Chapter 7

The finite element method

One of the major interests of variational methods is to provide both a theory for existence of solutions and numerical methods for computing accurate approximations of these solutions. Certainly, the most celebrated of these variational approximation methods is the finite element method. It is a Galerkin approximation scheme where the elements of the finite dimensional approximating subspaces V_n are piecewise polynomial functions. This method has been proved to be very successful, the main reason being that, because of the local character of several problems, by choosing a basis of the space V_n whose functions have small supports, one obtains approximated problems with sparse matrices, i.e., with most entries equal to zero. This is a decisive property in order to be able to solve numerically the corresponding linear system: one should notice that engineering problems involving systems of PDEs from continuum mechanics usually give rise to large linear systems (100×100 or 1000×1000 are quite frequent!).

Our scope is to introduce the main ideas of the finite element method and then to describe a typical example.

For simplicity of the exposition, we restrict ourselves to linear problems whose variational formulation enters into the abstract setting of the Lax–Milgram theorem: find $u \in V$ such that

$$a(u,v) = l(v) \quad \forall v \in V. \tag{7.1}$$

Let us first recall and make precise some aspects of the Galerkin method (which was introduced in Section 3.1.2).

7.1 ▪ The Galerkin method: Further results

Let us briefly recall the assumptions on the abstract variational problem (7.1): V is a Hilbert space, and $a : V \times V \to \mathbf{R}$ is a continuous coercive bilinear form, i.e., there exists some constants $M \in \mathbf{R}^+$ and $\alpha > 0$ such that

$$\forall u,v \in V, \quad |a(u,v)| \leq M\|u\|\|v\|, \tag{7.2}$$

$$\forall v \in V, \quad a(v,v) \geq \alpha\|v\|^2. \tag{7.3}$$

The linear form $l : V \to \mathbf{R}$ is supposed to be continuous. Then the Lax–Milgram theorem asserts the existence and uniqueness of a solution $u \in V$ of problem (7.1).

Typically, as in the boundary value problems which were studied in the previous sections, V is a Sobolev space, like $H^1(\Omega)$ or $H^1_0(\Omega)$. It is an infinite dimensional space; this is a common feature of all methods from functional analysis, and the numerical computation of the solution u requires a further step, which is the reduction to a finite dimensional problem.

The Galerkin method is based on the approximation of the infinite dimensional space V by a sequence of finite dimensional subspaces $(V_n)_{n\in\mathbf{N}}$. More precisely, for each $n \in \mathbf{N}$, V_n is a finite dimensional subspace of V and one supposes that the following approximation property holds:

$$\forall v \in V,\ \exists (v_n)_{n\in\mathbf{N}},\ v_n \in V_n\ \forall\, n \in \mathbf{N}, \text{ and } v_n \to v \text{ in } V. \tag{7.4}$$

For each $n \in \mathbf{N}$ the approximated variational problem is

$$\begin{cases} \text{find } u_n \in V_n \text{ such that} \\ a(u_n, v) = l(v)\ \forall\, v \in V_n. \end{cases} \tag{7.5}$$

One should notice that the approximated problem (7.5) is still a variational problem: it has the same structure as the initial variational problem (7.1), except now it is posed on a finite dimensional space V_n. Indeed, existence and uniqueness of the solution u_n of (7.5) follows, in a similar way, from the Lax–Milgram theorem. When a is symmetric, problem (7.5) reduces to a minimization problem, namely, u_n is the solution of

$$\min\left\{ \frac{1}{2}a(v,v) - l(v) : v \in V_n \right\}. \tag{7.6}$$

This alternate description of the finite dimensional approximation method is often called the Ritz method.

The term *variational approximation* is justified by the fact that the sequence of problems (7.5) does approximate the initial problem (7.1), in the sense that the sequence $(u_n)_{n\in\mathbf{N}}$ norm converges in V to u. More precisely, in Proposition 3.1.2 it was proved that

$$\|u - u_n\| \leq \frac{M}{\alpha}\,\mathrm{dist}(u, V_n), \tag{7.7}$$

and one can observe that, clearly, the approximation property (7.4) implies that for any $u \in V$, $\mathrm{dist}(u, V_n) \to 0$ as $n \to +\infty$.

Let us now make precise the structure of the approximated problem (7.5). Let us introduce a basis $(\varphi_1, \varphi_2, \dots, \varphi_{I(n)})$ of the vector space V_n with $I(n) = \dim V_n$. Let us write $u_n = \sum_{i=1}^{I(n)} \lambda_i \varphi_i$. Then (7.5) is equivalent to

$$\begin{cases} a(u_n, \varphi_j) = l(\varphi_j) \quad \forall\, j = 1, 2, \dots, I(n), \\ u_n = \displaystyle\sum_{i=1}^{I(n)} \lambda_i \varphi_i. \end{cases}$$

This, in turn, is equivalent to finding $\lambda = (\lambda_i)_{i=1,2,\dots,I(n)}$ in $\mathbf{R}^{I(n)}$ which is a solution of the linear system

$$\sum_{i=1}^{I(n)} \lambda_i a(\varphi_i, \varphi_j) = l(\varphi_j) \qquad \forall\, j = 1, 2, \dots, I(n). \tag{7.8}$$

Let us set $A_n = (a(\varphi_j, \varphi_i))_{1 \leq i,j \leq I(n)}$. It is an $I(n) \times I(n)$ square matrix which, by reference to the elasticity problem, is often called the stiffness matrix. Similarly, the vector $b_n = (l(\varphi_j))$ in $\mathbf{R}^{I(n)}$ is often called the load vector.

With this notation, one can write (7.8) in the following form:

$$A_n\lambda = b_n. \tag{7.9}$$

Let us now examine the properties of the matrix A_n: for any vector $\lambda \in \mathbf{R}^{I(n)}$ we have ($\langle\cdot,\cdot\rangle$ is the Euclidean scalar product in $\mathbf{R}^{I(n)}$ and $|\cdot|$ the Euclidean norm)

$$\begin{aligned}\langle A_n\lambda,\lambda\rangle &= \sum_{i=1}^{I(n)}(A_n\lambda)_i\lambda_i\\ &= \sum_{i=1}^{I(n)}\left(\sum_{j=1}^{I(n)}a(\varphi_j,\varphi_i)\lambda_j\right)\lambda_i\\ &= a\left(\sum_{j=1}^{I(n)}\lambda_j\varphi_j,\sum_{i=1}^{I(n)}\lambda_i\varphi_i\right)\\ &\geq \alpha\left\|\sum_{i=1}^{I(n)}\lambda_i\varphi_i\right\|^2.\end{aligned}$$

Since $(\varphi)_{i=1}^{I(n)}$ is a basis of V_n, one can easily verify that $\lambda \mapsto \|\sum_{i=1}^{I(n)}\lambda_i\varphi_i\|$ is a norm on $\mathbf{R}^{I(n)}$. All norms being equivalent on the finite dimensional space $\mathbf{R}^{I(n)}$, there exists some constant $c > 0$ such that

$$\forall\lambda\in\mathbf{R}^{I(n)} \qquad c|\lambda| \leq \left\|\sum_{i=1}^{I(n)}\lambda_i\varphi_i\right\|.$$

Hence

$$\forall\lambda\in\mathbf{R}^{I(n)} \qquad \langle A_n\lambda,\lambda\rangle \geq \alpha c^2|\lambda|^2. \tag{7.10}$$

From (7.10) it follows that A_n is one to one (that is, $\ker(A_n) = \{0\}$), which, in this finite dimensional setting, implies that for any load vector b_n, problem (7.9) has a unique solution. This is another elementary way (without using the Lax–Milgram theorem) to prove existence and uniqueness of the solution λ of the approximated problem (7.9). Let us also notice that when a is symmetric, so is the matrix A_n.

