F = \int_F \mathbf{q} \cdot \boldsymbol{\nu}_F \qquad \forall F \in \mathcal{F}_h.$$

The interpolant is well defined because the normal fluxes determine uniquely an element of RT_h^0. We can even give an explicit expression for the interpolant

$$\mathbf{q}_h = \sum_{F \in \mathcal{F}_h} \left(\int_F \mathbf{q} \cdot \boldsymbol{\nu}_F \right) \boldsymbol{\varphi}_F.$$

2.3 Piola transformations

A preliminary computation. We still need to relate the RT basis on the reference element with the basis on the physical element. This is going to be less trivial than just using the transformation

$$\mathrm{F}_K : \widehat{K} \to K, \qquad \mathrm{F}_K(\widehat{\mathbf{x}}) = \mathrm{B}_K \widehat{\mathbf{x}} + \mathbf{b}.$$

Vector fields are mapped to vector fields in a slightly different way in order to compensate that normals are not mapped to normals (affine maps do not preserve angles). One easy way to remember how to get the transformation of vector fields that we will be using is the following: we bring in a scalar field $u : K \to \mathbb{R}$ and a vector field $\mathbf{q} : K \to \mathbb{R}^d$ and we want to define $\widehat{\mathbf{q}} : \widehat{K} \to \mathbb{R}^d$ so that

$$\int_K \mathbf{q} \cdot \nabla u = \int_{\widehat{K}} \widehat{\mathbf{q}} \cdot \nabla \widehat{u}, \qquad \text{where, as usual, } \widehat{u} = -u \circ \mathrm{F}_K.$$

Recalling that $(\nabla u) \circ \mathrm{F}_K = \mathrm{B}_K^\top \nabla(u \circ \mathrm{F}_K)$, we can compute

$$\begin{aligned}\int_K \mathbf{q} \cdot \nabla u &= \int_{\widehat{K}} (\mathbf{q} \circ \mathrm{F}_K) \cdot ((\nabla u) \circ \mathrm{F}_K) |\det \mathrm{B}_K| \\ &= \int_{\widehat{K}} \underbrace{\big(|\det \mathrm{B}_K| \mathrm{B}_K^{-1} \mathbf{q} \circ \mathrm{F}_K\big)}_{\widehat{\mathbf{q}}} \cdot \nabla \widehat{u}.\end{aligned}$$

We thus reach the **Piola transformation**[6] for the vector field $\mathbf{q}$,

$$\widehat{\mathbf{q}} = |\det \mathrm{B}_K| \mathrm{B}_K^{-1} \mathbf{q} \circ \mathrm{F}_K : \widehat{K} \to \mathbb{R}^d,$$

which can be inverted to yield

$$\mathbf{q} = \frac{1}{|\det \mathrm{B}_K|} \mathrm{B}_K \widehat{\mathbf{q}} \circ \mathrm{F}_K^{-1} : K \to \mathbb{R}^d,$$

when we want to push forward a vector field from the reference element to the physical element.

More work on the Piola transformations. With some careful use of the chain rule (the vector notation is quite cumbersome here, and it's better to use index notation like physicists and engineers) it can be proved that

$$\nabla \cdot \widehat{\mathbf{q}} = |\det \mathrm{B}_K| (\nabla \cdot \mathbf{q}) \circ \mathrm{F}_K$$

and therefore

$$\int_K (\nabla \cdot \mathbf{q})\, u = \int_{\widehat{K}} (\nabla \cdot \widehat{\mathbf{q}})\, \widehat{u}.$$

Finally, the transformations of products of divergences by scalar fields and gradients by vector fields prove that

$$\begin{aligned}\int_{\partial K} (\mathbf{q} \cdot \boldsymbol{\nu})\, u &= \int_K (\nabla \cdot \mathbf{q})\, u + \int_K \mathbf{q} \cdot \nabla u && \text{(divergence theorem)} \\ &= \int_{\widehat{K}} \widehat{\mathbf{q}} \cdot \nabla \widehat{u} + \int_{\widehat{K}} (\nabla \cdot \widehat{\mathbf{q}})\, \widehat{u} && \text{(we just saw this)} \\ &= \int_{\partial \widehat{K}} (\widehat{\mathbf{q}} \cdot \widehat{\boldsymbol{\nu}})\, \widehat{u}. && \text{(divergence theorem again)}\end{aligned}$$

For this identity, we can infer (the mathematics needed for this are not complicated, but they need some careful treatment of regularity of functions, or good old plain hand-waving) that

$$\int_F (\mathbf{q} \cdot \boldsymbol{\nu})\, u = \int_{\widehat{F}} (\widehat{\mathbf{q}} \cdot \widehat{\boldsymbol{\nu}})\, \widehat{u},$$

where $\widehat{F} \in \mathcal{F}(\widehat{K})$ and F is the face of K that is transformed from $\widehat{F}$.

[6] Named after Gabrio Piola. If you have taken a class in continuum solid mechanics, you won't be surpirsed to find an Italian name here.

Piola, Raviart, and Thomas. Here's a computation that explains why the RT elements get along so well with the divergence operator. We start with an RT polynomial in the reference element

$$\widehat{\mathbf{p}}(\widehat{\mathbf{x}}) = \widehat{\mathbf{a}} + \widehat{b}\widehat{\mathbf{x}}.$$

Then, writing $\mathbf{x} = \mathrm{F}_K(\widehat{\mathbf{x}}) = \mathrm{B}_K\widehat{\mathbf{x}} + \mathbf{b}_K$ for the transformed variable, it follows that

$$\begin{aligned}
(\mathbf{p}\circ \mathrm{F}_K)(\widehat{\mathbf{x}}) &= \frac{1}{|\det \mathrm{B}_K|}\mathrm{B}_K\widehat{\mathbf{p}}(\widehat{\mathbf{x}}) && \text{(def. of Piola tr.)}\\
&= \frac{1}{|\det \mathrm{B}_K|}(\mathrm{B}_K\widehat{\mathbf{a}} + \widehat{b}\mathrm{B}_K\widehat{\mathbf{x}})\\
&= \underbrace{\left(\frac{1}{|\det \mathrm{B}_K|}(\mathrm{B}_K\widehat{\mathbf{a}} - \widehat{b}\mathbf{b}_K)\right)}_{\mathbf{a}} + \underbrace{\frac{\widehat{b}}{|\det \mathrm{B}_K|}}_{b}\mathrm{F}_K(\widehat{\mathbf{x}}),
\end{aligned}$$

or, equivalently,

$$\mathbf{p}(\mathbf{x}) = \mathbf{a} + b\mathbf{x} \in \mathrm{RT}_0(K).$$

This proves that RT fields in the reference element are mapped to RT fields in the physical element. Now, recall the 'Lagrange' basis associated to the RT degrees of freedom:

$$\mathbf{N}_\alpha^K \in \mathrm{RT}_0(K) \qquad \text{such that} \qquad \int_{F_\beta} \mathbf{N}_\alpha^K \cdot \boldsymbol{\nu}_{F_\beta} = \delta_{\alpha\beta}, \qquad \alpha,\beta = 1,\ldots,d+1.$$

Using the Piola transform and how it interacts with integrals on faces, it is clear that if $\widehat{\mathbf{N}}_\alpha$ is the basis on the reference element

$$\int_{\widehat{F}_\beta} \widehat{\mathbf{N}}_\alpha \cdot \widehat{\boldsymbol{\nu}}_{\widehat{F}_\beta} = \delta_{\alpha\beta},$$

then

$$\mathbf{N}_\alpha^K = \frac{1}{|\det \mathrm{B}_K|}\mathrm{B}_K\widehat{\mathbf{N}}_\alpha \circ \mathrm{F}_K, \qquad \alpha = 1,\ldots,d+1.$$

3 RT discretization of the mixed Laplacian

3.1 General ideas

Back to the model problem. We return to the weak formulation of our 'glorified Laplacian'[7], given in the form

$$\left[\begin{array}{lll}
(\mathbf{q},u) \in \mathbf{V}\times M, & & \\
\mathbf{q}\cdot\boldsymbol{\nu} = g_1 \text{ on } \Gamma_N, & & \\
a(\mathbf{q},\mathbf{p}) + b(\mathbf{p},u) & = \ell(\mathbf{p}) & \forall \mathbf{p}\in \mathbf{V}_0,\\
b(\mathbf{q},v) & = \chi(v) & \forall v \in M,
\end{array}\right.$$

[7] Some years ago, glorified Laplacian became a way of referring to any more or less trivial extension of the Laplacian that made it look much more general or 'applied.'

where

$$\begin{aligned}
\mathbf{V} &= \mathbf{H}(\mathrm{div},\Omega),\\
\mathbf{V}_0 &= \{\mathbf{p}\in\mathbf{V}\,:\,\mathbf{p}\cdot\boldsymbol{\nu}=0 \quad \text{on } \Gamma_N\},\\
M &= L^2(\Omega),
\end{aligned}$$

$$a(\mathbf{q},\mathbf{p}) := \int_\Omega (a\,\mathbf{q})\cdot\mathbf{p}, \qquad b(\mathbf{q},v) := \int_\Omega (\nabla\cdot\mathbf{q})\,v,$$

and

$$\ell(\mathbf{p}) := \int_{\Gamma_N} g_0(\mathbf{p}\cdot\boldsymbol{\nu}), \qquad \chi(v) = -\int_\Omega f\,v.$$

