CHAPTER 4

OTHER FINITE ELEMENT METHODS FOR SECOND-ORDER PROBLEMS

Introduction

Up to now, we have considered finite elements methods which are *conforming*, in the sense that the space V_h is a subspace of the space V, and the bilinear form and the linear form which are used in the definition of the discrete problem are identical to those of the original problem.

In this chapter, we shall analyze several ways of violating this "conformity", which are frequently used in everyday computations.

In Section 4.1, we assume, as before, that the domain $\bar{\Omega}$ is polygonal and that the inclusion $V_h \subset V$ still holds, but we consider the use of a *quadrature scheme* for computing the coefficients of the resulting linear system: each such coefficient being of the form

$$\sum_{K \in \mathcal{T}_h} \int_K \varphi(x)\,dx,$$

the integrals

$$\int_K \varphi(x)\,dx, \quad K \in \mathcal{T}_h,$$

are approximated by finite sums of the form

$$\sum_{l=1}^{L} \omega_{l,K} \varphi(b_{l,K}),$$

with weights $\omega_{l,K}$ and nodes $b_{l,K} \in K$, which are derived from a single *quadrature formula* defined over a reference finite element. This process results in an *approximate bilinear form* $a_h(.\,,.)$ and an *approximate linear form* $f_h(.)$ which are defined over the space V_h.

Our study of this approximation follows a general pattern that will also be the same for the two other methods to be described in this chapter:

First, we prove (the *first Strang lemma*; cf. Theorem 4.1.1) an abstract error estimate (which, as such, is intended to be valid in other situations; cf. Section 8.2). It is established under the critical assumption that the approximate bilinear forms are *uniformly V_h-elliptic*, i.e., that there exists a constant $\tilde{\alpha} > 0$, independent of h, such that $a_h(v_h, v_h) \geq \tilde{\alpha}\|v_h\|^2$ for all $v_h \in V_h$. This is why we next examine (Theorem 4.1.2) under which assumptions (on the quadrature scheme over the reference finite element) this property is true.

The abstract error estimate of Theorem 4.1.1 generalizes Céa's lemma: In the right-hand side of the inequality, there appear, in addition to the term $\inf_{v_h \in V_h} \|u - v_h\|$, two *consistency errors* which measure the quality of the approximation of the bilinear form and of the linear form, respectively.

We are then in a position to study the convergence of such methods. More precisely, we shall essentially concentrate on the following problem: *Find sufficient conditions which insure that the order of convergence in the absence of numerical integration is unaltered by the effect of numerical integration.* Restricting ourselves for simplicity to the case where $P_K = P_k(K)$ for all $K \in \mathcal{T}_h$, our main result in this direction (Theorem 4.1.6) is that one still has

$$\|u - u_h\|_{1,\Omega} = O(h^k),$$

provided the quadrature formula is exact for all polynomials of degree $(2k - 2)$. The proof of this result depends, in particular, on the *Bramble-Hilbert lemma* (Theorem 4.1.3), which is a useful tool for handling linear functionals which vanish on polynomial subspaces. In this particular case, it is repeatedly used in the derivation of the consistency error estimates (Theorems 4.1.4 and 4.1.5).

We next consider in Section 4.2 a first type of finite element method for which the spaces V_h are *not* contained in the space V. This violation of the inclusion $V_h \subset V$ results of the use of finite elements which are not of class $\mathscr{C}^0$ (i.e., which are not continuous across adjacent finite elements), so that the inclusion $V_h \subset H^1(\Omega)$ is not satisfied (Theorem 4.2.1). The terminology "*nonconforming finite element method*" is specifically reserved for this type of method (likewise, for fourth-order problems, nonconforming methods result from the use of finite elements which are not of class $\mathscr{C}^1$; cf. Section 6.2).

For definiteness, we assume through Section 4.2 that we are solving a homogeneous Dirichlet problem posed over a polygonal domain $\bar{\Omega}$. Then

the discrete problem consists in finding a function $u_h \in V_h$ such that, for all $v_h \in V_h$, $a_h(u_h, v_h) = f(v_h)$, where the *approximate bilinear form* $a_h(.\,,.)$ is defined by

$$a_h(.\,,.) = \sum_{K \in \mathcal{T}_h} \int_K \{. . .\}\, dx,$$

the integrand $\{. . .\}$ being the same as in the bilinear form which is used in the definition of the original problem. The linear form $f(.)$ need not be approximated since the inclusion $V_h \subset L^2(\Omega)$ holds.

Assuming that the mapping

$$v_h \in V_h \to \|v_h\|_h = \left(\sum_{K \in \mathcal{T}_h} |v_h|^2_{1,K} \right)^{1/2}$$

is a norm over the space V_h, we prove an abstract error estimate (the *second Strang lemma*; cf. Theorem 4.2.2) where the expected term $\inf_{v_h \in V_h} \|u - v_h\|_h$ is added a *consistency error*. Just as in the case of numerical integration, this result holds under the assumption that the approximate bilinear forms are *uniformly V_h-elliptic*, in the sense that there exists a constant $\tilde{\alpha} > 0$ independent of h such that $a_h(v_h, v_h) \geq \tilde{\alpha} \|v_h\|^2_h$ for all $v_h \in V_h$.

We then proceed to describe a three-dimensional "nonconforming" finite element, known as *Wilson's brick*, which has gained some popularity among engineers for solving the elasticity problem. Apart from being nonconforming, this finite element presents the added theoretical interest that some of its degrees of freedom are of a form not yet encountered. This is why we need to adapt to this finite element the standard interpolation error analysis (Theorem 4.2.3).

Next, using a "*bilinear lemma*" which extends the Bramble-Hilbert lemma to bilinear forms (Theorem 4.2.5), we analyze the consistency error (Theorem 4.2.6). In this fashion we prove that

$$\|u - u_h\|_h = \left(\sum_{K \in \mathcal{T}_h} |u - u_h|^2_{1,K} \right)^{1/2} = O(h),$$

if the solution u is in the space $H^2(\Omega)$. In passing, we establish the connection between the convergence of such nonconforming finite element methods and the *patch test* of B. Irons.

Another violation of the inclusion $V_h \subset V$ occurs in the approximation of a boundary value problem posed over a domain $\bar{\Omega}$ with a *curved* boundary Γ (i.e., the set $\bar{\Omega}$ is no longer assumed to be polygonal). In this

case, the set $\bar{\Omega}$ is usually approximated by two types of finite elements: The finite elements of the first type are *straight*, i.e., they have plane faces, and they are typically used "inside" $\bar{\Omega}$. The finite elements of the second type have at least one "curved" face, and they are especially used so as to approximate "as well as possible" the boundary Γ.

In Section 4.3, we consider one way of generating finite elements of the second type, the *isoparametric finite elements*, which are often used in actual computations. The key idea underlying their conception is the generalization of the notion of affine-equivalence: Let there be given a Lagrange finite element $(\hat{K}, \hat{P}, \{\hat{p}(\hat{a}_i), 1 \leq i \leq N\})$ in $\mathbb{R}^n$ and let $F: \hat{x} \in \hat{K} \to F(\hat{x}) = (F_i(\hat{x}))_{i=1}^n \in \mathbb{R}^n$ be a mapping such that $F_i \in \hat{P}$, $1 \leq i \leq n$. Then the triple

$$(K = F(\hat{K}), \quad P = \{p = \hat{p} \cdot F^{-1};\ \ \hat{p} \in \hat{P}\},$$
$$\{p(a_i = F(\hat{a}_i)), \quad 1 \leq i \leq N\})$$

is also a Lagrange finite element (Theorem 4.3.1), and two cases can be distinguished:

(i) The mapping F is *affine* (i.e., $F_i \in P_1(\hat{K})$, $1 \leq i \leq n$) and therefore the finite elements (K, P, Σ) and $(\hat{K}, \hat{P}, \hat{\Sigma})$ are affine-equivalent.

(ii) Otherwise, the finite element (K, P, Σ) is said to be *isoparametric*, and *isoparametrically equivalent* to the finite element $(\hat{K}, \hat{P}, \hat{\Sigma})$. If $(\hat{K}, \hat{P}, \hat{\Sigma})$ is a standard straight finite element, it is easily seen in the second case that the boundary of the set K is curved in general. This fact is illustrated by several examples.

We then consider the problem (particularly in view of Section 4.4) of developing an interpolation theory adapted to this type of finite element. In this analysis, however, we shall restrict ourselves to the isoparametric n-simplex of type (2), so as to simplify the exposition, yet retaining all the characteristic features of a general analysis. For an *isoparametric family* (K, P_K, Σ_K) of n-simplices of type (2), we show (Theorem 4.3.4) that the Π_K-interpolants of a function v satisfy inequalities of the form

$$|v - \Pi_K v|_{m,K} \leq C h_K^{3-m} \|v\|_{3,K}, \quad 0 \leq m \leq 3,$$

where $h_K = \operatorname{diam}(K)$. This result, which is the same as in the case of affine families (cf. Section 3.1) is established under the crucial assumption that the "isoparametric" mappings F_K do not deviate too much from affine mappings (of course the family is also assumed to be regular, in a sense that generalizes the regularity of affine families).

Even if we use isoparametric finite elements $K \in \mathcal{T}_h$ to "triangulate" a set $\bar{\Omega}$, the boundary of the set $\bar{\Omega}_h = \bigcup_{K \in \mathcal{T}_h} K$ is very close to, but not identical to, the boundary Γ. Consequently, since the domain of definition of the functions in the resulting finite element space V_h is the set $\bar{\Omega}_h$, the space V_h is *not* contained in the space V and therefore both the bilinear form and the linear form need to be approximated.

In order to be in as realistic a situation as possible we then study in Section 4.4 the simultaneous effects of such an approximation of the domain $\bar{\Omega}$ and of isoparametric numerical integration. As in Section 4.1, this last approximation amounts to use a quadrature formula over a reference finite element $\hat{K}$ for computing the integrals of the form $\int_K \varphi(x)\,dx$ (which appear in the linear system) via the isoparametric mappings $F_K : \hat{K} \to K$, $K \in \mathcal{T}_h$. Restricting ourselves again to isoparametric n-simplices of type (2) for simplicity, we show (Theorem 4.4.6) that, if the quadrature formula over the set $\hat{K}$ is exact for polynomials of degree 2, one has

$$\|\tilde{u} - u_h\|_{1,\Omega_h} = O(h^2),$$

where $\tilde{u}$ is an extension of the solution of the given boundary value problem to the set Ω_h (in general $\bar{\Omega}_h \not\subset \bar{\Omega}$), and $h = \max_{K \in \mathcal{T}_h} h_K$. This error estimate is obtained through the familiar process: We first prove an *abstract error estimate* (Theorem 4.4.1), under a *uniform V_h-ellipticity assumption* of the approximate bilinear forms. Then we use the interpolation theory developed in Section 4.3 for evaluating the term $\inf_{v_h \in V_h} \|\tilde{u} - v_h\|_{1,\Omega_h}$ (Theorem 4.4.3) and finally, we estimate the two *consistency errors* (Theorems 4.4.4 and 4.4.5; these results largely depend on related results of Section 4.1). It is precisely in these last estimates that a remarkable conclusion arises: *In order to retain the $O(h^2)$ convergence, it is not necessary to use more sophisticated quadrature schemes for approximating the integrals which correspond to isoparametric finite elements than for straight finite elements.*

4.1. The effect of numerical integration

Taking into account numerical integration.
Description of the resulting discrete problem.

Throughout this section, we shall assume that we are solving the second-order boundary value problem which corresponds to the follow-

ing data:

$$\begin{cases} V = H_0^1(\Omega), \\ a(u, v) = \displaystyle\int_\Omega \sum_{i,j=1}^n a_{ij}\partial_i u \partial_j v \, dx, \\ f(v) = \displaystyle\int_\Omega fv \, dx, \end{cases} \tag{4.1.1}$$

where $\bar{\Omega}$ is a polygonal domain in $\mathbf{R}^n$, the functions $a_{ij} \in L^\infty(\Omega)$ and $f \in L^2(\Omega)$ are assumed to be *everywhere* defined over $\bar{\Omega}$. We shall assume that the ellipticity condition is satisfied i.e.,

$$\begin{aligned} &\exists \beta > 0, \quad \forall x \in \bar{\Omega}, \quad \forall \xi_i, \quad 1 \leq i \leq n, \\ &\sum_{i,j=1}^n a_{ij}(x)\xi_i\xi_j \geq \beta \sum_{i=1}^n \xi_i^2, \end{aligned} \tag{4.1.2}$$

so that the bilinear form of (4.1.1) is $H_0^1(\Omega)$-elliptic.

This problem corresponds (cf. (1.2.28)) to the homogeneous Dirichlet problem for the operator

$$u \to -\sum_{i,j=1}^n \partial_j(a_{ij}\partial_i u),$$

i.e.,

$$\begin{cases} -\displaystyle\sum_{i,j=1}^n \partial_j(a_{ij}\partial_i u) = f \quad \text{in} \quad \Omega, \\ u = 0 \quad \text{on} \quad \Gamma. \end{cases} \tag{4.1.3}$$

The case of a more general operator of the form

$$u \to -\sum_{i,j=1}^n \partial_j(a_{ij}\partial_i u) + au$$

is left as a problem (Exercise 4.1.5).

We consider in the sequel a family of finite element spaces X_h made up of finite elements (K, P_K, Σ_K), $K \in \mathcal{T}_h$, where $\mathcal{T}_h$ are triangulations of the set $\bar{\Omega}$ (because the set $\bar{\Omega}$ is assumed to be polygonal, it can be exactly covered by triangulations). Then we define the spaces $V_h = \{v_h \in X_h; v_h = 0 \text{ on } \Gamma\}$.

The assumptions about the triangulations and the finite elements are the same as in Section 3.2. Let us briefly record these assumptions for convenience:

(H 1) The associated family of triangulations is regular.

(H 2) All the finite elements (K, P_K, Σ_K), $K \in \bigcup_h \mathcal{T}_h$, are affine-equivalent to a single reference finite element $(\hat{K}, \hat{P}, \hat{\Sigma})$.

(H 3) All the finite elements (K, P_K, Σ_K), $K \in \bigcup_h \mathcal{T}_h$, are of class $\mathscr{C}^0$.

As a consequence, the inclusions $X_h \subset H^1(\Omega)$ and $V_h \subset H_0^1(\Omega)$ hold, as long as the inclusion $\hat{P} \subset H^1(\hat{K})$ (which we will assume) holds.

Given a space V_h, solving the corresponding discrete problem amounts to finding the coefficients u_k, $1 \leq k \leq M$, of the expansion $u_h = \sum_{k=1}^{M} u_k w_k$ of the discrete solution u_h over the basis functions w_k, $1 \leq k \leq M$, of the space V_h. These coefficients are solutions of the linear system (cf. (2.1.4))

$$\sum_{k=1}^{M} a(w_k, w_m) u_k = f(w_m), \quad 1 \leq m \leq M, \tag{4.1.4}$$

where, according to (4.1.1),

$$a(w_k, w_m) = \sum_{K \in \mathcal{T}_h} \int_K \sum_{i,j=1}^{n} a_{ij} \partial_i w_k \partial_j w_m \, dx, \tag{4.1.5}$$

$$f(w_m) = \sum_{K \in \mathcal{T}_h} \int_K f w_m \, dx. \tag{4.1.6}$$

In practice, even if the functions a_{ij}, f have simple analytical expressions, the integrals $\int_K \ldots dx$ which appear in (4.1.5) and (4.1.6) are seldom computed exactly. Instead, they are approximated through the process of *numerical integration*, which we now describe:

Consider one of the integrals appearing in (4.1.5) or (4.1.6), let us say $\int_K \varphi(x)\, dx$, and let

$$F_K : \hat{x} \in \hat{K} \to F_K(\hat{x}) = B_K \hat{x} + b_K$$

be the invertible affine mapping which maps $\hat{K}$ onto K. Assuming, without loss of generality, that the (constant) Jacobian of the mapping F_K is positive, we can write

$$\int_K \varphi(x)\, dx = \det(B_K) \int_{\hat{K}} \hat{\varphi}(\hat{x})\, d\hat{x}, \tag{4.1.7}$$

the functions φ and $\hat{\varphi}$ being in the usual correspondence, i.e., $\varphi(x) = \hat{\varphi}(\hat{x})$ for all $x = F_K(\hat{x})$, $\hat{x} \in \hat{K}$. In other words, *computing the integral* $\int_K \varphi(x)\, dx$ *amounts to computing the integral* $\int_{\hat{K}} \hat{\varphi}(\hat{x})\, d\hat{x}$.

Then a *quadrature scheme* (over the set $\hat{K}$) consists in replacing the integral $\int_{\hat{K}} \hat{\varphi}(\hat{x})\, d\hat{x}$ by a finite sum of the form $\sum_{l=1}^{L} \hat{\omega}_l \hat{\varphi}(\hat{b}_l)$, an approxi-

mation process which we shall symbolically represent by

$$\int_{\hat{K}} \hat{\varphi}(\hat{x})\,d\hat{x} \sim \sum_{l=1}^{L} \hat{\omega}_l \hat{\varphi}(\hat{b}_l). \tag{4.1.8}$$

The numbers $\hat{\omega}_l$ are called the *weights* and the points $\hat{b}_l$ are called the *nodes* of the *quadrature formula* $\sum_{l=1}^{L} \hat{\omega}_l \hat{\varphi}(\hat{b}_l)$. For simplicity, we shall consider in the sequel only examples for which *the nodes belong to the set $\hat{K}$ and the weights are strictly positive* (nodes outside the set $\hat{K}$ and negative weights are not excluded in principle, but, as expected, they generally result in quadrature schemes which behave poorly in actual computations).

In view of (4.1.7) and (4.1.8), we see that *the quadrature scheme over the set $\hat{K}$ automatically induces a quadrature scheme over the set K, namely*

$$\int_K \varphi(x)\,dx \sim \sum_{l=1}^{L} \omega_{l,K} \varphi(b_{l,K}), \tag{4.1.9}$$

with *weights* $\omega_{l,K}$ and *nodes* $b_{l,K}$ defined by

$$\omega_{l,K} = \det(B_K)\hat{\omega}_l \quad \text{and} \quad b_{l,K} = F_K(\hat{b}_l), \quad 1 \leq l \leq L. \tag{4.1.10}$$

Accordingly, we introduce the *quadrature error functionals*

$$E_K(\varphi) = \int_K \varphi(x)\,dx - \sum_{l=1}^{L} \omega_{l,K} \varphi(b_{l,K}), \tag{4.1.11}$$

$$\hat{E}(\hat{\varphi}) = \int_{\hat{K}} \hat{\varphi}(\hat{x})\,d\hat{x} - \sum_{l=1}^{L} \hat{\omega}_l \hat{\varphi}(\hat{b}_l), \tag{4.1.12}$$

which are related by

$$E_K(\varphi) = \det(B_K)\hat{E}(\hat{\varphi}). \tag{4.1.13}$$

Remark 4.1.1. It is realized from the previous description that *one needs only a numerical quadrature scheme over the reference finite element.* This is again in accordance with the pervading principle that most of the analysis needs to be done on the reference finite element only, just as was the case for the interpolation theory (Section 3.1). □

Let us now give a few examples of often used quadrature formulas. Notice that *each scheme preserves some space of polynomials* and it is this polynomial invariance that will subsequently play a crucial role in the problem of estimating the error.

More precisely, given a space $\hat{\Phi}$ of functions $\hat{\varphi}$ defined over the set $\hat{K}$, we shall say that the quadrature scheme is *exact for the space* $\hat{\Phi}$, *or exact for the functions* $\hat{\varphi} \in \hat{\Phi}$, if $\hat{E}(\hat{\varphi}) = 0$ for all $\hat{\varphi} \in \hat{\Phi}$.

Let $\hat{K}$ be an n-simplex with barycenter

$$\hat{a} = \frac{1}{(n+1)} \sum_{i=1}^{n+1} \hat{a}_i.$$

(Fig. 4.1.1).

Then *the quadrature scheme*

$$\int_{\hat{K}} \hat{\varphi}(\hat{x})\, d\hat{x} \sim \text{meas}(\hat{K})\hat{\varphi}(\hat{a}) \tag{4.1.14}$$

is exact for polynomials of degree ≤ 1, i.e.,

$$\forall \hat{\varphi} \in P_1(\hat{K}), \quad \int_{\hat{K}} \hat{\varphi}(\hat{x})\, d\hat{x} - \text{meas}(\hat{K})\hat{\varphi}(\hat{a}) = 0. \tag{4.1.15}$$

To see this, let

$$\hat{\varphi} = \sum_{i=1}^{n+1} \hat{\varphi}(\hat{a}_i)\hat{\lambda}_i$$

be any polynomial of degree ≤ 1. Then using the equalities

$$\int_{\hat{K}} \hat{\lambda}_i(\hat{x})\, d\hat{x} = \text{meas}(\hat{K})/(n+1)$$

(Exercise 4.1.1), $1 \leq i \leq n+1$, we obtain

$$\int_{\hat{K}} \hat{\varphi}(\hat{x})\, d\hat{x} = \frac{\text{meas}(\hat{K})}{n+1} \sum_{i=1}^{n+1} \hat{\varphi}(\hat{a}_i) = \text{meas}(\hat{K})\hat{\varphi}(\hat{a}).$$

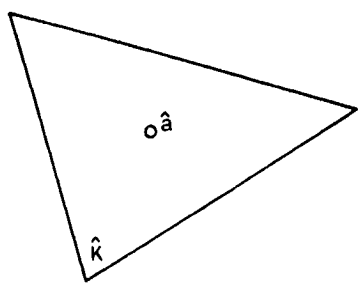


Fig. 4.1.1.

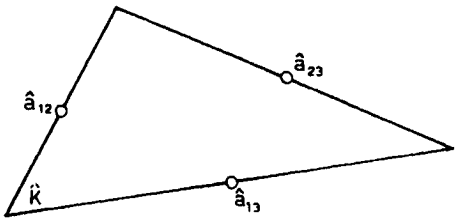


Fig. 4.1.2.

Let $n = 2$ and let $\hat{K}$ be a triangle with mid-points of the sides $\hat{a}_{ij}$, $1 \leq i < j \leq 3$ (Fig. 4.1.2).

Then *the quadrature scheme*

$$\int_{\hat{K}} \hat{\varphi}(\hat{x})\, d\hat{x} \sim \frac{\text{meas}(\hat{K})}{3} \sum_{1 \leq i < j \leq 3} \hat{\varphi}(\hat{a}_{ij}) \tag{4.1.16}$$

is exact for polynomials of degree ≤ 2 (cf. Exercise 4.1.1), i.e.,

$$\forall \hat{\varphi} \in P_2(\hat{K}), \quad \int_{\hat{K}} \hat{\varphi}(\hat{x})\, d\hat{x} - \frac{\text{meas}(\hat{K})}{3} \sum_{1 \leq i < j \leq 3} \hat{\varphi}(\hat{a}_{ij}) = 0. \tag{4.1.17}$$

Let $n = 2$ and let $\hat{K}$ be a triangle with vertices $\hat{a}_i$, $1 \leq i \leq 3$, with mid-points of the sides $\hat{a}_{ij}$, $1 \leq i < j \leq 3$, and with barycenter $\hat{a}_{123}$ (Fig. 4.1.3).

