

Chapter 7

The finite element approximation

”Pure mathematicians sometimes are satisfied with showing that the non-existence of a solution implies a logical contradiction, while engineers might consider a numerical result as the only reasonable goal. Such one sided views seem to reflect human limitations rather than objective values. In itself mathematics is an indivisible organism uniting theoretical contemplation and active application.”

Richard Courant (1888-1972)

The aim of this chapter is to introduce the basic theory of finite element methods. Nowadays, finite element methods are widely used in almost every field of engineering analysis. The German mathematician Richard Courant (1888-1972) shall probably be credited for formulating the essence of what is now called a *finite element* [Courant, 1943]. The development of these methods became effective with the advent of computers and is now recognized as one of the most powerful and versatile method for construction approximations of the solutions of boundary-value problems. We will give here a brief overview of the fundamental mathematical ideas that form the core of the method.

Contents

7.1	General principle	99
7.2	The one dimensional case	105
7.3	Triangular finite elements in higher dimensions	120
7.4	The finite element method for the Stokes problem	132

In Chapter 4 we have seen the principle of the variational approximation of elliptic problems. The main idea of the finite element method is to replace the Hilbert space V in which the variational formulation is posed by a finite dimensional subspace V_h . We will first briefly return to the internal variational approximation principle and present finite elements in one dimension.

7.1 General principle

We consider the variational abstract problem introduced in Chapter 4. More precisely, given a Hilbert space V , a bilinear continuous and V -elliptic form $a(\cdot, \cdot)$ defined on $V \times V$ and a continuous linear form

$l(\cdot)$ defined on V' , we consider the variational formulation of a problem:

$$(\mathcal{P}) : \quad \text{Find } u \in V \text{ such that } a(u, v) = l(v), \quad \text{for all } v \in V,$$

and we already have stated that this problem has a unique solution using Lax-Milgram Theorem 4.1.

7.1.1 Internal approximation of a variational problem

The *internal approximation* of problem (\mathcal{P}) consists then in replacing the Hilbert space V by a finite dimensional subspace, denoted V_h , in which to find the solution u_h . Here, $h > 0$ represents a parameter related to the discretization of the domain (in space or time) and intended to vanish in the theoretical results. We assume that for every $v \in V$, there exists an element $r_h(v) \in V_h$ such that

$$\lim_{h \rightarrow 0} \|r_h(v) - v\| = 0.$$

The bilinear form $a(\cdot, \cdot)$ and the linear form $l(\cdot)$ are defined on $V_h \times V_h$ and V_h , respectively and the problem (\mathcal{P}) is replaced by the following *discrete* problem:

$$(\mathcal{P}_h) : \quad \text{Find } u_h \in V_h \text{ such that } a(u_h, v_h) = l(v_h), \quad \text{for all } v_h \in V_h.$$

We introduce the keyword *internal* related to the approximation because we suppose here that $V_h \subset V$. If $N = \dim V_h$, we consider a basis $(\varphi_i)_{1 \leq i \leq N}$ of V_h . The decomposition of u_h in the basis of V_h , $u_h = \sum_{i=1}^N u_i \varphi_i$, leads to rewrite the problem (\mathcal{P}_h) in the following form:

$$\sum_{j=1}^N u_j a(\varphi_j, \varphi_i) = l(\varphi_i), \quad \forall 1 \leq i \leq N. \quad (7.1)$$

Introducing the *stiffness matrix* $A_h = (a_{ij}) \in \mathbb{R}^{N,N}$ of coefficients $a_{ij} = a(\varphi_j, \varphi_i)$, for all $1 \leq i, j \leq N$, the vector $U_h = (u_i)_{1 \leq i \leq N}$ and the vector $F_h = (f_i)_{1 \leq i \leq N}$ such that $f_i = l(\varphi_i)_{1 \leq i \leq N}$ allow us to conclude that we have the equivalence:

$$u_h \text{ solution of } (\mathcal{P}_h) \iff A_h U_h = F_h.$$

Obviously, the V -ellipticity assumption on the bilinear form a implies the existence and the uniqueness of the solution u_h to the problem (\mathcal{P}_h) . However, this assumption is too strong for our purposes. As we are considering a finite dimensional space V_h , it is sufficient to consider that:

$$(a(v_h, v_h) = 0) \implies (v_h = 0), \quad \forall v_h \in V_h,$$

Since the bilinear form a is V -elliptic, then the matrix A_h is positive definite (cf. Proposition 4.2).

7.1.2 A priori error estimates

It is interesting to evaluate the error related to the replacement of V by the finite dimensional subspace V_h . To this end, we assume that the problems (\mathcal{P}) and (\mathcal{P}_h) are both well-posed, in particular that the bilinear form $a(\cdot, \cdot)$ is continuous (with a continuity constant $M > 0$) and that there exists a constant $\alpha_h > 0$ such that:

$$\forall v_h \in V_h, \quad a(v_h, v_h) \geq \alpha_h \|v_h\|_V^2,$$

and we denote by u and u_h the respective solutions of problems (\mathcal{P}) and (\mathcal{P}_h) .

Proposition 7.1 *Under the previous assumptions, we have the orthogonality identity:*

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

Proof. Since $V_h \subset V$, we have trivially that $a(u, v_h) = l(v_h) = a(u_h, v_h)$, for all $v_h \in V_h$. \square

If the bilinear form $a(\cdot, \cdot)$ is symmetric, continuous and V -elliptic on $V \times V$ of constant α , then it defines an inner product and an *energy norm* associated to it as:

$$\|u\|_e = (a(u, u))^{1/2}, \quad \forall u \in V,$$

equivalent to the norm of V :

$$\sqrt{\alpha}\|u\|_V \leq \|u\|_e \leq \sqrt{\|a\|}\|u\|_V, \quad \forall u \in V.$$

The approximate solution u_h is thus the orthogonal projection for the inner product $a(\cdot, \cdot)$, of the solution u on the subspace V_h .

A strong advantage of the internal variational approximation is that it provides an optimal estimate of the error between the exact solution u of the problem (\mathcal{P}) and the approximate solution u_h of the problem (\mathcal{P}_h) . The error $\|u - u_h\|_V$ is comparable to the minimum of $\|u - v_h\|_V$ when v_h covers V_h . This error estimate in norm $\|\cdot\|_V$ is given by Céa's lemma (cf. Chapter 4) that we recall here.

Lemma 7.1 (Céa) *Under the previous hypothesis, we have the following error estimates.*

(i) *If the bilinear form is not V -elliptic, we have:*

$$\|u - u_h\|_V \leq \left(1 + \frac{\|a\|}{\alpha_h}\right) \inf_{v_h \in V_h} \|u - v_h\|_V, \quad \text{where } \|a\| = \sup_{v_h, w_h \in V_h} \frac{a(v_h, w_h)}{\|v_h\|_V \|w_h\|_V}.$$

(ii) *If the bilinear form $a(\cdot, \cdot)$ is continuous and V -elliptic with a coercivity constant α , then we have:*

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

(iii) *If in addition the bilinear form $a(\cdot, \cdot)$ is symmetric the previous estimate becomes:*

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

Proof. Let consider $w_h \in V_h$. From the orthogonality identity we deduce that

$$a(u_h - w_h, v_h) = a(u - w_h, v_h), \quad \forall v_h \in V_h.$$

and taking into account the continuity of a , we write:

$$\alpha_h \|u_h - w_h\|_V \leq \sup_{v_h \in V_h} \frac{a(u_h - w_h, v_h)}{\|v_h\|_V} = \sup_{v_h \in V_h} \frac{a(u - w_h, v_h)}{\|v_h\|_V} \leq \|a\| \|u - w_h\|_V.$$

Similarly, we have

$$a(u - u_h, u - u_h) = a(u - u_h, u - v_h), \quad \forall v_h \in V_h,$$

and we conclude with the V -ellipticity and the continuity of a . If the form a is symmetric, we can improve the estimate. Since u_h is the orthogonal projection of u onto V_h with respect to the inner product induced by a , we have the Pythagorean relation:

$$a(u - u_h, u - u_h) = \|u - u_h\|_e^2 \leq \|u - w_h\|_e^2 = a(u - w_h, u - w_h), \quad \forall w_h \in V_h,$$

and we deduce that for all $w_h \in V_h$

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

and the results follows. \square

Corollary 7.1 *Under the previous hypothesis, let $(V_h)_h$ denotes a set of finite dimensional subspaces of V and let us assume that*

$$\forall v \in V, \quad \inf_{v_h \in V_h} \|v - v_h\|_V \xrightarrow{h \rightarrow 0} 0.$$

Then, if $\inf_h \alpha_h > 0$, u_h converges toward u in V .

The objective for the “ideal” approximation method is to define suitable approximation spaces V_h to apply the Galerkin approach. To this end, we search for a compromise between the dimension N of V_h (and thus the dimension of the matrix A) and the accuracy of the numerical solution u_h . We shall also consider spaces V_h that allow to compute easily the quantities $a(\varphi_j, \varphi_i)$ and $l(\varphi_i)$. Finally, specific spaces V_h may result in *sparse* matrices A where the number of nonzero elements is small, or *well-conditioned* matrices with a small condition number and thus easy to solve. In this spirit, the finite element method tends to answer all these requirements. Before detailing the main concepts of the finite element methods, we give a few words about Ritz and Petrov-Galerkin methods.

7.1.3 Ritz and Petrov-Galerkin methods

The Ritz method

We consider that the hypothesis of the Lax-Milgram theorem are satisfied and we recall that, if the bilinear form $a(\cdot, \cdot)$ defined on $V \times V$ is symmetric, solving the problem:

$$(\mathcal{P}) \quad \text{find } u \in V, \text{ such that } a(u, v) = l(v), \quad \text{for all } v \in V$$

is indeed equivalent to solving the problem:

$$(\tilde{\mathcal{P}}) \quad \text{find } u \in V, \text{ such that } J(u) = \inf_{v \in V} J(v), \text{ where } J(v) = \frac{1}{2}a(v, v) - l(v).$$

In the Ritz method, the space V is replaced by a finite dimensional subspace $V_h \subset V$ such that $\dim V_h = N$ and the approximate solution u_h shall solve:

$$(\tilde{\mathcal{P}}_h) \quad \text{find } u_h \in V_h \text{ such that } J(u_h) = \inf_{v_h \in V_h} J(v_h).$$

Theorem 4.2 ensures the existence of a unique solution to this minimization problem as a consequence of Lax-Milgram theorem.

Since the dimension of the space V_h is N , there exists a basis $(\varphi_j)_{1 \leq j \leq N}$ of V_h and every $u_h \in V_h$ can be decomposed as $u_h = \sum_{j=1}^N u_j \varphi_j$ and we use the classical notation $U = (u_1, \dots, u_N)^t \in \mathbb{R}^N$.

We consider the one-to-one mapping $\xi : V_h \rightarrow \mathbb{R}^N, u_h \mapsto U$ and we pose $\mathcal{J} = J \circ \xi^{-1}$ such that for every $u_h \in V_h$ we have:

$$\mathcal{J}(U) = J(u_h),$$

or, when replacing $J(u_h)$ by its value:

$$\begin{aligned} J(u_h) &= \frac{1}{2}a\left(\sum_{j=1}^N u_j \varphi_j, \sum_{j=1}^N u_j \varphi_j\right) - l\left(\sum_{j=1}^N u_j \varphi_j\right) \\ &= \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N u_i u_j a(\varphi_i, \varphi_j) - \sum_{i=1}^N u_i l(\varphi_i). \end{aligned}$$

This leads to a matrix formulation of the minimization functional $J(u)$:

$$J(u_h) = \frac{1}{2}U^t A_h U - U^t F_h = \mathcal{J}(U),$$

where $A_h = (a_{ij}) \in \mathbb{R}^{N,N}$ with $a_{ij} = a(\varphi_i, \varphi_j)$ and $F_h = (f_i) \in \mathbb{R}^N$ is such that $f_i = l(\varphi_i)$. Hence, solving the minimization problem $(\widetilde{\mathcal{P}}_h)$ is equivalent to solving the following problem:

$$(\widetilde{\mathcal{P}}_{h,R}) \quad \text{find } U \in \mathbb{R}^N \text{ such that } \mathcal{J}(U) = \inf_{V \in \mathbb{R}^N} \mathcal{J}(V), \quad \text{where } \mathcal{J}(V) = \frac{1}{2} V^t A_h V - V^t F_h.$$

For obvious reasons, the stiffness matrix A_h is symmetric positive definite and thus the functional \mathcal{J} is quadratic on \mathbb{R}^N . This is sufficient to ensure the existence and uniqueness of $U \in \mathbb{R}^N$ solving the minimization problem $(\widetilde{\mathcal{P}}_{h,R})$. Furthermore, the solution U of the minimization problem $(\widetilde{\mathcal{P}}_{h,R})$ is also the solution of the linear system $A_h U = F_h$.

Remark 7.1 *When the bilinear form $a(\cdot, \cdot)$ is symmetric, the Galerkin and Ritz methods are strictly equivalent.*

The Petrov-Galerkin method

The principles of Petrov-Galerkin and Galerkin methods are very similar in the sense that they both will attempt to solve the problem (\mathcal{P}) . However, in the Petrov-Galerkin approach, we consider two finite-dimensional approximation subspaces V_h and W_h in V such that

$$\dim V_h = \dim W_h = N.$$

The approximate solution u_h is searched in the space V_h but the test functions in the variational formulation are now the shape functions of W_h . For these reasons, V_h is called the *approximation space* and W_h is the space of test functions. The problem to solve is now the following:

$$(\mathcal{P}_{h,PG}) \quad \text{find } u_h \in V_h, \quad \text{such that } a(u_h, v_h) = l(v_h), \quad \text{for all } v_h \in W_h.$$

Suppose $(\varphi_j)_{1 \leq j \leq N}$ is a basis of V_h and $(\psi_j)_{1 \leq j \leq N}$ a basis of W_h , then every $u_h \in V_h$ can be decomposed as $u_h = \sum_{j=1}^N u_j \varphi_j$ and we can rewrite the problem as follows:

$$\text{find } u_h \in V_h, \quad \text{such that } \sum_{j=1}^N u_j a(\varphi_j, \psi_i) = l(\psi_i), \quad i = 1, \dots, N,$$

And the linear system to solve is $A_h U = F_h$, where $A_h = (a_{ij}) \in \mathbb{R}^{N,N}$ with $a_{ij} = a(\varphi_j, \psi_i)$ and $F_h = (f_i)$ with $f_i = l(\psi_i)$, for all $i = 1, \dots, N$.

7.1.4 The finite element method

In the finite element method, the domain Ω is subdivided into a partition or a *mesh* T_h , *i.e.*, a (potentially large) collection of geometrically simple elements, and the approximation space V_h is composed of piecewise polynomial functions on each element K of the partition T_h . We will see in Chapter 9 how to construct a partition T_h for a domain Ω of arbitrary geometric shape. The parameter h represents here the grain of the discretization, *i.e.*, the elementary size of the elements K in T_h as defined by:

$$h = \max_{K \in T_h} \text{diam}(K).$$

Typically, a basis of V_h will be composed of functions whose support is restricted on one or a few elements of T_h and the polynomials are usually of low degree. Hence, when $h \rightarrow 0$ the space V_h will better and better approximate the space V and the stiffness matrix A will be sparse, most of its coefficients being zeros.

Finite elements vs. finite differences or finite volumes

The principle of finite difference or finite volume discretization methods is very similar. Both approaches consider a partition of the domain into a numerous small pieces, although none of them consider a variational formulation of the problem at hand. For instance, let consider the homogeneous Dirichlet boundary-value problem in two dimensions:

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

With the finite difference method, the domain Ω is covered by a regular uniform grid. At each internal node $x_{i,j} = (ih, jh)$, we search a discrete value $u_{i,j}$ to approximate $u(x_{i,j})$ and we assume for example that the Laplacian operator is approximated using a 5 points scheme, thus leading to write:

$$\frac{4u_{i,j} - u_{i,j+1} - u_{i,j-1} - u_{i+1,j} - u_{i-1,j}}{h^2} = f(x_{i,j}).$$

In the finite volume method, the partition T_h of the domain Ω is arbitrary and unknowns are associated with each element $K \in T_h$. Using Green's identity, we can rewrite the previous equation as follows:

$$-\int_{\partial K} \frac{\partial u}{\partial n} = \int_K f, \quad \forall K \in T_h$$

and we discretize the left-hand side term using a formula mixing the unknowns on K and on the neighboring elements. Considering a square domain, we have:

$$K_{i,j} = [(i - 1/2)h, (i + 1/2)h] \times [(j - 1/2)h, (j + 1/2)h],$$

and if $u_{i,j}$ is the approximation of u on $K_{i,j}$, the flux integrated on the interface $K_{i,j} \cap K_{i+1,j}$ is then discretized by $u_{i+1,l} - u_{k,l}$. Repeating the procedure for the other fluxes and approximating the term $\int_{K_{i,j}} f$ by $h^2 f(x_{i,j})$ yields the equation:

$$4u_{i,j} - u_{i,j+1} - u_{i,j-1} - u_{i+1,j} - u_{i-1,j} = h^2 f(x_{i,j}).$$

The resulting linear system is very similar to that obtained using a finite difference scheme.

Actually, the vast majority of finite difference methods can be deduced from finite element methods if the problem at hand has a variational formulation. It is less obvious for finite volume methods. Moreover, we have strong theoretical mathematical tools to study finite element methods. In addition, the latter have several intrinsic advantages:

1. the versatility of the formulation on arbitrarily complex geometries, and the possibility to locally refine the partition T_h to approximate solutions with singularities,
2. the boundary conditions are naturally taken into account in the space V in the variational formulation and in its internal approximation V_h ,
3. the general framework of the variational approximations is convenient for the error analysis.