We choose the following discrete spaces:

$$\begin{aligned}
\mathbf{V}_h &= \mathrm{RT}^0_h = \{\mathbf{q}_h\in\mathbf{H}(\mathrm{div},\Omega)\,:\,\mathbf{q}_h|_K\in\mathrm{RT}_0(K)\quad\forall K\in\mathcal{T}_h\},\\
\mathbf{V}_{h,0} &= \{\mathbf{p}_h\in\mathbf{V}_h\,:\,\mathbf{p}_h\cdot\boldsymbol{\nu}=0\quad\text{on }\Gamma_N\},\\
M_h &= \{v_h:\Omega\to\mathbb{R}\,:\,v_h|_K\in\mathbb{P}_0(K)\quad\forall K\in\mathcal{T}_h\}.
\end{aligned}$$

This means that we pair the (lowest order) Raviart-Thomas space for vector fields with the space of piecewise constant functions. Let's go for some fast bullet points:

- The dimension of $\mathbf{V}_h$ is the number of faces. If we number $\mathcal{F}_h=\{F_1,\ldots,F_N\}$ and we assign a unit normal vector $\boldsymbol{\nu}_F$ to each F, then the basis is determined by

  $$\int_{F_j}\boldsymbol{\varphi}_i\cdot\boldsymbol{\nu}_{F_j}=\delta_{ij}.$$

- Let Neu $\subset\{1,\ldots,N\}$ be the list of Neumann faces and Free $=\{1,\ldots,N\}\setminus$ Neu. Then

  $$\{\boldsymbol{\varphi}_i\,:\,i\in\text{Free}\}$$

  is a basis for $\mathbf{V}_{h,0}$. The reason is simple: the normal component on Γ_N of a function in RT^0_h vanishes if and only if the (constant) flux on each Neumann face vanishes. We then just need to get rid of the basis functions associated to Neumann faces.

- Let $\{K_1,\ldots,K_S\}$ be a numbering of $\mathcal{T}_h$. We can then define $\mu_j:\Omega\to\mathbb{R}$ to be the characteristic function of K_j

  $$\mu_j=\left\{\begin{array}{ll}1, & \text{in } K_j,\\ 0 & \text{elsewhere.}\end{array}\right.$$

  Then $\{\mu_i\,:\,i=1,\ldots,S\}$ is a basis for M_h and the dimension of M_h is the number of elements.

- It is clear that

  $$\mathbf{q}_h\in\mathbf{V}_h \qquad\Longleftrightarrow\qquad \nabla\cdot\mathbf{q}_h\in M_h.$$

  Moreover

  $$\int_\Omega(\nabla\cdot\mathbf{q}_h)\,v_h=b(\mathbf{q}_h,v_h)=\chi(v_h)=-\int_\Omega f\,v_h\qquad\forall v_h\in M_h$$

if and only if (take $v_h = \mu_i$ for all i)

$$(\nabla\cdot\mathbf{q}_h)|_K = \frac{1}{|K|}\int_K \nabla\cdot\mathbf{q}_h = -\frac{1}{|K|}\int_K f \qquad \forall K\in\mathcal{T}_h.$$

Thinking of the boundaries of the elements instead, we can write

$$\int_{\partial K}\mathbf{q}_h\cdot\boldsymbol{\nu} = \int_K \nabla\cdot\mathbf{q}_h = -\int_K f,$$

which means that the method is going to be locally conservative.

- Following element by element the arguments used in the local RT interpolation, it follows that if $\mathbf{q}\in\mathcal{C}^1(\overline{\Omega})^d$, and $\Pi_h\mathbf{q}$ is its RT interpolant, then

  $$\int_\Omega(\nabla\cdot\Pi_h\mathbf{q})v_h = \int_\Omega(\nabla\cdot\mathbf{q})v_h \qquad \forall v_h\in M_h.$$

  This implies that the divergence of the RT interpolant is the best $L^2(\Omega)$ approximation of $\nabla\cdot\mathbf{q}$ in M.

We will not deal here with the discrete inf-sup condition. Let us just be said that it is satisfied, so we now that the pair $\mathbf{V}_h\times M_h$ is a valid pair for the problem where we want to use it.

The discrete system. The Galerkin approximation of our model problem is

$$\left[\begin{array}{lll} (\mathbf{q}_h,u_h)\in\mathbf{V}_h\times M_h, & & \\ \int_F\mathbf{q}_h\cdot\boldsymbol{\nu} = \int_F g_1, \qquad F\subset\Gamma_N, & & \\ a(\mathbf{q}_h,\mathbf{p}_h)+b(\mathbf{p}_h,u_h) & =\ell(\mathbf{p}_h) & \forall\mathbf{p}_h\in\mathbf{V}_{h,0}, \\ b(\mathbf{q}_h,v_h) & =\chi(v_h) & \forall v_h\in M_h. \end{array}\right.$$

The key matrices and vectors are

$$\begin{aligned} a_{ij} &= \int_\Omega(a\boldsymbol{\varphi}_j)\cdot\boldsymbol{\varphi}_i, && i,j=1,\ldots,N,\\ b_{ij} &= \int_\Omega(\nabla\cdot\boldsymbol{\varphi}_j)\mu_i && i=1,\ldots,S,\quad j=1,\ldots,N,\\ \ell_i &= \int_{\Gamma_D} g_0(\boldsymbol{\varphi}_i\cdot\boldsymbol{\nu}) && i=1,\ldots,N\\ \chi_i &= -\int_\Omega f\,\mu_i && i=1,\ldots,S. \end{aligned}$$

Note that

$$b_{ij} = \int_{K_i}\nabla\cdot\boldsymbol{\varphi}_j = \int_{\partial K_i}\boldsymbol{\varphi}_j\cdot\boldsymbol{\nu},$$

which is going to simplify the computation of this matrix quite a lot. Also

$$\ell_i = \begin{cases} \dfrac{1}{|F_i|}\displaystyle\int_{F_i} g_0 \approx g_0(\mathbf{m}_{F_i}), & \text{if } F_i \subset \Gamma_D, \\ 0, & \text{otherwise,} \end{cases}$$

and

$$\chi_i = -\int_{K_i} f \approx -|K_i| f(\mathbf{b}_{K_i}).$$

Here we have used low order quadrature approximations for the right-hand sides of the equation, evaluating the Dirichlet condition in the midpoint/barycenter of the Dirichlet faces and the source term in the barycenter of the elements. We can also simplify the imposition of the essential (Neumann) condition by approximating

$$\int_F \mathbf{q}_h \cdot \boldsymbol{\nu} = \int_F g_1 \approx |F| g_1(\mathbf{m}_F).$$

With this simplication

$$\mathbf{q}_h = \sum_{j\in\mathrm{Free}} q_j \boldsymbol{\varphi}_j + \sum_{j\in\mathrm{Neu}} |F_j| g_1(\mathbf{m}_{F_j}) \boldsymbol{\varphi}_j.$$

3.2 Direct implementation

In this section, we are going to detail how to implement the RT approximation for the model mixed problem *in two dimensions* and assuming that the matrix $a = \kappa^{-1}$ is the identity matrix. Since we are back to the two dimensional case, let's call edges edges instead of faces.

Basic ideas. The kind of information that is needed for the lowest order RT element is included in what is needed to code the $\mathbb{P}_2$ FEM in a triangulation.

- We need to number edges. This is an $N_{\rm edg} \times 2$ list connecting pairs of vertices of the triangulation. The orientation of the edge is naturally induced by this list. When we go from the beginning to the end of the edge, we assume that the normal vector points to the right. Boundary edges should be numbered positively, that is, the normal should point outwards.

- We need to identify Neumann edges, giving a list Neu $\subset \{1, \ldots, N_{\rm edg}\}$ with the indices of edges contained in Γ_N. We then create Free $= \{1, \ldots, N_{\rm edg}\} \setminus$ Neu.

- We need a list of edges counted by element. This would be an $N_{\rm elt} \times 3$ list counting (in a preset order that is established in the reference element) the faces.

- We also need a list of orientations: the orientation of the ℓ-th face of the element K is positive when the outward pointing normal vector on this face coincides with the pre-established normal vector.

- For the right-hand side of the first equation, we need a list of Dirichlet edges.