We now come to the central point of this theory, which is the effective construction of the finite dimensional approximating subspaces V_n and the resolution of the linear system (7.9).

As we have already stressed, it is crucial, from a numerical point of view, that the matrix A_n possesses as many zeroes as possible. At this point, there are different strategies: in the next section, we shall describe an approach which uses spectral analysis in infinite dimensional spaces and a special basis whose elements are eigenfunctions.

The finite element method relies on a different strategy that we now describe.

7.2 ▪ Description of finite element methods

Let us now assume that V is a closed subspace of $H^1(\Omega)$ and the bilinear form $a : V \times V \to \mathbf{R}$ is of the type

$$a(u,v) = \int_\Omega(\nabla u\cdot\nabla v + a_0uv)\,dx \tag{7.11}$$

with $a_0 \in L^\infty(\Omega)$, $a_0 \geq 0$. This allows us to cover various situations like Dirichlet, Neumann, and mixed problems which were studied in the previous sections. A key property of a is that it is a *local* bilinear form, that is,

$$a(\varphi, \psi) = 0,$$

as soon as φ and ψ are two elements of V whose supports do not intersect, or more generally such that the Lebesgue measure of the intersection of their supports is zero. Let us recall that the stiffness matrix A_n is equal to $(a(\varphi_i, \varphi_j))$, where (φ_i), $i = 1, \dots, I(n)$, is a basis of V_n. The strategy is now clear: we have to choose V_n such that one can find a canonical basis in the space V_n whose corresponding functions have supports which are as small as possible.

This is made possible thanks to a *triangulation* of the set $\overline{\Omega}$. For simplicity of the exposition, we restrict ourselves to problems which are posed over sets $\overline{\Omega} \subset \mathbf{R}^2$ which are polyhedra; we also say that such set Ω is polygonal.

Definition 7.2.1. *A triangulation $\mathcal{T}$ of a polygonal set Ω of $\mathbf{R}^2$ is a finite decomposition of the set $\overline{\Omega}$ of the form*

$$\overline{\Omega} = \bigcup_{K \in \mathcal{T}} K$$

such that

(i) *each set $K \in \mathcal{T}$ is a triangle,*

(ii) *whenever K_1 and K_2 belong to $\mathcal{T}$, $K_1 \cap K_2$ is either empty or reduced to a common vertex or to a common face (edge).*

In particular, for each distinct $K_1, K_2 \in \mathcal{T}$, one has $\text{int}(K_1) \cap \text{int}(K_2) = \emptyset$. *Two triangles K_1 and K_2 which have a common face are said to be adjacent. The triangles $K \in \mathcal{T}$ are called finite elements.*

We set

$$h(\mathcal{T}) = \max_{K \in \mathcal{T}} \text{diam} K, \tag{7.12}$$

where $\text{diam} K = \sup\{|x - y| \ : \ x, y \in K\}$ is the diameter of K. By convention, we denote by $\mathcal{T}_h$ a triangulation $\mathcal{T}$ such that $h(\mathcal{T}) = h$. As we shall see, to each triangulation $\mathcal{T}_h$ will be associated a finite approximating dimensional subspace V_h. It is convenient to consider the family of approximating subspaces indexed by the positive parameter h, say, V_h with $h \to 0$. Of course, to reduce to an abstract Galerkin scheme as described in Section 7.1, one may take $V_h = V_{h_n}$ for some $h_n \to 0$. An example of triangulation is given in Figure 7.1.

One may already observe that one can use a fine triangulation in a subregion where a particular behavior of the solution is expected (for example, in special parts of airplanes). Figure 7.2 shows a forbidden situation where the intersection of two triangles K_1 and K_2 is not an edge of K_2.

Let us now describe the finite dimensional space V_h which is associated to a triangulation $\mathcal{T}_h$. At this point, we need to make precise the boundary condition; take, for example, the Dirichlet boundary condition $u = 0$ on $\partial\Omega$ and $V = H_0^1(\Omega)$. Then

$$V_h = \{v \in \mathbf{C}(\overline{\Omega}) : v \text{ is affine on each } K \in \mathcal{T}_h, \ v = 0 \text{ on } \partial\Omega\}.$$

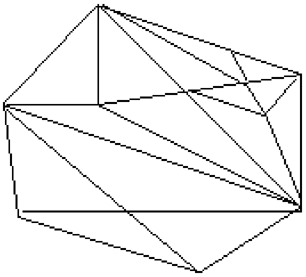

Figure 7.1. *Example of triangulation.*

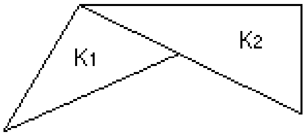

Figure 7.2. *A forbidden situation.*

In other words, V_h is the linear space of continuous functions on $\overline{\Omega}$ which are piecewise linear with respect to the triangulation $\mathcal{T}_h$ and which vanish on the boundary.

One can easily verify that an affine function on a triangle K is uniquely determined by its values at the vertices of K. Hence, any function $v \in V_h$ is uniquely determined by its values at the vertices (also called nodes) of the triangulation which are in the interior of $\overline{\Omega}$, i.e., in Ω. (On the vertices which are on the boundary $\partial\Omega$, v is prescribed to be equal to zero.)

For any vertex a_i of the triangulation, $i = 1, 2, \dots, I(h)$, which is in the interior set Ω, let us denote by φ_i the element of V_h which satisfies

$$\varphi_i(a_j) = \delta_{ij}, \qquad 1 \le i, j \le I(h).$$

Equivalently, φ_i is the function of V_h which is equal to one at the vertex a_i and is equal to zero at all other vertices a_j with $j \ne i$. It is usually called a hat function. Clearly, $(\varphi_1, \varphi_2, \dots, \varphi_{I(h)})$ is a basis of V_h and each element v of V_h can be uniquely written in the form

$$v = \sum_{i=1}^{I(h)} v(a_i)\varphi_i.$$

One should notice that each element φ_i of this basis has small support: more precisely, the support of φ_i is the union of all triangles K of $\mathcal{T}_h$ such that a_i is a vertex of K. The stiffness matrix $A_h = (a(\varphi_i, \varphi_j))_{1 \le i,j \le I(h)}$ is a sparse matrix, since $a(\varphi_i, \varphi_j) = 0$ except when a_i and a_j are two vertices of a same triangle K of the triangulation $\mathcal{T}_h$.

Let us now stress a technical but important point: the structure of A_h, i.e., the distribution of zeroes, for a given triangulation $\mathcal{T}_h$ highly depends on the enumeration of the vertices. Clearly, one has to use an enumeration to obtain A_h with a simple structure like, for example, tridiagonal matrices. Let us illustrate this in a concrete situation.

7.3 ▪ An example

Take $\Omega = (0,1) \times (0,1)$ the unit square in $\mathbf{R}^2$ and, given $f \in L^2(\Omega)$, let us consider the Dirichlet boundary value problem

$$\begin{cases} -\Delta u = f & \text{on } \Omega, \\ u = 0 & \text{on } \partial\Omega. \end{cases}$$

Its variational formulation is as follows: find $u \in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx \qquad \forall\, v \in H_0^1(\Omega).$$

Let us consider the triangulation of Ω in Figure 7.3.