Other variational methods have been developed, like *spectral methods*, that are especially adapted to the approximation of smooth solutions but are limited to simple geometries and methods using *wavelets basis*.

7.2 The one dimensional case

At first, we introduce the general principle of the Lagrange finite element method in one dimension of space. Without loss of generality, we can restrict our study to the unit domain $\Omega =]0, 1[$. To set the ideas, we will also consider the following boundary-value problem:

Given $f \in L^2(\Omega)$ and $c \in L^\infty(\Omega)$, find the function u solving:

$$\begin{cases} -u''(x) + c(x)u(x) = f(x), & x \in]0, 1[\\ u(0) = u(1) = 0. \end{cases} \quad (7.2)$$

Here, a *mesh* is simply a set of points $(x_j)_{0 \leq j \leq N+1}$ or intervals $K_j = [x_j, x_{j+1}]$ such that $0 = x_0 < x_1 < \dots < x_{N+1} = 1$. The mesh is said to be *uniform* if the points (x_j) are equidistributed along the segment $[0, 1]$, i.e. such that $x_j = jh$, with $h = 1/(N+1)$, $0 \leq j \leq N+1$. More generally, we denote by $h = \max |x_{j+1} - x_j|$ the size parameter.

7.2.1 Lagrange \mathbb{P}_1 elements

The finite element methods for Lagrange \mathbb{P}_1 elements involves the space of globally continuous affine functions on each interval:

$$V_h^1 = \{v_h \in C^0([0, 1]), v_h|_{K_j} \in \mathbb{P}_1, 0 \leq j \leq N\},$$

and the subspace of V_h^1 :

$$V_{0,h}^1 = \{v_h \in V_h^1, \text{ such that } v_h(0) = v_h(1) = 0\},$$

More generally, \mathbb{P}_k denotes the vector space of polynomials in one variable and of degree less than or equal to k :

$$\mathbb{P}_k = \left\{ p(x) = \sum_{j=0}^k \alpha_j x^j \mid \alpha_j \in \mathbb{R} \right\}.$$

The finite element method consists in applying the internal variational approximation approach to the spaces V_h^1 and $V_{0,h}^1$. In this context, the functions of V_h^1 can be represented using very simple *shape functions*.

Lemma 7.2 *The space V_h^1 is a subspace of $H^1(\Omega)$ of dimension $N+2$. Every function v_h of V_h^1 is uniquely determined by its values at the mesh vertices $(x_j)_{0 \leq j \leq N+1}$:*

$$v_h(x) = \sum_{j=0}^{N+1} v_h(x_j) \varphi_j(x), \quad \forall x \in [0, 1],$$

where $(\varphi_j)_{0 \leq j \leq N+1}$ is the basis of the shape functions φ_j with compact support in each interval $[x_{j-1}, x_{j+1}]$ defined as:

$$\varphi_j(x) = \begin{cases} \frac{x - x_{j-1}}{h} & x \in [x_{j-1}, x_j] \\ \frac{x_{j+1} - x}{h} & x \in [x_j, x_{j+1}] \end{cases} \quad \text{such that } \varphi_j(x_i) = \delta_{ij}. \quad (7.3)$$

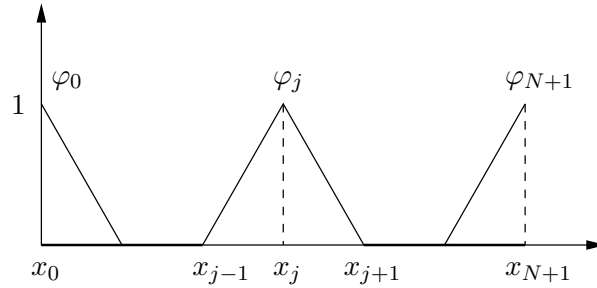


Figure 7.1: Global shape functions for the space V_h^1 .

Proof. We know that piecewise C^1 continuous functions belong to the space $H^1(\Omega)$. Hence, V_h is a subspace of $H^1(\Omega)$. Moreover, since we have $\varphi_j(x_i) = \delta_{ij}$, where δ_{ij} is the Kronecker symbol, the result follows. \square

Remark 7.2 Notice that the functions $(\varphi_j)_{0 \leq j \leq N+1}$ can be expressed using only two functions:

$$\omega_0(x) = 1 - x, \quad \omega_1(x) = x.$$

The basis functions are defined as the composition of a shape function of a reference finite element (i.e. that depends only of the polynomial approximation) and of an affine transformation (depending only on the discretization) as, for all $0 \leq j \leq N$:

$$\varphi_j(x) = \begin{cases} \omega_1\left(\frac{x - x_{j-1}}{x_j - x_{j-1}}\right), & x \in [x_{j-1}, x_j] \\ \omega_0\left(\frac{x - x_j}{x_{j+1} - x_j}\right), & x \in [x_j, x_{j+1}] \end{cases}$$

and $\varphi_{N+1}(x) = \omega_1((x - x_N)/h)$. This will be useful to compute the coefficients of the matrix in the linear system to solve.

Corollary 7.2 The space $V_{0,h}^1$ is a subspace of $H_0^1(\Omega)$ of dimension N and every function v_h of $V_{0,h}^1$ is uniquely determined by its values at the mesh vertices $(x_j)_{1 \leq j \leq N}$:

$$v_h(x) = \sum_{j=1}^N v_h(x_j) \varphi_j(x), \quad \forall x \in [0, 1],$$

Remark 7.3 Notice that functions $v_h \in V_h^1$ are not twice differentiable on Ω and thus it is meaningless to attempt solving the problem (7.2) as the second derivative of any $v_h \in V_h^1$ is a sum of Dirac masses at the mesh vertices. However, it is meaningful to solve a variational formulation of this problem with functions $v_h \in V_h^1$ since only the first derivatives are involved.

The variational formulation of problem (7.2) consists in finding $u \in H_0^1(\Omega)$, such that:

$$\int_{\Omega} u'(x)v'(x) dx + \int_{\Omega} c(x)u(x)v(x) dx = \int_{\Omega} f(x)v(x) dx, \quad \forall v \in H_0^1(\Omega), \quad (7.4)$$

and the variational formulation of the internal approximation consists in finding $u_h \in V_{0,h}$, such that:

$$\int_{\Omega} u_h'(x)v_h'(x) dx + \int_{\Omega} c(x)u_h(x)v_h(x) dx = \int_{\Omega} f(x)v_h(x) dx, \quad \forall v_h \in V_{0,h}. \quad (7.5)$$

Introducing the notation $u_h(x_j)_{1 \leq j \leq N}$ for the approximate value of the exact solution at the mesh vertex x_j , leads to the approximate problem:

find $u_h(x_1), \dots, u_h(x_N)$ such that for all $i = 1, \dots, N$

$$\sum_{j=1}^N \left(\int_{\Omega} \varphi'_j(x) \varphi'_i(x) dx + \int_{\Omega} c(x) \varphi_j(x) \varphi_i(x) dx \right) u_h(x_j) = \int_{\Omega} f(x) \varphi_i(x) dx .$$

And as expected, this formulation is equivalent to solving in \mathbb{R}^N the linear system:

$$A_h U_h = F_h$$

where $U_h = (u_h(x_j))_{1 \leq j \leq N}$, $F_h = (\int_{\Omega} f(x) \varphi_i(x) dx)_{1 \leq i \leq N}$ and the matrix A_h is defined as:

$$A_h = \left(\int_{\Omega} \varphi'_j(x) \varphi'_i(x) dx + \int_{\Omega} c(x) \varphi_j(x) \varphi_i(x) dx \right)_{1 \leq i, j \leq N} .$$

Coefficients of the matrix A_h

Actually, the matrix A_h appears as the sum of the *stiffness matrix* K_h defined by its coefficients $(k_{ij})_{1 \leq i, j \leq N}$:

$$k_{ij} = \int_{\Omega} \varphi'_j(x) \varphi'_i(x) dx = \sum_{k=0}^N \int_{x_k}^{x_{k+1}} \varphi'_j(x) \varphi'_i(x) dx ,$$

and the *mass matrix* M_h defined by its coefficients $(m_{ij})_{1 \leq i, j \leq N}$:

$$m_{ij} = \int_{\Omega} c(x) \varphi_j(x) \varphi_i(x) dx = \sum_{k=0}^N \int_{x_k}^{x_{k+1}} c(x) \varphi_j(x) \varphi_i(x) dx .$$

Since the shape functions φ_j have a small support, most of the coefficients in A_h are zeros. More precisely, for a given index i , there is only three consecutive values of j such that the coefficient a_{ij} is potentially not equal to zero. The structure of the matrix is then easy to deduce: A_h is a tridiagonal matrix. The coefficients of A_h are thus given by:

$$\begin{aligned} a_{jj} &= \int_{x_{j-1}}^{x_{j+1}} (\varphi'_j(x))^2 dx + \int_{x_{j-1}}^{x_{j+1}} c(x) (\varphi_j(x))^2 dx \\ a_{jj-1} &= \int_{x_{j-1}}^{x_j} \varphi'_j(x) \varphi'_{j-1}(x) dx + \int_{x_{j-1}}^{x_j} c(x) \varphi_j(x) \varphi_{j-1}(x) dx \\ a_{jj+1} &= \int_{x_j}^{x_{j+1}} \varphi'_j(x) \varphi'_{j+1}(x) dx + \int_{x_j}^{x_{j+1}} c(x) \varphi_j(x) \varphi_{j+1}(x) dx \end{aligned}$$

For the sake of simplicity, we consider here the function c as being constant, $c(x) = c_0$ for all $x \in \Omega$. Hence, we write:

$$\begin{aligned} m_{jj} &= c_0 \int_{x_{j-1}}^{x_j} \varphi_j(x) \varphi_{j-1}(x) dx = c_0 \int_{x_{j-1}}^{x_j} \omega_1 \left(\frac{x - x_{j-1}}{h} \right) \omega_0 \left(\frac{x - x_{j-1}}{h} \right) \\ &= c_0 h \int_0^1 \omega_1(y) \omega_0(y) dy = c_0 h \int_0^1 (1-y)y dy = c_0 \frac{h}{6} , \end{aligned}$$

where we posed $y = (x - x_{j-1})/h$. Finally, we find the coefficients of A_h :

$$a_{jj-1} = -\frac{1}{h} + c_0 \frac{h}{6} \quad a_{jj} = \frac{2}{h} + c_0 \frac{2h}{3} \quad \text{and} \quad a_{jj+1} = -\frac{1}{h} + c_0 \frac{h}{6} ,$$

Remark 7.4 Instead of regarding the node contributions, we could have analyzed the elements. Consider the element $K_j = [x_j, x_{j+1}]$; on this element there is only two non-zero shape functions:

$$\begin{aligned}\varphi_j|_{K_j} &= \frac{x_{j+1} - x}{x_{j+1} - x_j} = \frac{x_{j+1} - x}{h} & \varphi_{j+1}|_{K_j} &= \frac{x - x_j}{x_{j+1} - x_j} = \frac{x - x_j}{h} \\ \varphi'_j|_{K_j} &= \frac{-1}{x_{j+1} - x_j} = \frac{-1}{h} & \varphi'_{j+1}|_{K_j} &= \frac{1}{x_{j+1} - x_j} = \frac{1}{h}\end{aligned}$$

Then, we can arrange the elementary contributions of the element K_j to the stiffness matrix and to the mass matrix as 2×2 symmetric matrices EK_j and EM_j :

$$EK_j = \begin{pmatrix} k_{11}^j & k_{12}^j \\ k_{21}^j & k_{22}^j \end{pmatrix} \quad \text{and} \quad EM_j = \begin{pmatrix} m_{11}^j & m_{12}^j \\ m_{21}^j & m_{22}^j \end{pmatrix}$$

with

$$\begin{aligned}k_{11}^j &= \int_{x_j}^{x_{j+1}} (\varphi'_j(x))^2 dx, & k_{12}^j &= k_{21}^j = \int_{x_j}^{x_{j+1}} \varphi'_j(x) \varphi'_{j+1}(x) dx, & k_{22}^j &= \int_{x_j}^{x_{j+1}} (\varphi'_{j+1}(x))^2 dx \\ &= \int_{x_j}^{x_{j+1}} \frac{1}{h^2} dx = \frac{1}{h} & &= \int_{x_j}^{x_{j+1}} -\frac{1}{h^2} dx = -\frac{1}{h} & &= \int_{x_j}^{x_{j+1}} \frac{1}{h^2} dx = \frac{1}{h}\end{aligned}$$

$$m_{11}^j = \int_{x_j}^{x_{j+1}} c(x) (\varphi_j(x))^2 dx, \quad m_{12}^j = m_{21}^j = \int_{x_j}^{x_{j+1}} c(x) \varphi_j(x) \varphi_{j+1}(x) dx, \quad m_{22}^j = \int_{x_j}^{x_{j+1}} c(x) (\varphi_{j+1}(x))^2 dx$$

and thus to conclude that:

$$EK_j = \frac{1}{h} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \quad \text{and} \quad EM_j = c_0 \frac{h}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

We will see that this point of view is more practical when dealing with the matrix assembly, especially in higher dimensions.

Matrix assembly

The assembly of the matrix A_h is easy and is obtained algorithmically using a loop over all mesh elements K_j and adding their contributions to the right coefficients of the global system. Assuming $a(i, j)$ denotes the coefficients a_{ij} of A_h , a pseudo-code to perform this task would be:

```
for k=1, N+1 do // loop over all elements
  for i=1,2 do // local loop
    for j=1,2 do
      ig = k+i-2 // global indices
      jg = k+j-2

      A(ig,jg) = A(ig,jg) + a(i,j)
    end loop j
  end loop i
end loop k
```

The numerical resolution of the linear system is by far the most computationally expensive part of the method. We refer the reader to Chapter 5 for more details about the direct and indirect techniques to solve this system.

Coefficients of the right-hand side F_h

Each component f_i of the vector $F_h \in \mathbb{R}^N$ is obtained as:

$$f_i = \sum_{k=0}^N \int_{x_k}^{x_{k+1}} f(x) \varphi_i(x) dx.$$

Usually, the function f is not known analytically. Hence, we decompose f in the basis of the shape functions $(\varphi_j)_{1 \leq j \leq N}$:

$$f(x) = \sum_{j=1}^{N-1} f_j \varphi_j(x) dx$$

and the problem is reduced to the evaluation of the integrals:

$$\int_{x_k}^{x_{k+1}} \varphi_j(x) \varphi_i(x) dx.$$

We use for instance the trapeze formula:

$$\int_{x_k}^{x_{k+1}} \theta(x) dx = \frac{x_{k+1} - x_k}{2} (\theta(x_{k+1}) + \theta(x_k)),$$

that gives the exact result for polynomial of degree one and leads here $f_j = hf(x_j)$; or the Simpson formula:

$$\int_{x_k}^{x_{k+1}} \theta(x) dx = \frac{x_{k+1} - x_k}{6} (\theta(x_{k+1}) + 4\theta(x_{k+1/2}) + \theta(x_k)),$$

that gives the exact result for polynomials of degree lesser than or equal to 3, and leads here to

$$f_j = \frac{h}{6} (f(x_j) + 4f(x_{j+1/2}) + f(x_{j+1})).$$

Neumann boundary-value problem

The finite element method can be applied to solve the Neumann boundary-value problem:

Given $f \in L^2(\Omega)$, $c \in L^\infty(\Omega)$ such that $c(x) \geq c_0 > 0$ almost everywhere in Ω and $\alpha, \beta \in \mathbb{R}$, find the function u solving:

$$\begin{cases} -u''(x) + c(x)u(x) = f(x), & x \in \Omega \\ u'(0) = \alpha & u'(1) = \beta. \end{cases} \quad (7.6)$$

in a very similar manner. Recall that this problem has a unique solution $u \in H^1(\Omega)$ (cf. Chapter 4). The variational formulation of the internal approximation consists here in finding $u_h \in V_h^1$ such that

$$\int_{\Omega} u'_h(x) v'_h(x) dx + \int_{\Omega} c(x) u_h(x) v_h(x) dx = \int_{\Omega} f(x) v_h(x) dx - \alpha v_h(0) + \beta v_h(1), \quad \forall v_h \in V_h^1(\Omega). \quad (7.7)$$

The variational formulation consists in solving in \mathbb{R}^{N+2} the linear system:

$$A_h U_h = F_h$$

with $U_h = (u_h(x_j))_{0 \leq j \leq N+1}$ and the stiffness matrix A_h is defined as:

$$A_h = \left(\int_{\Omega} \varphi'_j(x) \varphi'_i(x) dx + \int_{\Omega} c(x) \varphi_j(x) \varphi_i(x) dx \right)_{0 \leq i, j \leq N+1}.$$

and $F_h = (f_j)_{0 \leq j \leq N+1}$ such that:

$$\begin{aligned} f_j &= \int_{\Omega} f(x) \varphi_j dx \quad 1 \leq j \leq N \\ f_0 &= \int_{\Omega} f(x) \varphi_0(x) dx - \alpha \\ f_{N+1} &= \int_{\Omega} f(x) \varphi_{N+1}(x) dx + \beta. \end{aligned}$$

7.2.2 Convergence of the Lagrange \mathbb{P}_1 finite element method

Definition 7.1 (Interpolation) *The linear mapping $\Pi_h : H^1(\Omega) \rightarrow V_h^1$ defined for every $v \in H^1(\Omega)$ as:*

$$(\Pi_h v)(x) = \sum_{j=0}^{N+1} v(x_j) \varphi_j(x), \quad \forall x \in [0, 1],$$

is called \mathbb{P}_1 interpolation operator. Furthermore, for every $v \in H^1(\Omega)$, the interpolation operator is such that:

$$\lim_{h \rightarrow 0} \|v - \Pi_h v\|_{H^1(\Omega)} = 0.$$

The \mathbb{P}_1 interpolate of a function v is the unique piecewise affine function that coincide with v at the mesh vertices x_j . The convergence of the finite element method is related to a series of results that we give here.