Assembly. The matrices $\mathbf{A}$ and $\mathbf{B}$ have to be prepared using local-to-global information and, in the case of $\mathbf{A}$, an assembly process. The matrix $\mathbf{B}$ is actually very simple to create. Assume that the element K_i has edges $\{n_1, n_2, n_3\} \subset \{1, \ldots, N_{\text{edg}}\}$ with orientations $\{s_1, s_2, s_3\}$ (here $s_i = \pm$). Then

$$\int_{K_i} \nabla \cdot \boldsymbol{\varphi}_{n_\alpha} = \int_{\partial K_i} \boldsymbol{\varphi}_{n_\alpha} \cdot \boldsymbol{\nu} = s_\alpha \int_{F_{n_\alpha}} \boldsymbol{\varphi}_{n_\alpha} \cdot \boldsymbol{\nu} = s_\alpha.$$

All other elements of the i-th row of $\mathbf{B}$ vanish, since they correspond to edges that are not edges of K_i. We then only need to place the orientations of the edges (counted by element) where the list of edges counted by elements tells us to. The matrix $\mathbf{A}$ requires some additional work. Note first that

$$a_{ij} = \int_K \boldsymbol{\varphi}_j \cdot \boldsymbol{\varphi}_i$$

vanishes if (i, j) are not vertices of the same element. We will start the assembly process by constructing local matrices

$$\mathbf{A}_K = \left[\int_K \mathbf{N}_\alpha^K \cdot \mathbf{N}_\beta^K \right]_{\alpha,\beta=1}^3 .$$

This matrix is not directly assembled, since we have to take into account the orientations of the edges. Let's try to clarify this issue with an example. Assume that the element K has edges $\{3, 8, 4\}$ with orientations $\{-, +, +\}$. Then

$$\boldsymbol{\varphi}_3|_K = -\mathbf{N}_1^K, \qquad \boldsymbol{\varphi}_8|_K = \mathbf{N}_2^K, \qquad \boldsymbol{\varphi}_4|_K = \mathbf{N}_3^K.$$

Instead of assembling the matrix

$$\begin{bmatrix} a_{11}^K & a_{12}^K & a_{13}^K \\ a_{21}^K & a_{22}^K & a_{23}^K \\ a_{31}^K & a_{32}^K & a_{33}^K \end{bmatrix},$$

we need to assemble

$$\begin{bmatrix} a_{11}^K & -a_{12}^K & -a_{13}^K \\ -a_{21}^K & a_{22}^K & a_{23}^K \\ -a_{31}^K & a_{32}^K & a_{33}^K \end{bmatrix},$$

that is, we change the sign of the rows and columns corresponding to edges that have a negative global orientation from the point of view of K. Alternatively, we can think of creating the matrix

$$s_\alpha^K s_\beta^K \int_K \int_K \mathbf{N}_\alpha^K \cdot \mathbf{N}_\beta^K,$$

which uses some global information (the orientations). After the signs have been corrected, the assembly is done in a very similar way as we do the assembly of the $\mathbb{P}_1$ FEM, but using the information for the list of edges counted by element instead of indexing vertices by element.

A computation in the reference element. The local matrix $\mathbf{A}_K$ can be computed in the reference element. Recall that

$$\mathbf{N}_\alpha \circ \mathrm{F}_K = \frac{1}{|\det \mathrm{B}_K|} \mathrm{B}_K \widehat{\mathbf{N}}_\alpha.$$

Then

$$\begin{aligned}\int_K \mathbf{N}^K_\alpha \cdot \mathbf{N}^K_\beta &= \frac{1}{|\det \mathrm{B}_K|} \int_{\widehat{K}} (\mathrm{B}_K \widehat{\mathbf{N}}_\alpha) \cdot (\mathrm{B}_K \widehat{\mathbf{N}}_\beta) \\ &= \int_{\widehat{K}} (\mathrm{C}_K \widehat{\mathbf{N}}_\alpha) \cdot \widehat{\mathbf{N}}_\beta \qquad\qquad \text{where} \quad \mathrm{C}_K = \frac{1}{|\det \mathrm{B}_K|} \mathrm{B}_K^\top \mathrm{B}_K.\end{aligned}$$

If we separate components

$$\widehat{\mathbf{N}}_\alpha = \left[\begin{array}{c} \widehat{N}^1_\alpha \\ \widehat{N}^2_\alpha \end{array}\right], \qquad \mathrm{C}_K = \left[\begin{array}{cc} c^K_{11} & c^K_{12} \\ c^K_{21} & c^K_{22} \end{array}\right],$$

we can write

$$\int_K \mathbf{N}^K_\alpha \cdot \mathbf{N}^K_\beta = c^K_{11} \int_{\widehat{K}} \widehat{N}^1_\alpha \widehat{N}^1_\beta + c^K_{12} \int_{\widehat{K}} \widehat{N}^2_\alpha \widehat{N}^1_\beta + c^K_{21} \int_{\widehat{K}} \widehat{N}^1_\alpha \widehat{N}^2_\beta + c^K_{22} \int_{\widehat{K}} \widehat{N}^2_\alpha \widehat{N}^2_\beta,$$

which reduces all the work to precomputing four (actually three) matrices in the reference element.

Imposition of essential BC. Similarly to what we did with the $\mathbb{P}_1$ FEM, we build the full matrices $\mathbf{A}$ and $\mathbf{B}$. The columns of

$$\left[\begin{array}{c} \mathbf{A} \\ \mathbf{B} \end{array}\right]$$

corresponding to Neu indices will be multiplied by the Neumann BC values and moved to the right-hand side. The rows of

$$\left[\begin{array}{cc} \mathbf{A} & \mathbf{B}^\top \end{array}\right]$$

(and the entries of $\boldsymbol{\ell}$) corresponding to Neu indices will be ignored in the final system.

4 Exercises

1. Compute the Lagrange basis for the lowest Raviart-Thomas space in the reference element in the plane. Use it to compute the matrices 3×3 matrices

$$\int_{\widehat{K}} \widehat{N}^l_\alpha \widehat{N}^m_\beta, \qquad \alpha, \beta = 1, 2, 3$$

for $l, m \in \{1, 2\}$. Here $(\widehat{N}^1_\alpha, \widehat{N}^2_\alpha)$ are the components of $\widehat{\mathbf{N}}_\alpha$.

2. Compute the Lagrange basis for $\mathrm{RT}_0(\widehat{K})$ in three dimensions.

3. Let $\{\boldsymbol{\varphi}_j\}$ be the global Lagrange basis for RT_h^0 in two or three dimensions. Show that the matrix

$$\int_{K_i} \nabla \cdot \boldsymbol{\varphi}_j \qquad i = 1, \ldots, N_{\mathrm{elt}}, \quad j = 1, \ldots, N_{\mathrm{faces}}$$

   does not have full row rank. Equivalently, show that its transpose has non-trivial nullspace. (Hint. $\mathbf{q}_h \equiv \mathbf{1} \in \mathrm{RT}_h^0$.)

Lesson 10

More on mixed elements

In this final lesson we continue with our study of mixed Finite Elements, introducing a different form of implementing the RT system for the mixed Laplacian and some alternative mixed Finite Elements for the same problem. The notation of this lesson follows that of the previous one and most concepts will not be reintroduced. Recall that we have written the generalized Laplace equation in the form of two first order PDE

$$a\,\mathbf{q} - \nabla u = 0 \qquad \nabla\cdot\mathbf{q} = -f \qquad \text{in } \Omega,$$

with associated boundary conditions

$$u = g_0 \quad \text{on } \Gamma_D, \qquad \mathbf{q}\cdot\boldsymbol{\nu} = g_1 \quad \text{on } \Gamma_N.$$

This led to a weak formulation of mixed type

$$\left[\begin{array}{lll} (\mathbf{q},u)\in\mathbf{V}\times M, & & \\ \mathbf{q}\cdot\boldsymbol{\nu} = g_1 \quad \text{on } \Gamma_N, & & \\ \displaystyle\int_\Omega (a\mathbf{q})\cdot\mathbf{p} + \int_\Omega (\nabla\cdot\mathbf{p})u & = \displaystyle\int_{\Gamma_D} g_0\,(\mathbf{p}\cdot\boldsymbol{\nu}) & \forall\mathbf{p}\in\mathbf{V}_0, \\ \displaystyle\int_\Omega (\mathbf{q}\cdot\boldsymbol{\nu})v & = -\displaystyle\int_\Omega f\,v & \forall v\in M, \end{array}\right.$$

using the spaces

$$\mathbf{V} = \mathbf{H}(\mathrm{div}\,\Omega), \quad \mathbf{V}_0 = \{\mathbf{p}\in\mathbf{V}\,:\,\mathbf{p}\cdot\boldsymbol{\nu} = 0 \quad \text{on } \Gamma_N\}, \quad M = L^2(\Omega).$$

1 Hybridized implementation

1.1 An extended system with more multipliers

Motivation. One of the main disadvantages of dealing with mixed FEM is the odd structure of the linear system, with a symmetric matrix that is highly indefinite and has a large block of zeros in a diagonal position. With the RT elements there is an alternative form of implementation that does the following:

- It decouples the system into element-by-element problems by adding some Lagrange multipliers that deal with the continuity of $\mathbf{q}_h \cdot \boldsymbol{\nu}$ across faces.
- It finds an equivalent system where only the Lagrange multipliers are solved.
- Finally, we solve element by element to find $\mathbf{q}_h$ and u_h.

This might seem a quite convoluted way of doing things (it is!), but it is very efficient and it has lead to the development of other very effective finite element methods in the category of Discontinuous Galerkin Methods. The hybridized formulation of the Raviart-Thomas mixed FEM is due to Douglas Arnold and Franco Brezzi. It has the additional advantage of offering a new quantity whose convergence properties are good as well.

Notation. We will divide the faces into three groups:

- Interior faces $\mathcal{F}_h^{\mathrm{int}}$,
- Dirichlet faces $\mathcal{F}_h^{\mathrm{dir}}$,
- Neumann faces $\mathcal{F}_h^{\mathrm{neu}}$.

Recall that Neumann faces play a separate role because it's on them that the essential boundary condition is imposed. Interior and Dirichlet faces go together at the time of counting Free degrees of freedom, although Dirichlet faces have to be separated at the time of computing the right hand side of the first equation.