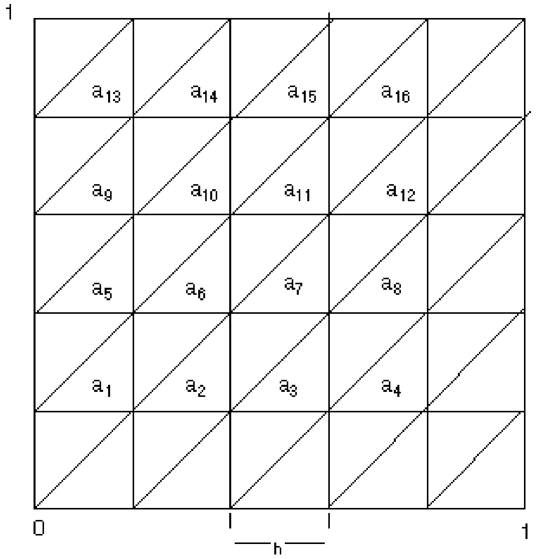

Figure 7.3. *Triangulation of Ω with indexed nodes.*

We denote by $a_{l,m} = (lh, mh)$, $0 \le l, m \le N+1$, the nodes of the triangulation. There are N^2 nodes which are in Ω, and the dimension of V_h is equal to N^2 with $h = 1/(N+1)$. Let us index the nodes of the triangulation as indicated in Figure 7.3, where N has been taken equal to 4. We draw, for example, the perspective of the hat function φ_6 which is an element of the finite element basis (Figure 7.4).

Recall that $a(\varphi_i, \varphi_j) \neq 0$ iff a_i and a_j are two vertices of the same triangle. Then, notice that a vertex a_i is connected to at most six other vertices a_j with $j \neq i$, which, a priori, yields a seven-point numerical scheme. The matrix A_h then has the following structure, where each cross represents an element $a(\varphi_i, \varphi_j)$ which is, a priori, not equal to zero.

It is a matrix with a band structure, that is,

$$a(\varphi_i, \varphi_j) = 0 \text{ for } |i - j| > d_{max},$$

where d_{max}, the width of the band, is small with respect to the size of the matrix.

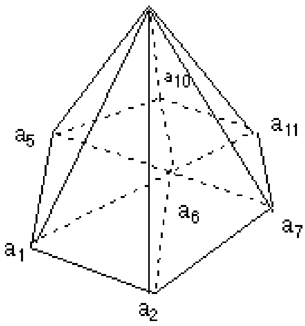

Figure 7.4. *The hat function* φ_6.

Indeed, for symmetry reasons

$$a(\varphi_{l,m},\varphi_{l+1,m+1})=0 \text{ and } a(\varphi_{l,m},\varphi_{l-1,m-1})=0$$

and we have a five-point scheme! Let $\lambda_{l,m}=u_h(\varphi_{l,m})$ be the component of u_h with respect to $\varphi_{l,m}$ so that $u_h=\sum\lambda_{l,m}\varphi_{l,m}$. An elementary computation yields

$$\begin{cases}-\lambda_{l-1,m}-\lambda_{l,m-1}+4\lambda_{l,m}-\lambda_{l,m+1}-\lambda_{l+1,m}=h^2 f_{l,m}, & 1\le l,m\le N,\\ \lambda_{l,0}=\lambda_{l,N+1}=0, \quad 0\le l\le N+1,\\ \lambda_{0,m}=\lambda_{N+1,m}=0, \quad 0\le m\le N+1.\end{cases}$$

This is the classical five-point scheme for the Laplacian. (Here $f_{l,m}=\int_\Omega f\varphi_{l,m}\,dx$ or, equivalently, an approximation of this integral.)

Remark 7.3.1. It is worth noticing that the tridiagonal block structure of A_h is intimately related to the enumeration of the elements of the basis. This structure may be lost by choosing a different enumeration!

7.4 ▪ Convergence of the finite element method

The convergence of the finite element method, which is a Galerkin approximation method, relies on Proposition 3.1.2:

$$||u-u_h||_{H_0^1(\Omega)}\le\frac{M}{\alpha}\,\mathrm{dist}(u,V_h).$$

Therefore, the estimate of the error $||u-u_h||$ (and showing that the error goes to zero as $h\to 0$) can be reduced to a problem in approximation theory: one has to evaluate (majorize) the distance for the $H_0^1(\Omega)$ norm between a function $u\in H_0^1(\Omega)$ and the subspace V_h of continuous functions which are piecewise affine relative to a given triangulation $\mathcal{T}_h$.

To that end, we need to make a geometrical assumption on the family of triangulations $(\mathcal{T}_h)_{h\to 0}$.

Definition 7.4.1. *A family of triangulations* $(\mathcal{T}_h)_{h>0}$ *is said to be regular if there exists a constant* σ *(*$\sigma\ge 0$*) such that for any* $h>0$ *and any* $K\in\mathcal{T}_h$

$$\frac{h_K}{\rho_K}\le\sigma, \tag{7.13}$$

where h_K is the diameter of K and ρ_K is the supremum of the diameters of the balls contained in K.

It can be easily shown that this condition is equivalent to the following. There exists a constant $\theta_0 > 0$ such that for any $h > 0$ and for any $K \in \mathcal{T}_h$,

$$\theta_K \geq \theta_0,$$

where θ_K denotes the smallest angle of the triangle K.

Thus the regularity of a family of triangulations $(\mathcal{T}_h)_{h>0}$ in the sense of Definition 7.4.1 prevents the triangles from becoming "flat" in the limit when $h \to 0$.

As we shall see, this is a key assumption to obtain the convergence of the method. For example, the situation with $\varepsilon_h \to 0$ in Figure 7.5 is not allowed in the context of a regular family of triangulation (as $\varepsilon_h \to 0$, K_h becomes flat).

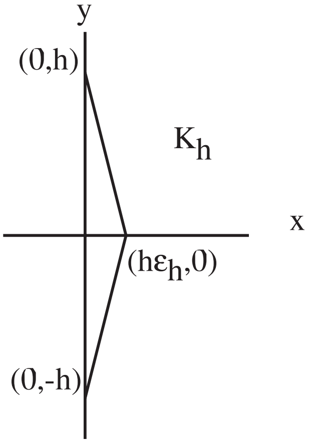

Figure 7.5. *A triangle K_h becoming flat.*

We shall return later to this example and show that in such a situation some of the following mathematical developments fail to be true.

The main result of this section, which is the convergence of the finite element method under the regularity assumption (7.13), is given below.

Theorem 7.4.1. *Let Ω be a polygon and let $(\mathcal{T}_h)_{h\to 0}$ be a regular family of triangulations of Ω. Then, the finite element method converges, i.e.,*

$$\lim_{h\to 0} ||u - u_h||_{H_0^1(\Omega)} = 0.$$

Moreover, if u belongs to $H^2(\Omega)$, the following estimate holds: there exists some constant $C > 0$ such that for all $h > 0$

$$||u - u_h||_{H_0^1(\Omega)} \leq Ch||u||_{H^2(\Omega)}.$$

PROOF. (a) Let us first assume that $u \in H^2(\Omega)$. Recalling that $N = 2$, by Sobolev embedding theorems (see Section 5.7) we have $u \in \mathbf{C}(\overline{\Omega})$. Indeed this is true under the assumption $N \leq 3$. This allows us to talk about the value of u at any point of $\overline{\Omega}$, and especially at the nodes $(a_i)_{i=1,\dots,I(h)}$ of the triangulation $\mathcal{T}_h$. Let us introduce the function $\Pi_h(u)$ which is the continuous affine interpolant of u at the nodes of $\mathcal{T}_h$:

$$\Pi_h(u)=\sum_{i=1}^{I(h)} u(a_i)\varphi_i.$$

Recall that φ_i is the hat function related to the node a_i and that the $(\varphi_i)_{i=1,\dots,I(h)}$ form a basis of V_h. The above formula is just the linear decomposition of $\Pi_h(u)$ in the basis $(\varphi_i)_{i=1,\dots,I(h)}$. Since $\Pi_h(u)\in V_h$, by definition of dist(u,V_h), we have

$$\operatorname{dist}(u,V_h)\le ||u-\Pi_h(u)||_{H_0^1(\Omega)}.$$

This inequality, when combined with Proposition 3.1.2 (Cea's lemma) yields

$$||u-u_h||_{H_0^1(\Omega)}\le \frac{M}{\alpha}||u-\Pi_h(u)||_{H_0^1(\Omega)}. \tag{7.14}$$

Let us now use the following approximation result that we admit for the moment (we shall return to this crucial result further): there exists a constant C independent of h such that for all $u\in H^2(\Omega)$

$$||u-\Pi_h(u)||_{H^1(\Omega)}\le Ch||u||_{H^2(\Omega)}. \tag{7.15}$$

Let us notice that this estimate makes use in an essential way of the regularity assumption (7.13) on the family of triangulations $(\mathcal{T}_h)_{h\to 0}$ and of the fact that $u\in H^2(\Omega)$.