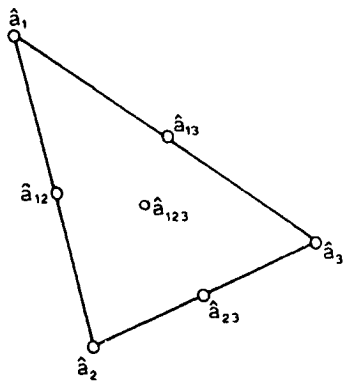


Fig. 4.1.3.

Then *the quadrature scheme*

$$\int_{\hat{K}} \hat{\varphi}(\hat{x})\,d\hat{x} \sim \frac{\text{meas}(\hat{K})}{60}\left(3\sum_{i=1}^{3}\hat{\varphi}(\hat{a}_i) + 8\sum_{1\leq i<j\leq 3}\hat{\varphi}(\hat{a}_{ij}) + 27\hat{\varphi}(\hat{a}_{123})\right) \tag{4.1.18}$$

is exact for polynomials of degree ≤ 3 (cf. Exercise 4.1.1), i.e., $\forall \hat{\varphi} \in P_3(\hat{K})$.

$$\int_{\hat{K}} \hat{\varphi}(\hat{x})\,d\hat{x} = \frac{\text{meas}(\hat{K})}{60}\left(3\sum_{i=1}^{3}\hat{\varphi}(\hat{a}_i) + 8\sum_{1\leq i<j\leq 3}\hat{\varphi}(\hat{a}_{ij}) + 27\hat{\varphi}(\hat{a}_{123})\right). \tag{4.1.19}$$

For examples of numerical quadrature schemes over rectangles, see Exercise 4.1.7.

Let us return to the definition of the discrete problem. Instead of solving the linear system (4.1.4) with the coefficients (4.1.5) and (4.1.6), all integrals $\int_K \ldots dx$ will be computed using a quadrature scheme given on the set $\hat{K}$. In other words, we are solving the *modified linear system*

$$\sum_{k=1}^{M} a_h(w_k, w_m)u_k = f_h(w_m), \quad 1 \leq m \leq M, \tag{4.1.20}$$

where (compare with (4.1.5) and (4.1.6) respectively)

$$a_h(w_k, w_m) = \sum_{K\in\mathcal{T}_h}\sum_{l=1}^{L}\omega_{l,K}\sum_{i,j=1}^{n}(a_{ij}\partial_i w_k \partial_j w_m)(b_{l,K}), \tag{4.1.21}$$

$$f_h(w_m) = \sum_{K\in\mathcal{T}_h}\sum_{l=1}^{L}\omega_{l,K}(fw_m)(b_{l,K}). \tag{4.1.22}$$

Remark 4.1.2. Conceivably, different quadrature formulas could be used for approximating the coefficients $a(w_k, w_m)$ on the one hand, and the coefficients $f(w_m)$ on the other hand. However, our final result (Theorem 4.1.6) will show that this is not necessary. □

For our subsequent analysis, rather than working with the linear system (4.1.20), it will be more convenient to consider the following equivalent formulation of the *discrete problem*: *We are looking for a discrete solution* $u_h \in V_h$ *which satisfies*

$$\forall v_h \in V_h, \quad a_h(u_h, v_h) = f_h(v_h), \tag{4.1.23}$$

where, for all functions $u_h, v_h \in V_h$, *the bilinear form* a_h *and the linear*

form f_h are respectively given by

$$a_h(u_h, v_h) = \sum_{K \in \mathcal{T}_h} \sum_{l=1}^{L} \omega_{l,K} \sum_{i,j=1}^{n} (a_{ij} \partial_i u_h \partial_j v_h)(b_{l,K}), \tag{4.1.24}$$

$$f_h(v_h) = \sum_{K \in \mathcal{T}_h} \sum_{l=1}^{L} \omega_{l,K} (f v_h)(b_{l,K}). \tag{4.1.25}$$

Looking at expressions (4.1.24) and (4.1.25), it is understood why the functions a_{ij} and f need to be defined everywhere over the set $\bar{\Omega}$. Also, in order that definition (4.1.21) make sense, it is necessary that, *over each finite element K, the first partial derivatives of the functions in the space X_h be unambiguously defined with unique extensions to the boundary of K*, should some node $b_{l,K}$ be situated on the boundary of K. If this node coincides with a node b_{l,K^*} corresponding to an adjacent finite element K^*, it should be clear that the values to be assigned to the derivatives $\partial_i v_h(b_{l,K})$ and $\partial_i v_h(b_{l,K^*})$ are generally different. Notice that, inasmuch as the definition of the discrete problem requires the knowledge of the values of the functions a_{ij} and f only at a finite number of points of $\bar{\Omega}$, it is quite reminiscent of finite-difference methods. In fact, this is true even to the point that *most classical finite difference schemes can be exactly interpreted as finite element methods with specific finite element spaces and specific quadrature schemes.* For results in this direction, see in particular Exercise 4.1.8. Conversely, a finite element method using Lagrange or Hermite finite elements (in which case one may always, at least theoretically, eliminate the unknowns which behave like derivatives) can be viewed as a finite difference method.

Abstract error estimate. The first Strang lemma

To sum up, we started out with a standard variational problem: Find $u \in V$ such that, for all $v \in V$, $a(u, v) = f(v)$, where the space V, the forms $a(\cdot, \cdot)$ and $f(\cdot)$ satisfy the assumptions of the Lax–Milgram lemma. Then given a finite-dimensional subspace V_h of the space V, the discrete problem consists in finding $u_h \in V_h$ such that, for all $v_h \in V_h$ $a_h(u_h, v_h) = f_h(v_h)$, where $a_h(\cdot, \cdot)$ is a bilinear form defined over the space V_h and $f_h(\cdot)$ is a linear form defined over the space V_h.

Notice that, in the present case, *the forms $a_h(\cdot, \cdot)$ and $f_h(\cdot)$ are not defined on the space V* (since the point values are not defined in general for functions in the space $H^1(\Omega)$).

Our first task is to prove an *abstract error estimate* adapted to the above abstract setting, but first we need some definitions.

For convenience, we shall refer to $a_h(\cdot,\cdot)$ as an *approximate bilinear form* and to $f_h(\cdot)$ as an *approximate linear form.* Denoting by $\|\cdot\|$ the norm of the space V, we shall say that approximate bilinear forms $a_h(\cdot,\cdot)$: $V_h \times V_h \to \mathbf{R}$, associated with a family of subspaces V_h of the space V, are *uniformly V_h-elliptic* if

$$\exists \tilde{\alpha} > 0, \quad \forall v_h \in V_h, \quad \tilde{\alpha}\|v_h\|^2 \leq a_h(v_h, v_h), \tag{4.1.26}$$

where the constant $\tilde{\alpha}$ is independent of the subspace V_h. Notice that such an assumption implies the existence of the discrete solutions.

Theorem 4.1.1 (*first Strang lemma*). *Consider a family of discrete problems for which the associated approximate bilinear forms are uniformly V_h-elliptic.*

Then there exists a constant C independent of the space V_h such that

$$\|u - u_h\| \leq C\Big(\inf_{v_h \in V_h} \Big\{ \|u - v_h\| + \sup_{w_h \in V_h} \frac{|a(v_h, w_h) - a_h(v_h, w_h)|}{\|w_h\|} \Big\} + \sup_{w_h \in V_h} \frac{|f(w_h) - f_h(w_h)|}{\|w_h\|} \Big). \tag{4.1.27}$$

Proof. Let v_h be an arbitrary element in the space V_h. With the assumption of uniform V_h-ellipticity, we may write:

$$\begin{aligned} \tilde{\alpha}\|u_h - v_h\|^2 &\leq a_h(u_h - v_h, u_h - v_h) \\ &= a(u - v_h, u_h - v_h) \\ &\quad + \{a(v_h, u_h - v_h) - a_h(v_h, u_h - v_h)\} \\ &\quad + \{f_h(u_h - v_h) - f(u_h - v_h)\}, \end{aligned}$$

so that, using the continuity of the bilinear from $a(\cdot,\cdot)$,

$$\begin{aligned} \tilde{\alpha}\|u_h - v_h\| &\leq M\|u - v_h\| + \frac{|a(v_h, u_h - v_h) - a_h(v_h, u_h - v_h)|}{\|u_h - v_h\|} \\ &\quad + \frac{|f_h(u_h - v_h) - f(u_h - v_h)|}{\|u_h - v_h\|} \\ &\leq M\|u - v_h\| + \sup_{w_h \in V_h} \frac{|a(v_h, w_h) - a_h(v_h, w_h)|}{\|w_h\|} \\ &\quad + \sup_{w_h \in V_h} \frac{|f_h(w_h) - f(w_h)|}{\|w_h\|}. \end{aligned}$$

Combining the above inequality with the triangular inequality

$$\|u - u_h\| \leq \|u - v_h\| + \|u_h - v_h\|$$

and taking the infimum with respect to $v_h \in V_h$ yields inequality (4.1.27). □

Remark 4.1.3. The abstract error estimate (4.1.27) generalizes the abstract error estimate established in Céa's lemma (Theorem 2.4.1) in the case of conforming finite element methods, since, in the absence of numerical integration, we would have $a_h(\cdot, \cdot) = a(\cdot, \cdot)$ and $f_h(\cdot) = f(\cdot)$. □

Sufficient conditions for uniform V_h-ellipticity

We now give sufficient conditions on a quadrature scheme which insure that the approximate bilinear forms are uniformly V_h-elliptic: Notice in particular that in the next theorem *assumptions* (i) *and* (ii) *exhibit the relationship which should exist between the reference finite element* $(\hat{K}, \hat{P}, \hat{\Sigma})$ *and the quadrature scheme defined on* $\hat{K}$ (for the case of negative weights, see Exercise 4.1.2).

Theorem 4.1.2. *Let there be given a quadrature scheme*

$$\int_{\hat{K}} \hat{\varphi}(\hat{x})\, d\hat{x} \sim \sum_{l=1}^{L} \hat{\omega}_l \hat{\varphi}(\hat{b}_l) \text{ with } \hat{\omega}_l > 0, \quad 1 \leq l \leq L,$$

over the reference finite element $(\hat{K}, \hat{P}, \hat{\Sigma})$, *for which there exists an integer* $k' \geq 1$ *such that*:

(i) *The inclusion* $\hat{P} \subset P_{k'}(\hat{K})$ *holds.*

(ii) *The union* $\bigcup_{l=1}^{L} \{\hat{b}_l\}$ *contains a* $P_{k'-1}(\hat{K})$*-unisolvent subset and/or the quadrature scheme is exact for the space* $P_{2k'-2}(\hat{K})$.

Then there exists a constant $\tilde{\alpha} > 0$ *independent of* h *such that, for all approximate bilinear forms of the form* (4.1.24) *and all spaces* V_h,

$$\forall v_h \in V_h, \quad \tilde{\alpha} |v_h|_{1,\Omega}^2 \leq a_h(v_h, v_h). \tag{4.1.28}$$

Proof. (i) Let us first assume that the union $\bigcup_{l=1}^{L} \{\hat{b}_l\}$ contains a $P_{k'-1}(\hat{K})$-unisolvent subset. Using the strict positivity of the weights, we find that

$$\hat{p} \in \hat{P} \quad \text{and} \quad \sum_{l=1}^{L} \hat{\omega}_l \sum_{i=1}^{n} (\partial_i \hat{p}(\hat{b}_l))^2 = 0 \Rightarrow \partial_i \hat{p}(\hat{b}_l) = 0,$$

$$1 \leq i \leq n, \quad 1 \leq l \leq L.$$

For each $i \in [1, n]$, the function $\partial_i \hat{p}$ is in the space $P_{k'-1}(\hat{K})$ by assumption (i), and thus it is identically zero since it vanishes on a $P_{k'-1}(\hat{K})$-unisolvent subset, by assumption (ii). As a consequence, the mapping

$$\hat{p} \to \left(\sum_{l=1}^{L} \hat{\omega}_l \sum_{i=1}^{n} (\partial_i \hat{p}(\hat{b}_l))^2 \right)^{1/2}$$

defines a norm over the quotient space $\hat{P}/P_0(\hat{K})$. Since the mapping $\hat{p} \to |\hat{p}|_{1,\hat{K}}$ is also a norm over this space and since this space is finite-dimensional, *there exists a constant* $\hat{C} > 0$ *such that*

$$\forall \hat{p} \in \hat{P}, \quad \hat{C}|\hat{p}|_{1,\hat{K}}^2 \leq \sum_{l=1}^{L} \hat{\omega}_l \sum_{i=1}^{n} (\partial_i \hat{p}(\hat{b}_l))^2. \tag{4.1.29}$$

If the quadrature scheme is exact for the space $P_{2k'-2}(\hat{K})$, the above inequality becomes an equality with $\hat{C} = 1$, since the function $\Sigma_{i=1}^{n} (\partial_i \hat{p})^2$ belongs to the space $P_{2k'-2}(\hat{K})$ for all $\hat{p} \in \hat{P}$ and since

$$\sum_{l=1}^{L} \hat{\omega}_l \sum_{i=1}^{n} (\partial_i \hat{p}(\hat{b}_l))^2$$

is precisely the quadrature formula which corresponds to the integral

$$\int_{\hat{K}} \sum_{i=1}^{n} (\partial_i \hat{p})^2 \, d\hat{x} = |\hat{p}|_{1,\hat{K}}^2.$$

(ii) Let us next consider the approximation of one of the integrals

$$\int_{\hat{K}} \sum_{i,j=1}^{n} a_{ij} \partial_i v_h \partial_j v_h \, dx.$$

Let $v_h|_K = p_K$ and let $\hat{p}_K \in \hat{P}$ be the function associated with p_K through the usual correspondence $\hat{x} \in \hat{K} \to F(\hat{x}) = B_K \hat{x} + b_K = x \in K$. We can write, using the ellipticity condition (4.1.2), and the positivity of the weights,

$$\sum_{l=1}^{L} \omega_{l,K} \sum_{i,j=1}^{n} (a_{ij} \partial_i v_h \partial_j v_h)(b_{l,K}) = \sum_{l=1}^{L} \omega_{l,K} \sum_{i,j=1}^{n} (a_{ij} \partial_i p_K \partial_j p_K)(b_{l,K})$$

$$\geq \beta \sum_{l=1}^{L} \omega_{l,K} \sum_{i=1}^{n} (\partial_i p_K(b_{l,K}))^2. \tag{4.1.30}$$

Observe that $\Sigma_{i=1}^{n} \partial_i p_K(b_{l,K}))^2$ is the square of the Euclidean norm $\|\cdot\|$ of the vector $Dp_K(b_{l,K})$. Since $\|D\hat{p}_K(\hat{b}_l)\| \leq \|B_K\| \, \|Dp_K(b_{l,K})\|$ (for all $\xi \in \mathbf{R}^n$, we have $D\hat{p}(\hat{b}_l)\xi = Dp(b_{l,K})(B_K \xi)$), we can write, using relations (4.1.10) and (4.1.29),

$$\sum_{l=1}^{L} \omega_{l,K} \sum_{i=1}^{n} (\partial_i p_K(b_{l,K}))^2 \geq \|B_K\|^{-2} \sum_{l=1}^{L} \omega_{l,K} \sum_{i=1}^{n} (\partial_i \hat{p}_K(\hat{b}_l))^2$$

$$= \det(B_K)\|B_K\|^{-2} \sum_{l=1}^{L} \hat{\omega}_l \sum_{i=1}^{n} (\partial_i \hat{p}_K(\hat{b}_l))^2$$

$$\geq \hat{C} \det(B_K)\|B_K\|^{-2} |\hat{p}_K|^2_{1,\hat{K}}$$

$$\geq \hat{C}(\|B_K\| \, \|B_K^{-1}\|)^{-2} |p_K|^2_{1,K}, \tag{4.1.31}$$

where we have also used Theorem 3.1.2. Since we are considering a regular family of triangulations, we have

$$\|B_K\| \, \|B_K^{-1}\| \leq \frac{\hat{h}}{\hat{\rho}} \frac{h_K}{\rho_K} \leq C, \tag{4.1.32}$$

for some constant C independent of $K \in \mathcal{T}_h$ and h. Combining inequalities (4.1.30), (4.1.31) and (4.1.32), we find that *there exists a constant* $\tilde{\alpha} > 0$ *independent of* $K \in \mathcal{T}_h$ *and* h *such that*

$$\forall v_h \in V_h, \quad \sum_{l=1}^{L} \omega_{l,K} \sum_{i,j=1}^{n} (a_{ij} \partial_i v_h \partial_j v_h)(b_{l,K}) \geq \tilde{\alpha} |v_h|^2_{1,K}. \tag{4.1.33}$$

(iii) It is then easy to conclude: Using inequalities (4.1.33) for all $K \in \mathcal{T}_h$, we obtain

$$\forall v_h \in V_h, \quad a_h(v_h, v_h) = \sum_{K \in \mathcal{T}_h} \sum_{l=1}^{L} \omega_{l,K} \sum_{i,j=1}^{n} (a_{ij} \partial_i v_h \partial_j v_h)(b_{l,K})$$

$$\geq \tilde{\alpha} \sum_{K \in \mathcal{T}_h} |v_h|^2_{1,K} = \tilde{\alpha} |v_h|^2_{1,\Omega}. \qquad \square$$

Remark 4.1.4. Notice that the expressions

$$\sum_{l=1}^{L} \hat{\omega}_l \sum_{i=1}^{n} (\partial_i \hat{p}_K(\hat{b}_l))^2$$

are exactly the approximations we get when we apply the quadrature scheme to the integrals $|\hat{p}_K|^2_{1,\hat{K}}$, which in turn correspond to the model problem $-\Delta u = f$ in Ω, $u = 0$ on Γ. Therefore it is natural to ask for assumptions (ii) which essentially guarantee that the mapping

$$\hat{p} \to \left(\sum_{l=1}^{L} \hat{\omega}_l \sum_{i=1}^{n} (\partial_i \hat{p}(\hat{b}_l))^2 \right)^{1/2}$$

is a norm over the quotient space $\hat{P}/P_0(\hat{K})$. $\square$

In view of this theorem, let us return to the examples of quadrature schemes given at the beginning of this section.

If $(\hat{K}, \hat{P}, \hat{\Sigma})$ is an n-simplex of type (1) ($\hat{P} = P_1(\hat{K})$ and thus $k' = 1$), we may use the quadrature scheme of (4.1.14) since $\{\hat{a}\}$ is a $P_0(\hat{K})$-unisolvent set.

If $(\hat{K}, \hat{P}, \hat{\Sigma})$ is a triangle of type (2) ($\hat{P} = P_2(\hat{K})$ and thus $k' = 2$), we may use the quadrature scheme of (4.1.16) since $\bigcup_{i<j}\{\hat{a}_{ij}\}$ is a $P_1(\hat{K})$-unisolvent set.

Notice that in both cases, the second assumption of (ii) is also satisfied.

If $(\hat{K}, \hat{P}, \hat{\Sigma})$ is a triangle of type (3) or (3') ($\hat{P} \subset P_3(\hat{K})$ and thus $k' = 3$), we may use the quadrature scheme of (4.1.18) since the set of numerical integration nodes (strictly) contains the $P_2(\hat{K})$-unisolvent subset $(\bigcup_i\{\hat{a}_i\}) \cup (\bigcup_{i<j}\{\hat{a}_{ij}\})$. However the quadrature scheme is not exact for the space $P_4(\hat{K})$ as the second assumption of (ii) would have required.

Consistency error estimates. The Bramble-Hilbert lemma

Now that the question of uniform V_h-ellipticity has been taken care of, we can turn to the problem of estimating the various terms appearing in the right-hand side of inequality (4.1.27). For the sake of clarity, we shall essentially concentrate on one special case (which nevertheless displays all the characteristic properties of the general case), namely the case where

$$\hat{P} = P_k(\hat{K})$$

for some integer $k \geq 1$ (the cases where $P_k(\hat{K}) \subset \hat{P} \subset P_{k'}(\hat{K})$ or where $P_k(\hat{K}) \subset \hat{P} \subset Q_k(\hat{K})$ are left as problems; cf. Exercises 4.1.6 and 4.1.7).

This being the case, if the solution is smooth enough so that it belongs to the space $H^{k+1}(\Omega)$, we have

$$\inf_{v_h \in V_h} \|u - v_h\|_{1,\Omega} \leq \|u - \Pi_h u\|_{1,\Omega} \leq Ch^k |u|_{k+1,\Omega},$$

assuming the X_h-interpolant of the solution u is well-defined, and thus, in the absence of numerical integration, we would have an $O(h^k)$ convergence. Then our basic objective is *to give sufficient conditions on the quadrature scheme which insure that the effect of numerical integration does not decrease this order of convergence.*

Remark 4.1.5. This criterion for appraising the required quality of the

quadrature scheme is perhaps arbitrary, but at least it is well-defined. Surprisingly, the results that shall be obtained in this fashion are nevertheless quite similar to the conclusions usually drawn by engineers through purely empirical criteria. □

Let us assume that the approximate bilinear forms are uniformly V_h-elliptic so that we may apply the abstract error estimate (4.1.27) of Theorem 4.1.1. Consequently, our aim is to obtain *consistency error estimates* of the form

$$\sup_{w_h \in V_h} \frac{|a(\Pi_h u, w_h) - a_h(\Pi_h u, w_h)|}{\|w_h\|_{1,\Omega}} \leq C(a_{ij}, u)h^k, \tag{4.1.34}$$

$$\sup_{w_h \in V_h} \frac{|f(w_h) - f_h(w_h)|}{\|w_h\|_{1,\Omega}} \leq C(f)h^k. \tag{4.1.35}$$

Notice that, in the usual terminology of numerical analysis, the uniform ellipticity condition appears as a *stability condition*, while the conditions (implied by the above error estimates)

$$\lim_{h \to 0} \sup_{w_h \in V_h} \frac{|a(\Pi_h u, w_h) - a_h(\Pi_h u, w_h)|}{\|w_h\|_{1,\Omega}} = 0,$$

$$\lim_{h \to 0} \sup_{w_h \in V_h} \frac{|f(w_h) - f_h(w_h)|}{\|w_h\|_{1,\Omega}} = 0,$$

appear as *consistency conditions*. This is why we call *consistency errors* the two terms of the form $\sup_{w_h \in V_h}(\ldots)$ appearing in the left-hand side of inequalities (4.1.34) and (4.1.35). By definition of the quadrature error functionals $E_K(\cdot)$ (cf.(4.1.11)), we have, for all $w_h \in V_h$,

$$a(\Pi_h u, w_h) - a_h(\Pi_h u, w_h) = \sum_{K \in \mathcal{T}_h} E_K\Big(\sum_{i,j=1}^{n} a_{ij}\partial_i(\Pi_h u)\partial_j w_h\Big), \tag{4.1.36}$$

$$f(w_h) - f_h(w_h) = \sum_{K \in \mathcal{T}_h} E_K(f w_h). \tag{4.1.37}$$

It turns out that we shall obtain (Theorems 4.1.4 and 4.1.5) "local" *quadrature error estimates* of the form

$$\forall p' \in P_K, \quad \forall p \in P_K, \quad |E_K(a_{ij}\partial_i p' \partial_j p)| \leq C(a_{ij|K}; \partial_i p')h_K^k |\partial_j p|_{0,K}, \tag{4.1.38}$$

$$\forall p \in P_K, \quad |E_K(fp)| \leq C(f_{|K})h_K^k \|p\|_{1,K}, \tag{4.1.39}$$

from which the "global" consistency error estimates (4.1.34) and (4.1.35) are deduced by an application of the Cauchy-Schwarz inequality (this is possible only because the constants $C(a_{ij|K}; \partial_i p')$ and $C(f_{|K})$ appearing in the above inequalities are of an oppropriate form).