Rethinking H(div)-conformity. Recall that we have the global RT space

$$\mathbf{V}_h = \{\mathbf{q}_h : \Omega \in \mathbb{R}^d \,:\, \mathbf{q}_h \in \mathbf{H}(\mathrm{div}, \Omega), \quad \mathbf{q}_h|_K \in \mathrm{RT}_0(K) \forall K \in \mathcal{T}_h\}.$$

This space can be understood as the subspace of

$$\mathbf{W}_h := \{\mathbf{q}_h : \Omega \in \mathbb{R}^d \,:\, \mathbf{q}_h|_K \in \mathrm{RT}_0(K) \forall K \in \mathcal{T}_h\}$$

consisting of functions such that

$$\int_F [\![\mathbf{q}_h \cdot \boldsymbol{\nu}]\!] = 0 \qquad \forall F \in \mathcal{F}_h^{\mathrm{int}}.$$

The quantity $[\![\mathbf{q}_h \cdot \boldsymbol{\nu}]\!]$ is the jump of $\mathbf{q}_h \cdot \boldsymbol{\nu}$ across F (remember that $\mathbf{q}_h \cdot \boldsymbol{\nu}|_F$ is constant). We also have

$$\int_F \mathbf{q}_h \cdot \boldsymbol{\nu} = \int_F g_1 \qquad \forall F \in \mathcal{F}_h^{\mathrm{neu}}.$$

Let us then introduce the skeleton of the triangulation

$$\partial\mathcal{T}_h := \cup\{F \,:\, F \in \mathcal{F}_h\},$$

and some auxiliary spaces defined on it

$$\begin{array}{rcl}\Xi_h & := & \{\xi_h : \partial\mathcal{T}_h \to \mathbb{R} \,:\, \xi_h|_{\in}\mathbb{P}_0(F) \quad \forall F \in \mathcal{F}_h\},\\ \Xi_h^{\mathrm{int}} & := & \{\xi_h \in \Xi_h \,:\, \xi_h = 0 \text{ on } \Gamma\},\\ \Xi_h^{0} & := & \{\xi_h \in \Xi_h \,:\, \xi_h = 0 \text{ on } \Gamma_D\},\\ \Xi_h^{\mathrm{dir}} & := & \{\xi_h \in \Xi_h \,:\, \xi_h|_F = 0 \quad F \in \mathcal{F}_h^{\mathrm{int}} \cup \mathcal{F}_h^{\mathrm{dir}}\}.\end{array}$$

We first write the conformity conditions (no jumps in normal component on internal faces) in the following unusual form

$$\sum_{K\in\mathcal{T}_h} \int_{\partial K} (\mathbf{q}_h \cdot \boldsymbol{\nu})\xi_h = 0 \qquad \forall \xi_h \in \Xi_h^{\mathrm{int}}.$$

In this formula, the normal vectors are exterior to each of the elements. To see why this formulation is equivalent to no jumping of normal components, isolate a single face $F \in \mathcal{F}_h^{\mathrm{int}}$ and define

$$\xi_F = \left\{\begin{array}{ll} 1 & \text{in } F,\\ 0 & \text{in all the other faces.}\end{array}\right.$$

Testing with this ξ_F isolates a single face in two elements, with opposite normals, which automatically creates the jump and forces its average to vanish. We can do slightly better and impose the no-jump condition and the Neumann (essential) condition in a single equation

$$\sum_{K\in\mathcal{T}_h} \int_{\partial K} (\mathbf{q}_h \cdot \boldsymbol{\nu})\xi_h = \int_{\Gamma_N} g_1 \qquad \forall \xi_h \in \Xi_h^{0}.$$

Uncoupling the discrete state equation. The first discrete equation is

$$\int_\Omega (a\,\mathbf{q}_h)\cdot\mathbf{p}_h + \int_\Omega (\nabla\cdot\mathbf{p}_h)u_h = \int_{\Gamma_D} g_0(\mathbf{p}_h\cdot\boldsymbol{\nu}) \quad \forall \mathbf{p}_h \in \mathbf{V}_{h,0}.$$

We can write it in this different way

$$\sum_{K\in\mathcal{T}_h} \int_K (a\,\mathbf{q}_h)\cdot\mathbf{p}_h + \sum_{K\in\mathcal{T}_h} \int_K (\nabla\cdot\mathbf{p}_h)u_h - \sum_{K\in\mathcal{T}_h} \int_{\partial K} \rho_h(\mathbf{p}_h\cdot\boldsymbol{\nu}) = 0,$$

where $\rho_h \in \Xi_h$ is still to be determined, although we start by imposing that

$$\int_F \rho_h = \int_F g_0 \quad \Longleftrightarrow \quad \int_F \rho_h(\mathbf{p}_h\cdot\boldsymbol{\nu}) = \int_F g_0(\mathbf{p}_h\cdot\boldsymbol{\nu}) \qquad \forall F \in \mathcal{F}_h^{\mathrm{dir}}.$$

In the sums of integrals over ∂K we have different types of faces:

- Internal faces appear twice, but the sum cancels because the test function $\mathbf{p}_h$ is div-conforming, that is, its normal component does not jump on internal faces.
- Dirichlet faces provide the integral over Γ_D that appeared in the original formulation.

- Neumann faces do not appear in the formulation since $\mathbf{p}_h \cdot \boldsymbol{\nu} = 0$ on them.

What we do now might seem quite bold, but there is some reason to be done. We impose the equation

$$\sum_{K\in\mathcal{T}_h} \int_K \mathbf{q}_h \cdot \mathbf{p}_h + \sum_{K\in\mathcal{T}_h} \int_K (\nabla \cdot \mathbf{p}_h) u_h - \sum_{K\in\mathcal{T}_h} \int_{\partial K} \rho_h (\mathbf{p}_h \cdot \boldsymbol{\nu}) = 0, \qquad \forall \mathbf{p}_h \in \mathbf{W}_h,$$

that is, we relax the need of $\mathbf{p}_h$ to be div-conforming and we relax the constraint $\mathbf{p}_h \cdot \boldsymbol{\nu} = 0$. Instead, we consider the global system with $\mathbf{q}_h$ (and its corresponding test function) uncoupled:

$$\left[\begin{array}{lll}
(\mathbf{q}_h, u_h, \rho_h) \in \mathbf{W}_h \times M_h \times \Xi_h, & & \\
\displaystyle\int_{\Gamma_D} \rho_h \xi_h = \int_{\Gamma_D} g_0 \xi_h \quad \forall \xi_h \in \Xi_h^{\mathrm{dir}}, & & \\
\displaystyle\sum_{K\in\mathcal{T}_h} \int_K (a\mathbf{q}_h) \cdot \mathbf{p}_h + \sum_{K\in\mathcal{T}_h} \int_K (\nabla \cdot \mathbf{p}_h) u_h + \sum_{K\in\mathcal{T}_h} \int_{\partial K} \rho_h (\mathbf{p}_h \cdot \boldsymbol{\nu}) & = 0, & \forall \mathbf{p}_h \in \mathbf{W}_h, \\
\displaystyle\sum_{K\in\mathcal{T}_h} (\nabla \cdot \mathbf{q}_h) v_h & = -\displaystyle\sum_{K\in\mathcal{T}_h} \int_K f\, v_h & \forall v_h \in M_h, \\
\displaystyle\sum_{K\in\mathcal{T}_h} \int_{\partial K} (\mathbf{q}_h \cdot \boldsymbol{\nu}) \xi_h & = \displaystyle\int_{\Gamma_N} g_1\, \xi_h & \forall \xi_h \in \Xi_h^0.
\end{array}\right.$$

This looks pretty bad. The system looks even worse than the original one, but there are several aspects to consider before we think how to implement it:

- The number of equations and unknowns matches. The dimensions of the trial spaces are

$$\mathrm{dim}\mathbf{W}_h = \underbrace{d+1}_{\mathrm{dimR}T_0} \#\mathcal{T}_h, \qquad \mathrm{dim}M_h = \#\mathcal{T}_h, \qquad \mathrm{dim}\Xi_h = \#\mathcal{F}_h,$$

  while for testing we separate Dirichlet faces and all other faces.

- Surprisingly enough, the Dirichlet condition has become now essential (again) and the Neumann foncition has found its place in the last equation.

- The second and third equations (state equations and conservation) can be written element by element because the sapces $\mathbf{W}_h$ and M_h are completely decoupled. We can then write instead

$$\begin{array}{lll}
\displaystyle\int_K (a\mathbf{q}_h) \cdot \mathbf{p}_K + \int_K (\nabla \cdot \mathbf{p}_K) u_h + \int_{\partial K} \rho_h (\mathbf{p}_K \cdot \boldsymbol{\nu}) & = 0, & \forall \mathbf{p}_K \in \mathrm{RT}_0(K), \\
\displaystyle\int_K (\nabla \cdot \mathbf{q}_h) v_K & = -\displaystyle\int_K f v_K & \forall v_k \in \mathbb{P}_0(K),
\end{array}$$

  adding that we impose this for all $K \in \mathcal{T}_h$. There's another way to see this. Forgetting about Dirichlet boundary conditions (take them to be zero so that $\rho_h \in \Xi_h^0$), and taking local bases for all the spaces we can find a matrix formulation with the form (the matrices $\mathbf{A}$ and $\mathbf{B}$ are not the same that before)

$$\begin{bmatrix} \mathbf{A} & \mathbf{B}^\top & \mathbf{C}^{-\top} \\ \mathbf{B} & \mathbf{O} & \mathbf{O} \\ \mathbf{C} & \mathbf{O} & \mathbf{O} \end{bmatrix}.$$

However, the matrices $\mathbf{A}$ and $\mathbf{B}$ are block diagonal, with many (as many as elements) little blocks corresponding to each of the elements. The **hybridization** is the system will consist in inverting those small blocks and deriving an equivalent system.