Let us now combine (7.14) and (7.15) to obtain

$$||u-u_h||_{H_0^1(\Omega)}\le \frac{M}{\alpha}Ch||u||_{H^2(\Omega)}. \tag{7.16}$$

Thus, in the case $u\in H^2(\Omega)$, we have convergence of the finite element method, that is, norm convergence in $H_0^1(\Omega)$ of the sequence $(u_h)_{h\to 0}$ to u as $h\to 0$. More precisely, the estimate (7.16) provides information about the rate of convergence of the method.

(b) In the general case, that is, $u\in H_0^1(\Omega)$, one completes the proof by a density argument: for any $\varepsilon>0$ let us introduce some $v_\varepsilon\in\mathcal{D}(\Omega)=\mathbf{C}_c^\infty(\Omega)$ such that

$$||u-v_\varepsilon||_{H_0^1(\Omega)}<\varepsilon. \tag{7.17}$$

Since $v_\varepsilon\in\mathcal{D}(\Omega)\subset H^2(\Omega)$, for each $\varepsilon>0$ we can use the previous argument and, by (7.15), we have

$$||v_\varepsilon-\Pi_h(v_\varepsilon)||_{H_0^1(\Omega)}\le Ch||v_\varepsilon||_{H^2(\Omega)}. \tag{7.18}$$

Let us now write the triangle inequality

$$||u-\Pi_h(v_\varepsilon)||_{H_0^1(\Omega)}\le ||u-v_\varepsilon||_{H_0^1(\Omega)}+||v_\varepsilon-\Pi_h(v_\varepsilon)||_{H_0^1(\Omega)}$$

and use inequalities (7.17) and (7.18) to obtain

$$||u-\Pi_h(v_\varepsilon)||_{H_0^1(\Omega)}\le \varepsilon+Ch||v_\varepsilon||_{H^2(\Omega)}.$$

Since $\Pi_h(v_\varepsilon)\in V_h$, this implies

$$\operatorname{dist}(u,V_h)\le \varepsilon+Ch||v_\varepsilon||_{H^2(\Omega)}.$$

Hence

$$\limsup_{h\to 0} \operatorname{dist}(u, V_h) \leq \varepsilon.$$

This being true for any $\varepsilon > 0$, we finally obtain

$$\lim_{h\to 0} \operatorname{dist}(u, V_h) = 0,$$

which, by Proposition 3.1.2 (Cea's lemma), implies the norm convergence in $H_0^1(\Omega)$ of u_h to u as $h \to 0$. □

Let us now give the proof of the piecewise affine interpolation inequality (7.15) which, together with the abstract Cea's lemma, is the key ingredient of the proof of Theorem 7.4.1. Because of its importance and its own interest let us state it independently.

Theorem 7.4.2. *Let Ω be a polygon and $(\mathcal{T}_h)_{h\to 0}$ a regular family of triangulations of Ω (i.e., (7.13) is supposed to be satisfied). Then, there exists a constant C, which is independent of h, such that for any $u \in H^2(\Omega)$*

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

We recall that $\Pi_h(u)$ is the piecewise affine interpolant of u relative to $\mathcal{T}_h$.

For pedagogical reasons, it is worthwhile to first prove Theorem 7.4.2 in one dimension, i.e., $\Omega = (a, b)$ is an interval of **R**. Indeed, the role of the norms $H^1(\Omega)$ and $H^2(\Omega)$ and of the assumption $u \in H^2(\Omega)$ already appear quite naturally in this situation, and the proof just requires elementary tools. We shall then consider the two-dimensional case and show how, in that case, one has to do some geometrical assumptions on $\mathcal{T}_h$. (This is where the regularity assumption (7.13) on $\mathcal{T}_h$ plays a central role.)

Proof of Theorem 7.4.2 in the one-dimensional case. Let $\Omega = (a, b)$ be an interval of **R** with $-\infty < a < b < +\infty$. Let

$$a = a_0 < a_1 < a_2 < \cdots < a_n = b$$

be a discretization of Ω, and set $h = \max_i |a_{i+1} - a_i|$.

(a) Let us first assume that u is smooth, say, $u \in \mathbf{C}^\infty([a, b])$, and let $x \in (a_j, a_{j+1})$. The Taylor–Lagrange formula at order one yields

$$\begin{aligned}\Big(\Pi_h(u)\Big)'(x) &= \frac{u(a_{j+1}) - u(a_j)}{a_{j+1} - a_j}\\ &= u'(a_j + \theta_j)\end{aligned}$$

for some $0 < \theta_j < h$. Hence, for any $x \in (a_j, a_{j+1})$

$$\begin{aligned}|u'(x) - \big[\Pi_h(u)\big]'(x)| &= |u'(x) - u'(a_j + \theta_j)|\\ &\leq \int_{a_j}^{a_{j+1}} |u''(s)|\,ds.\end{aligned}$$

Applying the Cauchy–Schwarz inequality, we obtain

$$|u'(x) - \big[\Pi_h(u)\big]'(x)|^2 \leq h \int_{a_j}^{a_{j+1}} |u''(s)|^2\,ds.$$

The above inequality holds for any $x \in (a_j, a_{j+1})$. After integration on (a_j, a_{j+1}) one obtains

$$\int_{a_j}^{a_{j+1}} |u'(x) - \left[\Pi_h(u)\right]'(x)|^2\, dx \le h^2 \int_{a_j}^{a_{j+1}} |u''(s)|^2\, ds.$$

Summing the above inequality with respect to $j = 0, 1, \dots, N-1$ finally yields

$$\int_a^b |u'(x) - \left[\Pi_h(u)\right]'(x)|^2\, dx \le h^2 \int_a^b |u''(s)|^2\, ds,$$

that is,

$$||u' - \Pi_h(u)'||_{L^2(\Omega)} \le h||u''||_{L^2(\Omega)}. \tag{7.19}$$

Let us prove that there exists some constant $C > 0$ such that

$$||u - \Pi_h(u)||_{L^2(\Omega)} \le Ch||u''||_{L^2(\Omega)}. \tag{7.20}$$

Using the same argument and notation as above, we can write for $x \in (a_j, a_{j+1})$

$$\begin{aligned} u(x) - \Pi_h(u)(x) &= u(x) - \left[u(a_j) + \frac{u(a_{j+1}) - u(a_j)}{a_{j+1} - a_j}(x - a_j) \right] \\ &= u(x) - u(a_j) - (x - a_j)u'(a_j + \theta_j) \\ &= (x - a_j)u'(a_j + \theta_{x,j}) - (x - a_j)u'(a_j + \theta_j), \end{aligned}$$

where $0 < \theta_{x,j} < x - a_j$. It follows that

$$|u(x) - \Pi_h(u)(x)| \le (x - a_j)\int_{a_j}^{a_{j+1}} |u''(t)|\, dt,$$

which by similar arguments as above yields

$$||u - \Pi_h(u)||_{L^2(\Omega)} \le \frac{h^2}{\sqrt{3}}||u''||_{L^2(\Omega)}.$$