Suppose that the function v is sufficiently smooth, i.e. $v \in H^2(\Omega)$. Since the derivative of the affine functions in V_h is constant on the intervals $K_j = [x_j, x_{j+1}]$, we have then:

$$(\Pi_h v)'(x) = \frac{v(x_{j+1}) - v(x_j)}{h} = \frac{1}{h} \int_{x_j}^{x_{j+1}} v'(t) dt, \quad \forall x \in [x_j, x_{j+1}].$$

Since we assumed $v \in H^2(\Omega)$ then $v' \in H^1(\Omega)$ and thus v is a continuous function. Using Rolle's theorem, we deduce that there exists a point $\theta_j \in [x_j, x_{j+1}]$ such that:

$$v'(\theta_j) = \frac{1}{h} \int_{x_j}^{x_{j+1}} v'(t) dt = (\Pi_h v)'(x), \quad \forall x \in [x_j, x_{j+1}].$$

We will search for an estimate on $\|v - \Pi_h v\|_{H^1(\Omega)}$. To this end, we write:

$$\|v - \Pi_h v\|_{H^1(\Omega)}^2 = \int_0^1 |v' - \Pi_h' v|^2 = \sum_{j=1}^{N-1} \int_{x_j}^{x_{j+1}} |v'(t) - v'(\theta_j)|^2 dt, \quad (7.8)$$

however, for all $t \in [x_j, x_{j+1}]$ we have:

$$v'(t) - v'(\theta_j) = \int_{\theta_j}^t v''(t) dt,$$

hence, using Cauchy-Schwarz's identity, we write:

$$\begin{aligned} |v'(t) - v'(\theta_j)|^2 &\leq \int_{\theta_j}^t |v''(t)|^2 dt |t - \theta_j| \\ &\leq \int_{x_j}^{x_{j+1}} |v''(t)|^2 dt |t - \theta_j|. \end{aligned}$$

By integrating on the interval $[x_j, x_{j+1}]$ yields:

$$\begin{aligned} \int_{x_j}^{x_{j+1}} |v'(t) - v'(\theta_j)|^2 dt &\leq \int_{x_j}^{x_{j+1}} |t - \theta_j| dt \left(\int_{x_j}^{x_{j+1}} |v''(t)|^2 dt \right) \\ &\leq \frac{(x_{j+1} - x_j)^2}{2} \int_{x_j}^{x_{j+1}} |v''(t)|^2 dt \\ &\leq \frac{h^2}{2} \int_{x_j}^{x_{j+1}} |v''(t)|^2 dt. \end{aligned}$$

And going back to equation (7.8), we have now:

$$\begin{aligned} \|v - \Pi_h v\|_{H^1(\Omega)} &\leq \frac{h^2}{2} \sum_{j=1}^{N-1} \int_{x_j}^{x_{j+1}} |v''(t)| dt \\ &\leq \frac{h^2}{2} \|v''\|_{L^2(\Omega)}. \end{aligned}$$

This leads to enounce the following result, that we already partly proved.

Lemma 7.3 (Interpolation error) *If $v \in H^2(\Omega)$ then, there exists two constant C_1 and C_2 independent of h such that:*

$$\|v - \Pi_h v\|_{H^1(\Omega)} \leq C_1 h^2 \|v''\|_{L^2(\Omega)} \quad \text{and} \quad \|v' - (\Pi_h v)'\|_{L^2(\Omega)} \leq C_2 h \|v''\|_{L^2(\Omega)}.$$

And we can establish the convergence of the finite element method for the Dirichlet boundary-value problem as follows.

Theorem 7.1 (Convergence) *Suppose $u \in H_0^1(\Omega)$ and $u_h \in V_{0,h}$ are the solutions of (7.2) and (7.5), respectively. Then, the Lagrange \mathbb{P}_1 finite element method converges, i.e. we have:*

$$\lim_{h \rightarrow 0} \|u - u_h\|_{H^1(\Omega)} = 0.$$

Furthermore, if $u \in H^2(\Omega)$ then, there exists a constant C independent of h such that:

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

Proof. Here, since the bilinear form $a(\cdot, \cdot)$ is $V_{0,h}$ -elliptic, we can consider the ellipticity constant $\alpha = 1$:

$$a(u, u) = \int_0^1 u'^2(x) + cu^2(x) dx \geq \int_0^1 u'^2(x) dx = \|u\|_{H_0^1(\Omega)}^2,$$

and regarding the continuity of $a(\cdot, \cdot)$ we write:

$$\begin{aligned} |a(u, v)| &\leq \left(\|u'\|_{L^2(\Omega)}^2 + c_1 \|u\|_{L^2(\Omega)}^2 \right)^{1/2} \\ &\leq (1 + c_1) \|u\|_{H_0^1(\Omega)} \|v\|_{H_0^1(\Omega)}, \end{aligned}$$

where we assumed that $0 \leq c_1 = \sup_{x \in [0,1]} c(x) \leq +\infty$. Thus, the continuity constant M is taken here as $1 + c_1$. Using Céa's lemma 7.1, we can easily conclude that:

$$\|u - u_h\|_{H_0^1(\Omega)} \leq \sqrt{1 + c_1} \inf_{v_h \in V_{0,h}} \|u - v_h\|_{H_0^1(\Omega)}. \quad (7.9)$$

If we assume $v \in H^2(\Omega)$, we can write, according to the previous lemma:

$$\begin{aligned} \|u - \Pi_h u\|_{H_0^1(\Omega)}^2 &\leq \frac{h^2}{2} \|u''\|_{L^2(\Omega)}^2 \\ &\leq \frac{h^2}{2} \|u\|_{H^2}^2 \leq C \frac{h^2}{2} \|f\|_{L^2(\Omega)}^2. \end{aligned}$$

Moreover, since $\Pi_h u \in V_{0,h}$, we have also:

$$\inf_{v_h \in V_{0,h}} \|u - v_h\|_{H_0^1(\Omega)} \leq \|u - \Pi_h u\|_{H_0^1(\Omega)}.$$

and using the inequality (7.9), we can conclude. \square

Lemma 7.4 *There exists a constant C independent of h such that for all $v \in H^1(\Omega)$:*

$$\|\Pi_h v\|_{H^1(\Omega)} \leq C \|v\|_{H^1(\Omega)}, \quad \text{and} \quad \|v - \Pi_h v\|_{L^2(\Omega)} \leq Ch \|v'\|_{L^2(\Omega)}.$$

Furthermore, for all $v \in H^1(\Omega)$, we have:

$$\lim_{h \rightarrow 0} \|v' - (\Pi_h v)'\|_{L^2(\Omega)} = 0.$$

Proof. Given $v \in H^1(\Omega)$, we have:

$$\|\Pi_h v\|_{L^2(\Omega)} \leq \sup_{x \in \Omega} |\Pi_h v(x)| \leq \sup_{x \in \Omega} |v(x)| \leq C \|v\|_{H^1(\Omega)}.$$

Moreover, since $\Pi_h v$ is an affine function, we have by Cauchy-Schwarz's identity:

$$\int_{x_j}^{x_{j+1}} |(\Pi_h v)'(t)|^2 dt = \frac{(v(x_{j+1}) - v(x_j))^2}{h} = \frac{1}{h} \left(\int_{x_j}^{x_{j+1}} v'(t) dt \right)^2 \leq \int_{x_j}^{x_{j+1}} |v'(t)|^2 dt.$$

and we obtain the first identity by summation over j . Similarly, we write:

$$|v(x) - \Pi_h v(x)| \leq 2 \int_{x_j}^{x_{j+1}} |v'(t)| dt.$$

We obtain the second identity by using Cauchy-Schwarz, by integrating with respect to x and then by summation over j .

Since $C^\infty(\Omega)$ is dense in $H^1(\Omega)$, for every $v \in H^1(\Omega)$ there exists $w \in C^\infty(\Omega)$ such that

$$\|v' - w'\|_{L^2(\Omega)} \leq \varepsilon, \quad \text{for } \varepsilon > 0.$$

Since Π_h is a linear mapping verifying the first identity, we have then:

$$\|(Pi_h v)' - (\Pi_h w)'\|_{L^2(\Omega)} \leq C \|v' - w'\|_{L^2(\Omega)} \leq C\varepsilon.$$

From Lemma 7.3, we deduce that, for h sufficiently small:

$$\|w' - (\Pi_h w)'\|_{L^2(\Omega)} \leq \varepsilon.$$

We can write, by adding the last identities:

$$\|v' - (\Pi_h v)'\|_{L^2(\Omega)} \leq \|v' - w'\|_{L^2(\Omega)} + \|w' - (\Pi_h w)'\|_{L^2(\Omega)} + \|(\Pi_h v)' - (\Pi_h w)'\|_{L^2(\Omega)} \leq C\varepsilon,$$

and the result follows. \square

7.2.3 Lagrange \mathbb{P}_2 elements

Before introducing the Lagrange \mathbb{P}_2 finite element method, we like to describe the advantage of considering higher-order polynomials on an example taken from [Šolin, 2005].

Motivation for high-order elements

Consider the simple homogeneous Poisson boundary-value problem in one dimension of space:

$$\begin{cases} -u''(x) = f(x), & \text{in } \Omega =]-1, 1[, \\ u(0) = u(1) = 0. \end{cases}, \quad \text{with } f(x) = \frac{\pi^2}{4} \cos\left(\frac{\pi x}{2}\right).$$

The exact solution to this problem has the form:

$$u(x) = \cos\left(\frac{\pi x}{2}\right).$$

The well-known variational formulation of this problem consists in: *finding* $u \in H_0^1(\Omega)$ such that:

$$\int_{\Omega} u'(x)v'(x) dx = \int_{\Omega} f(x)v(x) dx, \quad \forall v \in H_0^1(\Omega)$$

Suppose the domain is decomposed into two intervals $[-1, 0]$ and $[0, 1]$ and let consider the finite element space $V_{0,h}^1$ generated by a single piecewise affine function v_h defined as:

$$v_h(x) = \begin{cases} x + 1, & x \in [-1, 0] \\ 1 - x, & x \in [0, 1] \end{cases}$$

The exact solution u and the approximate solution u_h are given Figure 7.2, left. The approximation error in H^1 seminorm is:

$$|u - u_h|_{1,2} = \left(\int_{\Omega} |u'(x) - u_h'(x)|^2 dx \right)^{1/2} \approx 0.683667.$$

On the other hand, assume a single quadratic element covers the domain $[-1, 1]$. A basis of the finite element space $V_{0,h}^2$ is composed of the function $v_h(x) = 1 - x^2$. The exact solution u and the approximate solution u_h are given Figure 7.2, right. The approximation error is then:

$$|u - u_h(x)|_{1,2} \approx 0.20275.$$

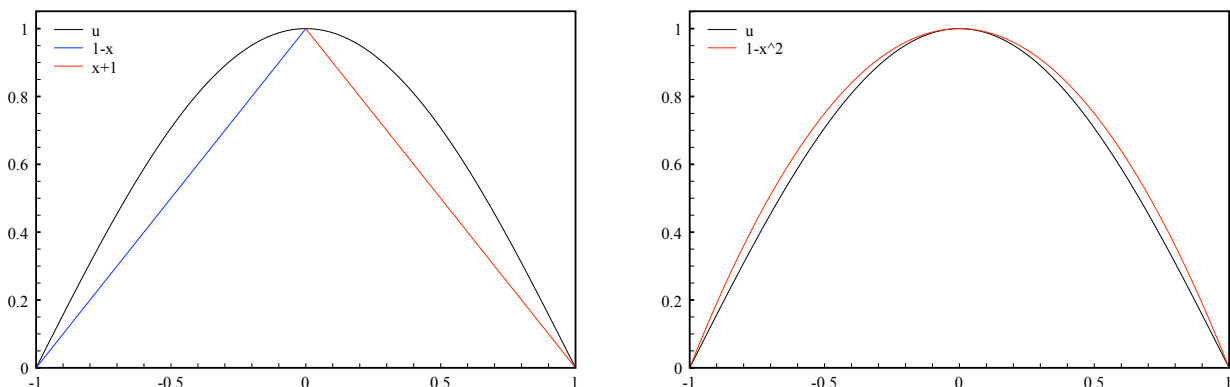


Figure 7.2: .Exact solution u with piecewise affine approximation u_h (left-hand side) and with quadratic approximation u_h (right-hand side).

This is a clear indication that high-order finite elements are better to approximate smooth functions. Conversely, less regular functions can be approximated accurately using lower degree finite elements. This will be emphasized by Theorem 7.2.

Lagrange \mathbb{P}_2 elements

We return to problem (7.2) and we consider a set of points $(x_j)_{0 \leq j \leq N+1}$ or intervals $K_j = [x_j, x_{j+1}]$ forming a uniform mesh of Ω . The finite element method for Lagrange \mathbb{P}_2 elements involves the discrete space:

$$V_h^2 = \{v_h \in C^0([0, 1]), v_h|_{K_j} \in \mathbb{P}_2, 0 \leq j \leq N\},$$

and its subspace:

$$V_{0,h}^2 = \{v_h \in V_h^2, \text{ such that } v_h(0) = v_h(1) = 0\}.$$

These spaces are composed of continuous, piecewise parabolic functions (polynomials of degree less than or equal to 2). The \mathbb{P}_2 finite element method consists in applying the internal variational approximation approach to these spaces.

Lemma 7.5 *The space V_h^2 is a subspace of $H^1(\Omega)$ of dimension $2N + 3$. Every function $v_h \in V_h^2$ is uniquely defined by its values at the mesh vertices $(x_j)_{0 \leq j \leq N+1}$ and at the midpoints $(x_{j+1/2})_{0 \leq j \leq N} = (x_j + \frac{h}{2})_{0 \leq j \leq N}$:*

$$v_h(x) = \sum_{j=0}^{N+1} v_h(x_j) \varphi_j(x) + \sum_{j=0}^N v_h(x_{j+1/2}) \varphi_{j+1/2}(x), \quad \forall x \in [0, 1].$$

where $(\varphi_j)_{0 \leq j \leq N+1}$ is the basis of the shape functions φ_j defined as:

$$\varphi_j(x) = \phi\left(\frac{x - x_j}{h}\right), 0 \leq j \leq N + 1 \quad \text{and} \quad \varphi_{j+1/2}(x) = \psi\left(\frac{x - x_{j+1/2}}{h}\right), 0 \leq j \leq N,$$

with

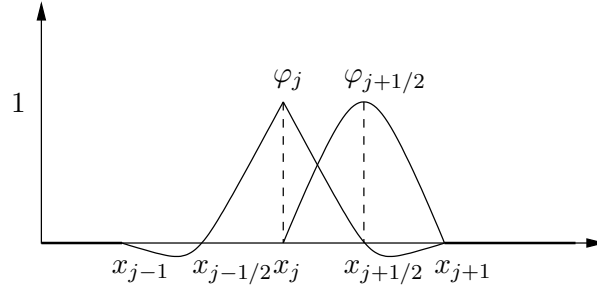
$$\phi(x) = \begin{cases} (1+x)(1+2x) & -1 \leq x \leq 0 \\ (1-x)(1-2x) & 0 \leq x \leq 1 \\ 0 & |x| > 1, \end{cases} \quad \text{and} \quad \psi(x) = \begin{cases} 1 - 4x^2 & |x| \leq 1/2 \\ 0 & |x| > 1/2 \end{cases}$$

Remark 7.5 *Notice that we have:*

$$\begin{aligned} \varphi_i(x_j) &= \delta_{ij} & \varphi_i(x_{j+1/2}) &= 0 \\ \varphi_{i+1/2}(x_j) &= 0 & \varphi_{i+1/2}(x_{j+1/2}) &= \delta_{ij} \end{aligned}$$

Corollary 7.3 *The space $V_{0,h}^2$ is a subspace of $H_0^1(\Omega)$ of dimension $2N + 1$ and every function $v_h \in V_{0,h}^2$ is uniquely defined by its values at the mesh vertices $(x_j)_{1 \leq j \leq N}$ and at the midpoints $(x_{j+1/2})_{0 \leq j \leq N}$:*

$$v_h(x) = \sum_{j=1}^N v_h(x_j) \varphi_j(x) + \sum_{j=0}^N v_h(x_{j+1/2}) \varphi_{j+1/2}(x), \quad \forall x \in [0, 1].$$

Figure 7.3: Global shape functions for the space V_h^2 .

The variational formulation of the internal approximation of the Dirichlet boundary-value problem (7.2) consists now in finding $u_h \in V_{0,h}^2$, such that:

$$\int_{\Omega} u_h'(x)v_h'(x) dx + \int_{\Omega} c(x)u_h(x)v_h(x) dx = \int_{\Omega} f(x)v_h(x) dx, \quad \forall v_h \in V_{0,h}^2. \quad (7.10)$$

Here, it is convenient to introduce the notation $(x_{k/2})_{1 \leq k \leq 2N+1}$ for the mesh points and $(\varphi_{k/2})_{1 \leq k \leq 2N+1}$ for the basis of $V_{0,h}^2$. Using these notations, we have:

$$u_h(x) = \sum_{k=1}^{2N+1} u_h(x_{k/2})\varphi_{k/2}(x).$$

This formulation leads to solve in \mathbb{R}^{2N+1} a linear system:

$$A_h U_h = F_h,$$

where $U_h = (u_h(x_{k/2}))_{1 \leq k \leq 2N+1}$ and it is easy to see that the matrix A_h of the linear system to solve is now defined as:

$$A_h = \left(\int_{\Omega} \varphi_{k/2}'(x)\varphi_{l/2}'(x) dx + \int_{\Omega} c(x)\varphi_{k/2}(x)\varphi_{l/2}(x) dx \right)_{1 \leq k, l \leq 2N+1},$$

and the right-hand side term becomes:

$$F_h = \left(\int_{\Omega} f(x)\varphi_{k/2}(x) dx \right)_{1 \leq k \leq 2N+1}.$$

Since the shape functions φ_j have a small support, the matrix A_h is mostly composed of zeros. However, the main difference with the Lagrange \mathbb{P}_1 finite element method, the matrix A_h is no longer a tridiagonal matrix.