Why is this system equivalent to the original one? The argument is not complicated, so let us have a look at it. First of all, let us check that the extended system (with $\mathbf{q}_h, u_h, \rho_h$ as unknowns) is uniquely solvable. It is a square system (we've already seen this), so we only need to verify that if the right-hand isde vanishes, then the solution vanishes. Note that if

$$\sum_{K\in\mathcal{T}_h} \int_{\partial K} (\mathbf{q}_h \cdot \boldsymbol{\nu})\xi_h = 0 \qquad \forall \xi_h \in \Xi_h^0,$$

then $\mathbf{q}_h \in \mathbf{V}_{h,0}$. We then test the equation

$$\sum_{K\in\mathcal{T}_h} \int_K (a\,\mathbf{q}_h)\cdot \mathbf{p}_h + \sum_{K\in\mathcal{T}_h} \int_K (\nabla\cdot \mathbf{p}_h)u_h + \sum_{K\in\mathcal{T}_h} \int_{\partial K} \rho_h(\mathbf{p}_h\cdot\boldsymbol{\nu}) = 0, \qquad \forall \mathbf{p}_h \in \mathbf{W}_h,$$

with $\mathbf{p}_h \in \mathbf{V}_{h,0}$ (no jumps, zero normal component on the Neumann faces) and use that

$$\int_{\Gamma_D} \rho_h \xi_h = 0 \quad \forall \xi_h \in \Xi_h^{\mathrm{dir}}.$$

What we get is

$$\int_\Omega (a\,\mathbf{q}_h)\cdot\mathbf{p}_h + \int_\Omega (\nabla\cdot\mathbf{q}_h)u_h = 0 \qquad \forall \mathbf{p}_h \in \mathbf{V}_{h,0}.$$

At the same time, we have assumed that

$$\int_\Omega (\nabla\cdot\mathbf{q}_h)v_h = 0 \qquad \forall v_h \in M_h.$$

These two equations are the ones of the original RT discretization of a homogeneous problem ($f = 0$, $g_0 = 0$, $g_1 = 0$) and, therefore, they imply that $\mathbf{q}_h = \mathbf{0}$ and $u_h = 0$. We then return to the state equation, knowing that $\mathbf{q}_h = 0$ and $u_h = 0$ to obtain

$$\sum_{K\in\mathcal{T}_h} \int_{\partial K} \rho_h(\mathbf{p}_h\cdot\boldsymbol{\nu}) = 0, \qquad \forall \mathbf{p}_h \in \mathbf{W}_h.$$

We just focus on a single element and take arbitrary constant values on the faces of the element to see that $\rho_h = 0$ on all the faces of that element. Since that can be done element by element, we have proved that $\rho_h = 0$. So far, we have only proved that the extended system is uniquely solvable. The final part of the argument is easy, but I'll leave it to the reader for practice. You show that if $(\mathbf{q}_h, u_h, \rho_h) \in \mathbf{W}_h \times M_h \times \Xi_h$ is a solution of the extended system, then $(\mathbf{q}_h, u_h) \in \mathbf{V}_h \times M_h$ (note the change in the first space) is the solution of the original system. The ideas are very much the same that the ones we used to prove uniqueness.

1.2 The hybridized form

Terminology. The expanded formulation we have given in the previous subsection is not the final goal of this section, but only an intermediate step. It is a **hybridizable** formulation that can become a **hybrid** formulation once we eliminate the unknowns $(\mathbf{q}_h, u_h)$ from the system and write an equivalent system where only $\rho_h \in \Xi_h$ appears as an unknown. The system will be symmetric and positive definite (like the ones for classical FEM) and it will have the same dimension as the number of faces. The variables $(\mathbf{q}_h, u_h)$ will be recovered in an element y element postprocessing step. The reason for calling the final formulation a hybrid formulation is because the resulting system looks like one of the Hybrid Finite Element Methods that were devised in the eighties. Hybrid methods, per se, do not seem to be very popular these days[1] though. For convenience we will write

$$N = \dim \Xi_h = \#\mathcal{F}_h.$$

The local systems. Consider the local basis for $\mathrm{RT}_0(K)$, $\{\mathbf{N}_\alpha^K : \alpha = 1, \ldots, d+1\}$ and the following matrices: $\mathbf{A}_K \in \mathbb{R}^{(d+1)\times(d+1)}$ given by

$$\int_K (a\mathbf{N}_\alpha^K) \cdot \mathbf{N}_\beta^K, \qquad \alpha, \beta = 1, \ldots, d+1,$$

$\mathbf{B}_K \in \mathbb{R}^{(d+1)\times 1}$ given by

$$\mathbb{A}_K = \left[\begin{array}{cc} \mathbf{A}_K & \mathbf{B}_K^\top \\ \mathbf{B}_K & 0 \end{array}\right] \in \mathbb{R}^{(d+2)\times(d+2)}.$$

We also need the ‘vector’ $\mathbf{f}_K \in \mathbb{R}$ whose only component is

$$-\int_K f_K$$

and the extended vector

$$\mathbf{b}_K = \left[\begin{array}{c} \mathbf{0} \\ \mathbf{f}_K \end{array}\right] \in \mathbb{R}^{d+2}.$$

A final group of large but very sparse global-local matrices are $\mathbf{C}_K \in \mathbb{R}^{N\times(d+1)}$ with entries

$$\int_{F_i} \mathbf{N}_\alpha^K \cdot \boldsymbol{\nu} \qquad i = 1, \ldots, N, \qquad \alpha = 1, \ldots, d+1,$$

where the normal vector is outward pointing from the point of view of K and the expanded

$$\mathbb{C}^K = \left[\begin{array}{cc} \mathbf{C}_K & \mathbf{0} \end{array}\right] \in \mathbb{R}^{N\times(d+2)}.$$

[1] Many people seem to have forgotten that they even existed. The celebrated monograph *Mixed and hybrid element methods* by Franco Brezzi and Michel Fortin —recently expanded with the contribution of Daniele Boffi— reminds mathematicians of the hybrid methods from time to time. I've asked about them to some colleagues and most of them seem not to have looked at that part of the book. However, Hybridizable Discontinuous Galerkin methods of all kinds and some modern methods based on Ultra Weak Variational Formulations are hybrid methods in their particular ways.

If

$$(\mathbf{q}_K, u_k) \in \mathbb{R}^{d+1} \times \mathbb{R} \longleftrightarrow (\mathbf{q}_h|_K, u_h|_K) \in \mathrm{RT}_0(K) \times \mathbb{P}_0(K),$$

and if $\boldsymbol{\rho} \in \mathbb{R}^N$ is the vector of values of $\rho_h \in \Xi_h$ in the faces $\mathcal{F}_h$, then the local state equation and conservation equations on the element K are equivalent to

$$\mathbb{A}_K \left[\begin{array}{c} \mathbf{q}_K \\ u_K \end{array} \right] + \mathbb{C}_K^\top \boldsymbol{\rho} = \mathbf{b}_K.$$

Note that $\mathbb{C}_K^\top \in \mathbb{R}^{(d+2)\times N}$ has very few non-zero columns, corresponding to the faces $F \in \mathcal{F}(K)$, the only ones that participate in the equations on the element K. These equations say that if we know $\boldsymbol{\rho}$, we can solve for $(\mathbf{q}_K, u_K)$ element by element.

The global hybridized system. Let us forget for the moment about any kind of Dirichlet boundary conditions. We construct the global vector $\mathbf{g} \in \mathrm{R}^N$ with components

$$g_i = \begin{cases} \displaystyle\int_{F_i} g_1 & \text{if } i \in \text{Neu, i.e., } F_i \in \mathcal{F}_h^{\text{neu}}, \\ 0, & \text{otherwise.} \end{cases}$$

The local systems derived in the previous paragraph are complemented with the following block of equations

$$\sum_{K\in\mathcal{T}_h} \mathbb{C}_K \left[\begin{array}{c} \mathbf{q}_K \\ u_K \end{array} \right] = \mathbf{g}.$$

The hybridized system consists of solving the local equations (in terms of $\boldsymbol{\xi}$, which is not know yet)

$$\left[\begin{array}{c} \mathbf{q}_K \\ u_K \end{array} \right] = -\mathbb{A}_K^{-1}\mathbb{C}_K^\top \boldsymbol{\rho} + \mathbb{A}_K^{-1}\mathbf{b}_K$$

and substitute in the newly obtained block of equations

$$\left(\sum_{K\in\mathcal{T}_h} \mathbb{C}_K \mathbb{A}_K^{-1} \mathbb{C}_K^\top \right) \boldsymbol{\rho} = -\mathbf{g} - \sum_{K\in\mathcal{T}_h} \mathbb{A}_K^{-1}\mathbf{b}_K.$$

This global system is written in terms of the variable $\boldsymbol{\rho} \leftrightarrow \rho_h$. We are going to leave it here with some final comments:

- Imposition of the Dirichlet boundary conditions is equivalent to fixing ρ_j for $j \in \mathrm{Dir}$ and sending that part of the system to the right-hand side. At the same time, the rows of the global hybridized system corresponding to Dirichlet faces are ignored.

- The matrices $\mathbb{C}_K \mathbb{A}_K^{-1} \mathbb{C}_K^\top \in \mathbb{R}^{N\times N}$ can be computed in a reduced $\mathbb{R}^{(d+1)\times(d+1)}$ form, corresponding to a local count of the faces of each element. Then they are assembled into the global system. That part of the code is very similar to the assembly of the matrix $\mathbf{A}$ in the RT system.