To begin with, we prove a useful preliminary result.

Theorem 4.1.3 (*Bramble-Hilbert lemma*). *Let Ω be an open subset of $\mathbf{R}^n$ with a Lipschitz-continuous boundary. For some integer $k \geq 0$ and some number $p \in [0, \infty]$, let f be a continuous linear form on the space $W^{k+1,p}(\Omega)$ with the property that*

$$\forall p \in P_k(\Omega), \quad f(p) = 0. \tag{4.1.40}$$

Then there exists a constant $C(\Omega)$ such that

$$\forall v \in W^{k+1,p}(\Omega), \quad |f(v)| \leq C(\Omega)\|f\|^{\star}_{k+1,p,\Omega}|v|_{k+1,p,\Omega}, \tag{4.1.41}$$

where $\|\cdot\|^{\star}_{k+1,p,\Omega}$ is the norm in the dual space of $W^{k+1,p}(\Omega)$.

Proof. Let v be any function in the space $W^{k+1,p}(\Omega)$. Since by assumption, $f(v) = f(v + p)$ for all $p \in P_k(\Omega)$, we may write

$$\forall p \in P_k(\Omega), \quad |f(v)| = |f(v + p)| \leq \|f\|^{\star}_{k+1,p,\Omega}\|v + p\|_{k+1,p,\Omega},$$

and thus

$$|f(v)| \leq \|f\|^{\star}_{k+1,p,\Omega} \inf_{p \in P_k(\Omega)} \|v + p\|_{k+1,p,\Omega}.$$

The conclusion follows by Theorem 3.1.1. □

In the sequel, we shall often use the following result: *Let the functions $\varphi \in W^{m,q}(\Omega)$ and $w \in W^{m,\infty}(\Omega)$ be given. Then the function φw belongs to the space $W^{m,q}(\Omega)$, and*

$$|\varphi w|_{m,q,\Omega} \leq C \sum_{j=0}^{m} |\varphi|_{m-j,q,\Omega}|w|_{j,\infty,\Omega}, \tag{4.1.42}$$

for some constant C solely dependent upon the integers m and n, i.e., it is in particular independent of the set Ω.

To prove this, use the formula

$$\forall \alpha, \quad |\alpha| = m, \quad \partial^{\alpha}(\varphi w) = \sum_{j=0}^{m} \sum_{\substack{|\beta|=j \\ \beta+\beta'=\alpha}} \partial^{\beta} w \partial^{\beta'} \varphi,$$

in conjunction with inequalities of the form

$$\left|\sum_{\lambda=1}^{\Lambda} \alpha_\lambda f_\lambda\right|_{0,q,\Omega} \leq \sum_{\lambda=1}^{\Lambda} |a_\lambda|_{0,\infty,\Omega} |f_\lambda|_{0,q,\Omega}.$$

Theorem 4.1.4. *Assume that, for some integer* $k \geq 1$,

$$\hat{P} = P_k(\hat{K}), \tag{4.1.43}$$

$$\forall \hat{\varphi} \in P_{2k-2}(\hat{K}), \quad \hat{E}(\hat{\varphi}) = 0. \tag{4.1.44}$$

Then there exists a constant C *independent of* $K \in \mathcal{T}_h$ *and* h *such that*

$$\begin{aligned} &\forall a \in W^{k,\infty}(K), \quad \forall p \in P_k(K), \quad \forall p' \in P_k(K), \\ &|E_k(a \partial_i p' \partial_j p)| \leq C h_K^k \|a\|_{k,\infty,K} \|\partial_i p'\|_{k-1,K} |\partial_j p|_{0,K} \\ &\qquad \leq C h_K^k \|a\|_{k,\infty,K} \|p'\|_{k,K} |p|_{1,K}. \end{aligned} \tag{4.1.45}$$

Proof. We shall get an error estimate for the expression $E_K(avw)$ for $a \in W^{k,\infty}(K)$, $v \in P_{k-1}(K)$, $w \in P_{k-1}$. From (4.1.13), we infer that

$$E_K(avw) = \det(B_K)\, \hat{E}(\hat{a}\hat{v}\hat{w}), \tag{4.1.46}$$

with $\hat{a} \in W^{k,\infty}(\hat{K})$, $\hat{v} \in P_{k-1}(\hat{K})$, $\hat{w} \in P_{k-1}(\hat{K})$. For a given $\hat{w} \in P_{k-1}(\hat{K})$ and any $\hat{\varphi} \in W^{k,\infty}(\hat{K})$, we have ($W^{k,\infty}(\hat{K}) \hookrightarrow \mathscr{C}^0(\hat{K})$ since $k \geq 1$)

$$\begin{aligned} |\hat{E}(\hat{\varphi}\hat{w})| &= \left|\int_{\hat{K}} \hat{\varphi}\hat{w}\, d\hat{x} - \sum_{l=1}^{L} \hat{\omega}_l (\hat{\varphi}\hat{w})(\hat{b}_l)\right| \\ &\leq \hat{C} |\hat{\varphi}\hat{w}|_{0,\infty,\hat{K}} \leq \hat{C} |\hat{\varphi}|_{0,\infty,\hat{K}} |\hat{w}|_{0,\infty,\hat{K}}, \end{aligned}$$

where, here and subsequently, the letter $\hat{C}$ represents various constants solely dependent upon the reference finite element. Since $|\hat{\varphi}|_{0,\infty,\hat{K}} \leq \|\hat{\varphi}\|_{k,\infty,\hat{K}}$, and since all norms are equivalent on the finite-dimensional space $P_{k-1}(\hat{K})$, we deduce that

$$|\hat{E}(\hat{\varphi}\hat{w})| \leq \hat{C} \|\hat{\varphi}\|_{k,\infty,\hat{K}} |\hat{w}|_{0,\hat{K}}.$$

Thus, *for a given* $\hat{w} \in P_{k-1}(\hat{K})$, *the linear from*

$$\hat{\varphi} \in W^{k,\infty}(\hat{K}) \to \hat{E}(\hat{\varphi}\hat{w})$$

is continuous with norm $\leq \hat{C} |\hat{w}|_{0,\hat{K}}$ on the one hand, and it vanishes over the space $P_{k-1}(\hat{K})$ on the other hand, by assumption (4.1.44). Therefore, using the Bramble-Hilbert lemma, there exists a constant $\hat{C}$

such that

$$\forall \hat{\varphi} \in W^{k,\infty}(\hat{K}), \quad \forall \hat{w} \in P_{k-1}(\hat{K}),$$
$$|\hat{E}(\hat{\varphi}\hat{w})| \leq \hat{C}|\hat{\varphi}|_{k,\infty,\hat{K}}|\hat{w}|_{0,\hat{K}}.$$

Next, let $\hat{\varphi} = \hat{a}\hat{v}$ with $\hat{a} \in W^{k,\infty}(\hat{K})$, $\hat{v} \in P_{k-1}(\hat{K})$. Using (4.1.42) and taking into account that $|\hat{v}|_{k,\infty,\hat{K}} = 0$, we get

$$|\hat{\varphi}|_{k,\infty,\hat{K}} = |\hat{a}\hat{v}|_{k,\infty,\hat{K}} \leq \hat{C}\sum_{j=0}^{k-1} |\hat{a}|_{k-j,\infty,\hat{K}}|\hat{v}|_{j,\infty,\hat{K}} \leq \hat{C}\sum_{j=0}^{k-1} |\hat{a}|_{k-j,\infty,\hat{K}}|\hat{v}|_{j,\hat{K}},$$

where, in the last inequality, we have again used the equivalence of norms over the finite-dimensional space $P_{k-1}(\hat{K})$. Therefore, we obtain

$$\forall \hat{a} \in W^{k,\infty}(\hat{K}), \quad \forall \hat{v} \in P_{k-1}(\hat{K}), \quad \forall \hat{w} \in P_{k-1}(\hat{K}),$$
$$|\hat{E}(\hat{a}\hat{v}\hat{w})| \leq \hat{C}\left(\sum_{j=0}^{k-1} |\hat{a}|_{k-j,\infty,\hat{K}}|\hat{v}|_{j,\hat{K}}\right)|\hat{w}|_{0,\hat{K}}. \tag{4.1.47}$$

Then it suffices to use the inequalities (cf. Theorems 3.1.2 and 3.1.3)

$$|\hat{a}|_{k-j,\infty,\hat{K}} \leq \hat{C}h_K^{k-j}|a|_{k-j,\infty,K}, \quad 0 \leq j \leq k-1,$$
$$|\hat{v}|_{j,\hat{K}} \leq \hat{C}h_K^j(\det(B_K))^{-1/2}|v|_{j,K}, \quad 0 \leq j \leq k-1,$$
$$|\hat{w}|_{0,\hat{K}} \leq \hat{C}(\det(B_K))^{-1/2}|w|_{0,K},$$

in conjunction with relations (4.1.46) and (4.1.47). We obtain in this fashion:

$$\forall a \in W^{k,\infty}(K), \quad \forall v \in P_{k-1}(K), \quad \forall w \in P_{k-1}(K),$$
$$|E_K(avw)| \leq Ch_K^k\left(\sum_{j=0}^{k-1} |a|_{k-j,\infty,K}|v|_{j,K}\right)|w|_{0,K}$$
$$\leq Ch_K^k\|a\|_{k,\infty,K}\|v\|_{k-1,K}|w|_{0,K},$$

and the conclusion follows by replacing v by $\partial_i p'$ and w by $\partial_i p$ in the last inequality. □

Remark 4.1.6. Let us indicate why a *direct* application of the Bramble-Hilbert lemma to the quadrature error functionals $E_K(\cdot)$ (in this direction, see also Exercise 4.1.4) would not have yielded the proper estimate. Let us assume that

$$\forall \hat{\varphi} \in P_l(\hat{K}), \quad \hat{E}(\hat{\varphi}) = 0,$$

for some integer $l \geq 0$, and let $r \in [1, \infty]$ be such that the inclusion

$W^{l+1,r}(\hat{K}) \hookrightarrow \mathscr{C}^0(\hat{K})$ holds, so that we have

$$\forall \hat{\varphi} \in W^{l+1,r}(\hat{K}), \quad |\hat{E}(\hat{\varphi})| \leq \hat{C}|\hat{\varphi}|_{0,\infty,\hat{K}} \leq \hat{C}\|\hat{\varphi}\|_{l+1,r,\hat{K}}.$$

Then assumption (4.1.44), together with the Bramble-Hilbert lemma, implies that

$$\forall \hat{\varphi} \in W^{l+1,r}(\hat{K}), \quad |\hat{E}(\hat{\varphi})| \leq \hat{C}|\hat{\varphi}|_{l+1,r,\hat{K}}.$$

Let us then replace $\hat{\varphi}$ by the product $\hat{a}\hat{v}\hat{w}$, with a sufficiently smooth function $\hat{a}$, $\hat{v} \in P_{k-1}(\hat{K})$, $\hat{w} \in P_{k-1}(\hat{K})$. Using inequalities of the form (4.1.42) and the equivalence of norms over the space $P_{k-1}(\hat{K})$, we would automatically get all the semi-norms $|w|_{j,K}$, $0 \leq j \leq \min\{l+1, k-1\}$, in the right-hand side of the final inequality, whereas only the semi-norm $|w|_{o,K}$ should appear. □

The reader should notice that the ideas involved in the proof of the previous theorem are very reminiscent of those involved in the proof of Theorem 3.1.4. In both cases, the central idea is to apply the fundamental result of Theorem 3.1.1 (in the disguised form of the Bramble-Hilbert lemma in the present case) over the reference finite element and then to use the standard inequalities to go from the finite element K to $\hat{K}$, and back. The same analogies also hold for our next result.

Theorem 4.1.5. *Assume that, for some integer $k \geq 1$,*

$$\hat{P} = P_k(\hat{K}), \tag{4.1.48}$$

$$\forall \hat{\varphi} \in P_{2k-2}(\hat{K}), \quad \hat{E}(\hat{\varphi}) = 0, \tag{4.1.49}$$

and let $q \in [1, \infty]$ be any number which satisfies the inequality

$$k - \frac{n}{q} > 0. \tag{4.1.50}$$

Then there exists a constant C independent of $K \in \mathcal{T}_h$ and h such that

$$\forall f \in W^{k,q}(K), \quad \forall p \in P_k(K), \tag{4.1.51}$$
$$|E_K(fp)| \leq Ch_K^k(\text{meas}(K))^{(1/2)-(1/q)}\|f\|_{k,q,K}\|p\|_{1,K}.$$

Proof. For any $f \in W^{k,q}(K)$ and any $p \in P_k(K)$, we have

$$E_K(fp) = \det(B_K)\hat{E}(\hat{fp}), \tag{4.1.52}$$

with $\hat{f} \in W^{k,q}(\hat{K})$, $\hat{p} \in P_k(\hat{K})$. Let us write

$$\hat{E}(\hat{f}\hat{p}) = \hat{E}(\hat{f}\hat{\Pi}\hat{p}) + \hat{E}(\hat{f}(\hat{p} - \hat{\Pi}\hat{p})), \tag{4.1.53}$$

where $\hat{\Pi}$ *is the orthogonal projection in the space* $L^2(\hat{K})$ *onto the subspace* $P_1(\hat{K})$.

(i) *Let us estimate* $\hat{E}(\hat{f}\hat{\Pi}\hat{p})$. For all $\hat{\psi} \in W^{k,q}(\hat{K})$, we have

$$|\hat{E}(\hat{\psi})| \leq \hat{C}|\hat{\psi}|_{0,\infty,\hat{K}} \leq \hat{C}\|\hat{\psi}\|_{k,q,\hat{K}},$$

since inequality (4.1.50) implies that the inclusion $W^{k,q}(\hat{K}) \hookrightarrow \mathscr{C}^0(\hat{K})$ holds, and, in addition, $\hat{E}(\hat{\psi}) = 0$ for all $\hat{\psi} \in P_{k-1}(\hat{K})$, by virtue of assumption (4.1.49) (therefore, this assumption is not fully used at this stage, unless $k = 1$). Using the Bramble-Hilbert lemma, we obtain

$$\forall \hat{\psi} \in W^{k,q}(\hat{K}), \quad |\hat{E}(\hat{\psi})| \leq \hat{C}|\hat{\psi}|_{k,q,\hat{K}}.$$

In particular, let $\hat{\psi} = \hat{f}\hat{\Pi}\hat{p}$ with $\hat{f} \in W^{k,q}(\hat{K})$, $\hat{p} \in P_k(\hat{K})$. Using inequality (4.1.42), we find:

$$|\hat{f}\hat{\Pi}\hat{p}|_{k,q,\hat{K}} \leq \hat{C}(|\hat{f}|_{k,q,\hat{K}}|\hat{\Pi}\hat{p}|_{0,\infty,\hat{K}} + |\hat{f}|_{k-1,q,\hat{K}}|\hat{\Pi}\hat{p}|_{1,\infty,\hat{K}}),$$

since all semi-norms $|\hat{\Pi}\hat{p}|_{l,\infty,\hat{K}}$ are zero for $l \geq 2$ ($\hat{\Pi}\hat{p} \in P_1(\hat{K})$). Using the equivalence of norms over the finite-dimensional space $P_1(\hat{K})$, we get

$$|\hat{f}\hat{\Pi}\hat{p}|_{k,q,\hat{K}} \leq \hat{C}(|\hat{f}|_{k,q,\hat{K}}|\hat{\Pi}\hat{p}|_{0,\hat{K}} + |\hat{f}|_{k-1,q,\hat{K}}|\hat{\Pi}\hat{p}|_{1,\hat{K}}).$$

Further we have

$$|\hat{\Pi}\hat{p}|_{0,\hat{K}} \leq |\hat{p}|_{0,\hat{K}},$$

since $\hat{\Pi}$ is a projection operator, and

$$|\hat{\Pi}\hat{p}|_{1,\hat{K}} \leq |\hat{p} - \hat{\Pi}\hat{p}|_{1,\hat{K}} + |\hat{p}|_{1,\hat{K}}.$$

Applying Theorem 3.1.4 to the operator $\hat{\Pi}$, which leaves the space $P_0(\hat{K})$ invariant, we find, for some constant $\hat{C}$,

$$|\hat{p} - \hat{\Pi}\hat{p}|_{1,\hat{K}} \leq \hat{C}|\hat{p}|_{1,\hat{K}}.$$

Thus, upon combining all our previous inequalities, we have found a constant $\hat{C}$ such that

$$\forall \hat{f} \in W^{k,q}(\hat{K}), \quad \forall \hat{p} \in P_k(\hat{K}), \tag{4.1.54}$$
$$|\hat{E}(\hat{f}\hat{\Pi}\hat{p})| \leq \hat{C}(|\hat{f}|_{k,q,\hat{K}}|\hat{p}|_{0,\hat{K}} + |\hat{f}|_{k-1,q,\hat{K}}|\hat{p}|_{1,\hat{K}}).$$

(ii) *Let us next estimate* $\hat{E}(\hat{f}(\hat{p} - \hat{\Pi}\hat{p}))$. Observe that if $k = 1$, the

difference $(\hat{p} - \hat{\Pi}\hat{p})$ vanishes and therefore, we may henceforth assume that $k \geq 2$. This being the case, there exists a number $\rho \in [1, +\infty]$ such that the inclusions

$$W^{k,q}(\hat{K}) \hookrightarrow W^{k-1,\rho}(\hat{K}) \hookrightarrow C^0(\hat{K})$$

hold.

To see this, consider first the case where $1 \leq q < n$, and define a number ρ by letting $(1/\rho) = (1/q) - (1/n)$, so that the inclusion $W^{1,q}(\hat{K}) \hookrightarrow L^\rho(\hat{K})$ (and consequently the inclusion $W^{k,q}(\hat{K}) \hookrightarrow W^{k-1,\rho}(\hat{K})$) holds. Then the inclusion $W^{k-1,\rho}(\hat{K}) \hookrightarrow \mathscr{C}^0(\hat{K})$ also holds because we have $k - 1 - (n/\rho) = k - (n/q) > 0$ by (4.1.50).

Consider next the case where $n \leq q$. Then either $n < q$ and the inclusion $W^{1,q}(\hat{K}) \hookrightarrow L^\rho(\hat{K})$ holds for all $\rho \in [1, \infty]$, or $n = q$ and the same inclusion holds for all (finite) $\rho \geq 1$, so that in both cases the inclusion $W^{k,q}(\hat{K}) \hookrightarrow W^{k-1,\rho}(\hat{K})$ holds for all $\rho \geq 1$. Since in this part (ii) we assume $k \geq 2$, it suffices to choose ρ large enough so that $k - 1 - (n/\rho) > 0$ and then the inclusion $W^{k-1,\rho}(\hat{K}) \hookrightarrow \mathscr{C}^0(\hat{K})$ holds.

Proceeding with the familiar arguments, we eventually find that

$$\begin{aligned}
&\forall \hat{f} \in W^{k-1,\rho}(\hat{K}) \hookrightarrow \mathscr{C}^0(\hat{K}),\ \forall \hat{p} \in P_k(\hat{K}),\\
&|\hat{E}(\hat{f}(\hat{p} - \hat{\Pi}\hat{p}))| \leq \hat{C}|\hat{f}(\hat{p} - \hat{\Pi}\hat{p})|_{0,\infty,\hat{K}} \leq \hat{C}|\hat{f}|_{0,\infty,\hat{K}}|\hat{p} - \hat{\Pi}\hat{p}|_{0,\infty,\hat{K}}\\
&\qquad\qquad \leq \hat{C}\|\hat{f}\|_{k-1,\rho,\hat{K}}|\hat{p} - \hat{\Pi}\hat{p}|_{0,\infty,\hat{K}}.
\end{aligned}$$

Thus for a given $\hat{p} \in P_k(\hat{K})$, the linear form

$$\hat{f} \in W^{k-1,\rho}(\hat{K}) \to \hat{E}(\hat{f}(\hat{p} - \hat{\Pi}\hat{p}))$$

is continuous with norm $\leq \hat{C}|\hat{p} - \hat{\Pi}\hat{p}|_{0,\infty,\hat{K}}$, and it vanishes over the space $P_{k-2}(\hat{K})$ (notice that, contrary to step (i), the "full" assumption (4.1.49) is used here). Another application of the Bramble-Hilbert lemma shows that

$$\begin{aligned}
&\forall \hat{f} \in W^{k-1,\rho}(\hat{K}),\quad \forall \hat{p} \in P_k(\hat{K}),\\
&|\hat{E}(\hat{f}(\hat{p} - \hat{\Pi}\hat{p}))| \leq \hat{C}|\hat{f}|_{k-1,\rho,\hat{K}}|\hat{p} - \hat{\Pi}\hat{p}|_{0,\hat{K}}.
\end{aligned}$$

Since the operator $\hat{\Pi}$ leaves the space $P_0(\hat{K})$ invariant we have, again by Theorem 3.1.4,

$$|\hat{p} - \hat{\Pi}\hat{p}|_{0,\hat{K}} \leq \hat{C}|\hat{p}|_{1,\hat{K}}.$$

Also, we have

$$\forall \hat{g} \in W^{1,q}(\hat{K}),\quad |\hat{g}|_{0,\rho,\hat{K}} \leq \hat{C}(|\hat{g}|_{0,q,\hat{K}} + |\hat{g}|_{1,q,\hat{K}}),$$

since the inclusion $W^{1,q}(\hat{K}) \hookrightarrow L^p(\hat{K})$ holds, and thus

$$\forall \hat{f} \in W^{k,q}(\hat{K}), \quad |\hat{f}|_{k-1,p,\hat{K}} \leq \hat{C}(|\hat{f}|_{k-1,q,\hat{K}} + |\hat{f}|_{k,q,\hat{K}}).$$

Combining all our previous inequalities, we obtain:

$$\forall \hat{f} \in W^{k,q}(\hat{K}), \quad \forall \hat{p} \in P_k(\hat{K}), \tag{4.1.55}$$
$$|\hat{E}(\hat{f}(\hat{p} - \hat{\Pi}\hat{p}))| \leq \hat{C}(|\hat{f}|_{k-1,q,\hat{K}} + |\hat{f}|_{k,q,\hat{K}})|\hat{p}|_{1,\hat{K}}.$$

(iii) The proof is completed by combining relations (4.1.52), (4.1.53), (4.1.54), (4.1.55), and

$$|\hat{f}|_{k-j,q,\hat{K}} \leq \hat{C}h_K^{k-j}(\det(B_K))^{-1/q}|f|_{k-j,q,K}, \quad j = 0, 1,$$
$$|\hat{p}|_{j,\hat{K}} \leq \hat{C}h_K^{j}(\det(B_K))^{-1/2}|p|_{j,K}, \quad j = 0, 1. \qquad \square$$

Remark 4.1.7. Several comments are in order about this proof.

(i) First, there always exists a number q which satisfies inequality (4.1.50). In particular, the choice $q = \infty$ is possible in all cases.

(ii) Just as in the case of Theorem 4.1.4, a direct application of the Bramble-Hilbert lemma would yield unwanted norms in the right-hand side of the final inequality, which should be of the form $|E_k(fp)| \leq \cdots \|p\|_{1,K}$ (cf. Remark 4.1.6).

(iii) Why did we have to introduce the projection $\hat{\Pi}$? otherwise (arguing as in part (ii) of the proof), we would find either

$$|\hat{E}(\hat{f}\hat{p})| \leq \hat{C}|\hat{f}|_{k-1,p,\hat{K}}|\hat{p}|_{0,\hat{K}}, \quad \text{or} \quad |\hat{E}(\hat{f}\hat{p})| \leq \hat{C}|\hat{f}|_{k-1,p,\hat{K}}\|\hat{p}\|_{1,\hat{K}}.$$

In both cases, there would be a loss of one in the exponent of h_K.