Finally, by combining (7.19) and (7.20), one obtains

$$||u - \Pi_h(u)||_{H^1(\Omega)} \le Ch||u||_{H^2(\Omega)}. \tag{7.21}$$

(b) Let us now extend the above inequality to an arbitrary $u \in H^2(\Omega)$. To that end one uses a density argument: noticing that $\mathbf{C}^\infty([a,b])$ is dense in $H^1(a,b)$ and $H^2(a,b)$, we just need to prove that the operator $\Pi_h : H^1(\Omega) \to H^1(\Omega)$ is continuous. More precisely, one can state the following result of independent interest, which concludes the proof in the one-dimensional case. □

Lemma 7.4.1. *Suppose $\Omega = (a,b)$ and $(\mathcal{T}_h)_{h\to 0}$ is a discretization of Ω. Then, there exists a constant $C > 0$ such that for any $h > 0$, for any $v \in H^1(\Omega)$,*

$$||\Pi_h(v)||_{H^1(\Omega)} \le C||v||_{H^1(\Omega)}.$$

PROOF. For any $x \in (a_j, a_{j+1})$

$$\Pi_h(v)'(x) = \frac{v(a_j+1) - v(a_j)}{a_{j+1} - a_j} = \frac{1}{a_{j+1} - a_j} \int_{a_j}^{a_{j+1}} v'(t)\,dt$$

(see Theorem 5.1.1). Hence

$$\left[\Pi_h(v)\right]^{'2}(x) \le \frac{1}{(a_{j+1} - a_j)^2} \left(\int_{a_j}^{a_{j+1}} |v'(t)|\,dt \right)^2,$$

which by the Cauchy–Schwarz inequality yields

$$\left[\Pi_h(v)\right]^{'2}(x) \le \frac{1}{a_{j+1} - a_j} \int_{a_j}^{a_{j+1}} |v'(t)|^2\,dt.$$

After integration on (a_j, a_{j+1}), and summation with respect to j, one obtains

$$||\Pi_h(v)'||_{L^2(\Omega)} \le ||v'||_{L^2(\Omega)}. \tag{7.22}$$

On the other hand,

$$\begin{aligned} ||\Pi_h(v)||_{L^2(\Omega)} &\le \sqrt{b-a}\,||\Pi_h(v)||_{L^\infty(\Omega)} \\ &\le \sqrt{b-a}\,||v||_{L^\infty(\Omega)}. \end{aligned} \tag{7.23}$$

In the one-dimensional case ($N = 1$) it was proved in Theorem 5.1.1 that each element of $H^1(a,b)$ has a unique continuous representative. Let us verify by some elementary computation that this canonical embedding $H^1(a,b) \subset \mathbf{C}([a,b])$ is continuous, i.e., there exists some constant $C > 0$ such that

$$\forall\, v \in H^1(a,b), \qquad ||v||_{L^\infty(a,b)} \le C||v||_{H^1(a,b)}. \tag{7.24}$$

Recall that we still denote $\tilde{v} = v$ the continuous representative of v and that for any $x_0, x \in [a,b]$

$$v(x_0) = v(x) + \int_{x_0}^{x} v'(t)\,dt.$$

Let us apply the Cauchy–Schwarz inequality to the above formula:

$$\begin{aligned} |v(x_0)| &\le |v(x)| + \int_{x_0}^{x} |v'(t)|\,dt \\ &\le |v(x)| + \sqrt{b-a} \left(\int_a^b |v'(t)|^2\,dt \right)^{1/2}. \end{aligned}$$

Let us now integrate this inequality with respect to $x \in [a,b]$:

$$\begin{aligned} (b-a)|v(x_0)| &\le \int_a^b |v(x)|\,dx + (b-a)^{3/2} \left(\int_a^b |v'(t)|^2\,dt \right)^{1/2} \\ &\le (b-a)^{1/2} \left(\int_a^b |v(x)|^2\,dx \right)^{1/2} + (b-a)^{3/2} \left(\int_a^b |v'(t)|^2\,dt \right)^{1/2}. \end{aligned}$$

The elementary inequality $(\alpha+\beta)^2 \le 2(\alpha^2+\beta^2)$ now yields

$$||v||_{L^\infty(a,b)} \le \sqrt{2}\left[\frac{1}{b-a}\int_a^b v(x)^2\,dx + (b-a)\int_a^b v'(x)^2\,dx\right]^{1/2}$$
$$\le \sqrt{2}\max\left\{\frac{1}{b-a}, b-a\right\}^{1/2} ||v||_{H^1(a,b)}.$$

Combining (7.23) and (7.24) finally yields

$$||\Pi_h(v)||_{L^2(\Omega)} \le C\sqrt{b-a}||v||_{H^1(\Omega)},$$

which, together with (7.22), gives

$$||\Pi_h(v)||_{H^1(\Omega)} \le C||v||_{H^1(\Omega)},$$

and the proof of Lemma 7.4.1 is complete. □

Remark 7.4.1. Indeed one can prove the following result: for all $v \in H^1(a,b)$

$$\lim_{h\to 0} ||v-\Pi_h(v)||_{H^1(a,b)} = 0. \tag{7.25}$$

This is slightly more precise than proving that for every $u \in H^1(a,b)$

$$\lim_{h\to 0} \mathrm{dist}(u, V_h) = 0,$$

which we have used in the proof of the convergence of the finite element method. The proof follows the lines of the previous arguments: given $v \in H^1(a,b)$, for any $\varepsilon > 0$, let $v_\varepsilon \in H^2(a,b)$ with $||v-v_\varepsilon||_{H^1(a,b)} < \varepsilon$. We have

$$||v-\Pi_h(v)||_{H^1(a,b)} \le ||v-v_\varepsilon||_{H^1(a,b)} + ||v_\varepsilon - \Pi_h(v_\varepsilon)||_{H^1(a,b)} + ||\Pi_h(v-v_\varepsilon)||_{H^1(a,b)}$$
$$\le C||v-v_\varepsilon||_{H^1(a,b)} + Ch||v_\varepsilon||_{H^2(a,b)}.$$

It follows that

$$\limsup_{h\to 0} ||v-\Pi_h(v)||_{H^1(a,b)} \le C\varepsilon.$$

This being true for any $\varepsilon > 0$, the conclusion follows.

Proof of Theorem 7.4.2 in the two-dimensional case. Suppose now that Ω is a polygon. The proof of the basic estimate for $u \in H^2(\Omega)$,

$$||u-\Pi_h(u)||_{H^1(\Omega)} \le Ch||u||_{H^2(\Omega)}, \tag{7.26}$$

is much more involved than in the one-dimensional case.

To establish this result one needs to make some geometrical assumptions on the triangulation $\mathcal{T}_h$; this is where the regularity assumption (7.13) plays a central role. Indeed, to establish (7.26) we first are going to argue with a single triangle (the key step) and show the following result.

Theorem 7.4.3. *There exists a constant* $C > 0$ *such that for any triangle* K, *for any* $u \in H^2(K)$,

$$||u-\Pi(u)||_{H^2(K)} \le Ch\left(h+\frac{h}{\rho}\right)||u||_{H^2(K)}, \tag{7.27}$$

where $\Pi(u)$ is the affine interpolant of u at the vertices of K, h is the diameter of K, and ρ is the diameter of the largest ball contained in K.

PROOF OF THEOREM 7.4.2 CONTINUED. The basic estimate (7.26) and so Theorem 7.4.2 can be easily deduced from (7.27), as shown in the following. By taking the square of each member of (7.27), writing the corresponding inequalities for all the triangles of the triangulation $\mathcal{T}_h$, and then summing these inequalities, one obtains

$$\|u - \Pi_h(u)\|^2_{H^1(\Omega)} \le Ch^2\left(h + \frac{h}{\rho}\right)^2 \|u\|^2_{H^2(\Omega)}.$$

Let us now use the regularity assumption (7.13) on the triangulation ($h/\rho \le \sigma$) and take $h \le 1$ to obtain

$$\|u - \Pi_h(u)\|_{H^1(\Omega)} \le Ch(1+\sigma)\|u\|_{H^2(\Omega)},$$

which proves (7.26). □

Thus our concern to complete the proof of the finite element method in the case $N = 2$ is to prove Theorem 7.4.3. A key idea in the process of getting the estimate (7.27) consists first in establishing such a formula for a fixed triangle $\widehat{K}$, which is used as a reference; take, for example, $\widehat{K}$ equal to the unit simplex.