Coefficients of A_h

The coefficients of the matrix A_h can be computed more easily by considering the following change of variables, for $t \in [-1, 1]$:

$$x = x_{j+1} + \frac{x_{j+2} - x_j}{2}t = x_{j+1} + \frac{h}{2}t, \quad \forall x \in [x_j, x_{j+2}], \quad 0 \leq j \leq 2N-1.$$

Hence, the shape functions can be reduced to only three basic shape functions (Figure 7.4):

$$\omega_{-1}(t) = \frac{t(t-1)}{2} \quad \omega_0(t) = -(t-1)(t+1) \quad \omega_1(t) = \frac{t(t+1)}{2},$$

and their respective derivatives:

$$\frac{d\omega_{-1}}{dt}(t) = \frac{2t-1}{2} \quad \frac{d\omega_0}{dt}(t) = -2t \quad \frac{d\omega_1}{dt}(t) = \frac{2t+1}{2}.$$

This approach consists in considering all computations on an interval $K_j = [x_j, x_{j+2}]$ on the *reference interval* $[-1, 1]$. Thus, we have:

$$\frac{d\varphi_j}{dx} = \frac{d\omega_k}{dt} \frac{dt}{dx},$$

where $\omega_k \in [-1, 1]$. In this case, the elementary contributions of the element K_j to the stiffness matrix and to the mass matrix are given by the 3×3 matrices EK_j and EM_j :

$$EK_j = \frac{1}{3h} \begin{pmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{pmatrix} \quad EM_j = c_0 \frac{h}{30} \begin{pmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4 \end{pmatrix}.$$

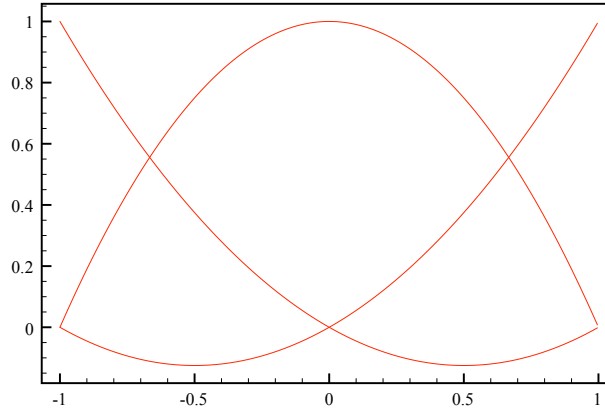


Figure 7.4: The three quadratic Lagrange \mathbb{P}_2 shape functions on the reference interval $[-1, 1]$.

Matrix assembly

```

for k=1, N do // loop over all elements
  for i=1,3 do // local loop
    for j=1,3 do
      ig = 2*k+i-3 // global indices
      jg = 2*k+j-3

      A(ig,jg) = A(ig,jg) + a(i,j)
    end loop j
  end loop i
end loop k

```

Coefficients of the right-hand side F_h

Usually, the function f is only known by its values at the mesh points $(x_j)_{0 \leq j \leq 2N}$ and thus, we use the decomposition of f in the basis of shape functions $(\varphi_j)_{0 \leq j \leq 2N}$:

$$f(x) = \sum_{j=0}^{2N} f(x_j) \varphi_j(x) dx.$$

Each component f_i of the right-hand side vector is obtained as:

$$f_i = \sum_{k=1}^N \int_{x_{2k-2}}^{x_{2k}} f(x) \varphi_i(x) dx.$$

Using the previous decomposition of f , we obtain:

$$f_i = \sum_{j=0}^{2N} f_j \left(\sum_{k=1}^N \int_{x_{2k-2}}^{x_{2k}} \varphi_j(x) \varphi_i(x) dx \right),$$

and the problem is reduced to computing the integrals:

$$\int_{x_{2k-2}}^{x_{2k}} \varphi_j(x) \varphi_i(x) dx,$$

It is easy to see that we obtain expressions very similar to that of the mass matrix. More precisely, the element $K_j = [x_i, x_{i+2}]$ will contribute to only three components of indices i , $i+1$ and $i+2$ as:

$$\begin{pmatrix} f_i^k \\ f_{i+1}^k \\ f_{i+2}^k \end{pmatrix} = \frac{h}{30} \begin{pmatrix} 4 & 2 & -1 \\ 2 & 16 & 2 \\ -1 & 2 & 4 \end{pmatrix} \begin{pmatrix} f_i \\ f_{i+1} \\ f_{i+2} \end{pmatrix},$$

where f_i^k denotes the contribution of element k to the component i .

7.2.4 Convergence of the Lagrange \mathbb{P}_2 finite element method

We rely on Céa's lemma that provides an estimate of the error:

$$\|u - u_h\|_{H^1(\Omega)} \leq \sqrt{\frac{M}{\alpha}} \inf_{v_h \in V_{0,h}^1} \|u - v_h\|_{H^1(\Omega)},$$

for a continuous and V -elliptic bilinear form $a(\cdot, \cdot)$ defined on $V_h^1 \times V_h^1$. Suppose now that $f \in H^1(\Omega)$, then $u \in H^3(\Omega)$ and if the function c is sufficiently smooth then:

$$\|u\|_{H^3(\Omega)} \leq C \|f\|_{H^1(\Omega)}.$$

In order to find an upper bound on the right-hand side of the previous estimate, we introduce a mapping $w_h \in V_{0,h}^2$ such that for all $1 \leq i \leq N$, $w_h(x_i) = u(x_i)$ and such that $w_h|_{[x_i, x_{i+1}]}$ is a polynomial of degree two or less. To this end, on each interval $[x_i, x_{i+1}]$, we consider the polynomial functions:

$$w_{i,1}(x) = \frac{\alpha_i}{2}(x - x_i)^2, \quad w_{i,2}(x) = w_{i,1}(x) + \beta_i(x - x_i),$$

where

$$\alpha_i = \frac{1}{h} \int_{x_i}^{x_{i+1}} u''(t) dt, \quad \beta_i = \frac{1}{h} \int_{x_i}^{x_{i+1}} (u - w_{i,1})'(t) dt$$

By definition, we have:

$$\int_{x_i}^{x_{i+1}} (u - w_{i,1})''(t) dt = 0, \quad \int_{x_i}^{x_{i+1}} (u - w_{i,2})''(t) dt = 0, \quad \int_{x_i}^{x_{i+1}} (u - w_{i,2})'(t) dt = 0.$$

Hence, from this relation, we deduce that, for every $0 \leq i \leq N$:

$$u(x_i) - w_{i,2}(x_i) = u(x_{i+1}) - w_{i,2}(x_{i+1}).$$

This allow us to define the polynomial function w_h on $[0, 1]$ as follows:

$$w_h(x) = w_{i,2}(x) + (u(x_i) - w_{i,2}(x_i)), \quad \forall 0 \leq i \leq N,$$

and the previous relations show that w_h is defined and continuous on $[0, 1]$, that $w_h(x_i) = u(x_i)$ for all $0 \leq i \leq N$ and that w_h is a polynomial of degree 2 on each $[x_i, x_{i+1}]$. We conclude easily that:

$$\|u - u_h\|_{H^1(\Omega)} \leq \sqrt{\frac{M}{\alpha}} \inf_{v_h \in V_{0,h}^1} \|u - v_h\|_{H^1(\Omega)} \leq \sqrt{\frac{M}{\alpha}} \|u - w_h\|_{H^1(\Omega)}.$$

Introducing the notation $r_h u = u - w_h$, we can see from the previous identities that $r_h u \in H^3(\Omega)$ and $r_h u|_{[x_i, x_{i+1}]} \in H_0^1([x_i, x_{i+1}])$. Furthermore, we have:

$$\int_{x_i}^{x_{i+1}} (r_h u)''(t) dt = 0, \quad \text{and} \quad \int_{x_i}^{x_{i+1}} (r_h u)'(t) dt = 0.$$

To achieve the estimate of $R_h u$, we introduce a result known as Poincaré-Wirtinger inequality.

Lemma 7.6 (Poincaré-Wirtinger inequality) *Given a bounded interval $[a, b]$ of \mathbb{R} , we pose*

$$W_{[a,b]} = \left\{ u \in H^1([a, b]), \int_a^b u(t) dt = 0 \right\}.$$

Then, we have:

$$\int_a^b u^2(t) dt \leq \frac{b-a}{2} \int_a^b u'(t) dt, \quad \forall u \in W_{[a,b]}.$$

Using this result and the previous identities, we deduce that $r_h'' u \in W_{[x_j, x_{j+1}]}$, for all $0 \leq i \leq N$ and thus:

$$\begin{aligned} \int_{x_i}^{x_{i+1}} |(r_h u)''(t)|^2 dt &\leq \frac{h^2}{2} \int_{x_i}^{x_{i+1}} |(r_h u)'''(t)|^2 dt \\ &\leq \frac{h^2}{2} \int_{x_i}^{x_{i+1}} |u'''(t)|^2 dt \end{aligned}$$

Similarly, we can deduce that:

$$\begin{aligned} \int_{x_i}^{x_{i+1}} |(r_h u)'(t)|^2 dt &\leq \frac{h^2}{2} \int_{x_i}^{x_{i+1}} |(r_h u)''(t)|^2 dt \\ &\leq \frac{h^4}{4} \int_{x_i}^{x_{i+1}} |u'''(t)|^2 dt \end{aligned}$$

Adding these last two results yields:

$$\|r_h u\|_{H_0^1(\Omega)}^2 = \int_{x_i}^{x_{i+1}} |(r_h u)'(t)| dt \leq \frac{h^4}{4} \|u'''\|_{L^2(\Omega)}.$$

Finally, we can enounce the convergence result as follows.

Theorem 7.2 (Convergence) *Suppose $u \in H_0^1(\Omega)$ and $u_h \in V_{0,h}^2$ are the solutions of (7.2) and (7.10), respectively. Then, the Lagrange \mathbb{P}_2 finite element method converges, i.e. we have:*

$$\lim_{h \rightarrow 0} \|u - u_h\|_{H^1(\Omega)} = 0.$$

Furthermore, if $u \in H^3(\Omega)$ (i.e. $f \in H^1(\Omega)$), then there exists a constant C independent of h such that:

$$\|u - u_h\|_{H^1(\Omega)} \leq C h^2 \|u'''\|_{L^2(\Omega)}.$$

Remark 7.6 *The convergence rate of the \mathbb{P}_2 finite element method is better than with the \mathbb{P}_1 finite element method. However, the data f must be sufficiently smooth, here $f \in H^1$.*

7.2.5 Lagrange \mathbb{P}_k elements

This section generalizes the concepts introduced in the previous sections to the interpolation of continuous and polynomials functions of degree $k \geq 1$. We consider a set of points $(x_j)_{0 \leq j \leq N+1}$ or intervals $K_j = [x_j, x_{j+1}]$ forming a uniform mesh of $\Omega =]0, 1[$.

Lagrange finite element spaces

For a given integer $k \geq 1$, we define the space of globally continuous functions on $[0, 1]$ whose restriction on each interval $K_j = [x_j, x_{j+1}]$ is a polynomial of degree k :

$$V_h^k = \{v_h \in C^0([0, 1]), v_h|_{K_j} \in \mathbb{P}_k, 0 \leq j \leq N\},$$

and the subspace of V_h^k :

$$V_{0,h}^k = \{v_h \in V_h^k, \text{ such that } v_h(0) = v_h(1) = 0\}.$$

In each interval K_j , a function of V_h^k is uniquely determined by its values at $k + 1$ distinct points along the segment. Hence, on each interval K_j , we introduce a set of *nodes*:

$$y_{j,l} = x_j + \frac{l}{k}(x_{j+1} - x_j) = x_j + \frac{l}{k}h_j, \quad 0 \leq l \leq k - 1,$$

and $y_{N+1,0} = x_{N+1}$.

Lemma 7.7 *The space V_h^k is a subspace of $H^1(\Omega)$ of dimension $k(N + 1) + 1$. Every function v_h of V_h^k is uniquely determined by its values at the mesh nodes $(y_{j,l})_{0 \leq j \leq N, 0 \leq l \leq k-1}$ and $y_{N+1,0}$. Furthermore, the shape functions are such that:*

$$\varphi_{j,l}(y_{j',l'}) = \delta_{jj'} \delta_{ll'}.$$

The space $V_{0,h}^k$ is a subspace of $H_0^1(\Omega)$ of dimension $k(N + 1) - 1$.

Convergence of the Lagrange \mathbb{P}_k finite element method

We can consider the internal approximation problem of finding $u_h \in V_{0,h}^k$ such that:

$$\int_{\Omega} u'_h(x) v'_h(x) dx + \int_{\Omega} c(x) u_h(x) v_h(x) dx = \int_{\Omega} f(x) v_h(x) dx, \quad \forall v_h \in V_{0,h}^k. \quad (7.11)$$

Following the same analysis than for the Lagrange \mathbb{P}_2 element, we have the following convergence result.

Theorem 7.3 (Convergence) *Suppose $u \in H_0^1(\Omega)$ and $u_h \in V_{0,h}^k$ are the solutions of (7.2) and (7.11), respectively. Then, the Lagrange \mathbb{P}_k finite element method converges, i.e. we have:*

$$\lim_{h \rightarrow 0} \|u - u_h\|_{H^1(\Omega)} = 0.$$

Furthermore, if $u \in H^{k+1}(\Omega)$ (i.e. $f \in H^{k-1}(\Omega)$), then there exists a constant C independent of h such that:

$$\|u - u_h\|_{H^1(\Omega)} \leq C h^k \|f\|_{H^{k-1}(\Omega)}.$$

Remark 7.7 1. *The approximate solution u_h converges toward the exact solution u in $H^1(\Omega)$ when $h \rightarrow 0$. The Lagrange \mathbb{P}_k method is of order k in h , if the function f is sufficiently smooth.*

2. *The matrix assembly becomes more and more difficult as the value of k increases. Since the size of the problem increases as well, this may result in additional difficulties in solving the resulting linear system.*

3. *The computation of the components of the right-hand side vector F_h must be carried out with a sufficiently accurate method.*

7.3 Triangular finite elements in higher dimensions

We consider here a boundary-value problem posed in an open bounded domain $\Omega \subset \mathbb{R}^d$ ($d = 2, 3$, in general). For the sake of simplicity, we restrict our study to domains with piecewise polygonal (resp. polyhedral when $d = 3$) boundaries, i.e. Ω can be exactly covered by a finite union of polygons (resp. polyhedra).

To set the ideas, we consider the homogeneous boundary-value problem (7.2) posed here in an open bounded domain Ω of \mathbb{R}^2 :

Given $f \in L^2(\Omega)$ and $c \in L^\infty(\Omega)$, find u such that:

$$\begin{cases} -\Delta u + cu = f, & \text{in } \Omega \\ u = 0, & \text{on } \partial\Omega \end{cases} \quad (7.12)$$

We already know that this problem has a unique solution $u \in H_0^1(\Omega)$.

7.3.1 Preliminary definitions

We proceed like in one dimension of space. The domain Ω is decomposed into a set of N finite elements, triangles $(K_j)_{1 \leq j \leq N}$ in dimension $d = 2$ and tetrahedra in dimension $d = 3$. These two types of elements belong to the generic class of *simplices*.

***d*-Simplices**

Definition 7.2 A *d*-simplex K is the convex hull (envelope) of $d + 1$ points $(a_j)_{1 \leq j \leq d+1}$ in \mathbb{R}^d , called the vertices of K , that are not all lying in the same hyperplane. It is the smallest convex passing through all these points.

Remark 7.8 Let consider $d + 1$ points $(a_j)_{1 \leq j \leq d+1}$ in \mathbb{R}^d and let denote $(a_{i,j})_{1 \leq i \leq d}$ the coordinates of vector (a_j) . These points are affinely independent, i.e. not lying in the same hyperplane, if the matrix

$$M = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,d+1} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,d+1} \\ \vdots & \vdots & \ddots & \vdots \\ a_{d,1} & a_{d,2} & \cdots & a_{d,d+1} \\ 1 & 1 & \cdots & 1 \end{pmatrix}$$

is invertible. In such case, the simplex is not degenerated. Any *d*-simplex has the same number of faces and vertices, each face being itself a $d - 1$ -simplex.

Furthermore, a few geometric parameters characterize a simplex K :

- (i) the *diameter* h_K : the length of the largest element edge,
- (ii) the *roundness* ρ_K : the diameter of the largest inscribed ball,
- (iii) the *aspect ratio* $\sigma_K = \frac{h_K}{\rho_K}$: a measure of the non-degeneracy of K .

Barycentric coordinates

Any simplex K can be represented by the *barycentric coordinates* $\{\lambda_j\}_{1 \leq j \leq d+1}$ of its vertices. Barycentric coordinates are a form of homogeneous coordinates.

Definition 7.3 For every $1 \leq j \leq d+1$, the barycentric coordinate λ_j of a point $x \in \mathbb{R}^d$ is the first-degree polynomial:

$$\lambda_j(x) = c_1 x_1 + \cdots + c_d x_d + c_{d+1},$$

such that for all $1 \leq i \leq d + 1$, $\lambda_j(a_i) = \delta_{ij}$.