(iv) Since in both steps (i) and (ii) of the proof, only the invariance of the space $P_0(\hat{K})$ through the projection operator is used, why did we not content ourselves with the orthogonal projection in the space $L^2(\hat{K})$ onto the subspace $P_0(\hat{K})$? Let us denote by $\hat{\Pi}_0$ such a projection mapping.

If $k \geq 2$, then the whole argument holds with $\hat{\Pi}_0$ instead of $\hat{\Pi}$. If $k = 1$ however, part (i) of the proof yields the inequality $|\hat{E}(\hat{f}\hat{\Pi}_0\hat{p})| \leq \hat{C}|\hat{f}|_{k,q,\hat{K}}|\hat{p}|_{0,\hat{K}}$, which is perfectly admissible, but then part (ii) of the proof is no longer empty and it is necessary to estimate the quantity $\hat{E}(\hat{f}(\hat{p} - \hat{\Pi}_0\hat{p}))$ for $\hat{p} \in P_1(\hat{K})$. But then it is impossible to find a space $W^{0,p}(\hat{K}) = L^p(\hat{K})$ which would be contained in the space $\mathscr{C}^0(\hat{K})$ with a continuous injection. Thus it is simply to avoid two distinct proofs (one with $\hat{\Pi}$ if $k = 1$, another one with $\hat{\Pi}_0$ if $k \geq 2$) that we have used the single mapping $\hat{\Pi}$.

(v) Why is it necessary to introduce the intermediate space $W^{k-1,\rho}(\hat{K})$? For all $\hat{p} \in P_k(\hat{K})$, the function $(\hat{p} - \hat{\Pi}\hat{p})$ is also a polynomial of degree $\leq k$. Since, on the other hand, the quadrature scheme is exact for polynomials of degree $\leq (2k-2)$, the application of the Bramble-Hilbert lemma to the linear form $\hat{f} \to \hat{E}(\hat{f}(\hat{p} - \hat{\Pi}\hat{p}))$ necessitates that the function $\hat{f}$ be taken in a Sobolev space which involves derivatives up to and including the order $(k-1)$, and no more. □

Estimate of the error $\|u - u_h\|_{1,\Omega}$

Combining the previous theorems, we obtain the main result of this section (compare with Theorem 3.2.2).

Theorem 4.1.6. *In addition to* $(H1)$, $(H2)$ *and* $(H3)$, *assume that there exists an integer* $k \geq 1$ *such that the following relations are satisfied*:

$$\hat{P} = P_k(\hat{K}), \tag{4.1.56}$$

$$H^{k+1}(\hat{K}) \hookrightarrow \mathscr{C}^s(\hat{K}), \tag{4.1.57}$$

where s *is the maximal order of partial derivatives occurring in the definition of the set* $\hat{\Sigma}$,

$$\forall \hat{\varphi} \in P_{2k-2}(\hat{K}), \quad \hat{E}(\hat{\varphi}) = 0. \tag{4.1.58}$$

Then if the solution $u \in H_0^1(\Omega)$ *of the variational problem corresponding to the data* (4.1.1) *belongs to the space* $H^{k+1}(\Omega)$, *if* $a_{ij} \in W^{k,\infty}(\Omega)$, $1 \leq i, j \leq n$, *and if* $f \in W^{k,q}(\Omega)$ *for some number* $q \geq 2$ *with* $k > (n/q)$, *there exists a constant* C *independent of* h *such that*

$$\|u - u_h\|_{1,\Omega} \leq Ch^k\Big(|u|_{k+1,\Omega} + \sum_{i,j=1}^{n} \|a_{ij}\|_{k,\infty,\Omega}\|u\|_{k+1,\Omega} + \|f\|_{k,q,\Omega}\Big), \tag{4.1.59}$$

where $u_h \in V_h$ *is the discrete solution.*

Proof. By virtue of the inclusion (4.1.57), we have (Theorem 3.2.1)

$$\|u - \Pi_h u\|_{1,\Omega} \leq Ch^k|u|_{k+1,\Omega},$$

where, here and subsequently, C stands for a constant independent of h.

Using (4.1.36), Theorem 4.1.4 and the Cauchy-Schwarz inequality, we

obtain for any $w_h \in V_h$,

$$\begin{aligned}|a(\Pi_h u, w_h) - a_h(\Pi_h u, w_h)| &\leq \sum_{K\in\mathcal{T}_h}\sum_{i,j=1}^{n} |E_K(a_{ij}\partial_i(\Pi_h u|_K)\partial_j(w_{h|K}))| \\ &\leq C\sum_{K\in\mathcal{T}_h} h_K^k \sum_{i,j=1}^{n} \|a_{ij}\|_{k,\infty,K}\|\Pi_h u\|_{k,K}|w_h|_{1,K} \\ &\leq Ch^k\Big(\sum_{i,j=1}^{n}\|a_{ij}\|_{k,\infty,\Omega}\Big) \\ &\quad\times\Big(\sum_{K\in\mathcal{T}_h}\|\Pi_h u\|_{k,K}^2\Big)^{1/2}|w_h|_{1,\Omega}.\end{aligned}$$

By Theorem 3.2.1, we have

$$\begin{aligned}\Big(\sum_{K\in\mathcal{T}_h}\|\Pi_h u\|_{k,K}^2\Big)^{1/2} &\leq \|u\|_{k,\Omega} + \Big(\sum_{K\in\mathcal{T}_h}\|u-\Pi_h u\|_{k,K}^2\Big)^{1/2} \\ &\leq \|u\|_{k,\Omega} + Ch|u|_{k+1,\Omega} \leq C\|u\|_{k+1,\Omega},\end{aligned}$$

and thus,

$$\begin{aligned}&\inf_{v_h\in V_h}\Big(\|u-v_h\|_{1,\Omega} + \sup_{w_h\in V_h}\frac{|a(v_h,w_h)-a_h(v_h,w_h)|}{\|w_h\|_{1,\Omega}}\Big) \\ &\leq \|u-\Pi_h u\|_{1,\Omega} + \sup_{w_h\in V_h}\frac{|a(\Pi_h u,w_h)-a_h(\Pi_h u,w_h)|}{\|w_h\|_{1,\Omega}} \\ &\leq Ch^k\Big(|u|_{k+1,\Omega} + \sum_{i,j=1}^{n}\|a_{ij}\|_{k,\infty,\Omega}\|u\|_{k+1,\Omega}\Big).\end{aligned}$$

Likewise, using (4.1.37) and Theorem 4.1.5, we obtain

$$\begin{aligned}|f(w_h) - f_h(w_h)| &\leq \sum_{K\in\mathcal{T}_h}|E_K(fw_h)| \\ &\leq C\sum_{K\in\mathcal{T}_h} h_K^k(\mathrm{meas}(K))^{(1/2)-(1/q)}\|f\|_{k,q,K}\|w_h\|_{1,K} \\ &\leq Ch^k\,\mathrm{meas}(\Omega)^{(1/2)-(1/q)}\|f\|_{k,q,\Omega}\|w_h\|_{1,\Omega},\end{aligned}$$

where, in the last inequality, we have made use of the inequality

$$\sum_K |a_K b_K c_K| \leq \Big(\sum_K |a_K|^\alpha\Big)^{1/\alpha}\Big(\sum_K |b_K|^\beta\Big)^{1/\beta}\Big(\sum_K |c_K|^\gamma\Big)^{1/\gamma}$$

valid for any numbers $\alpha \geq 1$, $\beta \geq 1$, $\gamma \geq 1$ which satisfy $(1/\alpha)+(1/\beta)+(1/\gamma) = 1$. Here, $(1/\alpha) = (1/2)-(1/q)$, $\beta = q$, $\gamma = 2$ (this is why the assumption $q \geq 2$ was needed).

Therefore we obtain

$$\sup_{w_h \in V_h} \frac{|f(w_h) - f_h(w_h)|}{\|w_h\|_{1,\Omega}} \leq Ch^k \operatorname{meas}(\Omega)^{(1/2)-(1/q)} \|f\|_{k,q,\Omega}.$$

To complete the proof, it suffices to use the abstract error estimate of Theorem 4.1.1 which we may indeed apply since, by virtue of assumptions (4.1.56) and (4.1.58), the approximate bilinear forms are uniformly V_h-elliptic (Theorem 4.1.2). □

Remark 4.1.8. When $\hat{P} = P_k(\hat{K})$, the condition that the quadrature scheme be exact for the space $P_{2k-2}(\hat{K})$ has a simple interpretation: It means that *all integrals* $\int_K a_{ij}\partial_i u_h \partial_j v_h \, dx$ *are exactly computed when all coefficients* a_{ij} *are constant functions.* To see this, notice that

$$\forall p', p \in P_K, \quad \int_K \partial_i p' \partial_j p \, dx = \int_{\hat{K}} \det B_K (\partial_i p')\hat{} \, (\partial_j p)\hat{} \, d\hat{x},$$

with

$$\det B_K = \text{constant}, \ (\partial_i p')\hat{} \in P_{k-1}(\hat{K}), \ (\partial_j p)\hat{} \in P_{k-1}(\hat{K}). \quad \square$$

To conclude, let us examine some applications of the last theorem:

If we are using n-simplices of type (1), then we still get $\|u - u_h\|_{1,\Omega} = O(h)$ provided we use a quadrature scheme exact for constant functions, such as that of (4.1.14).

If we use triangles of type (2), then we still get $\|u - u_h\|_{1,\Omega} = O(h^2)$ provided we use a quadrature scheme exact for polynomials of degree ≤ 2, such as that of (4.1.17).

If we use triangles of type (3), it would be necessary to use a quadrature scheme exact for polynomials of degree ≤ 4, in order to preserve the error estimate $\|u - u_h\|_{1,\Omega} = O(h^3)$, etc....

Exercises

4.1.1. (i) Let K be an n-simplex, and let $\lambda_i(x)$, $1 \leq i \leq n+1$, denote the barycentric coordinates of a point x with respect to the vertices of the n-simplex. Show that for any integers $\alpha_i \geq 0$, $1 \leq i \leq n+1$, one has

$$\int_K \lambda_1^{\alpha_1}(x)\lambda_2^{\alpha_2}(x) \ldots \lambda_{n+1}^{\alpha_{n+1}}(x) \, dx = \frac{\alpha_1!\alpha_2! \ldots \alpha_{n+1}! n!}{(\alpha_1 + \alpha_2 + \cdots + \alpha_{n+1} + n)!} \operatorname{meas}(K).$$

(ii) For $n = 2$, let $\hat{I}\hat{I}_2$ denote the $P_2(\hat{K})$-interpolation operator associated with the set $\hat{\Sigma} = \{p(\hat{a}_i),\ 1 \leq i \leq 3;\ \hat{p}(\hat{a}_{ij}),\ 1 \leq i < j \leq 3\}$. Using (i), show that the quadrature scheme of (4.1.16) can also be written

$$\int_{\hat{K}} \hat{\varphi}(\hat{x})\,d\hat{x} \sim \int_{\hat{K}} \hat{I}\hat{I}_2\hat{\varphi}(\hat{x})\,d\hat{x},$$

and consequently this scheme is exact for the space $P_2(\hat{K})$.

(iii) Show that in dimension 3, the same derivation would result in some strictly negative weights.

(iv) For $n = 2$, show that the set

$$\hat{\Sigma} = \{\hat{p}(\hat{a}_i),\quad 1 \leq i \leq 3;\quad \hat{p}(\hat{a}_{ij}),\quad 1 \leq i < j \leq 3;\quad \hat{p}(\hat{a}_{123})\}$$

is $\hat{P}$-unisolvent, where

$$\hat{P} = P_2(\hat{K}) \oplus \mathrm{V}\{\hat{\lambda}_1\hat{\lambda}_2\hat{\lambda}_3\}.$$

Using this fact combined with (i), show that the quadrature scheme of (4.1.18) can also be written

$$\int_{\hat{K}} \hat{\varphi}(\hat{x})\,d\hat{x} \sim \int_{\hat{K}} \hat{I}\hat{I}_2\hat{\varphi}(\hat{x})\,d\hat{x},$$

where $\hat{I}\hat{I}$ is the $\hat{P}$-interpolation operator.

(v) Show that the quadrature scheme of (4.1.18) is exact for the space $P_3(\hat{K})$, but not for the space $P_4(\hat{K})$.

4.1.2. Let there be given a quadrature scheme over the reference finite element for which the weights are not necessarily positive. Assume that there exists an integer k' such that the inclusion $\hat{P} \subset P_{k'}(\hat{K})$ holds and that the quadrature scheme is exact for the space $P_{2k'-2}(\hat{K})$.

(i) Show that there exists a constant C independent of $K \in \mathcal{T}_h$ and h such that

$$\forall p_K \in P_K,\quad \sum_{l=1}^{L} \omega_{l,K} \sum_{i,j=1}^{n} (a_{ij}\partial_i p_K \partial_j p_K)(b_{l,K}) \geq$$

$$\geq (\beta - C \max_{1 \leq i,j \leq n} \mathrm{osc}(a_{ij}; K))|p|^2_{1,K}.$$

(ii) Deduce from (i) that the approximate bilinear forms of the form (4.1.21) are uniformly V_h-elliptic for sufficiently small values of the parameter h, when the functions a_{ij}, $1 \leq i, j \leq n$, are continuous.

4.1.3. The purpose of this exercise is to obtain an abstract error estimate which generalizes that of Theorem 3.2.4 in the abstract setting of Theorem 4.1.1. Let H be a Hilbert space such that $\bar{V} = H$ with a continuous injection. With the same notations as in the text, show that

$$|u - u_h| \leq \sup_{g \in H} \frac{1}{|g|} \inf_{\varphi_h \in V_h} \{M\|u - u_h\| \, \|\varphi_g - \varphi_h\| + \\ + |a(u_h, \varphi_h) - a_h(u_h, \varphi_h)| + |f(\varphi_h) - f_h(\varphi_h)|\},$$

where $|\cdot|$ denotes the norm in H, and for each $g \in H$, the function $\varphi_g \in V$ is the unique solution of the variational problem

$$\forall v \in V, \quad a(v, \varphi_g) = g(v).$$

4.1.4. Let there be given a quadrature scheme over the reference finite element such that

$$\forall \hat{\varphi} \in P_l(\hat{K}), \quad \hat{E}(\hat{\varphi}) = 0$$

for some integer $l \geq 0$, and let $r \in [1, \infty]$ be such that the inclusion $W^{l+1,r}(\hat{K}) \hookrightarrow \mathscr{C}^0(\hat{K})$ holds.

Using the Bramble-Hilbert lemma, show that there exists a constant C independent of $K \in \mathscr{T}_h$ and h such that

$$\forall \varphi \in W^{l+1,r}(K), \quad |E_K(\varphi)| \leq C(\det B_K)^{1-(1/r)} h_K^{l+1} |\varphi|_{l+1,r,K}.$$

4.1.5. The purpose of this problem is to analyze the effect of numerical integration for the homogeneous Neumann problem corresponding to the following data:

$$\begin{cases} V = H^1(\Omega), \\ a(u, v) = \displaystyle\int_\Omega \left\{ \sum_{i,j=1}^n a_{ij} \partial_i u \partial_j v + auv \right\} dx, \\ f(v) = \displaystyle\int_\Omega fv \, dx, \end{cases}$$

where, in addition to the assumptions made at the beginning of the section, it is assumed that the function a is defined everywhere over the set $\bar{\Omega}$ and that

$$\exists a_0 > 0, \quad \forall x \in \bar{\Omega}, \quad a(x) \geq a_0 > 0.$$

Thus the discrete problem corresponds to the approximate bilinear

form

$$a_h(u_h, v_h) = \sum_{K\in\mathcal{T}_h}\sum_{l=1}^{L}\omega_{l,K}\sum_{i,j=1}^{n}(a_{ij}\partial_i u_h\partial_j v_h)(b_{l,K}) + \\ + \sum_{K\in\mathcal{T}_h}\sum_{l=1}^{L}\omega_{l,K}(au_h v_h)(b_{l,K}).$$

(i) With the same assumptions as in Theorem 4.1.2, show that there exists a constant $\tilde{\alpha} > 0$ such that

$$\forall v_h \in V_h, \quad \tilde{\alpha}\|v_h\|^2_{1,\Omega} \leq a_h(v_h, v_h).$$

(ii) Assume that, for some integer $k \geq 1$,

$$\hat{P} = P_k(\hat{K}),$$
$$\forall \hat{\varphi} \in P_{2k-2}(\hat{K}), \qquad \hat{E}(\hat{\varphi}) = 0.$$

Show that there exists a constant C independent of $K \in \mathcal{T}_h$ and h such that

$$\forall a \in W^{k,\infty}(K), \quad \forall p \in P_k(K), \quad \forall p' \in P_k(K),$$
$$|E_K(ap'p)| \leq Ch_K^k\|a\|_{k,\infty,K}\|p'\|_{k,K}\|p\|_{1,K}.$$

(iii) State and prove the analogue of Theorem 4.1.6 in this case.

4.1.6. The purpose of this problem is to consider the case where the space $\hat{P}$ satisfies the inclusions

$$P_k(\hat{K}) \subset \hat{P} \subset P_{k'}(\hat{K}).$$

In this case the question of V_h-ellipticity is already settled (cf. Theorem 4.1.2).

(i) Show that the analogues of Theorems 4.1.4 and 4.1.5 hold if the quadrature scheme is exact for the space $P_{k+k'-2}(\hat{K})$.

(ii) Deduce that the analogue of Theorem 4.1.6 holds if all the weights are positive, the union $\bigcup_{l=1}^{L}\{\hat{b}_l\}$ contains a $P_{k'-1}(\hat{K})$-unisolvent subset and the quadrature scheme is exact for the space $P_{k+k'-2}(\hat{K})$.

(iii) Deduce from this analysis that triangles of type (3′) may be used in conjunction with the quadrature scheme of (4.1.18). Could the quadrature scheme of (4.1.16) be used?

4.1.7. The purpose of this problem is to consider the case where the

space $\hat{P}$ satisfies the inclusions

$$P_k(\hat{K}) \subset \hat{P} \subset Q_k(\hat{K}),$$

i.e., essentially the case of rectangular finite elements.

(i) Let $n = 1$ and $K = [0, 1]$. It is well known that for each integer $k \geq 0$, there exist $(k+1)$ points $b_i \in [0, 1]$ and $(k+1)$ weights $\omega_i > 0$, $1 \leq i \leq k+1$, such that the quadrature scheme

$$\int_{[0,1]} \varphi(x)\,dx \sim \sum_{i=1}^{k+1} \omega_i \varphi(b_i)$$

is exact for the space $P_{2k+1}([0, 1])$. This particular quadrature formula is known as the *Gauss-Legendre formula.*

Then show that the quadrature scheme

$$\int_{[0,1]^n} \varphi(x)\,dx \sim \sum_{\substack{i_j=1 \\ 1 \leq j \leq n}}^{k+1} (\omega_{i_1}\omega_{i_2} \ldots \omega_{i_n})\varphi(b_{i_1}, b_{i_2}, \ldots, b_{i_n})$$

is exact for the space $Q_{2k+1}([0, 1]^n)$.

(ii) Assuming the positivity of the weights, show that the approximate bilinear forms are uniformly V_h-elliptic if the union $\bigcup_{l=1}^{L} \{\hat{b}_l\}$ contains a $Q_k(\hat{K}) \cap P_{nk-1}(\hat{K})$-unisolvent subset.

(iii) Show that the analogues of Theorems 4.1.4 and 4.1.5 hold if the quadrature scheme is exact for the space $Q_{2k-1}(\hat{K})$.

(iv) Deduce that the analogue of Theorem 4.1.6 holds if all the weights are positive, if the union $\bigcup_{l=1}^{L} \{\hat{b}_l\}$ contains a $Q_k(\hat{K}) \cap P_{nk-1}(\hat{K})$-unisolvent subset, and if the quadrature scheme is exact for the space $Q_{2k-1}(\hat{K})$.

As a consequence, and contrary to the case where $\hat{P} = P_k(\hat{K})$ (cf. Remark 4.1.8), it is no longer necessary to exactly compute the integrals

$$\int_K a_{ij}\partial_i u_h \partial_j v_h\,dx$$

when the coefficients a_{ij} are constant functions.

(v) Show that consequently one may use the Gauss-Legendre formula described in (i).

4.1.8. Let $\bar{\Omega} = [0, I\rho] \times [0, J\rho]$ where I and J are integers and ρ is a strictly positive number, and let $\mathcal{T}_h$ be a triangulation of the set $\bar{\Omega}$ made up of rectangles of type (1) of the form

$$[i\rho, (i+1)\rho] \times [j\rho, (j+1)\rho], \quad 0 \leq i \leq I-1, \quad 0 \leq j \leq J-1.$$

Let U_{ij} denote the unknown (usually denoted u_k) corresponding to the (k-th) node (ih, jh), $1 \leqslant i \leqslant I-1$, $1 \leqslant j \leqslant J-1$.

In what follows, we only consider nodes $(i\rho, j\rho)$ which are at least two squares away from the boundary of the set Ω, i.e., for which $2 \leqslant i \leqslant I-2$, $2 \leqslant j \leqslant J-2$.

Finally, we assume that the bilinear form is of the form

$$a(u, v) = \int_\Omega \sum_{l=1}^{n} \partial_l u \partial_l v \, dx,$$

i.e., the corresponding partial differential equation is the Poisson equation $-\Delta u = f$ in Ω.

(i) Show that, in the absence of numerical integration, the expression (usually denoted) $\Sigma_{k=1}^{M} a(w_k, w_m)u_k$ corresponding to the (m-th) node $(i\rho, j\rho)$ is, up to a constant factor, given by the expression

$$8U_{ij} - (U_{i+1,j} + U_{i+1,j+1} + U_{i,j+1} + U_{i-1,j+1} + U_{i-1,j} + U_{i-1,j-1} + U_{i,j-1} + U_{i+1,j-1}).$$

(ii) Assume that the quadrature scheme over the reference square $\hat{K} = [0, 1]^2$ is

$$\int_{[0,1]^2} \hat{\varphi}(\hat{x}) \, d\hat{x} \sim \frac{1}{4}(\hat{\varphi}(0, 0) + \hat{\varphi}(0, 1) + \hat{\varphi}(1, 1) + \hat{\varphi}(1, 0)).$$

Show that this quadrature scheme is exact for the space $Q_1(\hat{K})$. Since the set of nodes is $Q_1(\hat{K})$-unisolvent, the associated approximate bilinear forms are uniformly V_h-elliptic and therefore this scheme preserves the convergence in the norm $\|\cdot\|_{1,\Omega}$ (cf. Exercise 4.1.7). Show that the corresponding equality (usually denoted) $\Sigma_{k=1}^{M} a_h(w_k, w_m)u_k = f_h(w_m)$ is given by

$$4U_{ij} - (U_{i+1,j} + U_{i,j+1} + U_{i-1,j} + U_{i,j-1}) = \rho^2 f(i\rho, j\rho),$$

which is exactly *the standard five-point difference approximation to the equation*

$$-\Delta u = f.$$

(iii) Assume that the quadrature scheme over the reference square is

$$\int_{[0,1]^2} \hat{\varphi}(\hat{x}) \, d\hat{x} \sim \hat{\varphi}\left(\frac{1}{2}, \frac{1}{2}\right).$$

Show that this quadrature scheme is exact for the space $Q_1(\hat{K})$.