Lemma 7.4.2. *Let $\widehat{K}$ be a given triangle. Then there exists a constant $C > 0$ such that for any $v \in H^2(\widehat{K})$*

$$\|v - \Pi_{\widehat{K}}(v)\|_{H^1(\widehat{K})} \le C|D^2 v|_{L^2(\widehat{K})}, \tag{7.28}$$

where $|D^2 v|_{L^2(\widehat{K})} := \sum_{i_1+i_2=2} \int_{\widehat{K}} \left| \frac{\partial^2 v}{\partial x_1^{i_1} \partial x_2^{i_2}}(x)\right|^2 dx.$

This is the two-dimensional version of inequalities (7.19) and (7.20). At this stage, we don't need to know the precise value of the constant C, the point being just to know if such a constant exists.

Then, we shall pass from the reference triangle $\widehat{K}$ to a triangle K of $\mathcal{T}_h$ by using an affine transformation,

$$x = B\hat{x} + b = F(\hat{x}),$$

where B is an invertible matrix and $b \in \mathbf{R}^2$, which satisfies

$$K = F(\widehat{K}).$$

The geometrical properties of the triangulation will appear through this transformation.

PROOF OF LEMMA 7.4.2. To prove (7.28), and since we don't need to know the precise value of C, we follow an analysis similar to the one in the proof of general Poincaré inequalities (Theorem 5.4.3, with the Poincaré–Wirtinger inequality as an example). The idea is to argue by contradiction and use the Rellich–Kondrakov compact embedding theorem, Theorem 5.4.2. Without ambiguity, for simplicity of the notation let us write K instead of $\widehat{K}$.

Suppose the assertion (7.28) is false. Then we could find a sequence $(v_n)_{n\in\mathbf{N}}$ such that

$$\begin{cases} v_n \in H^2(K) \quad \forall\, n \in \mathbf{N}, \\ \dfrac{1}{|D^2 v_n|_{L^2(K)}} \|v_n - \Pi_K(v_n)\|_{H^1(K)} \ge n. \end{cases}$$

Noticing that $D^2(\Pi_K(v_n)) = 0$, one can rewrite the above inequality in the following form:

$$\left| D^2\left(\frac{v_n - \Pi_K(v_n)}{\|v_n - \Pi_K(v_n)\|_{H^1(K)}} \right) \right|_{L^2(K)} \leq \frac{1}{n}.$$

Let us introduce the function

$$u_n := \frac{v_n - \Pi_K(v_n)}{\|v_n - \Pi_K(v_n)\|_{H^1(K)}}.$$

We have

$$\begin{cases} u_n \in H^2(K), \\ \|u_n\|_{H^1(K)} = 1, \\ |D^2 u_n|_{L^2(K)} \leq 1/n, \\ u_n(a_j) = 0, \qquad j = 1,2,3, \text{ where } a_j \text{ are the vertices of } K. \end{cases}$$

Let us show how to obtain a contradiction from this set of properties. From $\|u_n\|_{H^1(K)} = 1$ and $|D^2 u_n|_{L^2(K)} \leq 1/n$ we obtain that the sequence $(u_n)_{n\in\mathbf{N}}$ is bounded in $H^2(K)$. Since K is bounded and piecewise $\mathbf{C}^1$ one can apply the Rellich–Kondrakov theorem, which gives that the sequence $(u_n)_{n\in\mathbf{N}}$ is relatively compact in $H^1(K)$. We can then extract a subsequence, which we still denote u_n, such that

$$u_n \to u \quad \text{in } H^1(K).$$

Hence $\|u\|_{H^1(K)} = 1$. On the other hand, from $|D^2 u_n|_{L^2(K)} \leq 1/n$, we obtain that

$$D^2 u_n \to 0 = D^2 u \quad \text{in } L^2(K).$$

Hence u is an affine function on K. The linear map $u \mapsto u(a_j)$ from $H^2(K)$ into $\mathbf{R}$ is continuous. Since $u_n \to u$ in $H^2(K)$ we have $u(a_j) = 0$, $j = 1,2,3$. The only affine function on K which is zero at the vertices is the function $u = 0$. This is a contradiction with $\|u\|_{H^1(K)} = 1$. This establishes (7.28) and concludes the proof of Lemma 7.4.2. □

Let us now consider an affine invertible transformation

$$F(\hat{x}) = B\hat{x} + b \tag{7.29}$$

with $K = F(\widehat{K})$ and examine how it affects the formula (7.28).

The following notation and definitions will be helpful. We use the mappings

$$\hat{x} \in \widehat{K} \overset{F}{\longmapsto} F(\hat{x}) = x \in K$$

and $F^{-1} : K \to \widehat{K}$, which is the inverse of F. To each function v defined on K one can associate the function $\hat{v} : \widehat{K} \to \mathbf{R}$, which is defined by

$$\hat{v}(\hat{x}) = v(F(\hat{x})),$$

which, with the above notation, gives

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

Recall that $F(\hat{x}) = B\hat{x} + b$ is an affine invertible map. The spectral norms of B and B^{-1} will play a central role in the following. Recall that these norms are defined by

$$\begin{aligned} ||B|| &= \sup\{|B\xi| \,:\, |\xi| = 1\}, \\ ||B^{-1}|| &= \sup\{|B^{-1}\xi| \,:\, |\xi| = 1\}, \end{aligned}$$

where $|\xi|$ is the Euclidean norm.

The geometrical characteristic properties of the triangulation (h and ρ) do appear naturally in the evaluation of these norms. Let us denote by h_K and ρ_K (respectively, $h_{\widehat{K}}$ and $\rho_{\widehat{K}}$) the geometrical characteristic numbers of K (respectively, $\widehat{K}$) as defined in (7.13).

Lemma 7.4.3. *The following estimates hold:*

$$||B|| \le \frac{h_K}{\rho_{\widehat{K}}}, \qquad ||B^{-1}|| \le \frac{h_{\widehat{K}}}{\rho_K}.$$

PROOF. We have

$$\begin{aligned} ||B|| &= \sup\{|B\xi| \,:\, |\xi| = 1\} \\ &= \frac{1}{\rho_{\widehat{K}}} \sup\{|B\xi| \,:\, |\xi| = \rho_{\widehat{K}}\}. \end{aligned}$$

By definition of $\rho_{\widehat{K}}$, for any ξ with $|\xi| = \rho_{\widehat{K}}$ one can find two points $\hat{x}_1$ and $\hat{x}_2$ in $\widehat{K}$ such that

$$\xi = \hat{x}_2 - \hat{x}_1.$$

Hence

$$\begin{aligned} B\xi &= B\hat{x}_2 - B\hat{x}_1 \\ &= F\hat{x}_2 - F\hat{x}_1 \\ &= x_2 - x_1, \end{aligned}$$

where x_2 and x_1 belong to $K = F(\widehat{K})$.