For each j , the $d + 1$ coefficients of the barycentric coordinate λ_j are the unknowns of a linear system of $d + 1$ equations. All $d + 1$ systems share the same matrix M^t and give a unique solution if the simplex K is not degenerated.

Proposition 7.2 For every point $x \in \mathbb{R}^d$, there exists a unique vector $(\lambda_j(x))_{1 \leq j \leq d+1}$ such that the following identities hold:

$$x = \sum_{j=1}^{d+1} a_j \lambda_j(x), \quad \text{and} \quad \sum_{j=1}^{d+1} \lambda_j(x) = 1.$$

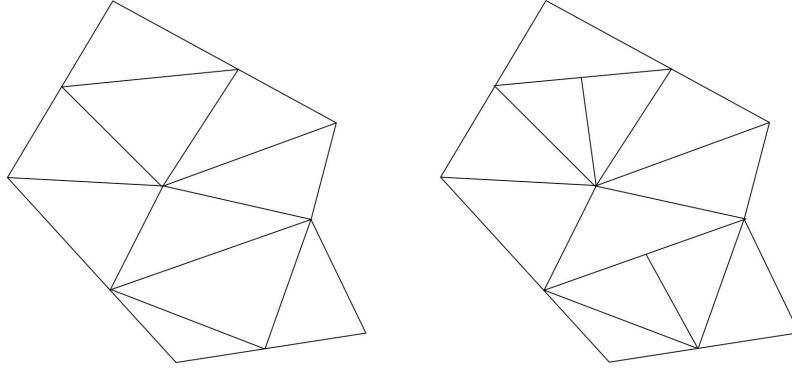


Figure 7.5: *Example (left) and counter-example (right) of conforming triangulation, in two dimensions.*

Proof. For each point $x = (x_i)_{1 \leq i \leq d}$, the scalar values $(\lambda_j(x))_{1 \leq j \leq d+1}$ are the solutions of a $d+1 \times d+1$ linear system that admits M as associated matrix. Hence, there is a unique solution to this system if the simplex K is non degenerated. Each function λ_j is an affine function and one can check easily that $\lambda_j(a_i) = \delta_{ij}$ will be a solution of this system. Since there is only one such affine function, it is then the barycentric coordinates function. \square

Since the λ_j are affine functions of x , then we can write:

$$K = \{x \in \mathbb{R}^d, 0 \leq \lambda_j(x) \leq 1, 1 \leq j \leq d+1\},$$

the faces of K are the intersections of K with the hyperplans $\lambda_j(x) = 0$, $1 \leq j \leq d+1$. We observe also that the change from Cartesian coordinates to barycentric coordinates is an affine transformation. Hence, a polynomial of total degree k in Cartesian coordinates can be expressed as a polynomial of total degree k in barycentric coordinates, and conversely. For a first-degree polynomial p we have:

$$p(x) = \sum_{j=1}^{d+1} p(a_j) \lambda_j(x). \quad (7.13)$$

Triangulations and meshes

Definition 7.4 A triangulation, also called a triangular mesh of $\bar{\Omega}$ is a set T_h of non degenerated d -simplices $(K_j)_{1 \leq j \leq N}$ such that:

$$(i) \quad K_j \subset \bar{\Omega} \text{ and } \bar{\Omega} = \bigcup_{j=1}^N K_j,$$

(ii) the intersection $K_i \cap K_j$ of any two simplices is a m -simplex, $0 \leq m \leq d-1$, such that all its vertices are also vertices of K_i and K_j .

This definition states that the intersection of two triangles, if it is not empty, shall be reduced to either a common vertex or an edge. Similarly, in three dimensions, the intersection of two tetrahedra can be either empty, or reduced to a single common entity (vertex, edge or face). Such a mesh is often called a *conforming* mesh (cf. Figure 7.5). The *vertices* or *nodes* of the mesh T_h are the vertices of the d -simplices K_j that compose the mesh. Algorithms to construct such triangulations will be described in Chapter 9 and we refer the reader to [Frey-George, 2000] for more information on this topic.

We introduce two conditions on the geometry of a triangulation, with respect to the diameter and the roundness of its elements.

Definition 7.5 Suppose $(T_h)_{h>0}$ is a sequence of meshes of Ω . This sequence is said to be a sequence of regular meshes, or a quasi-uniform sequence, if:

1. the sequence $h = \max_{K \in T_h} h_K$ tends toward 0,
2. there exists a constant $C \geq 1$ such that:

$$\forall h > 0, \quad \forall K \in T_h, \quad \frac{h}{\rho_K} \leq C. \quad (7.14)$$

Remark 7.9 In dimension two, if K is a triangle, the condition 7.14 is equivalent to the existence of an angle $\theta_0 > 0$ such that

$$\forall h > 0, \quad \forall K \in T_h, \quad \theta_K \geq \theta_0,$$

where θ_K is the smallest vertex angle in triangle K .

A set of points of a simplex K has a specific role as defined hereafter.

Definition 7.6 For every $k \in \mathbb{N}^*$, we call principal lattice of order k the set:

$$\Sigma_k = \left\{ x \in K, \lambda_j(x) \in \left\{ 0, \frac{1}{k}, \frac{2}{k}, \dots, \frac{k-1}{k}, 1 \right\}, \text{ for } 1 \leq j \leq d+1 \right\}. \quad (7.15)$$

For $k = 1$, the lattice is simply the set of vertices of K ; for $k = 2$, the principal lattice is composed of the vertices and of the midpoints of the edges (Figure 7.6). More generally, a lattice Σ_k is a finite set of points $(\sigma_j)_{1 \leq j \leq N_k}$.

Polynomial spaces

We introduce the set \mathbb{P}_k of the polynomials p with scalar coefficients of \mathbb{R}^d in \mathbb{R} of degree less than or equal to k :

$$\mathbb{P}_k = \left\{ p(x) = \sum_{\substack{i_1, \dots, i_d \geq 0 \\ i_1 + \dots + i_d \leq k}} \alpha_{i_1, \dots, i_d} x_1^{i_1} \dots x_d^{i_d}, \alpha_{i_j} \in \mathbb{R}, x = (x_1, \dots, x_d) \right\}.$$

Hence, in two and three dimensions of space, we will simply denote:

$$\mathbb{P}_k = \left\{ p(x, y) = \sum_{0 < i+j \leq k} \alpha_{ij} x^i y^j, \alpha_{ij} \in \mathbb{R} \right\},$$

$$\mathbb{P}_k = \left\{ p(x, y, z) = \sum_{0 \leq i+j+l \leq k} \alpha_{ijl} x^i y^j z^l, \alpha_{ijl} \in \mathbb{R} \right\}.$$

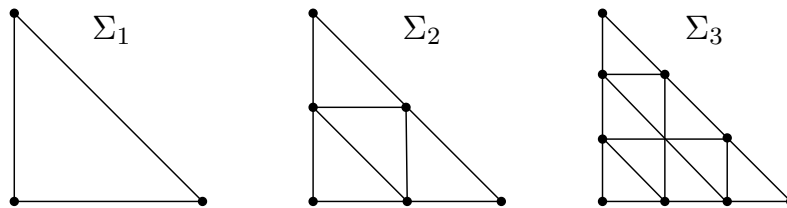


Figure 7.6: Principal lattice of order 1, 2 and 3 for a two-dimensional simplex.

It is easy to verify that \mathbb{P}_k is a vector space of dimension:

$$\dim(\mathbb{P}_k) = \sum_{l=0}^k \binom{d+l-1}{l} = \binom{d+k}{k} = \begin{cases} k+1 & d=1 \\ \frac{1}{2}(k+1)(k+2) & d=2 \\ \frac{1}{6}(k+1)(k+2)(k+3) & d=3 \end{cases}$$

The notion of lattice Σ_k of a simplex K allows to define a bijective mapping between a space of polynomials \mathbb{P}_k and a set of points $(\sigma_j)_{1 \leq j \leq N_k}$. The set Σ_k is said to be *unisolvent* for \mathbb{P}_k . We will use this property to define a finite element.

Lemma 7.8 *Given a simplex K . For $k \geq 1$, we consider the lattice Σ_k of order k whose points are denoted $(\sigma_j)_{1 \leq j \leq N_k}$. Then, every polynomial $p \in \mathbb{P}_k$ is uniquely determined by its values at the points $(\sigma_j)_{1 \leq j \leq N_k}$. There exists a basis $(\varphi_j)_{1 \leq j \leq N_k}$ of \mathbb{P}_k such that:*

$$\varphi_j(\sigma_i) = \delta_{ij}, \quad 1 \leq i, j \leq N_k.$$

Proof. At first, we notice that the cardinal of Σ_k and the dimension of the vector space \mathbb{P}_k coincide:

$$\text{card}(\Sigma_k) = \dim(\mathbb{P}_k) = \frac{(d+k)!}{d!k!}.$$

Indeed, we can write the elements of Σ_k as follows:

$$\Sigma_k = \left\{ \sum_{j=1}^d \frac{\alpha_j}{k} a_j + \left(1 - \sum_{j=1}^d \frac{\alpha_j}{k}\right) \alpha_0, \quad 0 \leq \alpha_1 + \dots + \alpha_d \leq k \right\},$$

where the $\alpha_j \in \mathbb{N}$. We know that the mapping associating to every polynomial \mathbb{P}_k its values on the lattice Σ_k is a linear mapping. Hence, it is sufficient to show that it is an injection to have the bijective property. We will prove by recurrence on the dimension d that if $p \in \mathbb{P}_k$ is such that $p(x) = 0$ for all $x \in \Sigma_k$ the $p = 0$ on \mathbb{R}^d . At first, notice that a polynomial of degree k that vanishes in $k+1$ points of \mathbb{R} is identically null. Suppose this is also true for the dimension $d-1$. We use a recurrence on the degree k . For $k=1$, an affine function that vanishes at the vertices of a non-degenerated simplex K is identically null according to the relation (7.13). Suppose this property is true for all polynomials of degree $k-1$ and let consider a degree k polynomial p that vanishes on Σ_k . We observe that Σ_k contains the subset

$$\Sigma'_k = \{x \in \Sigma_k, \lambda_0(x) = 0\},$$

that corresponds to the principal lattice of order k of the $d-1$ -simplex of vertices (a_1, \dots, a_d) . Since the restriction of p to the hyperplane generated by (a_1, \dots, a_d) is a polynomial of degree k in $d-1$ variables, then $p=0$ on this hyperplane, thanks to the recurrence hypothesis. If we introduce a system of coordinates (x_1, \dots, x_d) such that the hyperplane is now defined by $x_d = 0$, then

$$p(x_1, \dots, x_d) = x_d q(x_1, \dots, x_{d-1}),$$

where q is a polynomial of degree $d-1$ that vanishes on the set $\Sigma_k - \Sigma'_k$ since x_d is non null on this set. The set $\Sigma_k - \Sigma'_k$ is a principal lattice of order $k-1$ and thus the recurrence hypothesis leads to conclude that $q=0$ and consequently that $p=0$. The results follows. \square

In practice, we consider only polynomials of degree 1 or 2. Equation (7.13) provides the characterization of a polynomial of degree one. Given a d -simplex K of vertices $(a_j)_{1 \leq j \leq d+1}$, we define the edge midpoints $(a_{jj'})_{1 \leq j < j' \leq d+1}$ by their barycentric coordinates:

$$\lambda_j(a_{jj'}) = \lambda_{j'}(a_{jj'}) = \frac{1}{2}, \quad \lambda_l(a_{jj'}) = 0 \quad l \neq j, j',$$

The principal lattice Σ_2 is exactly composed of the vertices and the edge midpoints and every polynomial $p \in \mathbb{P}_2$ can be written as:

$$p(x) = \sum_{j=1}^{d+1} p(a_j) \lambda_j(x) (2\lambda_j(x) - 1) + \sum_{1 \leq j < j' \leq d+1} 4p(a_{jj'}) \lambda_j(x) \lambda_{j'}(x), \quad (7.16)$$

where the $(\lambda_j(x))_{1 \leq j \leq d+1}$ are the barycentric coordinates of $x \in \mathbb{R}^d$.

7.3.2 Triangular Lagrange \mathbb{P}_k finite elements

Suppose the domain Ω is covered by a simplicial mesh T_h . The finite element method for triangular Lagrange \mathbb{P}_k elements involves the discrete finite dimensional functional space:

$$V_h^k = \{v \in C^0(\Omega), v_h|_{K_j} \in \mathbb{P}_k, K_j \in T_h\},$$

and its subspace:

$$V_{0,h}^k = \{v_h \in V_h^k, v_h = 0 \text{ on } \partial\Omega\}.$$

Definition 7.7 A triangular Lagrange \mathbb{P}_k finite element is locally defined by a triad (K, P_k, Σ_k) , where:

- (i) K is a d -simplex associated with the mesh T_h ,
- (ii) P_k is a vector space of polynomials of degree less than or equal to k on K ,
- (iii) Σ_k is the principal lattice of order k of the simplex $K \in T_h$.

Σ_k is called the set of nodes of the degrees of freedom of the finite element (K, P_k, Σ_k) .

We consider the set of points $(a_i)_{1 \leq i \leq N_{dof}}$ of the principal lattices of order k of each of the simplices $K_i \in T_h$, where N_{dof} is the number of degrees of freedom of the \mathbb{P}_k finite element method. We call *degrees of freedom* of a function $v_h \in V_h^k$ the set of the values of v at the so-called *nodes* $(a_i)_{1 \leq i \leq N_{dof}}$.

Remark 7.10 We observe that the nodes of the degrees of freedom coincide exactly with the vertices of the simplices $K_i \in T_h$, when $k = 1$. The nodes of the degrees of freedom are composed by the mesh vertices and the edge midpoints, when $k = 2$.

Lemma 7.9 The space V_h^k is a subspace of the space $H^1(\Omega)$ of finite dimension corresponding to the number of degrees of freedom. Furthermore, there exists a basis $(\varphi_j)_{1 \leq j \leq N_{dof}}$ of V_h^k defined by:

$$\varphi_i(a_j) = \delta_{ij}, \quad 1 \leq i, j \leq N_{dof},$$

such that every function $v_h \in V_h^k$ can be uniquely written as

$$v_h(x) = \sum_{i=1}^{N_{dof}} v_h(a_i) \varphi_i(x).$$

Proof. It is easy to see that the elements of V_h^k belong to $H^1(\Omega)$. The Lemma 7.8 allows to conclude that each function $v_h \in V_h^k$ is exactly known by assembling on each $K_i \in T_h$ polynomials of degree k that coincide on the degrees of freedom of the d -faces. By assembling the basis of \mathbb{P}_k on each K_i , the basis $(\varphi_j)_{1 \leq j \leq N_{dof}}$ is defined. \square

Corollary 7.4 *The subspace $V_{0,h}^k$ is a subspace of $H_0^1(\Omega)$ of finite dimension corresponding to the number of internal degrees of freedom, i.e.] not taking the nodes on $\partial\Omega$ into account.*

Definition 7.8 *The triad (K, P_k, Σ_k) is said to be unisolvent if and only if the mapping $v \in P_k \mapsto (\varphi_1(v), \dots, \varphi_{N_{dof}}(v)) \in \mathbb{R}^{N_{dof}}$ is an isomorphism.*

The unisolvency property is equivalent to say that every function in the polynomial space P_k is entirely determined by its node values.

7.3.3 Finite element approximation of a boundary-value problem

We return to the numerical approximation of the solution of the homogeneous Dirichlet boundary-value problem (7.12). The variational formulation of the internal approximation reads:

find $u_h \in V_{0,h}^k$ such that

$$\int_{\Omega} (\nabla u_h \cdot \nabla v_h)(x) + \int_{\Omega} c(x) u_h(x) v_h(x) dx = \int_{\Omega} f(x) v_h(x) dx, \quad \forall v_h \in V_{0,h}^k, \quad (7.17)$$

By decomposing u_h on the canonical basis $(\varphi_j)_{1 \leq j \leq N_{dof}}$ and considering as test functions $v_h = \varphi_i$, we obtain:

$$\sum_{j=1}^{N_{dof}} u_h(a_j) \left(\int_{\Omega} (\nabla \varphi_j \cdot \nabla \varphi_i)(x) dx + \int_{\Omega} c(x) \varphi_j(x) \varphi_i(x) dx \right) = \int_{\Omega} f(x) \varphi_i(x) dx.$$

This formulation leads to solving a linear system in $\mathbb{R}^{N_{dof}}$:

$$A_h U_h = F_h,$$

where we have introduced the notations $U_h = (u_h(a_j))_{1 \leq j \leq N_{dof}}$ and $F_h = (\int_{\Omega} f \varphi_i dx)_{1 \leq i \leq N_{dof}}$ and

$$A_h = \left(\int_{\Omega} (\nabla \varphi_j \cdot \nabla \varphi_i)(x) dx + \int_{\Omega} c(x) \varphi_j(x) \varphi_i(x) dx \right)_{1 \leq i, j \leq N_{dof}}.$$

It is easy to see that the matrix A_h can be decomposed as a sum of a *stiffness* matrix K_h and a *mass* matrix M_h . Actually, this result is independent of the dimension.