Show that the associated approximate bilinear forms are not uniformly V_h-elliptic, however.

Show that the expression (usually denoted) $\Sigma_{k=1}^{M} a_h(w_h, w_m)u_k$ is, up to a constant factor, given by the expression

$$4U_{ij} - (U_{i+1,j+1} + U_{i-1,j+1} + U_{i-1,j-1} + U_{i+1,j-1}).$$

It is interesting to notice that the predictably poor performance of such a method is confirmed by the geometrical structure of the above finite difference scheme, which is subdivided in two distinct schemes!

4.2. A nonconforming method

Nonconforming methods for second-order problems. Description of the resulting discrete problem

Let us assume for definiteness that we are solving a second-order boundary value problem corresponding to the following data:

$$\begin{cases} V = H_0^1(\Omega), \\ a(u, v) = \displaystyle\int_\Omega \sum_{i,j=1}^{n} a_{ij}\partial_i u \partial_j v \, dx, \\ f(v) = \displaystyle\int_\Omega fv \, dx. \end{cases} \tag{4.2.1}$$

At this essentially descriptive stage, the only assumptions which we need to record are that

$$a_{ij} \in L^\infty(\Omega), \quad 1 \leq i,j \leq n, \quad f \in L^2(\Omega), \tag{4.2.2}$$

and that the set $\bar{\Omega}$ is *polygonal*. Just as in the previous section, this last assumption is made so as to insure that the set $\bar{\Omega}$ can be exactly covered with triangulations. Given such a triangulation $\bar{\Omega} = \bigcup_{K \in \mathcal{T}_h} K$, we construct *a finite element space X_h whose generic finite element is not of class $\mathscr{C}^0$*. Then the space X_h will not be contained in the space $H^1(\Omega)$, as we show in the next theorem, which is the converse of Theorem 2.1.1.

Theorem 4.2.1. *Assume that the inclusions $P_K \subset \mathscr{C}^0(K)$ for all $K \in \mathcal{T}_h$*

and $X_h \subset H^1(\Omega)$ hold. Then the inclusion

$$X_h \subset \mathscr{C}^0(\bar{\Omega})$$

holds.

Proof. Let us assume that the conclusion is false. Then there exists a function $v \in X_h$, there exist two adjacent finite elements K_1 and K_2, and there exists a non empty open set $\mathcal{O} \subset K_1 \cup K_2$ such that (for example)

$$(v_{|K_1} - v_{|K_2}) > 0 \quad \text{along} \quad K' \cap \mathcal{O}, \tag{4.2.3}$$

where K' is the face common to K_1 and K_2. Let then φ be a (non zero) positive function in the space $\mathscr{D}(\mathcal{O}) \subset \mathscr{D}(\Omega)$. Using Green's formula (1.2.4), we have (with standard notations)

$$\begin{aligned}\int_\Omega \partial_i v\varphi \, dx &= \sum_{\lambda=1,2} \int_{K_\lambda} \partial_i v\varphi \, dx \\ &= -\sum_{\lambda=1,2} \int_{K_\lambda} v\partial_i\varphi \, dx + \sum_{\lambda=1,2} \int_{\partial K_\lambda} v|_{K_\lambda} \varphi \nu_{K_\lambda} \, d\gamma \\ &= -\int_\Omega v\partial_i\varphi \, dx + \int_{K'} (v_{|K_1} - v_{|K_2})\varphi\nu_{K_1} \, d\gamma,\end{aligned}$$

and thus we reach a contradiction since the integral along K' should be strictly positive by (4.2.3). □

For the time being, we shall simply assume that the inclusions

$$\forall K \in \mathcal{T}_h, \quad P_K \subset H^1(K), \tag{4.2.4}$$

hold, so that, in particular, the inclusion

$$X_h \subset L^2(\Omega) \tag{4.2.5}$$

holds. Then one defines a subspace X_{0h} of X_h which takes as well as possible into account the boundary condition $v = 0$ along the boundary Γ of Ω. For example, if the generic finite element is a Lagrange element, all degrees of freedom are set equal to zero at the boundary nodes. But, again because the finite element is not of class $\mathscr{C}^0$ (cf. Remark 2.3.10), *the functions in the space X_{0h} will in general vanish only at the boundary nodes.*

In order to define a discrete problem over the space $V_h = X_{0h}$, we observe that, if the linear form f is still defined over the space V_h by

virtue of the inclusion (4.2.5), this is not the case of the bilinear form $a(\cdot, \cdot)$. To obviate this difficulty, we *define*, in view of (4.2.1) and (4.2.4), *the approximate bilinear form*

$$a_h(u_h, v_h) = \sum_{K \in \mathcal{T}_h} \int_K \sum_{i,j=1}^{n} a_{ij} \partial_i u_h \partial_j v_h \, dx, \tag{4.2.6}$$

and *the discrete problem consists in finding a functio* $_h \in V_h$ *such that*

$$\forall v_h \in V_h, \quad a_h(u_h, v_h) = f(v_h). \tag{4.2.7}$$

We shall say that such a process of constructing a finite element approximation of a second-order boundary value problem is a *nonconforming finite element method.* By extension, any generic finite element which is used in such method is often called a *nonconforming finite element.*

Abstract error estimate. The second Strang lemma

In view of our subsequent analysis, we need, of course, to equip the space V_h with a norm. In analogy with the norm $|\cdot|_{1,\Omega}$ of the space $V = H_0^1(\Omega)$, a natural candidate is the mapping

$$v_h \to \|v_h\|_h = \left(\sum_{K \in \mathcal{T}_h} |v_h|_{1,K}^2 \right)^{1/2}, \tag{4.2.8}$$

which is *a priori only a semi-norm over the space* V_h. Thus, given a specific nonconforming finite element, the first task is to check that the mapping of (4.2.8) is indeed a norm on the space V_h. Once this is done, we shall be interested in showing that, for a family of spaces V_h, the approximate bilinear forms of (4.2.6) are *uniformly* V_h*-elliptic* in the sense that

$$\exists \tilde{\alpha} > 0, \quad \forall V_h, \forall v_h \in V_h, \quad \tilde{\alpha} \|v_h\|_h^2 \leq a_h(v_h, v_h). \tag{4.2.9}$$

This is the case if the ellipticity condition (cf. (4.1.2)) is satisfied.

Apart from implying the existence and uniqueness of the solution of the discrete problem, this condition is essential in order to obtain the abstract error estimate of Theorem 4.2.2 below.

From now on, we shall consider that the domain of definition of both the approximate bilinear form of (4.2.6) and the semi-norm of (4.2.8) is the space $V_h + V$. This being the case, notice that

$$\forall v \in V, \quad a_h(v, v) = a(v, v) \quad \text{and} \quad \|v\|_h = |v|_{1,\Omega}. \tag{4.2.10}$$

Also, the first assumptions (4.2.2) imply that there exists a constant $\tilde{M}$ independent of the space V_h such that

$$\forall u, v \in (V_h + V), \quad |a_h(u, v)| \leq \tilde{M}\|u\|_h\|v\|_h. \tag{4.2.11}$$

Theorem 4.2.2 (*second Strang lemma*). *Consider a family of discrete problems for which the associated approximate bilinear forms are uniformly V_h-elliptic.*

Then there exists a constant C independent of the subspace V_h such that

$$\|u - u_h\|_h \leq C\Big(\inf_{v_h \in V_h} \|u - v_h\|_h + \sup_{w_h \in V_h} \frac{|a_h(u, w_h) - f(w_h)|}{\|w_h\|_h}\Big). \tag{4.2.12}$$

Proof. Let v_h be an arbitrary element in the space V_h. Then in view of the uniform V_h-ellipticity and continuity of the bilinear forms a_h (cf. (4.2.9) and (4.2.11)) and of the definition (4.2.7) of the discrete problem, we may write

$$\begin{aligned}\tilde{\alpha}\|u_h - v_h\|_h^2 &\leq a_h(u_h - v_h, u_h - v_h)\\ &= a_h(u - v_h, u_h - v_h) + \{f(u_h - v_h) - a_h(u, u_h - v_h)\},\end{aligned}$$

from which we deduce

$$\begin{aligned}\tilde{\alpha}\|u_h - v_h\|_h &\leq \tilde{M}\|u - v_h\|_h + \frac{|f(u_h - v_h) - a_h(u, u_h - v_h)|}{\|u_h - v_h\|_h}\\ &\leq \tilde{M}\|u - v_h\|_h + \sup_{w_h \in V_h} \frac{|f(w_h) - a_h(u, w_h)|}{\|w_h\|_h}.\end{aligned}$$

Then inequality (4.2.12) follows from the above inequality and the triangular inequality

$$\|u - u_h\|_h \leq \|u - v_h\|_h + \|u_h - v_h\|_h. \qquad \square$$

Remark 4.2.1. The error estimate (4.2.12) indeed generalizes the error estimate which was established in Céa's lemma (Theorem 2.4.1) for conforming methods, since the difference $f(w_h) - a_h(u, w_h)$ is identically zero for all $w_h \in V_h$ when the space V_h is contained in the space V. $\square$

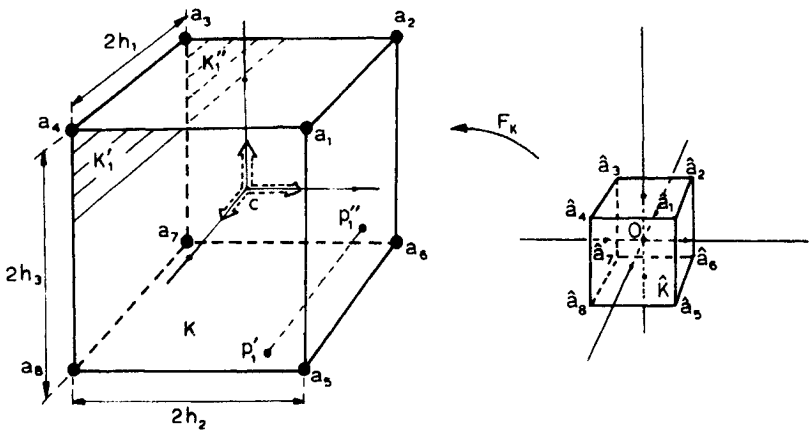
An example of a nonconforming finite element: Wilson's brick

Let us now describe a specific example of a nonconforming finite element known as *Wilson's brick*, which is used in particular in the approximation of problems of three-dimensional and two-dimensional elasticity posed over rectangular domains. We shall confine ourselves to the three-dimensional case, leaving the other case as a problem (Exercise 4.2.1).

Wilson's brick is an example of a *rectangular* finite element in $\mathbf{R}^3$, i.e., the set K is a 3-rectangle, whose vertices will be denoted a_i, $1 \leq i \leq 8$ (Fig. 4.2.1).

The space P_K is the space $P_2(K)$ to which are added linear combinations of the function $(x_1x_2x_3)$. Equivalently, we can think of the space P_K as being the space $Q_1(K)$ to which have been added linear combinations of the three functions x_j^2, $1 \leq j \leq 3$. We shall therefore record this definition by writing

$$P_K = P_2(K) \oplus \mathbf{V}\{x_1x_2x_3\} = Q_1(K) \oplus \mathbf{V}\{x_j^2, 1 \leq j \leq 3\}. \qquad (4.2.13)$$



Wilson's brick; $n = 3$
$P_K = Q_1(K) \oplus \mathbf{V}\{x_j^2, 1 \leq j \leq 3\}$; $\dim(P_K) = 11$; $\Sigma_K = \left\{p(a_i), 1 \leq i \leq 8; \dfrac{h_j^2}{h_1h_2h_3}\displaystyle\int_K \partial_{jj}p \, dx, 1 \leq j \leq 3\right\}$

Fig. 4.2.1

Notice that the inclusions

$$P_2(K) \subset P_K, \quad Q_1(K) \subset P_K \tag{4.2.14}$$

hold and that

$$\dim(P_K) = 11. \tag{4.2.15}$$

It is easily seen that *the values* $p(a_i)$, $1 \leq i \leq 8$, *at the vertices, together with the values of the* (constant) *second derivatives* $\partial_{jj}p$, $1 \leq j \leq 3$, *form a* P_K*-unisolvent set.* To see this, it suffices to check the validity of the following identity: For all functions $\hat{p} \in P_{\hat{K}}$, with $\hat{K} = [-1, +1]^3$, one has

$$\begin{aligned}\hat{p} = {} & \tfrac{1}{8}(1+x_1)(1+x_2)\{(1+x_3)\hat{p}(\hat{a}_1) + (1-x_3)\hat{p}(\hat{a}_5)\} \\ & + \tfrac{1}{8}(1-x_1)(1+x_2)\{(1+x_3)\hat{p}(\hat{a}_2) + (1-x_3)\hat{p}(\hat{a}_6)\} \\ & + \tfrac{1}{8}(1-x_1)(1-x_2)\{(1+x_3)\hat{p}(\hat{a}_3) + (1-x_3)\hat{p}(\hat{a}_7)\} \\ & + \tfrac{1}{8}(1+x_1)(1-x_2)\{(1+x_3)\hat{p}(\hat{a}_4) + (1-x_3)\hat{p}(\hat{a}_8)\} \\ & + \tfrac{1}{2}(x_1^2-1)\partial_{11}\hat{p} + \tfrac{1}{2}(x_2^2-1)\partial_{22}\hat{p} + \tfrac{1}{2}(x_3^2-1)\partial_{33}\hat{p}.\end{aligned} \tag{4.2.16}$$

Therefore if we denote by $c = \frac{1}{8}\sum_{i=1}^{8} a_i$ the center of the finite element K, one is naturally tempted to define the following set of degrees of freedom:

$$\Xi_K = \{p(a_i),\ 1 \leq i \leq 8;\ \partial_{jj}p(c),\ 1 \leq j \leq 3\}, \tag{4.2.17}$$

whose degrees of freedom are all in a familiar form. Of course, nothing obliges us to attach the last three degrees of freedom to the particular point c (except perhaps an aesthetical reason of symmetry), since the second derivatives $\partial_{jj}p$, $1 \leq j \leq 3$, are constant for any function $p \in P_K$.

Keeping this last property in mind, we may also choose for degrees of freedom the averages $\int_K \partial_{jj}p \, dx$, $1 \leq j \leq 3$, and we shall indeed show that this choice is more appropriate. For the time being, we observe that such degrees of freedom are of a new type, although they are still linear forms over the space $\mathscr{C}^2(K)$ as indeed they should be, to comply with the general definition given in Section 2.3.

Notice that since any function $p \in P_K$ satisfies

$$\partial_{jj}p(c) = \frac{1}{8h_1h_2h_3}\int_K \partial_{jj}p \, dx, \quad 1 \leq j \leq 3, \tag{4.2.18}$$

where $2h_j$, $1 \leq j \leq 3$, denote the lengths of the sides as indicated in Fig. 4.2.1, the two types of degrees of freedom are interchangeable over the space P_K. However, relations (4.2.18) do not hold in general for arbitrary

functions in the space $\mathscr{C}^2(K)$, and this is the basic reason why we obtain in this fashion two *different* finite elements (cf. Remark 4.2.2 below; also, this is an instance of a phenomenon that was mentioned in Remark 2.3.3).

Let us then equip Wilson's brick with degrees of freedom of the form (4.2.18). Our next objective is to extend the definition of affine-equivalence so that Wilson's bricks can be imbedded in an affine family, the reference finite element being in this case the hypercube $\hat{K} = [-1, +1]^3$. To do this, it suffices, according to Remark 2.3.5, to write the degrees of freedom in such a way that if we have the identity

$$\forall \hat{p} \in P_{\hat{K}}, \quad \hat{p} = \sum_{i=1}^{8} \hat{p}(\hat{a}_i)\hat{p}_i + \sum_{j=1}^{3} \hat{\phi}_j(\hat{p})\hat{q}_j, \tag{4.2.19}$$

then we also have the identity

$$\forall p \in P_K, \quad p = \sum_{i=1}^{8} p(a_i)p_i + \sum_{j=1}^{3} \phi_j(p)q_j, \tag{4.2.20}$$

where the basis functions $\hat{p}_i$ and p_i, resp. $\hat{q}_j$ and q_j, are in the usual correspondence (2.3.18), and $\hat{\phi}_j$ and ϕ_j, $1 \leq j \leq 3$, denote the degrees of freedom of the form $\int_K \partial_{jjj} p \, dx$, attached to the sets $\hat{K}$ and K, respectively. Using (4.2.16), we easily deduce that any function p in the space P_K satisfies the following identity, where c_i, $1 \leq i \leq 3$, denote the coordinates of the point c:

$$\begin{aligned} p = \frac{1}{8}\left(1 + \frac{(x_1 - c_1)}{h_1}\right)\left(1 + \frac{(x_2 - c_2)}{h_2}\right)\left\{\left(1 + \frac{(x_3 - c_3)}{h_3}\right)p(a_1)\right. \\ \left. + \left(1 - \frac{(x_3 - c_3)}{h_3}\right)p(a_5)\right\} + \cdots \\ + \sum_{j=1}^{3} \frac{1}{16}\left\{\left(\frac{x_j - c_j}{h_j}\right)^2 - 1\right\}\frac{h_j^2}{h_1 h_2 h_3}\int_K \partial_{jjj} p \, dx. \end{aligned} \tag{4.2.21}$$

Upon comparing (4.2.20) and (4.2.21) we find that the proper choices for ϕ_j and q_j are:

$$\begin{aligned} \phi_j(p) &= \frac{h_j^2}{h_1 h_2 h_3}\int_K \partial_{jjj} p \, dx, \\ q_j &= \frac{1}{16}\left(\left(\frac{x_j - c_j}{h_j}\right)^2 - 1\right), \quad 1 \leq j \leq 3. \end{aligned} \tag{4.2.22}$$

These choices insure that the following relations hold:

$$\begin{aligned} &p_i(a_k) = \delta_{ik}, \quad 1 \leq i, k \leq 8, \\ &q_j(a_i) = 0, \quad 1 \leq i \leq 8, \quad 1 \leq j \leq 3, \\ &\phi_j(p_i) = 0, \quad 1 \leq i \leq 8, \quad 1 \leq j \leq 3, \\ &\phi_l(q_j) = \delta_{lj}, \quad 1 \leq j, l \leq 3. \end{aligned} \tag{4.2.23}$$

Consequently, *we shall henceforth consider that the set of degrees of freedom of Wilson's brick is*

$$\Sigma_K = \left\{ p(a_i), \quad 1 \leq i \leq 8; \qquad \frac{h_j^2}{h_1 h_2 h_3} \int_K \partial_{jj} p \, dx, \quad 1 \leq j \leq 3 \right\}. \tag{4.2.24}$$

Notice that we could drop the multiplicative factors $h_j^2/h_1h_2h_3$ in the last degrees of freedom without changing the definition of the finite element.

Following definition (2.3.6), the associated operator Π_K is such that, for any sufficiently smooth function $v: K \to \mathbf{R}$, the function $\Pi_K v$ belongs to the space P_K and is uniquely determined by the conditions

$$\begin{aligned} &\Pi_K v(a_i) = v(a_i), \quad 1 \leq i \leq 8, \\ &\text{and} \quad \phi_j(\Pi_K v) = \phi_j(v), \; 1 \leq j \leq 3. \end{aligned} \tag{4.2.25}$$

Notice that the last three conditions can also be written as

$$\int_K \partial_{jj}(\Pi_K v) \, dx = \int_K \partial_{jj} v \, dx, \quad 1 \leq j \leq 3. \tag{4.2.26}$$

By construction, the P_K-interpolation operator satisfies

$$\widehat{(\Pi_K v)} = \Pi_{\hat{K}} \hat{v} \tag{4.2.27}$$

for functions v and $\hat{v}$ in the usual correspondence. Also, by virtue of the first relation (4.2.13), we have

$$\forall \hat{p} \in P_2(\hat{K}), \quad \Pi_{\hat{K}} \hat{p} = \hat{p}. \tag{4.2.28}$$

Remark 4.2.2. According to definition (2.3.9), the finite elements (K, P_K, Ξ_K) and (K, P_K, Σ_K) (cf. (4.2.17) and (4.2.24)) are *not* identical since the associated interpolation operators do not coincide over the space $\mathscr{C}^2(K)$ (ignoring momentarily that the domain of the interpolation operator corresponding to the set Σ_K is wider, as we next indicate). □

We are now in a position to explain the definite advantage of choosing the forms ϕ_j as degrees of freedom, rather than the point values $\partial_{ij}p(c)$. On the one hand the basic properties (4.2.27) and (4.2.28) of the interpolation operator are unaltered, but on the other hand, *the interpolation operator Π_K has a wider domain*: Whereas in the first case, one is led to assume that the function $v: K \to \mathbf{R}$ is twice differentiable over K in order to define its P_K-interpolant, in the second case the P_K-interpolant is well-defined for functions "only" in the space $H^2(K)$ (which is contained in the space $\mathscr{C}^0(K)$ for $n = 3$). This property will later avoid unnecessary restrictions on the smoothness of the solution u of our original problem (cf. Theorem 4.2.6).

Although the larger Sobolev space over which the P_K-interpolant is defined is the space $W^{2,p}(K)$ for $p > \frac{3}{2}$, we shall consider for simplicity that

$$\text{dom } \Pi_K = H^2(K). \tag{4.2.29}$$

In the next theorem, we shall estimate the interpolation errors $|v - \Pi_K v|_{m,K}$. The notations h_K and ρ_K represent the usual geometrical parameters (cf. (3.1.40)).

Theorem 4.2.3. *There exists a constant C such that, for all Wilson's bricks,*

$$\forall v \in H^l(\Omega), \quad |v - \Pi_K v|_{m,K} \leq C \frac{h_K^l}{\rho_K^m} |v|_{l,K}, \quad 0 \leq m \leq l, \quad l = 2, 3. \tag{4.2.30}$$

Proof. Using an argument similar to that used in the proof of Theorem 3.1.5, it can be checked that the mapping

$$\Pi_{\hat{K}}: H^l(\hat{K}) \subset H^2(\hat{K}) = \text{dom } \Pi_{\hat{K}} \to H^m(\hat{K})$$

is continuous for $0 \leq m \leq l$, $l = 2$ or 3. Combining this fact with relations (4.2.27) and (4.2.28), it only remains to apply Theorem 3.1.4. □

Let us assume that the set $\bar{\Omega}$ is rectangular so that it may be covered by triangulations $\mathscr{T}_h$ composed of 3-rectangles.

We then let X_h denote the finite element space whose functions v_h have the following properties: (i) For each $K \in \mathscr{T}_h$, the restrictions $v_{h|K}$ belong to the space P_K defined in (4.2.13). (ii) Each function $v_h \in X_h$ is

defined by its values at all the vertices and by the averages $\int_K \partial_{jj} v_h|_K \, dx$, $1 \leq j \leq 3$, $K \in \mathcal{T}_h$.

Since the basis functions q_j given in (4.2.22) do not vanish on the boundary of Wilson's brick, *this element is not of class* $\mathscr{C}^0$ *and the space* X_h *is not contained in the space* $H^1(\Omega)$, by Theorem 4.2.1. Continuity is however guaranteed at the vertices of the triangulations, since the functions q_j vanish at all nodes of Wilson's brick (cf. (4.2.23)).

Finally, we let $V_h = X_{0h}$, where X_{0h} denotes the space of all functions $v_h \in X_h$ which vanish at the boundary nodes. For the same reasons as before, the functions in the space X_{0h} do not vanish along the boundary Γ, but they vanish at the boundary nodes.