By definition of h_K, which is the diameter of K, we have

$$|B\xi| = |x_2 - x_1| \le h_K.$$

This inequality being true for any ξ with $|\xi| = \rho_{\widehat{K}}$, we deduce

$$||B|| \le \frac{h_K}{\rho_{\widehat{K}}}.$$

The other inequality is obtained in a similar way, by reversing the role of K and $\widehat{K}$. □

The other basic ingredient of the proof is the change of variables in the integrals which are equal to Sobolev norms. Let us write, for a given integer $m \ge 0$

$$|v|_{m,K} = \left(\sum_{|\alpha|=m} \int_K |\partial^\alpha v(x)|^2 \, dx \right)^{1/2}. \tag{7.30}$$

Lemma 7.4.4. *Let K and $\widehat{K}$ be two finite elements which are affine equivalent, that is, $K = F(\widehat{K})$ with $F(\hat{x}) = B\hat{x} + b$ and B affine invertible. If a function v belongs to the space $H^m(K)$ for some integer $m \geq 0$, then the function $\hat{v} = v \circ F$ belongs to $H^m(\widehat{K})$ and there is a constant $C(m) > 0$ such that*

$$\forall v \in H^m(K) \qquad |\hat{v}|_{m,\widehat{K}} \leq C(m)\|B\|^m |\det B|^{-1/2} |v|_{m,K}.$$

Analogously, one has

$$\forall \hat{v} \in H^m(\widehat{K}) \qquad |v|_{m,K} \leq C(m)\|B^{-1}\|^m |\det B|^{1/2} |\hat{v}|_{m,K}.$$

PROOF. By standard density arguments, one just needs to argue with $v \in \mathbf{C}^\infty(\overline{K})$. Hence $\hat{v} \in \mathbf{C}^\infty(\overline{\widehat{K}})$. It is convenient to introduce the first and second derivatives of v: then $Dv(x)$ is a linear form and

$$\frac{\partial v}{\partial x_i}(x) = Dv(x) \cdot e_i,$$

where (e_i) are the vectors of the canonical basis in $\mathbf{R}^N$ (here $N = 2$). Similarly, $D^2 v(x)$ is the bilinear symmetric form associated to the Hessian matrix and

$$\frac{\partial^2 v}{\partial x_i \partial x_j}(x) = D^2 v(x) \cdot (e_i, e_j).$$

One can unify these two situations (and much more) by writing for any multi-index $\alpha = (\alpha_1, \alpha_2)$ with length $|\alpha| = m$

$$\partial^\alpha v(x) = D^m v(x)(e_1, \ldots, e_1, e_2, \ldots, e_2),$$

where e_1 is repeated α_1 times, and e_2 is repeated α_2 times. Recall that $\partial^\alpha v(x) = (\partial^{|\alpha|} v / \partial x_1^{\alpha_1} \partial x_2^{\alpha_2})(x)$. Set

$$\|D^m v(x)\| = \sup\left\{ |D^m v(x)(\xi_1, \ldots, \xi_m)| \ : \ |\xi_i| \leq 1,\ 1 \leq i \leq m \right\}.$$

Then

$$|\partial^\alpha v(x)| \leq \|D^m v(x)\|$$

and

$$\begin{aligned} |v|_{m,K} &= \left(\int_K \sum_{|\alpha|=m} |\partial^\alpha v(x)|^2 \, dx \right)^{1/2} \\ &\leq C_1 \left(\int_K \|D^m v(x)\|^2 \, dx \right)^{1/2}, \end{aligned} \tag{7.31}$$

where C_1^2 is the cardinal of the set of indices α such that $|\alpha| = m$, i.e., $C_1 = C_1(m)$. We can now perform the differentiation rule for composition of functions. Recalling that $\hat{v}(\hat{x}) = v(F(\hat{x})) = v(B\hat{x} + b)$ we have

$$D^m \hat{v}(\hat{x})(\xi_1, \ldots, \xi_m) = D^m v(x)(B\xi_1, \ldots, B\xi_m)$$

so that

$$\|D^m \hat{v}(\hat{x})\| \leq \|D^m v(x)\| \|B\|^m.$$

Taking the square and after integration on $\widehat{K}$, one obtains

$$\int_{\widehat{K}} \|D^m \hat{v}(\hat{x})\|^2 \, d\hat{x} \leq \|B\|^{2m} \int_{\widehat{K}} \|D^m v(F(\hat{x}))\|^2 \, d\hat{x}.$$

Using the formula of change of variables in multiple integrals we get

$$\int_{\widehat{K}} \|D^m \hat{v}(\hat{x})\|^2 \, d\hat{x} \leq \|B\|^{2m} |\det(B^{-1})| \int_K \|D^m v(x)\|^2 \, dx. \tag{7.32}$$

Combining (7.31) and (7.32) we obtain

$$|\hat{v}|_{m,\widehat{K}} \leq C_1(m)\|B\|^m |\det B|^{-1/2} \Big(\int_K \|D^m v(x)\|^2 \, dx \Big)^{1/2}.$$

Since conversely there exists a constant $C_2(m)$ such that

$$\Big(\int_K \|D^m v(x)\|^2 \, dx \Big)^{1/2} \leq C_2(m) |v|_{m,K}, \tag{7.33}$$

we finally obtain

$$|\hat{v}|_{m,\widehat{K}} \leq C(m)\|B\|^m |\det B|^{-1/2} |v|_{m,K}$$

with $C(m) = C_1(m)C_2(m)$.

Reversing the role of K and $\widehat{K}$ yields the other inequality. □

End of the proof of Theorem 7.4.3. We now have all the ingredients of the proof of the basic estimate (7.27) in Theorem 7.4.3:

$$\|u - \Pi(u)\|_{H^1(K)} \leq Ch\left(h + \frac{h}{\rho}\right)\|u\|_{H^2(K)}.$$

Let $u \in H^2(K)$. By Lemma 7.4.4 we have for $m = 0, 1$

$$|u - \Pi(u)|_{m,K} \leq C\|B^{-1}\|^m |\det B|^{1/2} |\hat{u} - \widehat{\Pi(u)}|_{m,\widehat{K}}. \tag{7.34}$$

The estimate (7.28) on $\widehat{K}$ yields for $m = 0, 1$

$$|\hat{u} - \widehat{\Pi(u)}|_{m,\widehat{K}} \leq C|\hat{u}|_{2,\widehat{K}}. \tag{7.35}$$

Applying again Lemma 7.4.4 we have

$$|\hat{u}|_{2,\widehat{K}} \leq C\|B\|^2 |\det B|^{-1/2} |u|_{2,K}. \tag{7.36}$$

Combining (7.34), (7.35), and (7.36) we obtain

$$|u - \Pi(u)|_{m,K} \leq C\|B^{-1}\|^m \|B\|^2 |u|_{2,K}. \tag{7.37}$$

Take the square of inequality (7.37) and sum over $m = 0, 1$ to obtain

$$\|u - \Pi(u)\|_{H^1(K)} \leq C\|B\|^2 \big(1 + \|B^{-1}\|\big) |u|_{2,K}.$$

Using Lemma 7.4.3 we finally get

$$\begin{aligned}\|u-\Pi(u)\|_{H^1(K)} &\le C\frac{h_K^2}{\rho_{\widehat{K}}^2}\left(1+\frac{h_{\widehat{K}}}{\rho_K}\right)|u|_{2,K}\\ &\le C\left(h_K^2+\frac{h_K^2}{\rho_K}\right)|u|_{2,K},\end{aligned}$$

where $1/\widehat{\rho_K}$ and $\widehat{h_K}$ have been included in the constant C. Recall that $\widehat{K}$ is a fixed reference triangle. Noticing that $|u_{2,K}|\le\|u\|_{H^2(K)}$, the proof is complete. □

Remark 7.4.2. Note that we have obtained a slightly more precise result than (7.27); indeed, we proved that for every $u\in H^2(K)$

$$\|u-\Pi_h(u)\|_{H^1(K)}\le Ch\left(h+\frac{h}{\rho}\right)|u|_{2,K},$$

where $|u|_{2,K}=|D^2u|_{L^2(K)}$ just involves the L^2 norm of the second-order partial derivatives of u. Consequently, in Theorem 7.4.2, we have that for any $u\in H^2(\Omega)$

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

Similarly, in Theorem 7.4.1, we have that if $u\in H^2(\Omega)$,

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

7.5 ▪ Complements

7.5.1 ▪ Flat triangles

Let us return to the situation described in Figure 7.5, which illustrates a family of triangulations $(\mathcal{T}_h)_{h\to 0}$ involving triangles $K_h\in\mathcal{T}_h$ becoming flat as $h\to 0$. Let us show that on such triangles the affine interpolate can lead to significant errors.