- The variable ρ_h has been introduced as a sort of Lagrange multiplier compensanting from the fact that we impose the continuity of $\mathbf{q}_h \cdot \boldsymbol{\nu}$ (and the Neumann BC) with an equation instead of a reduction of the number of unknowns. It is remarkable, however, that $\rho_h|_F \approx u|_F$, that is, the value of ρ_h of F is an approximation of u on that face.

2 Higher order div-conforming elements

2.1 Second order Raviart-Thomas elements

Motivation. We haven't really talked about convergence of the mixed method (based on RT elements of the lowest order and piecewise constant functions) for the generalized Laplacian. We have an abstract theory that gives (after assuming a certain amount of hypotheses, among which the not-easy-to-grasp inf-sup condition) a Céa-type estimate for the solution. After that, what is left is using scaling arguments and the Bramble-Hilbert lemma, using the Raviart-Thomas interpolation operator. Long story short, the order of convergence for both variables in the natural norms is one

$$\|\mathbf{q}_h - \mathbf{q}\|_{\mathrm{div},\Omega} + \|u - u_h\|_\Omega = \mathcal{O}(h),$$

assuming that $\mathbf{q} \in H^1(\Omega)^d$, that $\nabla \cdot \mathbf{q} \in H^1(\Omega)$ and that $u \in H^1(\Omega)$. In this section we present a second order RT element. Similar to what happens to Lagrange finite elements, there is a family of RT elements of all orders.

Notation. In addition to the $(d+1)$-dimensional space of polynomials of degree at most one, $\mathbb{P}_1(K)$, we will need to consider the d-dimensional space of homogeneous polynomials of degree one

$$\widetilde{\mathbb{P}}_1(K) = \{\mathbf{a}\cdot\mathbf{x} \,:\, \mathbf{a} \in \mathbb{R}^d\} = \begin{cases} \mathrm{span}\{x_1, x_2\} & \text{when } d = 2,\\ \mathrm{span}\{x_1, x_2, x_3\} & \text{when } d = 3,\end{cases}$$

and the d-dimensional space of polynomials of degree at most one defined on a face $F \in \mathcal{F}(K)$

$$\mathbb{P}_1(F) = \{p : F \to \mathbb{R} \,:\, p = q|_F, \quad q \in \mathbb{P}_1(K)\}.$$

For the arguments that follow, we will assume to have been given a basis of $\mathbb{P}_1(F)$ for every face of K

$$\mathbb{P}_1(F) = \mathrm{span}\{\psi_\alpha^K \,:\, \alpha = 1, \ldots, d\}.$$

The new local RT space. We define

$$\begin{aligned}\mathrm{RT}_1(K) &= \{\mathbf{p}(\mathbf{x}) = \mathbf{p}_1(\mathbf{x}) + \widetilde{p}_1(\mathbf{x})\mathbf{x} \,:\, \mathbf{p}_1 \in \mathbb{P}_1(K)^d,\, \widetilde{p}_1 \in \widetilde{\mathbb{P}}_1(K)\}\\ &= \{\mathbf{p}(\mathbf{x}) = \mathbf{p}_1(\mathbf{x}) + p_1(\mathbf{x})\mathbf{x} \,:\, \mathbf{p}_1 \in \mathbb{P}_1(K)^d,\, p_1 \in \mathbb{P}_1(K)\}.\end{aligned}$$

Note that while the second description of the space is simpler (it doesn't need the introduction of the space of homogeneous polynomials, it allows for the construction of the same polynomial with the two different parts of the construction, since $a\mathbf{x} \in \mathbb{P}_1(K)^d$ when $a \in \mathbb{R}$. Just to be sure that we know what we are talking about, here is a general element of $\mathrm{RT}_1(K)$ when $d = 2$

$$\begin{bmatrix} a_0 + a_1 x_1 + a_2 x_2 \\ b_0 + b_1 x_1 + b_2 x_2 \end{bmatrix} + (c_1 x_1 + c_2 x_2)\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Easy properties.

1. It is easy to check that

$$\begin{aligned}\dim \mathrm{RT}_1(K) &= d\dim \mathbb{P}_1(K) + \dim \widetilde{\mathbb{P}}_1(K)\\ &= d(d+1)+d = d(d+2)\\ &= \begin{cases} 8, & \text{for } d=2\\ 15, & \text{for } d=3.\end{cases}\end{aligned}$$

2. It is quite obvious that

$$\mathbb{P}_1(K)^d \subset \mathrm{RT}_1(K) \subset \mathbb{P}_2(K)^d.$$

3. If $\mathbf{p} \in \mathrm{RT}_1(K)$, then

$$\mathbf{p}\cdot\boldsymbol{\nu}_F|_F \in \mathbb{P}_1(F) \qquad \forall F \in \mathcal{F}(K).$$

   The reason is the same than in the case of the lowest order RT elements: on F, $\mathbf{x}\cdot\boldsymbol{\nu}_F$ is a constant.

4. If $\mathbf{p} = \mathbf{p}_1 + \widetilde{p}_1\,\mathbf{x}$ (with $\mathbf{p}_1 \in \mathbb{P}_1(K)^d$ and $\widetilde{p}_1 \in \widetilde{\mathbb{P}_1}^d$), then

$$\begin{aligned}\nabla\cdot\mathbf{p} &= \nabla\cdot\mathbf{p}_1 + \nabla\widetilde{p}_1\cdot\mathbf{x} + \widetilde{p}_1(\nabla\cdot\mathbf{x})\\ &= \nabla\cdot\mathbf{p}_1 + \widetilde{p}_1 + \widetilde{p}_1\,d = \nabla\cdot\mathbf{p}_1 + (d+1)\widetilde{p}_1.\end{aligned}$$

   (We have applied Euler's homogeneous function theorem, or the very simple argument that if $\widetilde{p}_1 = \mathbf{a}\cdot\mathbf{x}$, then $\nabla\widetilde{p}_1 = \mathbf{a}$ and therefore $\nabla\widetilde{p}_1\cdot\mathbf{x} = \mathbf{a}\cdot\mathbf{x} = \widetilde{p}_1$.) This formula has some consequences. Note that the divergence of the two parts of $\mathbf{p}$ is neatly separated into a constant function (coming for $\mathbf{p}_1$) and a multiple of the homogeneous polynomial $\widetilde{p}_1$ used in the construction of $\mathbf{p}$. Therefore, if $\nabla\cdot\mathbf{p} = 0$, it follows that $\mathbf{p} = \mathbf{p}_1 \in \mathbb{P}_1(K)^d$ and $\nabla\cdot\mathbf{p}_1 = 0$.

The local degrees of freedom. Since we are using these elements to build a Finite Element space where we need the normal component to be continuous on the interelement faces, it is natural to start by controling $\mathbf{p}\cdot\boldsymbol{\nu}_F \in \mathbb{P}_1(F)$ on all the faces. This is equivalent to imposing the DOF

$$\int_F (\mathbf{p}\cdot\boldsymbol{\nu}_F)\psi_\alpha^F \qquad \alpha = 1,\ldots,d, \quad F \in \mathcal{F}(K).$$

This gives us $d(d+1)$ ($d+1$ faces and d moments on each face) degrees of freedom. So we are still short by d conditions from having a well defined set of degrees of freedom. The additional conditions are defined in the interior of the element. We impose the value of

$$\int_K \mathbf{p}.$$

Is this like a proof? We are next going to shot that these DOF are valid for the space $\mathrm{RT}_1(K)$. This can be asserted as the ability to define a local RT interpolation operator: for arbitrary c_α^F and $Ↄ \in \mathbb{R}^d$, there exists a unique $\mathbf{p} \in \mathrm{RT}_1(K)$ such that

$$\int_F (\mathbf{p} \cdot \boldsymbol{\nu}_F)\psi_\alpha^F = c_\alpha^F \qquad \alpha = 1, \ldots, d, \quad F \in \mathcal{F}(K),$$

and

$$\int_K \mathbf{p} = \mathbf{a}.$$

This is equivalent to (please, make sure that you understand why) proving that if $\mathbf{p} \in \mathrm{RT}_1(K)$ satisfies

$$\int_F (\mathbf{p} \cdot \boldsymbol{\nu}_F)\psi_\alpha^F = o \qquad \alpha = 1, \ldots, d, \quad F \in \mathcal{F}(K),$$

and

$$\int_K \mathbf{p} = \mathbf{0},$$

then $\mathbf{p} = \mathbf{0}$. Note that if

$$\int_F (\mathbf{p} \cdot \boldsymbol{\nu}_F)\psi_\alpha^F = o \qquad \alpha = 1, \ldots, d,$$

then $\mathbf{p} \cdot \boldsymbol{\nu}_F = 0$ on F (this follows from the fact that $\mathbf{p} \cdot \boldsymbol{\nu}_F|_F \in \mathbb{P}_1(F)$. Therefore, the cancellation of the boundary DOF implies that $\mathbf{p} \cdot \boldsymbol{\nu} = 0$ on ∂K. At the same time, we know that

$$\int_K \mathbf{p} \cdot \mathbf{b} = 0 \qquad \forall \mathbf{b} \in \mathbb{R}^d.$$

Therefore, for any $q \in \mathbb{P}_1(K)$,

$$\begin{aligned}
0 &= \int_{\partial K} (\mathbf{p} \cdot \boldsymbol{\nu}) q && (\mathbf{p} \cdot \boldsymbol{\nu} = 0 \text{ on } \partial K) \\
&= \int_K \mathbf{p} \cdot \nabla q + \int_K (\nabla \cdot \mathbf{p}) q && (\text{divergence theorem}) \\
&= \int_K (\nabla \cdot \mathbf{p}) q. && (\nabla q \text{ is constant})
\end{aligned}$$