According to the analysis made at the beginning of this section, we need first to verify that the mapping $\|\cdot\|_h$ defined in (4.2.8) is indeed a norm over the space V_h.

Theorem 4.2.4. *The mapping*

$$v_h \to \|v_h\|_h = \left(\sum_{K \in \mathcal{T}_h} |v_h|^2_{1,K} \right)^{1/2} \tag{4.2.31}$$

is a norm over the space V_h.

Proof. Let v_h be a function in the space V_h which satisfies

$$\|v_h\|_h = \left(\sum_{K \in \mathcal{T}_h} |v_{h|K}|^2_{1,K} \right)^{1/2} = 0.$$

Then each polynomial $v_{h|K}$ is a constant so that one has $\partial_{jj}(v_{h|K}) = 0$, $1 \leq j \leq 3$, $K \in \mathcal{T}_h$, on the one hand. On the other, the function $v_h : \bar{\Omega} \to \mathbf{R}$ is a single constant since it is continuous at all the vertices and thus, it is identically zero since it vanishes at the boundary nodes. □

In order to simplify the exposition, *we shall henceforth assume that the bilinear form of* (4.2.1) *is*

$$a(u, v) = \int_\Omega \sum_{i=1}^{3} \partial_i u \partial_i v \, d\hat{x}, \tag{4.2.32}$$

i.e., the corresponding boundary value problem is a homogeneous Dirichlet problem for the operator $-\Delta$. In this particular case, the uniform V_h-ellipticity of the approximate bilinear forms is a consequence

of the identity

$$\forall v_h \in V_h, \quad \|v_h\|_h^2 = a_h(v_h, v_h). \tag{4.2.33}$$

This being the case, we may apply the abstract error estimate of Theorem 4.2.2. The first term, $\inf_{v_h \in V_h} \|u - v_h\|_h$, is easily taken care of: Assuming that we consider a family of discrete problems associated with a regular family of triangulations, and assuming that the solution u is in the space $H^2(\Omega)$, we deduce from Theorem 4.2.3 that

$$\inf_{v_h \in V_h} \|u - v_h\|_h \leq \left(\sum_{K \in \mathcal{T}_h} |u - \Pi_K u|_{1,K}^2 \right)^{1/2} \leq Ch|u|_{2,\Omega}. \tag{4.2.34}$$

Notice that the derivation of this interpolation error estimate uses in an essential manner the familiar implication (cf. (2.3.38))

$$v \in \operatorname{dom} \Pi_h = H^2(\Omega) \quad \text{and} \quad v_{|\Gamma} = 0 \Rightarrow \Pi_h v \in X_{0h},$$

where Π_h is the X_h-interpolation operator.

Remark 4.2.3. Of course, we could assume that $u \in H^3(\Omega)$, thus getting an $O(h^2)$ estimate instead of (4.2.34). However the eventual gain is nil because the other term in the right-hand side of inequality (4.2.12) is of order h, whatever the additional smoothness of the solution may be. Besides, we recall that the assumption $u \in H^2(\Omega)$ is realistic: One does not have a smoother solution in general on convex polygonal domains. □

Consistency error estimate. The bilinear lemma

Thus it remains to evaluate the other term, $\sup_{w_h \in V_h} |a_h(u, w_h) - f(w_h)|/\|w_h\|_h$, appearing in inequality (4.2.12) and this will be achieved through a careful analysis of the difference

$$D_h(u, w_h) = a_h(u, w_h) - f(w_h), \quad w_h \in V_h \tag{4.2.35}$$

(the consideration of the simplified bilinear form of (4.2.32) will allow for shorter computations in this process).

Since $-\Delta u = f$, we can write for any function $w_h \in V_h$,

$$\begin{aligned} D_h(u, w_h) &= \sum_{K \in \mathcal{T}_h} \int_K \sum_{i=1}^3 \partial_i u \partial_i w_h \, dx - \int_\Omega f w_h \, dx \\ &= \sum_{K \in \mathcal{T}_h} \int_K \left\{ \sum_{i=1}^3 \partial_i u \partial_i w_h \, dx + \Delta u w_h \right\} dx, \end{aligned} \tag{4.2.36}$$

i.e., we have obtained *one* decomposition of the form

$$D_h(u, w_h) = \sum_{K \in \mathcal{T}_h} D_K(u_{|K}, w_{h|K}), \tag{4.2.37}$$

where, for each $K \in \mathcal{T}_h$, the mapping $D_K(.\,,.)$ appears as a bilinear form over the space $H^2(K) \times P_K$. Ignoring for the time being that such a decomposition is not unique (we shall return to this crucial point later), let us assume that, for one decomposition of the form (4.2.37), we can show that there exists a constant C independent of $K \in \mathcal{T}_h$ and h such that

$$\forall v \in H^2(K), \quad \forall p \in P_K, \quad |D_K(v, p)| \leq Ch_K |v|_{2,K} |p|_{1,K}. \tag{4.2.38}$$

Then an application of Cauchy–Schwarz inequality yields

$$|D_h(u, w_h)| \leq Ch |u|_{2,\Omega} \|w_h\|_h, \tag{4.2.39}$$

and therefore we obtain

$$\sup_{w_h \in V_h} \frac{|a_h(u, w_h) - f(w_h)|}{\|w_h\|_h} \leq Ch |u|_{2,\Omega}, \tag{4.2.40}$$

i.e., *an estimate similar to that of* (4.2.34).

Remark 4.2.4. The term

$$\sup_{w_h \in V_h} \frac{|a_h(u, w_h) - f(w_h)|}{\|w_h\|_h}$$

is a *consistency error* term due to the "non conformity" of the method. Consequently, a sufficient condition for convergence is the *consistency condition*:

$$\lim_{h \to 0} \sup_{w_h \in V_h} \frac{|a_h(u, w_h) - f(w_h)|}{\|w_h\|_h} = 0. \qquad \square$$

For proving estimates such as (4.2.38), the following result turns out to be useful. It plays with respect to bilinear forms the role played by the Bramble–Hilbert lemma (Theorem 4.1.3) with respect to linear forms. For this reason, we shall at times refer to this result as the "*bilinear lemma*".

Theorem 4.2.5. *Let Ω be an open subset of $\mathbf{R}^n$ with a Lipschitz-continuous boundary. Let b be a continuous bilinear form over the space*

$W^{k+1,p}(\Omega) \times W$, where the space W satisfies the inclusions

$$P_l(\Omega) \subset W \subset W^{l+1,q}(\Omega), \tag{4.2.41}$$

and is equipped with the norm $\|\cdot\|_{l+1,q,\Omega}$. We assume that

$$\forall p \in P_k(\Omega), \quad \forall w \in W, \quad b(p, w) = 0, \tag{4.2.42}$$

$$\forall v \in W^{k+1,p}(\Omega), \quad \forall q \in P_l(\Omega), \quad b(v, q) = 0. \tag{4.2.43}$$

Then there exists a constant $C(\Omega)$ such that

$$\forall v \in W^{k+1,p}(\Omega), \quad \forall w \in W, \quad |b(v, w)| \leq \leq C(\Omega)\|b\|\,|v|_{k+1,p,\Omega}|w|_{l+1,q,\Omega}, \tag{4.2.44}$$

where $\|b\|$ is the norm of the bilinear form b in the space

$$\mathscr{L}_2(W^{k+1,p}(\Omega) \times W; \mathbf{R}).$$

Proof. Given a function $w \in W$, the linear form $b(.\,, w)$: $v \in W^{k+1,p}(\Omega) \to b(v, w)$ is continuous and it vanishes over the space $P_k(\Omega)$, by (4.2.42). Thus, by the Bramble–Hilbert lemma, there exists a constant $C_1(\Omega)$ such that

$$\forall v \in W^{k+1,p}(\Omega), \quad |b(v, w)| \leq C_1(\Omega)\|b(.\,, w)\|^{\star}_{k+1,p,\Omega}|v|_{k+1,p,\Omega}. \tag{4.2.45}$$

Using (4.2.43), we may write $b(v, w) = b(v, w + q)$ for all $q \in P_l(\Omega)$ so that we get

$$|b(v, w)| = |b(v, w + q)| \leq \|b\|\,\|v\|_{k+1,p,\Omega}\|w + q\|_{l+1,q,\Omega}.$$

Therefore,

$$\forall v \in W^{k+1,p}(\Omega), \quad \forall w \in W,$$

$$|b(v, w)| \leq \|b\|\,\|v\|_{k+1,p,\Omega} \inf_{q \in P_l(\Omega)} \|w + q\|_{l+1,q,\Omega}$$

$$\leq C_2(\Omega)\|b\|\,\|v\|_{k+1,p,\Omega}|w|_{l+1,q,\Omega},$$

as an application of Theorem 3.1.1 shows. Consequently,

$$\|b(.\,, w)\|^{\star}_{k+1,p,\Omega} = \sup_{v \in W^{k+1,p}(\Omega)} \frac{|b(v, w)|}{\|v\|_{k+1,p,\Omega}} \leq C_2(\Omega)\|b\|\,|w|_{l+1,q,\Omega}, \tag{4.2.46}$$

and inequality (4.2.44) follows by combining inequalities (4.2.45) and (4.2.46). □

Estimate of the error $(\Sigma_{K\in\mathcal{T}_h}|u-u_h|^2_{1,K})^{1/2}$

We now prove our main result.

Theorem 4.2.6. *Assume that the solution u is in the space* $H^2(\Omega)$. *Then for any regular family of triangulations there exists a constant C independent of h such that*

$$\|u-u_h\|_h = \left(\sum_{K\in\mathcal{T}_h}|u-u_h|^2_{1,K}\right)^{1/2} \leq Ch|u|_{2,\Omega}. \tag{4.2.47}$$

Proof. The central idea of the proof is to apply the bilinear lemma to each term $D_K(u, w_h)$ occuring in a decomposition of the expression $D_h(u, w_h)$ of the form (4.2.37). Some care has to be exercised, however: From (4.2.36), an obvious choice for the bilinear forms D_K is

$$v\in H^2(K),\quad p\in P_K \to \int_K \left\{\sum_{i=1}^3 \partial_i v\partial_i p + \Delta v p\right\} dx = \int_{\partial K}\partial_{\nu_K} v p\, d\gamma,$$

where ν_K denotes the outer normal along the boundary ∂K of the element K. However, there are not "enough" polynomial invariances at our disposal in such bilinear forms D_K in order to eventually obtain estimates of the form (4.2.38) (the reader should check this statement). Fortunately, there are other choices for a decomposition of the form (4.2.37) which will yield the right estimates. The key idea is *to obtain the desired additional "local" polynomial invariances from a "global" polynomial invariance*, as we now show.

Let Y_h denote the finite element space whose generic finite element is the rectangle of type (1). In other words:

(i) For each $K\in\mathcal{T}_h$, the restrictions $v_h|_K$ span the space $Q_1(K)$.

(ii) Each function $v_h\in Y_h$ is defined by its values at all the vertices of the triangulation. Then we let $W_h = Y_{0h}$ denote the space of all functions $v_h\in Y_h$ which vanish at the boundary nodes. Therefore the *inclusion*

$$W_h \subset \mathscr{C}^0(\bar{\Omega})\cap H^1_0(\Omega)$$

holds, and consequently (cf. Remark 4.2.1), we have

$$\forall v\in H^2(\Omega),\quad \forall w_h\in W_h,\quad D_h(v, w_h) = 0, \tag{4.2.48}$$

where it is henceforth understood that the function $D_h: H^2(\Omega)\times X_h$ is given by the second expression of (4.2.36), i.e.,

$$D_h(v, w_h) = \sum_{K\in\mathcal{T}_h}\int_K \left\{\sum_{i=1}^3 \partial_i v\partial_i w_h + \Delta v w_h\right\} dx. \tag{4.2.49}$$

Notice that the second inclusion of (4.2.14) implies that the inclusions

$$Y_h \subset X_h \quad \text{and} \quad Y_{0h} = W_h \subset X_{0h} = V_h \tag{4.2.50}$$

hold.

For any function $w_h \in X_h$, let $\Lambda_h w_h$ denote the unique function in the space Y_h which takes the same values as w_h at all the vertices of the triangulation. Notice that, for each $K \in \mathcal{T}_h$, $\Lambda_h w_{h|K} = \Lambda_K(w_{h|K})$, where Λ_K denotes the corresponding $Q_1(K)$-interpolation operator, and that the function $\Lambda_h w_h$ belongs to the space $W_h = Y_{0h}$ if the function w_h belongs to the space $V_h = X_{0h}$. Using relations (4.2.49), we deduce that

$$\forall v \in H^2(\Omega), \quad \forall w_h \in V_h, \quad D_h(v, w_h) = D_h(v, w_h - \Lambda_h w_h), \tag{4.2.51}$$

so that another possible decomposition of the difference $D_h(.,.)$ of (4.2.49) consists in writing

$$\forall v \in H^2(\Omega), \quad \forall w_h \in V_h, \quad D_h(v, w_h) = \sum_{K \in \mathcal{T}_h} D_K(v, w_h),$$

where the bilinear forms $D_K(.,.)$ are now given by

$$\forall v \in H^2(K), \quad \forall p \in P_K, \quad D_K(v, p) = \int_{\partial K} \partial_{\nu_K} v (p - \Lambda_K p)\, d\gamma. \tag{4.2.52}$$

We observe that, by definition of the operator Λ_K, we have

$$\forall v \in H^2(K), \quad \forall p \in Q_1(K), \quad D_K(v, p) = 0, \tag{4.2.53}$$

and thus we get a *first polynomial invariance.*

To obtain the other polynomial invariance, assume that the function v belongs to the space $P_1(K)$. Then the expression $D_K(v, p)$ is a sum of three terms, each of which is, up to a constant multiplicative factor, the difference between integrals of the expression $(p - \Lambda_K p)$ over opposite faces. Consider one such term, say (with the notations of Fig. 4.2.1):

$$\delta_1 = \int_{K_1'} (p - \Lambda_K p)\, dx_2\, dx_3 - \int_{K_1} (p - \Lambda_K p)\, dx_2\, dx_3. \tag{4.2.54}$$

Using the properties of the interpolation operator Λ_K, the identity (4.2.21), and the equations $\partial_{ij}(\Lambda_h p) = 0$, $1 \leq j \leq 3$, we deduce that

$$p - \Lambda_K p = \sum_{j=1}^{3} \frac{1}{16} \left(\left(\frac{x_j - c_j}{h_j} \right)^2 - 1 \right) \frac{h_j^2}{h_1 h_2 h_3} \int_K \partial_{jj} p \, dx. \tag{4.2.55}$$

Since the function $(((x_1 - c_1)/h_1)^2 - 1)$ vanishes along the faces K_1' and

K_1'', and since the functions $(((x_j - c_j)/h_j)^2 - 1)$, $j = 2, 3$, take on the same values at the points P_1' and P_1'' (cf. Fig. 4.2.1), we conclude that $\delta_1 = 0$. Likewise, the other similar terms vanish. Consequently, we obtain a *second polynomial invariance*:

$$\forall v \in P_1(K), \quad \forall p \in P_K, \quad D_K(v, p) = 0. \tag{4.2.56}$$

Each expression $D_K(v, p)$ found in (4.2.52) is of the form

$$D_K(v, p) = \sum_{j=1}^{3} \Delta_{j,K}(v, p), \tag{4.2.57}$$

where

$$\Delta_{1,K}(v, p) = \int_{K_1'} \partial_1 v(p - \Lambda_K p)\, dx_2\, dx_3 - \int_{K_1''} \partial_1 v(p - \Lambda_K p)\, dx_2\, dx_3, \tag{4.2.58}$$

and the expressions $\Delta_{2,K}(v, p)$ and $\Delta_{3,K}(v, p)$ are analogously defined.

Using the standard correspondences $\hat{v} \to v$ between the functions $\hat{v}: \hat{K} \to \mathbf{R}$ and $v: K \to \mathbf{R}$, we obtain

$$\Delta_{1,K}(v, p) = \frac{h_2 h_3}{h_1} \Delta_{1,\hat{K}}(\hat{v}, \hat{p}). \tag{4.2.59}$$

The previous analysis implies that, for each $j \in \{1, 2, 3\}$,

$$\begin{cases} \forall \hat{v} \in H^2(\hat{K}), & \forall \hat{p} \in P_0(\hat{K}), & \Delta_{j,\hat{K}}(\hat{v}, \hat{p}) = 0, \\ \forall \hat{v} \in P_1(\hat{K}), & \forall \hat{p} \in P_{\hat{K}}, & \Delta_{j,\hat{K}}(\hat{v}, \hat{p}) = 0, \end{cases} \tag{4.2.60}$$

so that, by the bilinear lemma, there exists a constant $\hat{C}$ such that

$$\forall \hat{v} \in H^2(\hat{K}), \quad \forall \hat{p} \in P_{\hat{K}}, \quad |\Delta_{j,\hat{K}}(\hat{v}, \hat{p})| \leq \hat{C} |\hat{v}|_{2,\hat{K}} |\hat{p}|_{1,\hat{K}}. \tag{4.2.61}$$

Using Theorem 3.1.2 and the regularity assumption, there exist constants C such that

$$|\hat{v}|_{2,\hat{K}} \leq C \|B_K\|^2 |\det(B_K)|^{-1/2} |v|_{2,K} \leq C h_K^{1/2} |v|_{2,K}, \tag{4.2.62}$$

$$|\hat{p}|_{1,\hat{K}} \leq C \|B_K\| \, |\det(B_K)|^{-1/2} |p|_{1,K} \leq C h_K^{-1/2} |p|_{1,K}, \tag{4.2.63}$$

so that, upon combining (4.2.57), (4.2.59), (4.2.61), (4.2.62), (4.2.63), we find that there exists a constant C such that

$$\forall K \in \mathcal{T}_h, \quad \forall v \in H^2(K), \quad \forall p \in P_K,$$
$$|D_K(v, p)| \leq C h_K |v|_{2,K} |p|_{1,K}.$$

This last inequality is of the form (4.2.38) and therefore the proof is complete. □

Remark 4.2.5. Loosely speaking, one may think of the space W_h introduced in the above proof as representing the "conforming" part of the otherwise "nonconforming" space V_h. □

Remark 4.2.6. Adding up equations (4.2.56), we find that

$$\forall p \in P_1(\bar{\Omega}), \quad \forall w_h \in V_h, \quad D_h(p, w_h) = 0.$$

In particular if we restrict ourselves to a basis function $w_i \in V$, whose support is a *patch* $\mathcal{P}_i$, i.e., a union of finite elements $K \in \mathcal{T}_h$, we find that

$$\forall p \in P_1(\mathcal{P}_i), \quad D_h(p, w_i) = 0.$$

This is an instance of *Irons patch test*, which B. Irons was the first to (empirically) recognize as a condition for getting convergence of a nonconforming finite element method. For further details about the patch test, see STRANG & FIX (1973, Section 4.2). □

Exercises

4.2.1. Describe the analog of Wilson's brick in dimension 2, which is known as *Wilson's rectangle* (there should be six degrees of freedom). For the application of this element to the system of plane elasticity, see LESAINT (1976).

4.2.2. Extend the analysis carried out in the text to the case of more general bilinear forms such as $a(u, v) = \int_\Omega \{\sum_{i,j=1}^n a_{ij}\partial_i u \partial_j v + auv\}\, dx$.

4.2.3. (i) Let H be a Hilbert space such that $\bar{V} = H$, $V \hookrightarrow H$, and $V_h \subset H$ for all h, and let, for all $u, v \in V_h + V$,

$$D_h(u, v) = a_h(u, v) - f(v).$$

Finally, assume that the bilinear form is symmetric.

Show that the estimate of the Aubin–Nitsche lemma (Theorem 3.2.4) is replaced in the present situation by

$$|u - u_h| \leq \sup_{g \in H} \frac{1}{|g|} \inf_{\varphi_h \in V_h} \{\tilde{M}\|u - u_h\| \, \|\varphi_g - \varphi_h\| + \\ + |D_h(u, \varphi_g - \varphi_h)| + |D_h(\varphi_g, u - u_h)|\},$$

where $|\cdot|$ denotes the norm in H, and for each $g \in H$, $\varphi_g \in V$ denotes the

unique solution of the variational problem

$$\forall v \in V, \quad a(v, \varphi_g) = g(v).$$

This abstract error estimate is found in NITSCHE (1974) and LASCAUX & LESAINT (1975).

(ii) Using part (i), show that, if the solution u is in the space $H^2(\Omega)$, one has (LESAINT (1976))

$$|u - u_h|_{0,\Omega} \leq Ch^2|u|_{2,\Omega}.$$

It is worth pointing out that, by contrast with (4.2.60), the "full" available polynomial invariances are used in the derivation of the above error estimate in the norm $|\cdot|_{0,\Omega}$.

4.3. Isoparametric finite elements

Isoparametric families of finite elements

Our first task consists in extending the notions of affine-equivalence and affine families which we discussed in Section 2.3. There, we saw how to generate finite elements through affine maps, a construction that will be generalized in Theorem 4.3.1 below. For simplicity we shall restrict ourselves in this section to Lagrange finite elements, leaving the case of Hermite finite elements as a problem (Exercise 4.3.1).

Theorem 4.3.1. *Let $(\hat{K}, \hat{P}, \hat{\Sigma})$ be a Lagrange finite element in $\mathbf{R}^n$ with $\hat{\Sigma} = \{\hat{p}(\hat{a}_i),\ 1 \leq i \leq N\}$ and let there be given a one-to-one mapping $F: \hat{x} \in \hat{K} \to (F_j(\hat{x}))_{j=1}^n \in \mathbf{R}^n$ such that*

$$F_j \in \hat{P}, \quad 1 \leq j \leq n. \tag{4.3.1}$$

Then if we let

$$\begin{cases} K = F(\hat{K}), \\ P = \{p: K \to \mathbf{R};\ \ p = \hat{p} \cdot F^{-1}, \quad \hat{p} \in \hat{P}\}, \\ \Sigma = \{p(F(\hat{a}_i));\ \ 1 \leq i \leq N\}, \end{cases} \tag{4.3.2}$$

the set Σ is P-unisolvent. Consequently, if K is a closed subset of $\mathbf{R}^n$ with a non-empty interior, the triple (K, P, Σ) is a Lagrange finite element.

Proof. Let us establish the bijections:

$$\hat{x} \in \hat{K} \to x = F(\hat{x}) \in K,$$
$$\hat{p} \in \hat{P} \to p = \hat{p} \cdot F^{-1} \in P.$$

If $\hat{p}_i$, $1 \leq i \leq N$, denote the basis functions of the finite element $(\hat{K}, \hat{P}, \hat{\Sigma})$, we have for all $p \in P$ and all $x \in K$,

$$p(x) = \hat{p}(\hat{x}) = \sum_{i=1}^{N} \hat{p}(\hat{a}_i)\hat{p}_i(\hat{x}) = \sum_{i=1}^{N} p(a_i)p_i(x),$$

i.e.,

$$\forall p \in P, \quad p = \sum_{i=1}^{N} p(a_i)p_i.$$

The functions p_i, $1 \leq i \leq N$, are linearly independent since $\Sigma_{i=1}^{N} \lambda_i p_i = 0$ implies $\Sigma_{i=1}^{N} \lambda_i \hat{p}_i = 0$ and therefore $\lambda_i = 0$, $1 \leq i \leq N$. In other words, we have shown that the set Σ is P-unisolvent, which completes the proof. □

We shall henceforth use the following notation: To indicate that a mapping $F: \hat{x} \in \hat{K} \to F(\hat{x}) = (F_j(\hat{x}))_{j=1}^{n} \in \mathbf{R}^n$ satisfies relations (4.3.1), we shall write:

$$F \in (\hat{P})^n \Leftrightarrow F_j \in \hat{P}, \quad 1 \leq j \leq n.$$

Notice that the construction of Theorem 4.3.1 is indeed a generalization of the construction which led to affine-equivalent finite elements, because the inclusion $P_1(\hat{K}) \subset \hat{P}$ is satisfied by all the finite elements hitherto considered.