Take a simple function which is not affine, for example, a quadratic function

$$u(x,y)=y^2.$$

Let us compute the affine interpolate $\Pi_h(u)$ of u on the triangle K_h whose vertices are $(0,\frac{h}{2})$, $(0,-\frac{h}{2})$, and $(h\varepsilon_h,0)$. We have that $\Pi_h(u)$ vanishes at $(h\varepsilon_h,0)$ and is equal to $h^2/4$ at the two other vertices. Hence

$$\frac{\partial}{\partial x}\Pi_h(u)=-\frac{h^2/4}{h\varepsilon_h}=-\frac{h}{4\varepsilon_h}.$$

Since $\frac{\partial u}{\partial x}=0$ we obtain $|\frac{\partial}{\partial x}(u-\Pi_h(u))|=\frac{h}{4\varepsilon_h}$ and

$$\left(\int_{K_h}\left|\frac{\partial}{\partial x}(u-\Pi_h(u))\right|^2(x)\,dx\right)^{1/2}=\frac{h^2}{4\sqrt{2}\sqrt{\varepsilon_h}}.$$

On the other hand, we have $\frac{\partial^2u}{\partial x^2}=\frac{\partial^2u}{\partial x\partial y}=0$ and $\frac{\partial^2u}{\partial y^2}=2$, which give

$$|u|_{2,K_h}=\left(\int_{K_h}4\,dx\,dy\right)^{1/2}=\sqrt{2}h\sqrt{\varepsilon_h}.$$

An inequality of the type

$$\|u-\Pi_h(u)\|_{H^1(K_h)} \le Ch|u|_{2,K_h}$$

would then imply

$$\left\|\frac{\partial}{\partial x}(u-\Pi_h(u))\right\|_{L^2(K_h)} \le Ch|u|_{2,K},$$

that is,

$$\frac{h^2}{4\sqrt{2}\sqrt{\varepsilon_h}} \le Ch\sqrt{2}h\sqrt{\varepsilon_h},$$

which is equivalent to

$$\inf_{h>0} \varepsilon_h > 0.$$

Thus, the convergence analysis developed in this chapter fails to be true without any geometrical assumption on the family of triangulations preventing the triangles from becoming flat.

7.5.2 ▪ $H^2(\Omega)$ regularity of the solution of the Dirichlet problem on a convex polygon

In the model situation studied in this section, we chose to take as Ω a polygon in $\mathbf{R}^2$, to make as simple as possible the description of the triangulation in the finite element method. (Otherwise, for general Ω one has to approximate it by such polygonal sets Ω_h.)

Conversely, we have a difficulty, which is to know if the solution u of the Dirichlet problem with $f \in L^2(\Omega)$

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

satisfies the property $u \in H^2(\Omega)$. Indeed the estimate

$$\|u-u_h\|_{H_0^1(\Omega)} \le h$$

has been established under the assumption $u \in H^2(\Omega)$.

We are in a situation where Ω is a polygon, its boundary is not smooth (it is only piecewise $\mathbf{C}^1$ or Lipschitz continuous), and the classical Agmon–Douglis–Nirenberg theorem which asserts that $u \in H^2(\Omega)$ under the assumption that Ω is of class $\mathbf{C}^2$ does not apply.

The answer to this question is quite involved. It was studied by Grisvard in [233], [234], who proved in particular that if Ω is a polygon which is supposed to be convex, then $u \in H^2(\Omega)$.

7.5.3 ▪ Finite element methods of type P_2

The method which has been developed in $\mathbf{R}^2$ with Ω a polygon and finite elements which are triangles can be naturally extended to $\mathbf{R}^3$ when replacing triangles by tetrahedrons. Functions of the approximating subspaces V_h are continuous and piecewise affine. This is what we call a finite element method of type P_1 (by reference to the degree of the polynomial functions which are used). To improve the quality of the approximation, one may naturally think to enrich the approximating subspaces and make them contain more functions. This can be done, for example, by considering functions which are piecewise polynomial of degree less than or equal to two. Let us briefly describe an example of such a finite element method of type P_2.

Take Ω a polygon in $\mathbf{R}^2$ and a given triangulation $\mathcal{T}_h$ of Ω. We introduce the space

$$V_h = \left\{ v \in \mathbf{C}(\overline{\Omega}) \;:\; v\lfloor K \in P_2 \text{ for every } K \in \mathcal{T}_h \right\},$$

where P_2 is the family of polynomial functions on $\mathbf{R}^2$ of degree less than or equal to two. The general form of an element $p \in P_2$ is then

$$p(x,y) = a + bx + cy + dx^2 + exy + fy^2,$$

and one can verify that P_2 is a vector space of dimension equal to 6. Then, to fix an element $p \in P_2$, it is not sufficient to give its values at the vertices of a triangle (as was the case for $p \in P_1$): we need to give its values at six points carefully chosen. Take, for example, the case of Figure 7.6, where a_1, a_2, a_3 are the vertices of K and $a_{ij} = \frac{1}{2}(a_i + a_j)$ are the midpoints of the edges of K.

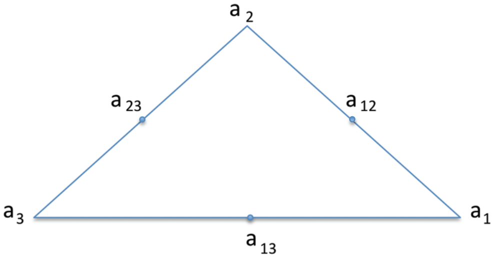

Figure 7.6. *Six points on the triangle K.*

This choice leads to triangulations whose nodes are the vertices of the triangles and the midpoints of all the edges. As an illustration consider the case of Figure 7.7.

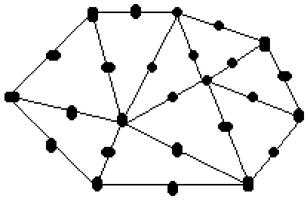

Figure 7.7. *A triangulation for finite element method of type P_2.*

Let us denote by (N_j) the nodes of $\mathcal{T}_h$, $j = 1, \ldots, I(h)$ (vertices + midpoints of edges). It is quite elementary to verify that

(a) V_h is a subspace of dimension $I(h)$ of $H^1(\Omega)$ and any function v of V_h is uniquely determined by its values at the nodes of the triangulation,

(b) a basis of V_h is given by the family of functions $(p_j)_{j=1,\ldots,I(h)}$ which is defined by

$$\begin{aligned} & p_j \in V_h, \\ & p_j(N_i) = \delta_{ij} \quad \text{for } i,j = 1,\ldots,I(h). \end{aligned}$$

For any element $v \in V_h$ one has

$$v(x) = \sum_{j=1}^{I(h)} v(N_j) p_j(x).$$

The finite element method can now be developed in a way parallel to what we did before. Indeed, as expected, one can get a better order in the approximation by piecewise P_2 functions. Let us denote by

$$\Pi_h(v) = \sum_{j=1}^{I(h)} v(N_j) p_j$$

the element of V_h obtained by interpolation of v on the nodes of the triangulation ($\Pi_h(v) = v$ on the nodes). Then, one can show the following result (which we do not prove): if the family of triangulations $(\mathscr{T}_h)_{h\to 0}$ is regular, then there exists a constant $C > 0$ such that for any $u \in H^3(\Omega)$,

$$||u - \Pi_h(u)||_{H^1(\Omega)} \leq C h^2 |u|_{3,\Omega}.$$