Coefficients of A_h

Since the shape functions φ_j have a small support around a node a_i , the intersection of the supports of φ_j and φ_i is often the empty set and thus the resulting matrix A_h will contain a lot of zero coefficients. It is a *sparse* matrix. The coefficients of A_h can be computed via an exact integration formula. Let denote $(\lambda_i(x))_{1 \leq i \leq d+1}$ the barycentric coordinates of the point x in a simplex $K \in T_h$. For every $(\alpha_i)_{1 \leq i \leq d+1}$ we have:

$$\int_K \lambda_1(x)^{\alpha_1} \dots \lambda_{d+1}(x)^{\alpha_{d+1}} dx = |K| \frac{\alpha_1 + \dots + \alpha_{d+1}! d!}{\left(\sum_{1 \leq j \leq d+1} \alpha_j + d \right)!},$$

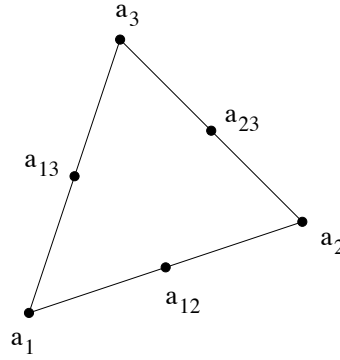


Figure 7.7: Nodes of the degrees of freedom on a finite element $(K, \mathbb{P}_2, \Sigma_2)$.

where $|K| = \text{mes}(K)$ denotes the "volume" of the simplex K . The integrals of the right-hand side term F_h are approximated using quadrature formulas in each simplex $K \in T_h$. For instance the generalization of the trapeze formula for a simplex K is:

$$\int_K \varphi_j(x) dx \approx \frac{|K|}{d+1} \sum_{i=1}^{d+1} \varphi_j(a_i),$$

where $(a_i)_{1 \leq i \leq d+1}$ represent the vertices of K . These formulas are exact for affine functions. For instance, considering a triangle $K \in \mathbb{R}^2$ of vertices $(a_i)_{1 \leq i \leq 3}$ and denoting by $(a_{ij})_{1 \leq i, j \leq 3}$ the midpoints of the edges $a_i a_j$ (cf. Figure 7.7), the following quadrature formula is exact for $\varphi_k \in \mathbb{P}_2$:

$$\int_K \varphi_k(x) \approx \frac{|K|}{3} \sum_{1 \leq i, j \leq 3} \varphi_k(a_{ij}),$$

where $|K|$ represents the area of simplex K , while the following formula is exact for $\varphi_k \in \mathbb{P}_3$:

$$\int_K \varphi_k(x) dx \approx \frac{|K|}{60} \left(\sum_{1 \leq i \leq 3} \varphi_k(a_i) + 8 \sum_{1 \leq i < j \leq 3} \varphi_k(a_{ij}) + 27 \varphi_k(a_0) \right),$$

where the point $a_0 = \frac{1}{d+1} \sum_{1 \leq i \leq d+1} a_i$ denotes here the barycenter of K .

Remark 7.11 *As we mentioned in one dimension of space, the analysis can be simplified by considering an affine transformation allowing to consider any d -simplex $K \in T_h$ as the image of a reference element \hat{K} . Hence, all computations can be performed on this reference simplex.*

7.3.4 The reference finite element

By convention, the vertices of the *reference simplex* \hat{K} are given by the *origin* $\hat{a}_0 = (0, \dots, 0)$ and the points $\hat{a}_i = (0, \dots, 1, \dots, 0)$, for which all coordinates are equal to zero except the d^{th} coordinate that is equal to one. For a non-degenerated simplex K , we denote by $F_K : \mathbb{R}^d \rightarrow \mathbb{R}^d$ the unique affine transformation that maps \hat{a}_i on a_i for all $i = 0, \dots, d$. Hence, we write;

$$F_K(x) = a_0 + B_K x,$$

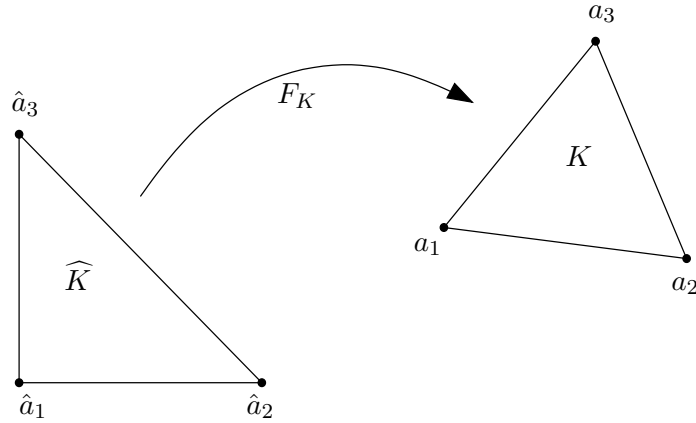


Figure 7.8: Affine transformation of the reference linear triangle in two dimensions.

where B_K is a $d \times d$ matrix such that the column i is given by the coordinates of $a_i - a_0$. Since the simplex K is non-degenerated, B_K is invertible and F_K is a one-to-one mapping that maps \widehat{K} on K and we have:

$$|K| = |\det(B_K)| |\widehat{K}| = \frac{|\det(B_K)|}{d!}.$$

We use the notation $\hat{\cdot}$ in the reference element. Moreover, to simplify, we denote \hat{q} any quantity obtained by transporting a quantity q using the transformation F_K . Hence, we denote:

$$\hat{x} = F_K^{-1}(x) = B_K^{-1}(x - a_0) \Leftrightarrow x = F_K(\hat{x}).$$

For every function v defined on K , we define \hat{v} defined on \widehat{K} as:

$$\hat{v}(\hat{x}) = v(x) \Leftrightarrow \hat{v} = v \circ F_K = v(F_K(\hat{x})),$$

and we remove the $\hat{\cdot}$ notation when dealing with F_K^{-1} , for every function \hat{v} defined on \widehat{K} , we write:

$$v(x) = \hat{v} \circ F_K^{-1}(x) = \hat{v}(F_K^{-1}(x)).$$

Similarly, if ψ is a linear form acting on the functions defined on K , we define the transported linear form $\hat{\psi}$ acting on the functions defined on \widehat{K} as:

$$\hat{\psi}(\hat{v}) = \psi(v).$$

Notice that the barycentric coordinates are preserved by the affine transformation F_K :

$$\hat{\lambda}_i(\hat{x}) = \lambda_i(x).$$

This leads to conclude that the principal lattice of order k of the simplex K is the image by F_K of the principal lattice of order k of \widehat{K} , denoted by $\widehat{\Sigma}_k$. Finally, we observe that the space of polynomials of degree k is left invariant by the transformation F_K .

Proposition 7.3 We denote by $\|\cdot\|$ the Euclidean norm of \mathbb{R}^d and its subordinate norm. Hence, we have:

$$\|B_K\| \leq \frac{h_K}{\rho_{\widehat{K}}}, \quad \|B_K^{-1}\| \leq \frac{h_{\widehat{K}}}{\rho_K}, \quad |\det(B_K)| = \frac{|K|}{|\widehat{K}|}.$$

Proof. by definition,

$$\|B_K\| = \sup_{v \in \mathbb{R}^d} \frac{\|B_K v\|}{\|v\|} = \sup_{\|v\|=\rho_{\widehat{K}}} \frac{\|B_K v\|}{\rho_{\widehat{K}}},$$

and the sup is attained. Let $v \in \mathbb{R}^d$ be a vector of norm $\rho_{\widehat{K}}$ that correspond to the sup. Hence, there exists two points \hat{y} and \hat{z} on the boundary of the inscribed sphere of \widehat{K} such that $v = \hat{y} - \hat{z}$. Thus,

$$B_K v = B_K(\hat{y} - \hat{z}) = F_K(\hat{y}) - F_K(\hat{z}) = y - z$$

where y, z are tzo points in K . Hence,

$$\|B_K v\| = \|y - z\| \leq h_K.$$

The second inequality is deduced from the first by interchanging the roles of K and \widehat{K} . The third equality is obtained by noticing that $|\det(B_K)|$ is the Jacobian of F_K . Hence,

$$|K| = \int_K dx = \int_{\widehat{K}} |\det(B_K)| d\hat{x} = |\det(B_K)| |\widehat{K}|.$$

□

As a consequence, we deduce that there exists two constants C_1 and C_2 , independent of K such that:

$$C_2 \rho_K^d \leq |\det(B_K)| \leq C_1 h_K^d.$$

Transformation of the derivatives

We have the following identities:

$$\begin{aligned} \|v\|_{L^p(K)} &= \int_K |v(x)|^p dx = \int_{\widehat{K}} |v(F_K(\hat{x}))|^p |\det(B_K)| d\hat{x} \\ &= d! |K| \int_{\widehat{K}} |\hat{v}(\hat{x})|^p d\hat{x} \end{aligned}$$

thus yielding the identity:

$$\|v\|_{L^p(K)} = C |K|^{1/p} \|\hat{v}\|_{L^p(\widehat{K})}, \quad (7.18)$$

with $C = (d!)^{1/p}$. Moreover, we observe that if $v(x) = \hat{v}(\hat{x})$, *i.e.* if $\hat{v} = v \circ F_K$, then we have:

$$\nabla \hat{v}(\hat{x}) = B_K^t \nabla v(x),$$

and we obtain the identity:

$$\begin{aligned} |v|_{H^1(K)}^2 &= \int_K \|\nabla v(x)\|^2 dx = \int_{\widehat{K}} \|(B_K^t)^{-1} \nabla \hat{v}(\hat{x})\|^2 |\det(B_K)| d\hat{x} \\ &= C_2 d! \frac{|K|}{\rho_K^2} \int_{\widehat{K}} \|\nabla \hat{v}(\hat{x})\|^2 d\hat{x} \end{aligned}$$

and then:

$$|v|_{H^1(K)} \leq C \frac{h_K}{|K|^{1/2} \rho_K} |\hat{v}|_{H^1(\widehat{K})}, \quad (7.19)$$

where C depends only on the dimension d . By interchanging the role of K and \widehat{K} , we obtain similarly:

$$|\hat{v}|_{H^1(\widehat{K})} \leq C \frac{|K|^{1/2}}{\rho_K} |v|_{H^1(K)}. \quad (7.20)$$

We give the following generalization result.

Theorem 7.4 *If $v \in H^m(K)$, then $\hat{v} \in H^m(\hat{K})$ and there exists a constant C_1 , independent of K , such that:*

$$\forall v \in H^m(K), \quad |\hat{v}|_{H^m(\hat{K})} \leq C_1 \|B_K\|^m |\det(B_K)|^{-1/2} |v|_{H^m(K)}. \quad (7.21)$$

If $\hat{v} \in H^m(\hat{K})$, then $v \in H^m(K)$ and there exists a constant C_2 , independent of K , such that:

$$\forall \hat{v} \in H^m(\hat{K}), \quad |v|_{H^m(K)} \leq C_2 \|B_K^{-1}\|^m |\det(B_K)|^{1/2} |\hat{v}|_{H^m(\hat{K})}. \quad (7.22)$$

Proof. Result admitted here. □

Remark 7.12 *Now, we are able to construct finite elements from a unisolvent triad $(\hat{K}, \hat{P}_k, \hat{\Sigma}_k)$ by defining directly the finite element (K, P_k, Σ_k) as the transport of these quantities by the transformation F_K . It is easy to see that the new finite element is unisolvent.*

7.3.5 Convergence of the finite element method

We introduce a preliminary result.

Proposition 7.4 *Let consider $(T_h)_{h>0}$ a sequence of regular triangulations of $\bar{\Omega} \subset \mathbb{R}^d$ ($d \leq 3$). Then, for every $v \in H^{k+1}(\Omega)$, the interpolate $r_h v$ is defined and there exists a constant C , independent of h such that:*

$$\|v - r_h v\|_{H^1(\Omega)} \leq C h^k \|v\|_{H^{k+1}(\Omega)}.$$

The following result states the convergence of the \mathbb{P}_k finite element method.

Theorem 7.5 *Let consider $(T_h)_{h>0}$ a sequence of regular triangulations of $\bar{\Omega} \subset \mathbb{R}^d$ ($d \leq 3$) where all elements in all triangulations are affine equivalent to a same reference element $(\hat{K}, \hat{P}_k, \hat{\Sigma}_k)$ of class C^0 for a given $k \geq 1$ such that $P_k \subset \hat{P} \subset H^1(\hat{K})$. Then, the \mathbb{P}_k finite element method converges, i.e. the approximate solution u_h of the problem (7.17) converges toward the solution u of problem (7.12) in $H^1(\Omega)$:*

$$\lim_{h \rightarrow 0} \|u - u_h\|_{H^1(\Omega)} = 0.$$

Furthermore, if $u \in H^{k+1}(\Omega)$, then there exists a constant C independent of h such that:

$$\|u - u_h\|_{H^1(\Omega)} \leq C h^k \|u\|_{H^{k+1}(\Omega)},$$

Proof. To show the convergence, we use Theorem (4.11) where we consider $\mathcal{V} = C_c^\infty(\Omega) \subset H^{k+1}(\Omega)$ dense in $H^1(\Omega)$. The estimate in the previous proposition allows to verify the assumption of Theorem (4.11). Then, we use Céa's lemma to write:

$$\|u - u_h\|_{H^1(\Omega)} \leq C \inf_{v_h \in V_{0,h}^k} \|u - v_h\|_{H^1(\Omega)} \leq C \|u - r_h u\|_{H^1(\Omega)},$$

if $r_h u \in H^1(\Omega)$. The previous proposition allows to conclude. □

Remark 7.13 *This result involves the exact solution u_h of the internal approximation problem in $V_{0,h}^k$. This requires computing all integrals in A_h and F_h exactly. Since numerical integration formulas are used in practice, this result may not be valid. Nonetheless, if these integration formulas are exact or at least very accurate, the order of convergence of the Lagrange finite element method is preserved,*

The Theorem 7.5 shows that the regularity of the exact solution has a direct impact on the order of convergence of the finite element approximation. Hence, it is often important to have an *a priori* knowledge of its regularity. The following result gives an indication for a convex domain.

Theorem 7.6 *Let Ω be a convex polygon and let $f \in \mathbb{R}^2$. Then, the solution u of the homogeneous Dirichlet problem:*

find u such that:

$$\int_{\Omega} \nabla u \cdot \nabla v = \int_{\Omega} f v, \quad \forall v \in H_0^1(\Omega),$$

is in $H^2(\Omega)$ and we have the following estimate:

$$\forall f \in L^2(\Omega), \quad \|u\|_{H^2(\Omega)} \leq \|f\|_{L^2(\Omega)}.$$

7.3.6 Numerical resolution of the linear system

We remind here that the finite element approximation of the boundary-value problem (7.12) implies the numerical solving of a linear system in $\mathbb{R}^{N_{dof}}$:

$$A_h U_h = (K_h + M_h) U_h = F_h,$$

where $U_h = (u_h(a_j))_{1 \leq j \leq N_{dof}}$ and $F_h = (\int_{\Omega} f \varphi_i dx)_{1 \leq i \leq N_{dof}}$. The generic term of the stiffness matrix K_h is of the form:

$$k_{ij} = a(\varphi_j, \varphi_i), \quad 1 \leq i, j \leq N_{dof}.$$

where $a(\cdot, \cdot)$ is the bilinear form defined on the functional space H_0^1 and $(\varphi_j)_{1 \leq j \leq N_{dof}}$ is the canonical basis of the approximation space. We have mentioned already that the stiffness matrix K_h inherits of the properties of the bilinear form a . More precisely, if a is V-elliptic, then K_h is positive definite and if a is symmetric then K_h is symmetric. We have also indicated that the matrix A_h is a *sparse* matrix, *i.e.* it contains a large number of zeroes, as well as the matrices K_h and M_h . Hence, iterative methods are preferred to solve this linear system (cf. Chapter 5).

Let denote $0 < \lambda_1(K_h) < \dots < \lambda_{N_{dof}}(K_h)$ the eigenvalues of K_h and $0 < \lambda_1(M_h) < \dots < \lambda_{N_{dof}}(M_h)$ the eigenvalues of M_h . We introduce the notations:

$$\kappa(K_h) = \frac{\lambda_{N_{dof}}(K_h)}{\lambda_1(K_h)}, \quad \text{and} \quad \kappa(M_h) = \frac{\lambda_{N_{dof}}(M_h)}{\lambda_1(M_h)},$$

and we have the following result.

Proposition 7.5 *Suppose $(T_h)_{h>0}$ is a quasi-uniform sequence. Then, there exists two constants $C_1 > 0$, $C_2 > 0$ and a constant $C_3 > 0$ such that:*

$$C_1 \leq h^2 \kappa(K_h) \leq C_2, \quad \text{and} \quad \kappa(M_h) \leq C_3.$$

We observe that the *mass* matrix M_h is always well-conditioned since its condition number $\kappa(M_h)$ is independent of the mesh size h , under the condition that the refinement is quasi-uniform. On the other hand, the *stiffness* matrix K_h becomes *ill*-conditioned, *i.e.* $\kappa(K_h) \gg 1$, when the mesh size h is small.

Remark 7.14 *We have shown that the approximation error between the exact solution u and the Galerkin finite element solution u_h is bounded by above by a constant times the "distance" between the spaces V and V_h . The smaller the distance, the better the approximation error will be. It seems then natural to increase the dimension of the approximation space V_h in order to improve the accuracy of the solution. To this end, we can either:*

1. increase the number of elements and thus increase the global number of degrees of freedom while preserving the local reference element. This strategy is known as *h-refinement*,
2. increase the degree of the polynomials and thus modifying the local reference element while preserving the total number of elements. Obviously, this choice can be only be applied if the exact solution u is sufficiently smooth. This strategy is known as *p-refinement*.

We will discuss these adaptation options and others in the chapter 9.

7.4 The finite element method for the Stokes problem

To illustrate the application of the finite element method to solve a system of partial differential equations, we consider the Stokes problem in two dimensions of space.