With Theorem 4.3.1 in mind, we proceed to give several definitions: First, any finite element (K, P, Σ) constructed from another finite element $(\hat{K}, \hat{P}, \hat{\Sigma})$ through the process given in this theorem will be called an *isoparametric finite element*, and the finite element (K, P, Σ) will be said to be *isoparametrically equivalent* to the finite element $(\hat{K}, \hat{P}, \hat{\Sigma})$. Observe that this is *not* a symmetric relation in general, by contrast with the definition of affine-equivalence.

Next, we shall say that the family of finite elements (K, P_K, Σ_K) is an *isoparametric family* if all its elements are isoparametrically equivalent to a single finite element $(\hat{K}, \hat{P}, \hat{\Sigma})$, called the *reference finite element* of the family. In other words, for each K, there exists an *isoparametric*

mapping $F_K: \hat{K} \to \mathbb{R}^n$, i.e., which satisfies the relations

$$F_K \in (\hat{P})^n, \quad \textit{and} \quad F_K \textit{ is one-to-one}, \tag{4.3.3}$$

such that

$$\begin{cases} K = F_K(\hat{K}), \\ P_K = \{p: K \to \mathbb{R}; \quad p = \hat{p} \cdot F_K^{-1}; \ \hat{p} \in \hat{P}\}, \\ \Sigma_K = \{p(F_K(\hat{a}_i)); \quad 1 \leq i \leq N\}. \end{cases} \tag{4.3.4}$$

As exemplified by the special case of affine-equivalent finite elements, one may consider families of isoparametric finite elements for which the associated mappings F_K belong to some space $(\hat{Q})^n$, where $\hat{Q}$ is a strict subspace of the space $\hat{P}$. Such finite elements are sometimes called *subparametric finite elements*. For examples, see in particular Fig. 4.3.4.

Remark 4.3.1. The prefix "iso" in the adjective "isoparametric" refers to the fact that it is precisely the space $\hat{P}$ of the finite element $(\hat{K}, \hat{P}, \hat{\Sigma})$ which is used in the definition of the mapping F_K (and, consequently, in the definitions of the set K, the space P_K, and the set Σ_K, which use in turn the mapping F_K). □

It is worth pointing out that, by contrast with affine-equivalent finite elements, *the space* P_K *defined* in (4.3.4) *generally contains functions which are not polynomials, even when the space* $\hat{P}$ *consists of polynomials only* (see Exercise 4.3.3). However *this complication is ignored in practical computation*, inasmuch as all the computations are performed on the set $\hat{K}$, not on the set K. All that is needed is the knowledge of the mapping F_K, as we shall see in the next section.

In practice, an isoparametric finite element is not directly determined by a mapping F but, rather, by the data of N distinct points a_i, $1 \leq i \leq N$, which in turn uniquely determine a mapping F satisfying

$$F \in (\hat{P})^n \quad \text{and} \quad F(\hat{a}_i) = a_i, \quad 1 \leq i \leq N. \tag{4.3.5}$$

Such a mapping is given by

$$F: \hat{x} \in \hat{K} \to F(\hat{x}) = \sum_{i=1}^{N} \hat{p}_i(\hat{x}) a_i, \tag{4.3.6}$$

as it is readily verified, and it is uniquely defined since for each

$j \in \{1, 2, \ldots, n\}$, we must have, with $a_i = (a_{ji})_{j=1}^n$,

$$F_j \in \hat{P} \quad \text{and} \quad F_j(\hat{a}_i) = a_{ji}, \quad 1 \leq i \leq N,$$

and the set $\hat{\Sigma}$ is $\hat{P}$-unisolvent. However, *in the absence of additional assumptions, nothing guarantees that the mapping* $F: \hat{K} \to F(\hat{K})$ *is invertible,* and indeed this property will require a verification for each example.

Notice that the points a_i are the *nodes* of the finite element (K, P, Σ).

The main interest of isoparametric finite elements is that *the freedom in the choice of the points* a_i *yields more general geometric shapes of sets* K than the polygonal shapes considered up to now. As we shall show in the next section, this property is crucial for getting a good approximation of curved boundaries.

Examples of isoparametric finite elements

Let us next examine several instances of commonly used isoparametric finite elements. For brevity, we shall give a detailed discussion only for our first example, the *isoparametric n-simplex of type* (2), i.e., for which the finite element $(\hat{K}, \hat{P}, \hat{\Sigma})$ is the n-simplex of type (2). Such an isoparametric finite element is determined by the data of $(n+1)$ *vertices* a_i, $1 \leq i \leq n+1$, and $n(n+1)/2$ points which we shall denote by a_{ij}, $1 \leq i < j \leq n+1$. Then (cf. (4.3.5)) there exists a unique mapping F such that

$$\begin{cases} F \in (P_2(\hat{K}))^n, \\ F(\hat{a}_i) = a_i, \quad 1 \leq i \leq n+1, \\ F(\hat{a}_{ij}) = a_{ij}, \quad 1 \leq i < j \leq n+1. \end{cases}$$

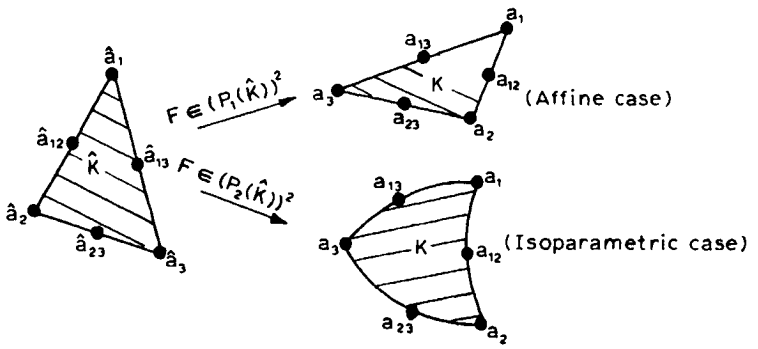
This mapping is given by (cf. (2.2.9) and (4.3.6))

$$F: \hat{x} \in \hat{K} \to F(\hat{x}) = \sum_i \lambda_i(\hat{x})(2\lambda_i(\hat{x}) - 1)a_i + \sum_{i<j} 4\lambda_i(\hat{x})\lambda_j(\hat{x})a_{ij}. \tag{4.3.7}$$

Observe that *if it so happened that the points* a_{ij} *were exactly the mid-points* $(a_i + a_j)/2$, *then,* by virtue of the uniqueness of the mapping F, *the mapping* F *would "degenerate" and become affine.*

These considerations are illustrated in Fig. 4.3.1 for $n = 2$, i.e., in the case of the *isoparametric triangle of type* (2).

It is only later (Theorem 4.3.3) that we shall give sufficient conditions



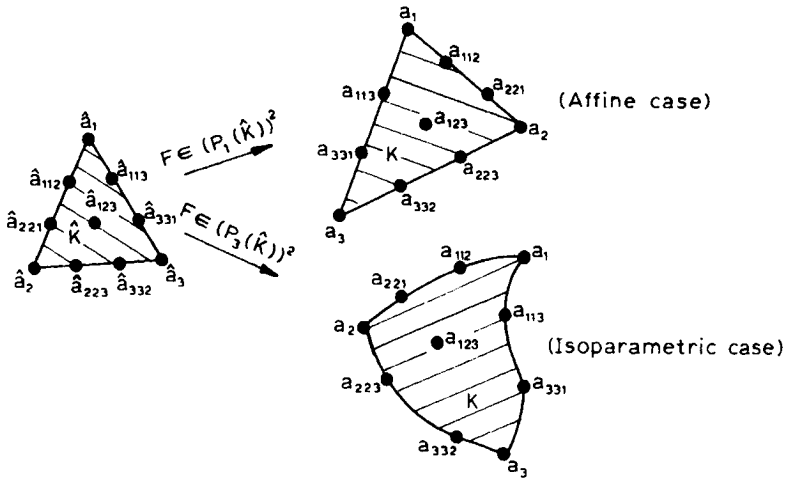
Isoparametric triangle of type (2)
Fig. 4.3.1

which guarantee the invertibility of the mapping F of (4.3.7), but at least we can already indicate that these conditions proceed from a natural idea: When $n = 2$ (cf. Fig. 4.3.1), let us assume that the three vertices a_i, $1 \leq i \leq 3$, are the vertices of a nondegenerate triangle $\hat{K}$. Then the mapping $F: \hat{K} \to K$ is invertible if the points a_{ij} are not "too far" from the actual mid-points $(a_i + a_j)/2$ of the triangle $\hat{K}$ (for a counter-example, see Exercise 4.3.4).

The *boundary* of the set $K = F(\hat{K})$ is composed of *faces*, i.e., the images $F(\hat{K}')$ of the faces $\hat{K}'$ of the n-simplex $\hat{K}$. Since each basis function $\hat{\varphi}$ of the n-simplex $\hat{K}$ of type (2) vanishes along any face of $\hat{K}$ which does not contain the node associated with $\hat{\varphi}$ (cf. Remark 2.3.10), we conclude that each face of the isoparametric n-simplex of type (2) is solely determined by the nodes through which it passes (see also Exercise 4.3.4). This property, which is true of all isoparametric finite elements considered in the sequel (as the reader may check) allows the construction of triangulations made up of isoparametric finite elements (cf. Section 4.4).

We can similarly consider the *isoparametric n-simplex of type* (3) (cf. Fig. 4.3.2 for $n = 2$), for which the mapping F is given by (cf. (2.2.10)):

$$F: \hat{x} \in \hat{K} \to F(\hat{x}) = \sum_i \frac{\lambda_i(\hat{x})(3\lambda_i(\hat{x}) - 1)(3\lambda_i(\hat{x}) - 2)}{2} a_i + \sum_{i \neq j} \frac{9\lambda_i(\hat{x})\lambda_j(\hat{x})(3\lambda_i(\hat{x}) - 1)}{2} a_{iij} + \sum_{i<j<k} 27\lambda_i(\hat{x})\lambda_j(\hat{x})\lambda_k(\hat{x}) a_{ijk}. \tag{4.3.8}$$



Isoparametric triangle of type (3).
Fig. 4.3.2

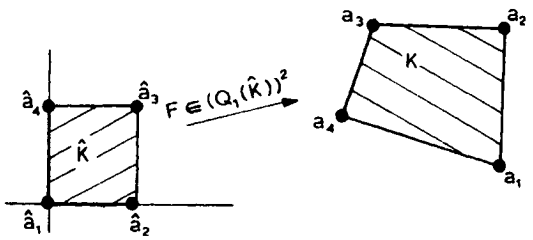
Observe in this case that, for $n = 2$, even if *the point* a_{123} *plays no role in the definition of the boundary of the set* K, the space P_K still depends on its position. We leave it to the reader to similarly define the *isoparametric n-simplex of type* (3'), and the *isoparametric n-simplex of type* (k) for any integer $k \geq 1$. All these isoparametric finite elements are instances of *simplicial* (or *triangular* if $n = 2$, or *tetrahedral* if $n = 3$) *isoparametric finite elements* in the sense that they are isoparametrically equivalent to a finite element for which the set $\hat{K}$ is an n-simplex.

We next describe some examples of *quadrilateral finite elements*, in the sense that they are isoparametrically equivalent to a finite element for which the set $\hat{K}$ is an n-rectangle, for example the unit hypercube $\hat{K} = [0, 1]^n$. In this fashion we obtain the *quadrilateral of type* (1) (cf. Fig. 4.3.3 for $n = 2$).

For $n = 2$, this is an example of a true isoparametric finite element whose sides are not curved! This is so because the functions in the space $Q_1([0, 1]^2)$ are affine in the direction of each coordinate axis. However, this is special to dimension 2. If $n = 3$ for instance, the faces of the set K are portions of hyperbolic paraboloids and are therefore generally curved.

Another example of a quadrilateral finite element is the *quadrilateral of type* (2). In Fig. 4.3.4, we have indicated various subparametric cases of interest for this element, when $n = 2$.

Given a finite element (K, P, Σ) isoparametrically equivalent to a finite



Quadrilateral of type (1) for $n = 2$
Fig. 4.3.3

element $(\hat{K}, \hat{P}, \hat{\Sigma})$ through a mapping F, we shall use the usual correspondences

$$\hat{x} \in \hat{K} \to x = F(\hat{x}) \in K, \tag{4.3.9}$$

$$\hat{p} \in \hat{P} \to p = \hat{p} \cdot F^{-1} \in P, \tag{4.3.10}$$

between the points in the sets $\hat{K}$ and K, and between the functions in the spaces $\hat{P}$ and P, respectively. We shall extend the correspondence (4.3.10) to functions defined over the sets $\hat{K}$ and K by letting

$$(\hat{v}: \hat{K} \to \mathbf{R}) \to (v = \hat{v} \cdot F^{-1}: K \to \mathbf{R}). \tag{4.3.11}$$

Then it is an easy matter to see that the associated $\hat{P}$-interpolation and P-interpolation operators $\hat{\Pi}$ and Π are such that

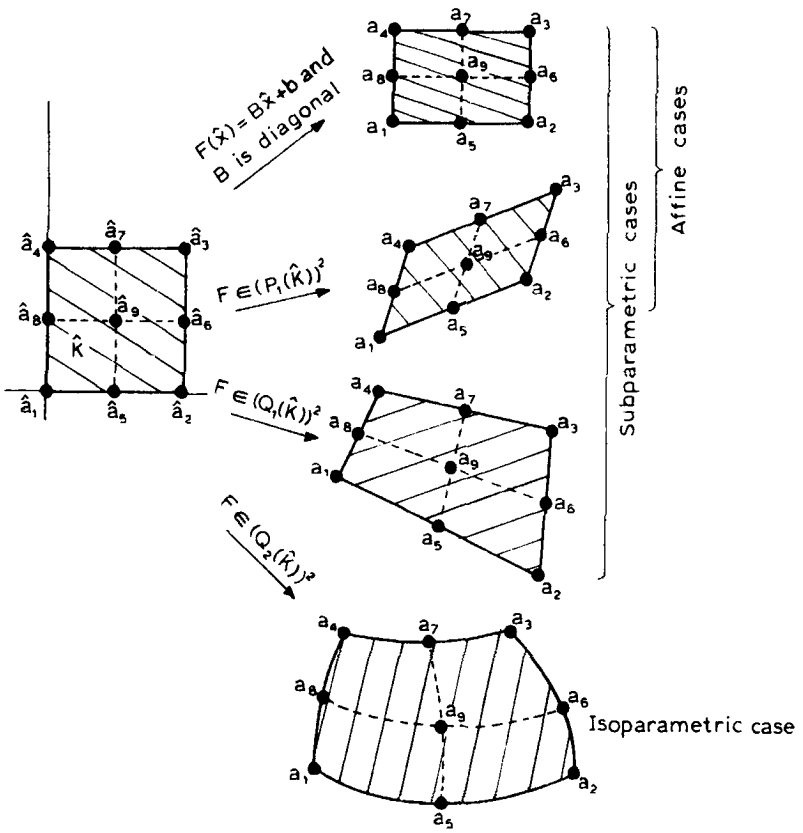
$$\forall \hat{v} \in \operatorname{dom} \hat{\Pi} = \mathscr{C}^0(\hat{K}), \quad (\Pi v)\hat{} = \hat{\Pi}\hat{v}, \tag{4.3.12}$$

provided $\hat{v} \in \operatorname{dom} \hat{\Pi} \Rightarrow v = \hat{v} \cdot F^{-1} \in \operatorname{dom} \hat{\Pi} = \mathscr{C}^0(K)$ (this condition excludes situations where the mapping F^{-1} would not be continuous).

Estimates of the interpolation errors $|v - \Pi_K v|_{m,q,K}$

The remainder of this section will be devoted to the derivation of an interpolation theory for isoparametric finite elements, i.e., we shall estimate the *interpolation errors* $|v - \Pi_K v|_{m,q,K}$ for finite elements (K, P, Σ) isoparametrically equivalent to a reference finite element $(\hat{K}, \hat{P}, \hat{\Sigma})$. This analysis is carried out in three stages, which parallel those used for affine-equivalent finite elements:

(i) Assuming the $\hat{P}$-interpolation operator $\hat{\Pi}$ leaves the space $P_k(\hat{K})$ invariant, an argument similar to that used in Theorem 3.1.4 yields



Quadrilateral of type (2) for $n = 2$
Fig. 4.3.4

inequalities of the form

$$|\hat{v} - \hat{\Pi}\hat{v}|_{m,q,\hat{K}} \leq C(\hat{K}, \hat{P}, \hat{\Sigma})|\hat{v}|_{k+1,p,\hat{K}}. \tag{4.3.13}$$

Thus this step is the same as before.

(ii) We then examine how the semi-norms occuring in (4.3.13) are transformed from $\hat{K}$ to K and vice versa. Recall that for affine families, we found inequalities of the form (cf. Theorem 3.1.2):

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

with $F: \hat{x} \in \hat{K} \to F(\hat{x}) = B\hat{x} + b$. In the present case, we shall find for example (Theorem 4.3.2) that the semi-norms $|v|_{m,p,K}$ are bounded above not only in terms of the semi-norm $|\hat{v}|_{m,p,\hat{K}}$, but instead in terms of all the semi-norms $|\hat{v}|_{l,p,\hat{K}}$, $1 \leq l \leq m$.

(iii) Just as the quantities $\|B\|$, $|\det(B)|$, etc..., which appeared in the affine case were eventually expressed in terms of the geometrical parameters meas(K), h and ρ_K (cf. Theorem 3.1.3), we shall subsequently turn (Theorem 4.3.3) to the problem of estimating analogous quantities (found in step (ii)) in terms of simple geometrical parameters attached to the finite element K.

Thus there are essentially two new steps ((ii) and (iii)) to develop, and rather than giving the general theory (for which the reader is referred to CIARLET & RAVIART (1972b)), we shall concentrate on one example, the *isoparametric n-simplex of type* (2).

We shall use the following notations:

$$\begin{cases} J_F(\hat{x}) = \text{Jacobian of } F \text{ at } \hat{x} = \det(\partial_j F_i(\hat{x})), \\ J_{F^{-1}}(x) = \text{Jacobian of } F^{-1} \text{ at } x = (J_F(\hat{x}))^{-1}, \end{cases} \tag{4.3.14}$$

$$\begin{cases} |F|_{l,\infty,\hat{K}} = \sup_{\hat{x} \in \hat{K}} \|D^l F(\hat{x})\|_{\mathcal{L}_l(\mathbf{R}^n;\mathbf{R}^n)}, \\ |F^{-1}|_{l,\infty,K} = \sup_{x \in K} \|D^l F^{-1}(x)\|_{\mathcal{L}_l(\mathbf{R}^n;\mathbf{R}^n)}, \end{cases} \tag{4.3.15}$$

whenever $F: \hat{K} \to K = F(\hat{K})$ is a sufficiently smooth mapping defined on any subset $\hat{K}$ of $\mathbf{R}^n$ with a sufficiently smooth inverse $F^{-1}: K \to \hat{K}$. Notice that when the mapping F is of the form $F: \hat{x} \to F(\hat{x}) = B\hat{x} + b$, then

$$J_F = \det(B), \quad J_{F^{-1}} = \det(B^{-1}), \quad |F|_{1,\infty,\hat{K}} = \|B\|, \quad \|F^{-1}\|_{1,\infty,K} = \|B^{-1}\|.$$

Since we are considering n-simplices of type (2), we need apply inequality (4.3.13) with the values $m = 0, 1, 2, 3$ and $k + 1 = 3$ only and thus we shall restrict ourselves to the semi-norms $|\cdot|_{l,p,\Omega}$ with $0 \leq l \leq 3$ in the next theorem. Notice that the following result is valid for general mappings F, i.e., it is irrelevant that the mapping F be in the space $(\hat{P})^n$ for some finite element $(\hat{K}, \hat{P}, \hat{\Sigma})$.

Theorem 4.3.2. *Let Ω and $\hat{\Omega}$ be two bounded open subsets of $\mathbf{R}^n$ such that $\Omega = F(\hat{\Omega})$, where F is a sufficiently smooth one-to-one mapping with a sufficiently smooth inverse $F^{-1}: \Omega \to \hat{\Omega}$.*

Then if a function $\hat{v}: \hat{\Omega} \to \mathbf{R}$ belongs to the space $W^{l,p}(\hat{\Omega})$ for some

integer $l \geq 0$ and some number $p \in [1, \infty]$, the function $v = \hat{v} \cdot F^{-1}: \Omega \to \mathbf{R}$ belongs to the space $W^{l,p}(\Omega)$ and, in addition, there exist constants C such that

$$\forall \hat{v} \in L^p(\hat{\Omega}), \quad |v|_{0,p,\Omega} \leq |J_F|_{0,\infty,\hat{\Omega}}^{1/p} |\hat{v}|_{0,p,\hat{\Omega}}, \tag{4.3.16}$$

$$\forall \hat{v} \in W^{1,p}(\hat{\Omega}), \quad |v|_{1,p,\Omega} \leq C |J_F|_{0,\infty,\hat{\Omega}}^{1/p} |F^{-1}|_{1,\infty,\Omega} |\hat{v}|_{1,p,\hat{\Omega}}, \tag{4.3.17}$$

$$\begin{aligned} \forall \hat{v} \in W^{2,p}(\hat{\Omega}), \quad & |v|_{2,p,\Omega} \leq \\ & \leq C |J_F|_{0,\infty,\hat{\Omega}}^{1/p} (|F^{-1}|_{1,\infty,\Omega}^2 |\hat{v}|_{2,p,\hat{\Omega}} + |F^{-1}|_{2,\infty,\Omega} |\hat{v}|_{1,p,\hat{\Omega}}), \end{aligned} \tag{4.3.18}$$

$$\begin{aligned} \forall \hat{v} \in W^{3,p}(\hat{\Omega}), \quad & |v|_{3,p,\Omega} \leq \\ & \leq C |J_F|_{0,\infty,\hat{\Omega}}^{1/p} (|F^{-1}|_{1,\infty,\Omega}^3 |\hat{v}|_{3,p,\hat{\Omega}} + |F^{-1}|_{1,\infty,\Omega} |F^{-1}|_{2,\infty,\Omega} |\hat{v}|_{2,p,\hat{\Omega}} \\ & \quad + |F^{-1}|_{3,\infty,\Omega} |\hat{v}|_{1,p,\hat{\Omega}}). \end{aligned} \tag{4.3.19}$$

Proof. As in Theorem 3.1.2, it suffices for $p < \infty$ to prove inequalities (4.3.16) through (4.3.19) for smooth functions (the case $p = \infty$ is left to the reader).

Using the formula for change of variables in multiple integrals, we obtain

$$|v|_{0,p,\Omega}^p = \int_\Omega |v(x)|^p \, dx = \int_\Omega |\hat{v}(F^{-1}(x))|^p \, dx = \int_{\hat{\Omega}} |J_F(\hat{x})| \, |\hat{v}(\hat{x})|^p \, d\hat{x},$$

for which we deduce inequality (4.3.16).