This implies that $\nabla \cdot \mathbf{p} \in \mathbb{P}_1(K)$ vanishes (take $q = \nabla \cdot \mathbf{p}$ in the previous argument) and therefore $\mathbf{p} \in \mathbb{P}_1(K)^d$ (this was one of the easy properties). Take a single face F and its normal vector $\boldsymbol{\nu}_F$. Consider the polynomial $\mathbf{p} \cdot \boldsymbol{\nu}_F \in \mathbb{P}_1(K)$ (defined in the element, not only on the face F). This polynomial, of degree one at most, vanishes on the plane (line when $d = 2$) containing F, and therefore, has constant sign in K, which lies on one side of this plane. However, since $\boldsymbol{\nu}_F \in \mathbb{R}^d$, we know that

$$\int_K \mathbf{p} \cdot \boldsymbol{\nu}_F = 0.$$

Therefore $\mathbf{p}\cdot\boldsymbol{\nu}_F \equiv 0$. Since we can prove the same result for all the faces, we have shown that =0 (with d of the normal vectors we can build a basis for $\mathbb{R}^d$). We have not given many proofs of this type in these notes. I have decided to include it in order to show how 'nonlinear' these arguments are. You start with some of the hypotheses, conclude something, start using another group of hypotheses and a side argument (an easy property) to get something else, go back to the first set of conclusions and use the hypotheses in another way, etc. And this is just a simple particular case. I want to emphasize that these nifty arguments are prevalent in construction of FE spaces of all kinds.

The global spaces and the mixed Laplacian. Gluing the local RT spaces we can build

$$\mathbf{V}_h = \mathrm{RT}^1_h = \{\mathbf{p}_h : \Omega \to \mathbb{R}^d \,:\, \mathbf{p}_h \in \mathbf{H}(\mathrm{div},\Omega),\, \mathbf{p}_h|_K \in \mathrm{RT}_1(K) \quad \forall K \in \mathcal{T}_h\}.$$

Since a function in $\mathrm{RT}_1(K)$ can be reconstructed by its local degrees of freedom and all the DOF that control the value of the normal component on a common face are shared by adjacent elements, it is not difficult to convince yourself (and even prove it!) that

$$\dim \mathrm{RT}^1_h = d\,\#\mathcal{F}_h + d\,\#\mathcal{T}_h = d(\#\{\text{faces}\} + \#\{\text{elements}\}).$$

This space of discrete vector fields is paired with the space

$$M_h = \{u_h : \Omega \to \mathbb{R} \,:\, u_h|_K \in \mathcal{P}_1(K) \quad \forall K \in \mathcal{T}_h\},$$

containing all piecewise linear (not globally continuous) functions. The dimension of this space is $(d+1)$ times the number of elements. Note that if $\mathbf{p}_h \in \mathbf{V}_h$ satisfies

$$\int_\Omega (\nabla\cdot\mathbf{p}_h)v_h = 0 \qquad \forall v_h \in M_h,$$

then (take $v_h = \nabla\cdot\mathbf{p}_h$) $\nabla\cdot\mathbf{p}_h = 0$. (This was one of the hypotheses that we required for the pair $\mathbf{V}_h \times M_h$ to work as a discretization pair for the mixed Laplacian.) The uniform discrete inf-sup condition is also satisfied, but we will not show why. Beyond this (nontrivial) detail, here's no need to repeat again what we did for the lowest order RT element. The order of convergence for smooth solutions of the problem is $\mathcal{O}(h^2)$.

2.2 The lowest order BDM elements

3 Exercises

1. Compute the Lagrange basis for $\mathrm{RT}_1(\widehat{K})$ in the plane.

2. Let $\mathbf{p}_h \in \mathrm{RT}_1(K)$ be the RT interpolant of a sufficiently smooth $\mathbf{p} : K \to \mathbb{R}^d$. Show that

$$\int_K (\nabla\cdot\mathbf{p}_h)q = \int_K (\nabla\cdot\mathbf{p})\,q \qquad \forall q \in \mathbb{P}_1(K).$$

(**Hint.** In case of doubt, integrate by parts!)

3. Describe the supports of the global basis functions for RT_h^1.

4. Define a global RT interpolation operator for second order RT elements. Show that the divergence of the interpolant is the best $L^2(\Omega)$ approximation of the divergence on the space M_h.

Appendices

1 Bookkeeping for $\mathbb{P}_1$ elements

A important aspect of the implementation of the Finite Element Method (and of any other non-trivial numerical method) is the choice of good data structures to store the necessary information. We are going to detail here one usual option for these structures. Keep in mind that we are going to see much more complicated methods in the sequel, so part of what is done here is preparing the ground for more complicated situations.

We assume that the boundary is divided in sides with a numbering of boundary subdomains. It is necessary to know what the boundary condition is on each sumdomain and how to evaluate the corresponding function for the boundary condition. The following data are needed for the implementation of the $\mathbb{P}_1$ finite element method:

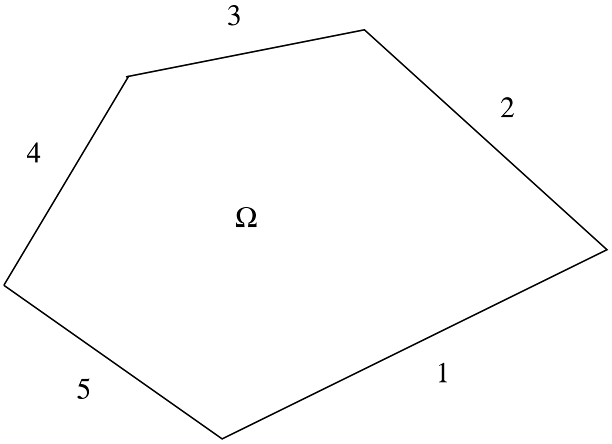


Figure 10.1: Numbering of the sides of the boundary

- The global number of nodes nNod.
- A numbering of the nodes and their coordinates. This can simply be done by giving a double vector with the coordinates of all nodes. Numbering is done by component:

$$\begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \end{bmatrix}$$

- The number of triangles.
- A relation of nodes elementwise:

$$\begin{bmatrix} n_{11} & n_{12} & n_{13} \\ n_{21} & n_{22} & n_{23} \\ \vdots & \vdots & \vdots \end{bmatrix}$$

  (for instance, the 4th row of this matrix are the global indices for the 1st, 2nd and 3rd vertices of the fourth triangle). It is common to expect from a mesh-generator that the ordering of the nodes is counterclockwise.

- A list of Dirichlet nodes (Dir), mentioning on what boundary subdomain they are so that we know which function to evaluate.
- The number of Neumann edges (edges that lie on the Neumann boundary)
- A list of the Neumann edges, indicating what their vertices are and on which boundary subdomain they lie.

Usually grid generators give Dirichlet edges instead of nodes, i.e.,

- A list of Dirichlet edges (edges on the Dirichlet boundary), indicating what their vertices are and on which boundary subdomain they lie.

From this list, the construction of the list Dir and the complementary list Ind is a simple preprocess that has to be performed before the assembly process is begun.

Let us now detail the example of Lessons 1 and 2. Figure 10.2 gives you a numbering of the vertices. The triangulation is described with the following data:

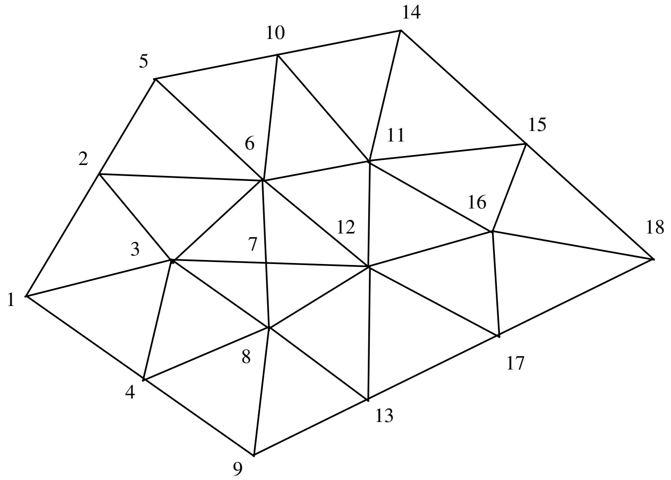


Figure 10.2: Global numbering of nodes.

- 18 nodes (of which 6 are Dirichlet nodes)

- 23 triangles
- 6 Neumann edges
- Relation between local and global numbering of vertices for triangles:

$$\text{23 rows}\left\{\begin{bmatrix} 1 & 3 & 2 \\ 1 & 4 & 3 \\ 4 & 8 & 3 \\ 3 & 6 & 2 \\ 3 & 7 & 6 \\ 3 & 8 & 7 \\ 7 & 12 & 6 \\ 8 & 12 & 7 \\ 4 & 9 & 8 \\ \vdots & \vdots & \vdots \end{bmatrix}\right.$$

- A list of Dirichlet nodes, indicating the boundary subdomain (the side of Γ) where they are

$$\begin{bmatrix} 9 & 1 \\ 13 & 1 \\ 17 & 1 \\ 18 & 2 \\ 15 & 2 \\ 14 & 2 \end{bmatrix}$$

  (in this list it is not relevant that the order is increasing). Node number 18 could be placed on the 2nd or the 1st side. Since the Dirichlet condition cannot be discontinuous in this formulation, it is immaterial which choice is taken.