We consider the flow of a fluid inside a bounded domain $\Omega \subset \mathbb{R}^2$ and subjected to an external force field f . The flow is considered as stationary and inertial forces are supposed negligible here, hence the Stokes equations introduced in Chapter 4 describe this viscous flow problem:

Find (u, p) in appropriate Hilbert spaces such that:

$$\begin{cases} -\nu \Delta u + \nabla p = f & \text{in } \Omega \\ \operatorname{div} u = g & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases} \quad (7.23)$$

where the unknowns u and p represent the *velocity* and the *pressure* of the fluid, respectively and $\nu > 0$ is the *viscosity*. Usually, in the applications, the flow is incompressible and the function g is equal to zero.

7.4.1 Mixed formulation

We have already seen the mixed formulation of this problem in Chapter 4. We introduce the functional space:

$$L_0^2(\Omega) = \{q \in L^2(\Omega), \int_{\Omega} q \, dx = 0\},$$

and the mixed formulation reads:

Given $f \in L^2(\Omega)$ and $g \in L_0^2(\Omega)$, find $u \in H_0^1(\Omega)^2$ and $p \in L_0^2(\Omega)$ such that:

$$\begin{cases} a(u, v) + b(v, p) = f(v), & \forall v \in H_0^1(\Omega)^2, \\ b(u, q) = g(q), & \forall q \in L_0^2(\Omega) \end{cases} \quad (7.24)$$

where we posed $a(u, v) = \nu \int_{\Omega} \nabla u : \nabla v$, $b(v, p) = - \int_{\Omega} p \operatorname{div} v$, $f(v) = \int_{\Omega} f v$ and $g(q) = - \int_{\Omega} g q$. The functional spaces $H_0^1(\Omega)^2$ and $L^2(\Omega)^2$ are endowed with the canonical norms:

$$\|u\|_{H^1(\Omega)} = \left(\sum_{i=1,2} \|u_i\|_{H^1(\Omega)}^2 \right)^{1/2}, \quad \|f\|_{L^2(\Omega)} = \left(\sum_{i=1,2} \|f_i\|_{L^2(\Omega)}^2 \right)^{1/2},$$

for $u = (u_i)_{1 \leq i \leq 2} \in H_0^1(\Omega)^2$ and $f = (f_i)_{1 \leq i \leq 2} \in L^2(\Omega)^2$.

Remark 7.15 If $f \in C^0(\bar{\Omega})^2$ and $g \in C^0(\bar{\Omega})$, if Ω is of class C^2 and if the solution (u, p) of the problem (7.24) is such that $u \in C^2(\bar{\Omega})^2$ and $p \in C^1(\bar{\Omega})$, then (u, p) is the classical solution of the Stokes problem.

The formulation above is not the only weak formulation possible for the Stokes problem. An alternate view consists in including the divergence constraint (for instance the incompressibility constraint $\operatorname{div} u = 0$) directly in the space of test functions. To this end, we introduce the space:

$$V_0 = \{v \in H_0^1(\Omega)^2, \operatorname{div} v = 0\}.$$

Since the divergence operator is surjective, there exists a function $u_g \in H_0^1(\Omega)^2$ such that for every $g \in L_0^2(\Omega)$, we have:

$$\operatorname{div} u_g = g, \quad \text{and} \quad \|u_g\|_{H^1(\Omega)^2} \leq c \|g\|_{L^2(\Omega)},$$

where $c > 0$ is a constant (see [Brezis, 2005] for more details). Introducing the change of variables $u' = u - u_g$ leads to the condition $\operatorname{div} u' = 0$ and to consider the following weak formulation:

find $u' \in V_0$ such that:

$$a(u', v) = \langle f, v \rangle - a(u_g, v), \quad \forall v \in V_0. \quad (7.25)$$

This weak formulation is called a *constrained* formulation, since the restriction on the test functions to be in the space V_0 leads to discard the term $b(v, p)$ that vanishes here.

Proposition 7.6 *The problem (7.25) is a well-posed problem.*

Proof. It is a direct application of Lax-Milgram theorem. \square Although this new formulation has some theoretical advantages, many practical drawbacks prevent its application for solving a discrete problem. It is indeed difficult to construct divergence-free finite elements. The reader interested is referred to the book [Brezzi-Fortin, 1991] for more details on this topic.

7.4.2 Numerical approximation

We will now investigate the condition required for the approximation spaces in velocity and pressure to define a well-posed discrete problem.

Let consider X_h a subspace of the space $H_0^1(\Omega)^2$ and M_h a subspace of the space $L_0^2(\Omega)$. We introduce the discrete Stokes problem:

Find $u_h \in X_h$ and $p_h \in M_h$ such that:

$$\begin{cases} \nu \int_{\Omega} \nabla u_h : \nabla v_h - \int_{\Omega} p_h \operatorname{div} v_h = \int_{\Omega} f \cdot v_h, & \forall v_h \in X_h, \\ \int_{\Omega} q_h \operatorname{div} u_h = \int_{\Omega} g q_h, & \forall q_h \in M_h. \end{cases} \quad (7.26)$$

We introduce the space

$$\operatorname{Ker}(B_h) = \{v_h \in X_h, \int_{\Omega} q_h \operatorname{div} v_h = 0, \forall q_h \in M_h\},$$

and we enounce the fundamental compatibility result.

Theorem 7.7 *The discrete problem (7.26) is a well-posed problem if and only if the spaces X_h and M_h satisfy the following compatibility condition:*

$$\exists \beta_h > 0, \quad \inf_{q_h \in M_h} \sup_{v_h \in X_h} \frac{\int_{\Omega} q_h \operatorname{div} v_h}{\|q_h\|_{L^2(\Omega)} \|v_h\|_{H^1(\Omega)}} \geq \beta_h. \quad (7.27)$$

Furthermore, under this condition, we have the following estimates:

$$\|u - u_h\|_{H^1(\Omega)} \leq c_{1h} \inf_{v_h \in X_h} \|u - v_h\|_{H^1(\Omega)} + c_{2h} \inf_{q_h \in M_h} \|p - q_h\|_{L^2(\Omega)} \quad (7.28)$$

$$\|p - p_h\|_{L^2(\Omega)} \leq c_{3h} \inf_{v_h \in X_h} \|u - v_h\|_{H^1(\Omega)} + c_{4h} \inf_{q_h \in M_h} \|p - q_h\|_{L^2(\Omega)} \quad (7.29)$$

where the constants $(c_{ih})_{1 \leq i \leq 4}$ are given by Theorem 4.12.

Proof. We know that the bilinear form $a(\cdot, \cdot)$ is V -elliptic on $H_0^1(\Omega)^2 \times H_0^1(\Omega)^2$ and the result is then a direct consequence of Theorem 4.12. \square

The following result provides a practical method for checking whether the compatibility condition is satisfied or not.

Lemma 7.10 (Fortin) *Let consider the spaces $X = H_0^1(\Omega)^2$ and $M = L_0^2(\Omega)$ and let $b \in \mathcal{L}(X, M)$. Suppose there exists $\beta > 0$ such that:*

$$\inf_{q \in M} \sup_{v \in X} \frac{b(v, q)}{\|v\|_X \|q\|_M} \geq \beta.$$

Then, there exists $\beta_h > 0$ such that:

$$\inf_{q_h \in M_h} \sup_{v_h \in X_h} \frac{b(v_h, q_h)}{\|v_h\|_{X_h} \|q_h\|_{M_h}} \geq \beta_h,$$

if and only if there exists $\gamma_h > 0$ such that for every $v \in X$, there exists a restriction operator $r_h \in \mathcal{L}(X, X_h)$, i.e. $r_h(v) \in X_h$, such that:

$$\forall q_h \in M_h, \quad b(v, q_h) = b(r_h(v), q_h) \quad \text{and} \quad \|r_h(v)\|_X \leq \gamma_h \|v\|_X.$$

Proof. If the criterion is satisfied, for every $q_h \in M_h$ we write

$$\begin{aligned} \sup_{v_h \in X_h} \frac{b(v_h, q_h)}{\|v_h\|_X} &\geq \sup_{v \in X} \frac{b(r_h(v), q_h)}{\|r_h(v)\|_X} \\ &\geq \frac{1}{\gamma_h} \sup_{v \in X} \frac{b(v, q_h)}{\|v\|_X} \\ &\geq \frac{\beta}{\gamma_h} \|q_h\|_M. \end{aligned}$$

Conversely, if the discrete condition is satisfied, then the operator $B_h : X_h \rightarrow M_h$ defined by $b(v_h, q_h) = \langle B_h v_h, q_h \rangle$ defines an isomorphism (admitted here). Hence, for every $v \in X$, there exists a unique $v_h = r_h(v)$ such that

$$\forall q_h \in M_h, \quad b(r_h(v), q_h) = b(v, q_h).$$

The mapping $v \mapsto r_h(v)$ is linear and we have:

$$\begin{aligned} \|r_h(v)\|_X &\leq \frac{1}{\beta} \|B_h r_h(v)\|_{M'} \\ &\leq \frac{1}{\beta} \sup_{q_h \in M_h} \frac{b(r_h(v), q_h)}{\|q_h\|_M} \\ &\leq \frac{1}{\beta} \sup_{q_h \in M_h} \frac{b(v, q_h)}{\|q_h\|_M} \\ &\leq \frac{\gamma_h}{\beta} \|v\|_X. \end{aligned}$$

The operator r_h has the desired properties. \square

This proof provides a technique to verify the discrete compatibility condition; it consists in constructing the operator r_h using the solution of the system:

$$b(r_h(v), q_h) = b(v, q_h), \quad \forall q_h \in M_h,$$

and to check if it is bounded in the space X . This is clearly a non-trivial task. On the other hand, it is possible to construct a simpler operator for specific approximation spaces and finite elements and to check if the compatibility condition is satisfied.

It is easy to construct an internal variational approximation of the Stokes problem, as we have already seen:

find $u_h \in X_h$ and $p_h \in M_h$ such that:

$$\begin{cases} a(u_h, v_h) + b(v_h, p_h) = f(v_h), & \forall v_h \in X_h, \\ b(u_h, q_h) = g(q_h), & \forall q_h \in M_h, \end{cases} \quad (7.30)$$

If we denote by n_u (resp. n_p) the dimension of the space X_h (resp. M_h), we introduce the basis $(\varphi_j)_{1 \leq j \leq n_u}$ of X_h and the basis $(\psi_j)_{1 \leq j \leq n_p}$ of M_h defined using the basis shape functions of the finite elements and we decompose u_h and p_h on these basis:

$$u_h = \sum_{j=1}^{n_u} u_h(a_j) \varphi_j(x), \quad \text{and} \quad p_h(x) = \sum_{j=1}^{n_p} p_h(a'_j) \psi_j(x).$$

We obtain the following linear system to solve:

$$\begin{pmatrix} A_h & B_h^t \\ B_h & 0 \end{pmatrix} \begin{pmatrix} U_h \\ P_h \end{pmatrix} = \begin{pmatrix} F_h \\ G_h \end{pmatrix}. \quad (7.31)$$

where $U_h = (u_h(a_j))_{1 \leq j \leq n_u}$ and $P_h = (p_h(a'_j))_{1 \leq j \leq n_p}$ and the matrices $A_h = (a_{ij}) \in \mathbb{R}^{n_u, n_u}$ and $B_h = (b_{ki}) \in \mathbb{R}^{n_p, n_u}$ are given by:

$$a_{ij} = \nu \int_{\Omega} \nabla \varphi_i : \nabla \varphi_j \, dx, \quad \text{and} \quad b_{ki} = - \int_{\Omega} \psi_i \operatorname{div} \varphi_k \, dx.$$

The vectors $F_h = (f_j)_{1 \leq j \leq n_u}$ and $G_h = (g_k)_{1 \leq k \leq n_p}$ are defined as:

$$f_i = \int_{\Omega} f \cdot \varphi_i \, dx, \quad \text{and} \quad g_k = \int_{\Omega} g \psi_k \, dx.$$

Lemma 7.11 *The vector $\mathbf{1} \in \mathbb{R}^{n_p}$ is always contained in the kernel $\operatorname{Ker}(B_h^t)$. The discrete pressure p_h is defined up to an additive constant.*

Proof. Let $r_h \in M_h$ and $w_h \in X_h$. By definition,

$$W_h \cdot B_h^t R_h = B_h W_h \cdot R_h = \int_{\Omega} r_h \operatorname{div} w_h \, dx.$$

Since $r_h = 1$ is in M_h and since

$$\forall w_h \in X_h, \quad \int_{\Omega} \operatorname{div} w_h \, dx = \int_{\partial\Omega} w_h \cdot n \, ds = 0.$$

we conclude that $R_h = \mathbf{1} \in \operatorname{Ker}(B_h^t)$. \square

Finding a pair of approximation spaces (X_h, M_h) for the Stokes problem has been an active research field in numerical analysis. We present here a few examples of such pairs.

The $\mathbb{P}_1/\mathbb{P}_1$ finite element

This seems the obvious first choice for implementing the finite element method for the Stokes problem. We consider the unit domain $\Omega =]0, 1[^2$ and we suppose a regular Cartesian mesh is defined on Ω (Figure 7.9). We introduce the approximation spaces:

$$X_h = \{u_h \in C^0(\bar{\Omega})^2, \forall K \in T_h, u_h|_K \in \mathbb{P}_1^2, u_h|_{\partial\Omega} = 0\},$$

$$M_h = \{p_h \in L_0^2(\Omega) \cap C^0(\bar{\Omega}), \forall K \in T_h, p_h|_K \in \mathbb{P}_1\}.$$

For a given element $K \in T_h$, we denote $(a_{iK})_{1 \leq i \leq 3}$ its vertices. Then, we have:

$$\begin{aligned} \forall v_h \in X_h, \quad \int_{\Omega} p_h \operatorname{div} v_h &= \sum_{K \in T_h} (\operatorname{div} v_h)|_K \int_K p_h, \\ &= \sum_{K \in T_h} (\operatorname{div} v_h)|_K \frac{|K|}{3} \sum_{j=1}^3 p_h(a_{j,K}), \end{aligned}$$

Now, let us consider the pressure field p_h defined on the triangulation T_h by its values 0, -1, 1 at the three vertices of every triangle $K \in T_h$. In this case, we have

$$\sum_{j=1}^3 p_h(a_{j,K}) = 0,$$

and it is easy to see that:

$$\forall v_h \in X_h, \quad \int_{\Omega} p_h \operatorname{div} v_h = 0.$$

This is clearly in contradiction with the requirement of the inf-sup condition. Hence, the $\mathbb{P}_1/\mathbb{P}_1$ finite element shall be discarded for solving the Stokes problem.

When the dimension of $\operatorname{Ker}(B_h^t)$ is strictly larger than one, the finite element method is unstable. In such case, solving the linear system using an iterative method will result in oscillations in the pressure values. On the other hand, if $\dim(\operatorname{Ker}(B_h^t)) = 1$, the discrete pressure can be uniquely determined by fixing its value at a mesh node or by setting its average value on the domain Ω .

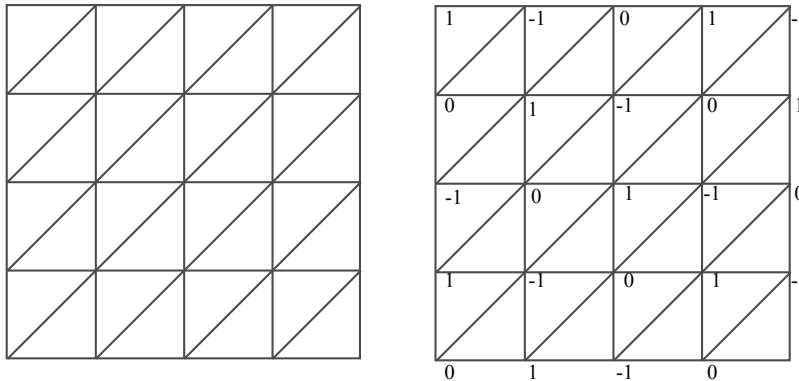


Figure 7.9: A regular Cartesian mesh of the unit domain $\Omega =]0, 1[^2$ and the unstable pressure mode for the $\mathbb{P}_1/\mathbb{P}_1$ finite element.

The $\mathbb{P}_1/\mathbb{P}_0$ finite element

This finite element presents the *a priori* advantage of being very simple to implement. Given a triangulation of the domain Ω , we define a piecewise continuous approximation of the velocity and a piecewise constant (and thus discontinuous) approximation for the pressure. Since the velocity is piecewise affine, its divergence is piecewise constant and thus by testing the divergence using constant functions, the divergence of the discrete field is strongly forced to be null. Hence, this finite element allows to obtain exactly a divergence free velocity field. Nonetheless, the $\mathbb{P}_1/\mathbb{P}_0$ finite element does not fulfill the *inf-sup* condition and Fortin's criterion.

Again, we provide a counter-example in two dimensions. Suppose the triangulation T_h of the simply connected polygonal domain Ω contains t triangles, v_i internal vertices and v_b boundary vertices. We shall have $2v_i$ degrees of freedom for the space X_h , as the velocity must vanish on the boundary and t degrees of freedom for the pressure leading to $t - 1$ independent divergence-free constraints. The Euler formulas for a simple polygon indicate that:

$$t = 2v_i + v_b - 2,$$

from which we deduce that:

$$(t - 1) \geq 2(v_i - 1).$$

We observed that $\dim(M_h) = t - 1$ and that $\dim(X_h) = 2v_i$. The rank theorem allows to write:

$$\begin{aligned} \dim(\text{Ker}(B_h^t)) &= \dim(M_h) - \dim(\text{Im}(B_h^t)) \geq (t - 1) - \dim(X_h) \\ &= (t - 1) - 2v_i = v_b - 3 \end{aligned}$$

This last relation indicates that there is at least $v_b - 3$ spurious pressure modes. In other words, the space M_h is too rich to impose the $B_h u_h = 0$ constraint. A function $u_h \in X_h$ is thus overconstrained and a *locking phenomenon* occurs. In general, $u_h = 0$ is the only discrete divergence-free function of X_h such that $B_h u_h = 0$; *i.e.* B_h is not surjective.

The \mathbb{P}_1 -bubble/ \mathbb{P}_1 (mini) finite element

We have seen that with the $\mathbb{P}_1/\mathbb{P}_1$ finite element, the dimension of the space of the degrees of freedom of the velocity is not large enough. To overcome this problem, it seems natural to attempt enriching the velocity space rather than impoverishing the pressure space. However, it is not necessary to consider a space of polynomials of degree two for approximating the velocity. Crouzeix and Raviart suggested (in 1973) to add one extra degree of freedom to the barycenter of every simplex of the triangulation T_h of the domain Ω . Hence, we define in two dimensions:

$$P_1 = (\mathbb{P}_1(K) \oplus \text{Span}\{b_K, K \in T_h\})^2,$$

with the bubble function b_K defined in two dimensions as:

$$b_K = 27 \prod_{i=1}^3 \lambda_i, \quad \text{and generalized in any dimension as} \quad b_K = (d+1)^{d+1} \prod_{i=1}^{d+1} \lambda_i,$$

where $(\lambda_i)_{1 \leq i \leq 3}$ denote the barycentric coordinates on K . The degrees of freedom associated with P_1 are defined as:

$$\Sigma_k = \{v \mapsto v(a_i), 1 \leq i \leq 3\} \cup \{v \mapsto \int_{e_i} v \cdot n_i d\sigma\}.$$

Then, we consider the approximation spaces:



Figure 7.10: The \mathbb{P}_1 -bubble/ \mathbb{P}_1 (mini) element for velocity (left) and pressure (right) approximation.

$$\begin{aligned} X_h &= \{v_h \in C^0(\bar{\Omega})^2, \forall K \in T_h, v_h|_K \in P_1, u_h|_{\partial\Omega} = 0\}, \\ M_h &= \{p_h \in L_0^2(\Omega) \cap C^0(\bar{\Omega}), \forall K \in T_h, p_h|_K \in \mathbb{P}_1\}. \end{aligned} \quad (7.32)$$

It is easy to see that $X_h \subset (H_0^1(\Omega))^2$.

Before giving a compatibility result for this finite element, we need to introduce an interpolation operator r_h . Since $v \in H_0^1(\Omega)^2$ is not necessarily continuous, its Lagrange interpolate might not be defined. Therefore, in each triangle, we define r_h as:

$$r_h(v) = C_h(v) + \alpha_K b_K,$$

where $C_h(v) = (C_h(v_1), C_h(v_2))$ is the Clément interpolate defined hereafter and $\alpha_K \in \mathbb{R}^2$ is chosen so as to satisfy the equation:

$$\int_K (r_h(v) - v) = 0.$$

To this end, we define:

$$\alpha_K = \left(\int_K b_K \right)^{-1} \int_K (v - C_h(v)).$$

For every $v \in X_h$ and $q_h \in M_h$, we write:

$$\int_{\Omega} q_h \operatorname{div}(v - r_h(v)) = \int_{\Omega} \nabla q_h (r_h(v) - v) = \sum_{K \in T_h} \int_K \nabla q_h (r_h(v) - v) = 0,$$

since ∇q_h is piecewise constant on every $K \in T_h$.

Definition 7.9 Let T_h be a triangulation of Ω , Γ_h be the set of the internal vertices of T_h . We introduce Θ_h the space of continuous piecewise linear functions on T_h which are zero on the boundary and we denote $(\varphi_{\gamma})_{\gamma \in \Gamma_h}$ the basis of \mathbb{P}_1 finite elements with homogeneous boundary conditions. We introduce the operator $C_h \in \mathcal{L}(H_0^1(\Omega), \Theta_h)$, proposed by Clément in 1975, and defined as follows:

$$C_h(v) = \sum_{\gamma \in \Gamma_h} \left(\frac{1}{|\omega_{\gamma}|} \int_{\omega_{\gamma}} v \right) \varphi_{\gamma},$$

where the set ω_{γ} is defined as the union of all triangles $K \in T_h$ for which γ is a vertex.

In this definition, the nodal value $v(\gamma)$ is replaced in the definition of the interpolant r_h by the local average value around γ . The Clément operator can be applied to any function in $L^2(\Omega)$. We have the following result.

Lemma 7.12 If the family of triangulations $(T_h)_{h>0}$ is quasi-uniform, then the operator C_h satisfies the local estimates:

$$\forall v \in H_0^1(\Omega), \quad \|v - C_h(v)\|_{L^2(K)} \leq C h_K |v|_{H^1(\omega_K)},$$

$$\forall v \in H_0^1(\Omega), \quad |C_h(v)|_{H^1(K)} \leq C |v|_{H^1(\omega_K)},$$

where ω_K is the union of all triangles sharing at least one of the vertices of K .

Theorem 7.8 *If the family of triangulations $(T_h)_{h>0}$ is quasi-uniform, the approximation spaces X_h and M_h are uniformly compatible with respect to h , i.e. there exists a constant β_h , independent of h , satisfying the condition (7.27).*

Proof. Let consider $v \in H_0^1(\Omega)^2$. We construct the operator $r_h(v) \in X_h$ such that

$$\forall q_h \in M_h, \quad \int_{\Omega} q_h \operatorname{div}(r_h(v)) = \int_{\Omega} q_h \operatorname{div}(v).$$

Since $M_h \subset H^1(\Omega)$, the previous identity is equivalent to writing:

$$\forall q_h \in M_h, \quad \sum_{K \in T_h} \int_K r_h(v) \cdot \nabla q_h = \sum_{K \in T_h} \int_K v \cdot \nabla q_h.$$

Since $\nabla q_h \in \mathbb{P}_0^d$, the previous identity leads to conclude that:

$$\forall K \in T_h, \quad \int_K r_h(v) = \int_K v.$$

Since the operator r_h is composed of two terms, we analyze them separately. Regarding the Clément interpolate C_h , we invoke the previous estimate:

$$\forall v \in H_0^1(\Omega), \quad |C_h(v)|_{H^1(K)} \leq C |v|_{H^1(\omega_K)}.$$

For the term $\alpha_K b_K$, we write:

$$|\alpha_K b_K|_{H^1(K)} \leq C |\alpha_K|,$$

with the constant C independent of h_K and K . Then, we write

$$\begin{aligned} |\alpha_K| &= \left(\int_K b_K \right)^{-1} \left| \int_K (v - C_h(v)) \right| \\ &\leq C \frac{1}{|K|} \left| \int_K (v - C_h(v)) \right| \\ &\leq |K|^{-1/2} \|v - C_h(v)\|_{L^2(K)}. \end{aligned}$$

We use the estimate $\|v - C_h(v)\|_{L^2(K)} \leq C h_K |v|_{H^1(\omega_K)}$, to obtain:

$$|\alpha_K| \leq C |v|_{H^1(\omega_K)},$$

and finally:

$$|r_h(v)|_{H^1(K)} \leq C |v|_{H^1(\omega_K)},$$

By taking the square of this expression and summing over all $K \in T_h$ we conclude that:

$$\|r_h(v)\|_{H^1(\Omega)} \leq C \|v\|_{H^1(\Omega)}.$$

The final result is obtained by invoking Fortin's lemma. \square

Proposition 7.7 *Suppose the solution (u, p) of the problem (7.24) is sufficiently smooth, i.e. $u \in H^2(\Omega)^2 \cap H_0^1(\Omega)^2$ and $p \in H^1(\Omega) \cap L_0^2(\Omega)$. Then, the solution (u_h, p_h) of the problem (7.26) with the spaces defined by (7.32) is such that:*

$$\|u - u_h\|_{H^1(\Omega)} + \|p - p_h\|_{L^2(\Omega)} \leq C h (\|u\|_{H^2(\Omega)} + \|p\|_{H^1(\Omega)}). \quad (7.33)$$

7.4.3 Numerical resolution

We have seen that the discrete problem (7.30) leads to solve the linear system (7.31). We assume here the natural hypothesis that the bilinear form $a(\cdot, \cdot)$ is V -elliptic on X with a constant α and that the form $b(\cdot, \cdot)$ satisfies on $x_h \times M_h$ a *inf-sup* condition uniformly in h , and we denote the constant β .

The resolution of the linear system using a direct method is likely to be costly because of the number of degrees of freedom and thus the size of the linear system. Direct methods are appropriate up to $N_{dof} = 10^4$. For this reason, iterative methods are favored. However, the matrix of the system (7.31) is neither positive nor definite. It is indeed an *indefinite* matrix (but symmetric) with that zero block on the diagonal. It will have positive and also negative eigenvalues. It is possible to make it positive by solving the system:

$$\begin{cases} A_h U + B_h^t P = F_h \\ -B_h U = -G_h \end{cases} . \quad (7.34)$$

However, the new matrix corresponding to this system is still indefinite and is no longer symmetric. We know from Chapter 5 that gradient methods are not efficient on this type of matrix. We present two commonly used methods for solving such mixed problems.

Resolution of the saddle point by penalization

The following method is gaining popularity for solving the Stokes problem (and saddle point problems in general) because of its versatility and facility to be implemented. It consists in replacing the problem (7.31) by the following problem:

$$\begin{cases} A_h U_\varepsilon + B_h^t P_\varepsilon = F_h \\ -B_h U_\varepsilon + \varepsilon S_h P_\varepsilon = -G_h \end{cases} , \quad (7.35)$$

where the matrix $S_h \in \mathbb{R}^{n_p, n_p}$ is such that:

$$(S_h P, Q) = \langle p_h, q_h \rangle_M ,$$

where (\cdot, \cdot) and $\langle \cdot, \cdot \rangle_M$ denote the standard Euclidean scalar product in \mathbb{R}^{n_p} and the inner product in M , respectively. In the system (7.35), ε is a penalization coefficient sufficiently small.

The main idea is simple. Suppose A_h is a positive definite symmetric matrix, the system (7.31) is then equivalent to:

$$\inf_{B_h V_h = G_h} \left\{ \frac{1}{2} (A_h V_h, V_h) - (F_h, V_h) \right\} .$$

If S_h is any positive definite matrix in \mathbb{R}^{n_p, n_p} , the previous expression can be replaced by:

$$\inf_{V_h} \left\{ \frac{1}{2} (A_h V_h, V_h) + \frac{1}{2\varepsilon} (S_h^{-1} (B_h V_h - G_h), B_h V_h - G_h) - (F_h, V_h) \right\} ,$$

or equivalently, by writing $P_\varepsilon = (1/\varepsilon) S_h^{-1} (B_h V_h - G_h)$:

$$\begin{cases} A_h U_\varepsilon + B_h^t P_\varepsilon = F_h \\ B_h U_\varepsilon - G_h = \varepsilon S_h P_\varepsilon \end{cases} . \quad (7.36)$$

By replacing the quantity P_ε in the first equation, we obtain the linear system:

$$\left(A_h + \frac{1}{\varepsilon} B_h^t S_h^{-1} B_h \right) U_\varepsilon = F_h + \frac{1}{\varepsilon} B_h^t S_h^{-1} G_h . \quad (7.37)$$

If the matrix S_h^{-1} is a sparse matrix, this provides an efficient technique to solve our problem. This system can be solved by an iterative technique like the conjugate gradient, since the matrix is now positive definite and symmetric. Notice however that the term $1/\varepsilon B_h^t S_h^{-1} B_h$ has a strong negative impact on the condition number of the linear system (7.4.3).

We now investigate the effect of changing the constrained problem to a penalized problem. We have the following error estimate.

Proposition 7.8 *Let consider $\varepsilon > 0$ and let (U, P) and $(U_\varepsilon, P_\varepsilon)$ denote the solutions of system (7.34) and (7.36), respectively. Then, we have the error estimate:*

$$\alpha\beta\|U - U_\varepsilon\|_X + \alpha\beta^2\|P - P_\varepsilon\|_M \leq C\varepsilon\|P\|_M. \quad (7.38)$$

Proof. By difference, we obtain the system:

$$\begin{cases} A_h(U - U_\varepsilon) + B_h^t(P - P_\varepsilon) = 0 \\ -B_h(U - U_\varepsilon) - \varepsilon S_h P_\varepsilon = 0 \end{cases}.$$

We introduce on \mathbb{R}^{n_u} the norm $\|\cdot\|_*$ defined by:

$$\forall U \in \mathbb{R}^{n_u}, \quad \|U\|_* = \sup_{V \in \mathbb{R}^{n_u}} \frac{(U, V)}{\|V\|_X}.$$

The continuity and the V -ellipticity of the bilinear form a implies that:

$$\forall U \in \mathbb{R}^{n_u}, \quad \|A_h U\|_* \leq \|a\| \|U\|_X, \quad \text{and} \quad \forall U \in \mathbb{R}^{n_u}, \quad \alpha \|U\|_X \leq \|A_h U\|_*.$$

The matrix $B_h \in \mathbb{R}^{n_p, n_p}$ satisfies the *inf-sup* condition:

$$\min_{\|P\|_M \neq 0} \max_{\|U\|_X \neq 0} \frac{(B_h^t P, U)}{\|P\|_M \|U\|_X} \geq \beta,$$

or equivalently:

$$\forall P \in \mathbb{R}^{n_p}, \quad \beta \|P\|_M \leq \|B_h^t P\|_*.$$

The continuity of the form b implies that:

$$\forall P \in \mathbb{R}^{n_p}, \quad \|B_h^t P\|_* \leq \|b\| \|P\|_M.$$

Using these inequalities leads to write the first equation of the system as:

$$\begin{aligned} \|P - P_\varepsilon\|_M &\leq \frac{1}{\beta} \|B_h^t(P - P_\varepsilon)\|_* \\ &= \frac{1}{\beta} \|A_h(U - U_\varepsilon)\|_* \leq \frac{C}{\beta} \|U - U_\varepsilon\|_X. \end{aligned}$$

By multiplying the first equation by $(U - U_\varepsilon)$ and using the V -ellipticity of the form a we obtain for the second equation:

$$\begin{aligned} \alpha \|U - U_\varepsilon\|_X^2 &\leq (A_h(U - U_\varepsilon), U - U_\varepsilon) = (B_h^t(P_\varepsilon - P), U - U_\varepsilon) = (P_\varepsilon - P, B_h(U - U_\varepsilon)) = -\varepsilon(P_\varepsilon - P, S_h P_\varepsilon) \\ &= -\varepsilon(P_\varepsilon - P, S_h(P_\varepsilon - P)) - \varepsilon(P_\varepsilon - P, S_h P) \\ &\leq -\varepsilon(P_\varepsilon - P, S_h P) \\ &\leq \varepsilon \|P_\varepsilon - P\|_M \|P\|_M. \end{aligned}$$

The error estimate is obtained by combining the two previous inequalities. \square

Remark 7.16 *Notice that discretizing a penalized problem is not necessarily equivalent to penalize a discrete problem [Brezzi-Fortin, 1991]. The penalty method is considered in this last case as a procedural (algorithmic) technique, since a choice of spaces $X_h \subset X$ and $M_h \subset M$ has been done. On the other hand, discetizing the penalized problem consists in choosing $Q_h = B_h V_h$ that is in general a poor choice.*

Uzawa's operator

To conclude this section, we introduce an iterative, although very efficient, algorithm to solve such problem when the operator A_h associated with the bilinear form a is invertible. We know since Chapter 4 that the Stokes equations are equivalent to solving a problem of energy minimization. More precisely, the resolution of the linear system (7.31) is equivalent to the following minimization:

$$J(U_h) = \min_{V_h \in \text{Ker} B_h} J(V_h) \quad \text{with} \quad J(V_h) = \frac{1}{2}(A_h V_h, V_h) - (F_h, V_h). \quad (7.39)$$

In the optimization terminology, the pressure is the Lagrange multiplier that imposes the constraint $\text{div } u = g$ when the energy is minimized.

First, we eliminate the variable U_h from the linear system (7.31), if the matrix A_h is invertible, thus leading to:

$$U_h = A_h^{-1}(F_h - B_h^t P)$$

and then by replacing U_h in the second equation $B_h U_h = G_h$, we obtain the system:

$$B_h A_h^{-1} B_h^t P = B_h A_h^{-1} F_h - G_h. \quad (7.40)$$

We consider the *Uzawa matrix* $\mathcal{U} = B_h A_h^{-1} B_h$.

Proposition 7.9 *If matrix A_h is positive definite, this matrix is also positive definite.*

Proof. One has:

$$(B_h A_h^{-1} B_h^t P, P) = (A_h^{-1} B_h^t P, B_h^t P) \geq \alpha \|B_h^t P\|.$$

It is positive definite if $\text{Ker} B_h^t = \{0\}$. □

Hence, problem (7.40) is more easy to solve than the original problem (7.31) as efficient methods exist for positive definite systems. Unfortunately, even if the matrix A_h is a sparse matrix (*i.e.* contains lot of zero values), its inverse A_h^{-1} is likely to be a full matrix. The resolution of (7.40) must be performed using an iterative method since inverting the matrix A_h^{-1} is extremely computationally costly. The convergence rate of iterative methods is strongly related to the condition number of the matrix.

Proposition 7.10 *If the matrix \mathcal{U} is symmetric, we have the estimate:*

$$\kappa(\mathcal{U}) \leq \frac{C}{\alpha^2 \beta^2} \kappa(S_h).$$

Remark 7.17 *If the family of mesh $(T_h)_{h>0}$ is quasi-uniform, the condition number $\kappa(S_h)$ is a constant independent of h . Hence, the previous estimate shows that the conditioning of Uzawa's operator is of order one (if α and β are of order one, which is a reasonable hypothesis). The iterative techniques converge fast without any preconditioning.*