- The list Ind is obtained from the list $\{1, 2, \ldots, 18\}$ by erasing everything that appears on the firs column of Dir

$$\text{Ind} = (1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 16)$$

- List of Neumann edges (the third column is the number of the side of Γ where they are:

$$\begin{bmatrix} 14 & 10 & 3 \\ 10 & 5 & 3 \\ 5 & 2 & 4 \\ 2 & 1 & 4 \\ 1 & 4 & 5 \\ 4 & 9 & 5 \end{bmatrix}$$

  As given, all edges are numbered with positive orientation, leaving the exterior domain outside.

- Instead of the list of Dirichlet nodes with their associated boundary side, we could be given a list of Dirichlet edges with the boundary side (third column)

  $$\begin{bmatrix} 9 & 13 & 1 \\ 13 & 17 & 1 \\ 17 & 18 & 1 \\ 18 & 15 & 2 \\ 15 & 14 & 2 \end{bmatrix}$$

  and build therewith both Dir and Ind.

For $\mathbb{P}_2$ elements we also need a list of edges. Unless I made a mistake in counting, we have 39 edges. We then create a list of edges:

$$39 \text{ rows} \left\{ \begin{bmatrix} 1 & 4 \\ 4 & 9 \\ 4 & 8 \\ 4 & 3 \\ 3 & 8 \\ 9 & 8 \\ 9 & 13 \\ 1 & 3 \\ 2 & 1 \\ 2 & 3 \\ \vdots & \vdots \end{bmatrix} \right.$$

Note that boundary edges are positively oriented. We also need a list with the edges organized by element, plus another one with the orientations. Here are the beginning of these lists:

$$23 \text{ rows} \left\{ \begin{bmatrix} 10 & 9 & 8 \\ 4 & 8 & 1 \\ 5 & 4 & 3 \\ \vdots & \vdots & \vdots \end{bmatrix} \right. \qquad 23 \text{ rows} \left\{ \begin{bmatrix} - & + & + \\ + & - & + \\ - & - & + \\ \vdots & \vdots & \vdots \end{bmatrix} \right.$$

Let's have a look at the second element. Its nodes are $[4, 8, 3]$. If we count locally the first edge to be facing the first vertex and continue counterclockwise, then the edges are the fifth one (move from node 8 to node 3), the fourth one (3 to 4), and the third one (4 to 8). However, the edge 5 is numbered 3-to-8 and hence we get a $-$ for the orientation. Similarly, the edge 4 is numbered 4-to-3 and we get another minus sign. For the final edge the element and the edge orientations coincide, and we get a plus sign.

2 The one dimensional problem

Here we deal with the simplest model problem in one dimension:

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

For the homogeneous Dirichlet problem, the variational formulation is given in the space

$$H_0^1(0,1) := \Big\{ u \in L^2(0,1) \,\Big|\, u' \in L^2(0,1), \quad u(0) = u(1) = 0 \Big\},$$

whereas the Neumann problem is set in the full Sobolev space

$$H^1(0,1) := \Big\{ u \in L^2(0,1) \,:\, u' \in L^2(0,1) \Big\}.$$

A very distinct feature of these spaces in the one-dimensional case is the fact that all functions in $H^1(0,1)$ are continuous functions. The variational formulation for the homogeneous Dirichlet problem, that is, demanding $u(0) = u(1) = 0$ is

$$\left[\begin{array}{l} u \in H_0^1(0,1), \\ \displaystyle\int_0^1 u'\,v' + \int_0^1 u\,v = \int_0^1 f\,v, \qquad \forall v \in H_0^1(0,1). \end{array}\right.$$

The homogeneous Neumann problem is obtained by taking $H^1(0,1)$ as space in the formulation above.

Let us choose a positive integer n and

$$h := \frac{1}{n+1}, \qquad 0 = x_0 < x_1 < x_2 < \ldots < x_n < x_{n+1} = 1, \qquad x_j := j\,h.$$

The $\mathbb{P}_1$ finite element space

$$V_h := \Big\{ u_h \in \mathcal{C}[0,1] \,:\, u_h|_{(x_i,x_{i+1})} \in \mathbb{P}_1, \quad \forall i \Big\}$$

has dimension equal to $n+2$. A basis for the space is given by the functions

$$\varphi_i(x) := \begin{cases} \dfrac{x - x_{i-1}}{h}, & x_{i-1} \le x \le x_i, \\ \dfrac{x_{i+1} - x}{h}, & {}_i \le x \le x_{i+1}, \\ 0, & \text{otherwise}, \end{cases}$$

with the obvious modifications for the cases $j = 0$ and $j = n+1$. A basis for $V_h \cap H_0^1(0,1) = \{u_h \in V_h \,|\, u_h(0) = u_h(1) = 0\}$ is given by the functions $\{\varphi_1, \ldots, \varphi_n\}$.

Simple computations show that the mass and stiffness matrices (including all degrees of freedom of the Neumann problem) are respectively

$$\mathbf{M} = \frac{h}{6} \begin{bmatrix} 2 & 1 & & & \\ 1 & 4 & 1 & & \\ & 1 & \ddots & \ddots & \\ & & \ddots & 4 & 1 \\ & & & 1 & 2 \end{bmatrix}, \qquad \mathbf{W} = \frac{1}{h} \begin{bmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & 2 & -1 \\ & & & -1 & 1 \end{bmatrix}.$$

The blocks obtained by dropping the first and last row and columns of both matrices are tridiagonal

$$\mathbf{M}^{\mathrm{Dir}} = \frac{h}{6}\,\mathrm{tridiag}(1,4,1), \qquad \mathbf{W}^{\mathrm{Dir}} = \frac{1}{h}\,\mathrm{tridiag}(-1,2,-1).$$

The eigenvalues for these two matrices are respectively

$$\lambda^k_{M,h} = \frac{h}{3}\Big(1 + 2\sin^2(\pi\,h\,k/2)\Big), \qquad \lambda^k_{W,h} = \frac{4}{h}\sin^2(\pi\,h\,k/2), \qquad k = 1,\ldots,n.$$

It is then clear that the mass matrix is well-conditioned since its eigenvalues can be ordered as

$$\frac{h}{3} < \lambda^1_{M,h} < \ldots < \lambda^n_{M,h} < h.$$

On the other hand

$$\frac{\max_k \lambda^k_{W,h}}{\min_k \lambda^k_{W,h}} = \frac{\sin^2\left(\frac{\pi}{2}\,\frac{n}{n+1}\right)}{\sin^2\left(\frac{\pi}{2}\,\frac{1}{n+1}\right)} \overset{n\to\infty}{\longrightarrow} \infty,$$

which shows that the stiffness matrix tends to be ill-conditioned, especially as we discretize with finer grids. Another computation that can be done explicitly in this simple case is the one of the discrete Dirichlet eigenvalues and eigenvectors. The solutions to the generalized eigenvalue problem

$$\mathbf{W}^{\mathrm{Dir}}\mathbf{u} = \lambda\,\mathbf{M}^{\mathrm{Dir}}\mathbf{u}$$

are the values

$$\lambda^k_h := \frac{6}{h^2}\,\frac{2-2\cos(\pi k h)}{4+2\cos(\pi k h)} \geq k^2\pi^2 = \lambda_k, \qquad k = 1,\ldots,n$$

that overestimate the exact eigenvalues. The eigenvectors are

$$\Big(\sin(\pi\,h\,k), \sin(\pi\,h\,k\,2), \ldots, \sin(\pi\,h\,k\,n)\Big) = \Big(\sin(\pi\,k\,x_j)\Big)_{j=1}^n$$

that are exact nodal values of the corresponding eigenfunctions.

3 Bibliography

The literature on Finite Element Methods is huge. Let me mention here some interesting reference texts, that you should have a look if you try to deepen on either the theory or the practice of the method.

General books The book

> S. C. Brenner and L. R. Scott, *The mathematical theory of finite element methods*, 2nd ed., Springer, 2002

is a modern reference for a full account of the theory (numerical analysis) of finite element methods. The classical reference for mathematicians has been for many years

P.G. Ciarlet, *The Finite Element Method for Elliptic Problems, Series Studies in Mathematics and its Applications*, North-Holland, Amsterdam, 1978 (also SIAM, 2002)

although it is a much more difficult book to read. Another modern account of theoretical (and some practical) aspects of the method, including applications to solid mechanics is

D. Braess, *Finite Elements; Theory, Fast Solvers, and Applications in Solid Mechanics*, Cambridge University Press, 1997 (3rd edition, 2007)

If you want to insist on more practical aspects of the method, applications and implementation, and you are happy with a strong engineering flavor in a numerical book, a good choice is

T.J.R. Hughes, *Finite Element Method - Linear Static and Dynamic Finite Element Analysis*, Prentice-Hall, Englewood Cliffs, 1987 (also Dover, 2000)

Finally, the all-time classical engineering reference for the finite element method is

O.C. Zienkiewicz, R.L. Taylor, *The finite element method. Vol. I. Basic formulations and linear problems.* McGraw-Hill, London, 1989

Particular topics. For a posteriori error estimation, see

Mark Ainsworth and John Tinsley Oden, *A posteriori error estimation in finite element analysis.* Wiley-Interscience, 2000.

or

Rüdiger Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques.* Wiley-Teubner, 1996.

Contents