Since $v = \hat{v} \cdot F^{-1}$, we infer that

$$\forall x = F(\hat{x}), \quad Dv(x) = D\hat{v}(\hat{x}) \cdot DF^{-1}(x),$$

and thus,

$$\forall x \in \Omega, \quad \|Dv(x)\| \leq |F^{-1}|_{1,\infty,\Omega} \|D\hat{v}(F^{-1}(x))\|.$$

Consequently,

$$\begin{aligned} \int_\Omega \|Dv(x)\|^p \, dx & \leq |F^{-1}|_{1,\infty,\Omega}^p \int_\Omega \|D\hat{v}(F^{-1}(x))\|^p \, dx \\ & = |F^{-1}|_{1,\infty,\Omega}^p \int_{\hat{\Omega}} |J_F(\hat{x})| \, \|D\hat{v}(\hat{x})\|^p \, d\hat{x} \\ & \leq |F^{-1}|_{1,\infty,\Omega}^p |J_F|_{0,\infty,\hat{\Omega}} \int_{\hat{\Omega}} \|D\hat{v}(\hat{x})\|^p \, d\hat{x}, \end{aligned}$$

from which we deduce inequality (4.3.17), using the equivalence between the semi-norms (cf. (3.1.22) and (3.1.25))

$$v \to |v|_{m,p,\Omega} \quad \text{and} \quad v \to \left(\int_\Omega \|D^m v(x)\|^p \, dx\right)^{1/p}.$$

Likewise, we have for all $x \in \Omega$, $\xi_1 \in \mathbf{R}^n$, $\xi_2 \in \mathbf{R}^n$,

$$D^2 v(x)(\xi_1, \xi_2) = D\hat{v}(\hat{x})(D^2 F^{-1}(x)(\xi_1, \xi_2)) + \\ + D^2\hat{v}(\hat{x})(DF^{-1}(x)\xi_1, DF^{-1}(x)\xi_2),$$

so that we obtain, for all $x = F(\hat{x}) \in \Omega$,

$$\|D^2 v(x)\| = \sup_{\substack{|\xi_i|\leq 1 \\ i=1,2}} |D^2 v(x)(\xi_1, \xi_2)| \leq \\ \leq |F^{-1}|_{2,\infty,\Omega} \|D\hat{v}(\hat{x})\| + |F^{-1}|^2_{1,\infty,\Omega} \|D^2\hat{v}(\hat{x})\|.$$

Therefore,

$$\left(\int_\Omega \|D^2 v(x)\|^p \, dx\right)^{1/p} \leq |F^{-1}|_{2,\infty,\Omega} \left(\int_\Omega \|D\hat{v}(F^{-1}(x))\|^p \, dx\right)^{1/p} + \\ + |F^{-1}|^2_{1,\infty,\Omega} \left(\int_\Omega \|D^2\hat{v}(F^{-1}(x))\|^p \, dx\right)^{1/p}.$$

Arguing as before, we find that, for any integer l,

$$\left(\int_\Omega \|D^l \hat{v}(F^{-1}(x))\|^p \, dx\right)^{1/p} = \left(\int_{\hat{\Omega}} |J_F(\hat{x})| \, \|D^l\hat{v}(\hat{x})\|^p \, d\hat{x}\right)^{1/p} \\ \leq |J_F|^{1/p}_{0,\infty,\Omega} \left(\int_{\hat{\Omega}} \|D^l\hat{v}(\hat{x})\|^p \, d\hat{x}\right)^{1/p},$$

and thus inequality (4.3.18) is proved.

Inequality (4.3.19) is proved analogously by using the following inequality:

$$\|D^3 v(x)\| \leq |F^{-1}|_{3,\infty,\Omega} \|D\hat{v}(\hat{x})\| + 3|F^{-1}|_{1,\infty,\Omega} |F^{-1}|_{2,\infty,\Omega} \|D^2\hat{v}(\hat{x})\| \\ + |F^{-1}|^3_{1,\infty,\Omega} \|D^3\hat{v}(\hat{x})\|$$

which the reader may easily establish for all $x = F(\hat{x}) \in \Omega$. □

To apply the previous theorem, we must next obtain estimates of the

following quantities:

$$|J_F|_{0,\infty,\hat{K}};\ |J_{F^{-1}}|_{0,\infty,K};\ |F|_{l,\infty,\hat{K}},\ l = 1, 2, 3;$$

$$|F^{-1}|_{l,\infty,K},\ l = 1, 2,$$

for an isoparametric n-simplex of type (2). To do this, the key idea is the following: Since the affine case is a special case of the isoparametric case, we may expect the same type of error bounds, provided the mapping F is not "too far" from the unique affine mapping $\tilde{F}$ which satisfies

$$\tilde{F}(\hat{a}_i) = a_i, \quad 1 \leq i \leq n+1. \tag{4.3.20}$$

Therefore we are naturally led to introduce the n-simplex

$$\tilde{K} = \tilde{F}(\hat{K}) \tag{4.3.21}$$

and the points

$$\tilde{a}_{ij} = \tilde{F}(\hat{a}_{ij}), \quad 1 \leq i < j \leq n+1. \tag{4.3.22}$$

Looking at Fig. 4.3.5 (where we have represented the case $n = 2$), we expect the vectors $(a_{ij} - \tilde{a}_{ij})$ to serve as a good measure of the discrepancy between the mappings F and $\tilde{F}$: Indeed, if we let $\hat{p}_{ij}$ denote the basis functions of the n-simplex of type (2) attached to the node $\hat{a}_{ij}$, we have

$$F = \tilde{F} + \sum_{i<j} \hat{p}_{ij}(a_{ij} - \tilde{a}_{ij}). \tag{4.3.23}$$

To see this, it suffices to verify that the mapping

$$G = \tilde{F} + \sum_{i<j} \hat{p}_{ij}(a_{ij} - \tilde{a}_{ij})$$

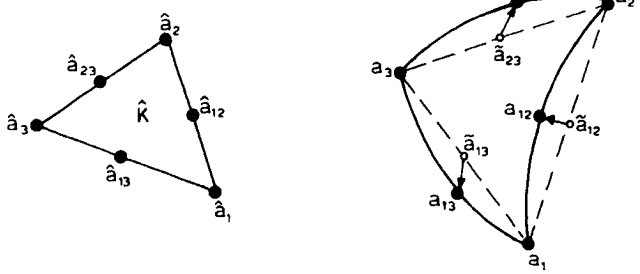


Fig. 4.3.5

satisfies the relations

$$\begin{cases} G \in (P_2(\hat{K}))^n, \\ G(\hat{a}_i) = a_i, \quad 1 \leq i \leq n+1, \\ G(\hat{a}_{ij}) = a_{ij}, \quad 1 \leq i < j \leq n+1, \end{cases}$$

which precisely characterize the unique isoparametric mapping F.

Let there be given an isoparametric family of n-simplices K of type (2), each of which is determined by the data of vertices $a_{i,K}$, $1 \leq i \leq n+1$, and points $a_{ij,K}$, $1 \leq i < j \leq n+1$. Then in view of (4.3.20) and (4.3.21), we let $\tilde{F}_K$ denote for each K the unique affine mapping which satisfies $\tilde{F}_K(\hat{a}_i) = a_{i,K}$, $1 \leq i \leq n+1$, and we define the n-simplex $\tilde{K} = \tilde{F}_K(\hat{K})$. Finally, we let, for each $\tilde{K}$,

$$h_K = \text{diam}(\tilde{K}), \tag{4.3.24}$$

$$\rho_K = \text{diameter of the sphere inscribed in } \tilde{K}. \tag{4.3.25}$$

We shall say that *an isoparametric family of n-simplices K of type* (2) *is regular* if the following three conditions are satisfied:

(i) There exists a constant σ such that

$$\forall K, \quad \frac{h_K}{\rho_K} \leq \sigma. \tag{4.3.26}$$

(ii) The quantities h_K approach zero.

(iii) We have

$$\|a_{ij,K} - \tilde{a}_{ij,K}\| = O(h_K^2), \quad 1 \leq i < j \leq n+1, \tag{4.3.27}$$

where, for each K, we let $\tilde{a}_{ij,K} = \tilde{F}_K(\hat{a}_{ij})$.

Remark 4.3.2. In the special case of an affine family, condition (4.3.27) is automatically satisfied ($a_{ij,K} = \tilde{a}_{ij,K}$), so that the above definition contains the definition of a regular affine family which was given in Section 3.1. □

Although it is clear that condition (4.3.27) does insure that the mappings F_K and $\tilde{F}_K$ do not differ too much, the reason the vectors $(a_{ij,K} - \tilde{a}_{ij,K})$ have to be precisely of order $O(h_K^2)$ may seem arbitrary at

this stage. As we shall show later (cf. Theorem 4.3.4), the basic justification of this assumption is that it yields the same interpolation error estimates as in the affine case.

To begin with, we show that this assumption allows us to obtain upper bounds of the various quantities found in the inequalities of Theorem 4.3.2.

We are also able to show in this particular case that the mappings F_K are invertible (the invertibility of the mapping F_K is part of the definition of an isoparametric family).

Theorem 4.3.3. *Let there be given a regular isoparametric family of n-simplices of type* (2). *Then, provided h_K is small enough, the mappings $F_K: \hat{K} \to K = F_K(\hat{K})$ are one-to-one, their Jacobians J_{F_K} do not vanish, and there exist constants C such that*

$$|F_K|_{1,\infty,\hat{K}} \leq Ch_K, \quad |F_K|_{2,\infty,\hat{K}} \leq Ch_K^2, \quad |F_K|_{3,\infty,\hat{K}} = 0, \tag{4.3.28}$$

$$|F_K^{-1}|_{1,\infty,K} \leq \frac{C}{h_K}, \quad |F_K^{-1}|_{2,\infty,K} \leq \frac{C}{h_K}, \tag{4.3.29}$$

$$|J_{F_K}|_{0,\infty,\hat{K}} \leq C \operatorname{meas}(\hat{K}), \quad |J_{F_K^{-1}}|_{0,\infty,K} \leq \frac{C}{\operatorname{meas}(\hat{K})}. \tag{4.3.30}$$

Proof. For notational convenience, we shall drop the index K throughout the proof. Using the decomposition (4.3.23) of the mapping F, we deduce that, for all $\hat{x} \in \hat{K}$,

$$DF(\hat{x}) = D\tilde{F}(\hat{x}) + E(\hat{x}) = B + E(\hat{x}), \tag{4.3.31}$$

where

$$E(\hat{x}) = \sum_{i<j} (a_{ij} - \tilde{a}_{ij}) D\hat{p}_{ij}(\hat{x}).$$

Therefore, by virtue of assumption (4.3.27), and since the basis functions $\hat{p}_{ij}$ are independent of K, we find that

$$\sup_{\hat{x} \in \hat{K}} \|E(\hat{x})\| \leq Ch^2 \tag{4.3.32}$$

(as usual the same letter C stands for various constants). Thus we have

$$|F|_{1,\infty,\hat{K}} = \sup_{\hat{x} \in \hat{K}} \|DF(\hat{x})\| \leq \|B\| + \sup_{\hat{x} \in \hat{K}} \|E(\hat{x})\| \leq Ch,$$

since $\|B\| \leq Ch$ (cf. Theorem 3.1.3). Likewise, we have

$$D^2F(\hat{x}) = DE(\hat{x}),$$

since $D^2\tilde{F} = 0$, and, arguing as before, we find that

$$\sup_{\hat{x}\in\hat{K}} \|DE(\hat{x})\| \leq Ch^2,$$

so that

$$|F|_{2,\infty,\hat{K}} = \sup_{\hat{x}\in\hat{K}} \|D^2F(\hat{x})\| \leq Ch^2.$$

Hence all relations (4.3.28) are proved, the last one being obvious since $F \in (P_2(\hat{K}))^n$.

Considered as a function of its column vectors $\partial_j F(\hat{x})$, $1 \leq j \leq n$, the determinant $J_F(\hat{x}) = \det(DF(\hat{x}))$ is a continuous multilinear mapping and therefore there exists a constant $C = C(n)$ such that

$$\forall \hat{x} \in \hat{K}, \quad J_F(\hat{x}) \leq C \sum_{j=1}^{n} \|\partial_j F(\hat{x})\|.$$

Since the inequality $|F|_{1,\infty,\hat{K}} \leq Ch$ proved above implies the similar inequalities $\sup_{\hat{x}\in\hat{K}} \|\partial_j F(\hat{x})\| \leq Ch$, $1 \leq j \leq n$, we deduce that

$$|J_F|_{0,\infty,\hat{K}} = \sup_{\hat{x}\in\hat{K}} |J_F(\hat{x})| \leq Ch^n \leq C \text{ meas } \hat{K},$$

and the first inequality of (4.3.30) is proved.

Because of assumption (4.3.26), the matrices B are all invertible so that we can write (4.3.31) in the form

$$DF(\hat{x}) = B(I + B^{-1}E(\hat{x})).$$

Using inequality $\|B^{-1}\| \leq C/h$ (cf. Theorem 3.1.3 and assumptions (4.3.26)) and inequality (4.3.32), we deduce that $\sup_{\hat{x}\in\hat{K}} \|B^{-1}E(\hat{x})\| \leq Ch$. Let then γ be a fixed number in the interval $]0, 1[$. There exists $h_0 > 0$ such that

$$\forall h \leq h_0, \quad \sup_{\hat{x}\in\hat{K}} \|B^{-1}E(\hat{x})\| \leq \gamma,$$

so that, for $h \leq h_0$, the operator $(I + B^{-1}E(\hat{x}))$ is invertible for each $\hat{x} \in \hat{K}$, and

$$\sup_{\hat{x}\in\hat{K}} \|(I + B^{-1}E(\hat{x}))^{-1}\| \leq \frac{1}{1-\gamma}. \tag{4.3.33}$$

This shows that the derivative $DF(\hat{x})$ is invertible for all $\hat{x} \in \hat{K}$, with

$$(DF(\hat{x}))^{-1} = (I + B^{-1}E(\hat{x}))^{-1}B^{-1}. \tag{4.3.34}$$

We next prove that the mapping $F: \hat{K} \to K$ is invertible (by the implicit function theorem, we can only deduce that the mapping F is invertible *locally*, i.e., in a sufficiently small neighborhood of each point of $\hat{K}$; this is why the global invertibility requires an additional analysis; for a more general approach, see Exercise 4.3.6). Let $\hat{x}, \hat{y} \in \hat{K}$ be such that $F(\hat{x}) = F(\hat{y})$. Since the set $\hat{K}$ is convex, we may apply Taylor formula:

$$F(\hat{y}) = F(\hat{x}) + DF(\hat{x})(\hat{y} - \hat{x}) + \frac{A}{2}(\hat{y} - \hat{x})^2,$$

where $A \in \mathscr{L}_2(\mathbf{R}^n; \mathbf{R}^n)$ is the constant second derivative of the mapping F. We deduce that

$$DF(\hat{x})(\hat{y} - \hat{x}) = -\frac{A}{2}(\hat{y} - \hat{x})^2 = -\frac{A}{2}(\hat{x} - \hat{y})^2 = DF(\hat{y})(\hat{x} - \hat{y}),$$

and consequently,

$$(DF(\hat{x}) + DF(\hat{y}))(\hat{y} - \hat{x}) = 0.$$

Each component F_i of the mapping F is in the space $P_2(\hat{K})$, so that we have $((\hat{x} + \hat{y})/2 \in \hat{K})$:

$$\partial_j F_i \in P_1(\hat{K}) \Rightarrow \partial_j F_i(\hat{x}) + \partial_j F_i(\hat{y}) = 2\partial_j F_i\left(\frac{\hat{x} + \hat{y}}{2}\right), \quad 1 \leq i, j \leq n,$$

i.e.,

$$0 = (DF(\hat{x}) + DF(\hat{y}))(\hat{y} - \hat{x}) = 2DF\left(\frac{\hat{x} + \hat{y}}{2}\right)(\hat{y} - \hat{x}).$$

Since the derivative $DF((\hat{x} + \hat{y})/2)$ is an invertible mapping in $\mathscr{L}(\mathbf{R}^n)$, we conclude that $\hat{x} = \hat{y}$.

We can write

$$\forall \hat{x} \in \hat{K}, \quad (DF(\hat{x}))^{-1} = DF^{-1}(x),$$

and thus, by relations (4.3.33) and (4.3.34),

$$|F^{-1}|_{1,\infty,K} = \sup_{x \in K} \|DF^{-1}(x)\| \leq \frac{C}{h},$$

which proves the first inequality of (4.3.29).

Given functions $F: \mathbf{R}^n \to \mathbf{R}^n$ and $G: \mathbf{R}^n \to \mathbf{R}^n$, the function $H = G \cdot F: \mathbf{R}^n \to \mathbf{R}^n$ is such that, for all vectors $\xi_1, \xi_2 \in \mathbf{R}^n$,

$$D^2H(\hat{x})(\xi_1, \xi_2) = DG(x)(D^2F(\hat{x})(\xi_1, \xi_2)) +$$

$$+ D^2G(x)(DF(\hat{x})\xi_1, DF(\hat{x})\xi_2).$$

If we apply this formula with $G = F^{-1}$, so that $H = I$, we obtain, for all $x = F(\hat{x}) \in K$,

$$D^2F^{-1}(x)(DF(\hat{x})\xi_1, DF(\hat{x})\xi_2) = -DF^{-1}(x)(D^2F(\hat{x})(\xi_1, \xi_2)).$$

Since for each $x = F(\hat{x}) \in K$, the mapping $DF(\hat{x}): \mathbf{R}^n \to \mathbf{R}^n$ is invertible, we deduce that for all vectors $\eta_1, \eta_2 \in \mathbf{R}^n$,

$$D^2F^{-1}(x)(\eta_1, \eta_2) = -DF^{-1}(x)(D^2F(\hat{x})(DF^{-1}(x)\eta_1, DF^{-1}(x)\eta_2)),$$

and thus,

$$\|D^2F^{-1}(x)\| = \sup_{\substack{|\eta_l| \leq 1 \\ l=1,2}} \|D^2F^{-1}(x)(\eta_1, \eta_2)\| \leq \|D^2F(\hat{x})\| \, \|DF^{-1}(x)\|^3,$$

so that, using the second inequality of (4.3.28) and the first inequality (4.3.29),

$$|F^{-1}|_{2,\infty,K} = \sup_{x \in K} \|D^2F^{-1}(x)\| \leq |F|_{2,\infty,\hat{K}} |F^{-1}|^3_{1,\infty,K} \leq \frac{C}{h},$$

and the second inequality of (4.3.29) is proved.

Using (4.3.34), we can write

$$\forall \hat{x} \in \hat{K}, \quad B = DF(\hat{x})(I + B^{-1}E(\hat{x}))^{-1},$$

and thus, by (4.3.33),

$$\forall \hat{x} \in \hat{K}, \ |\det(B)| = |J_F(\hat{x})| |\det(I + B^{-1}E(\hat{x}))^{-1}| \leq \frac{|J_F(\hat{x})|}{(1-\gamma)^n}.$$

Therefore, we deduce that

$$\frac{1}{|J_{F^{-1}}|_{0,\infty,K}} = \frac{1}{\sup_{x \in K} |J_{F^{-1}}(x)|} = \inf_{\hat{x} \in \hat{K}} J_F(\hat{x}) \geq$$

$$\geq (1-\gamma)^n |\det(B)| \geq C \operatorname{meas}(\hat{K}),$$

and the second inequality of (4.3.30) is proved, which completes the proof. □

Combining Theorems 4.3.2 and 4.3.3, we are in a position to prove our main result (compare with Theorems 3.1.6).

Theorem 4.3.4. *Let there be given a regular isoparametric family of n-simplices K of type* (2) *and let there be given an integer $m \geq 0$ and two numbers $p, q \in [1, \infty]$ such that the following inclusions hold*:

$$W^{3,p}(\hat{K}) \hookrightarrow \mathscr{C}^0(\hat{K}), \tag{4.3.35}$$

$$W^{3,p}(\hat{K}) \hookrightarrow W^{m,q}(\hat{K}), \tag{4.3.36}$$

where $\hat{K}$ is the reference n-simplex of type (2) *of the family.*

Then provided the diameters h_K are small enough, there exists a constant C such that, for all finite elements in the family, and all functions $v \in W^{3,p}(K)$,

$$\|v - \Pi_K v\|_{m,q,K} \leq C(\text{meas}(\hat{K}))^{1/q-1/p} h_K^{3-m}(|v|_{2,p,K} + |v|_{3,p,K}). \tag{4.3.37}$$

Proof. The inclusion (4.3.35) guarantees the existence of the interpolation operators $\hat{\Pi}$ and Π_K, which satisfy the relation (4.3.12). Combining the inequalities of Theorems 4.3.2 and 4.3.3, we obtain, if $m = 0$, 1 or 2, respectively,

$$\begin{aligned}
|v - \Pi_K v|_{0,q,K} &\leq |J_{F_K}|_{0,\infty,\hat{K}}^{1/q} |\hat{v} - \hat{\Pi}\hat{v}|_{0,q,\hat{K}} \\
&\leq C(\text{meas}(\hat{K}))^{1/q} |\hat{v} - \hat{\Pi}\hat{v}|_{0,q,\hat{K}}, \\
|v - \Pi_K v|_{1,q,K} &\leq C|J_{F_K}|_{0,\infty,\hat{K}}^{1/q} |F_K^{-1}|_{1,\infty,K} |\hat{v} - \hat{\Pi}\hat{v}|_{1,q,\hat{K}} \\
&\leq C(\text{meas}(\hat{K}))^{1/q} \left(\frac{1}{h_K} |\hat{v} - \hat{\Pi}\hat{v}|_{1,q,\hat{K}} \right), \\
|v - \Pi_K v|_{2,q,K} &\leq C|J_{F_K}|_{0,\infty,\hat{K}}^{1/q} (|F_K^{-1}|_{1,\infty,K}^2 |\hat{v} - \hat{\Pi}\hat{v}|_{2,q,\hat{K}} \\
&\quad + |F_K^{-1}|_{2,\infty,K} |\hat{v} - \hat{\Pi}\hat{v}|_{1,q,\hat{K}}) \\
&\leq C(\text{meas}(\hat{K}))^{1/q} \left(\frac{1}{h_K^2} \left(|\hat{v} - \hat{\Pi}\hat{v}|_{2,q,\hat{K}} \right. \right. \\
&\quad \left. \left. + \frac{1}{h_K} |\hat{v} - \hat{\Pi}\hat{v}|_{1,q,\hat{K}} \right) \right).
\end{aligned}$$

By virtue of the inclusions (4.3.35) and (4.3.36), we may argue as in Theorem 3.1.4 and infer that there exists a constant C depending only on

the set $\hat{K}$ such that for all $\hat{v} \in W^{3,p}(\hat{K})$,

$$|\hat{v} - \hat{\Pi}\hat{v}|_{l,q,\hat{K}} \leq C|\hat{v}|_{3,p,\hat{K}}, \quad l \leq m.$$

Upon combining the above inequalities, we obtain

$$|v - \Pi_K v|_{m,q,K} \leq C(\text{meas}(\tilde{K}))^{1/q} \frac{1}{h_K^m} |\hat{v}|_{3,p,\hat{K}},$$

and another application of Theorems 4.3.2 and 4.3.3 yields:

$$\begin{aligned} |\hat{v}|_{3,p,\hat{K}} &\leq C|J_{F_K^{-1}}|_{0,\infty,K}^{1/p}(|F_K|_{1,\infty,\hat{K}}^3 |v|_{3,p,K} + \\ &\quad + |F_K|_{1,\infty,\hat{K}} |F_K|_{2,\infty,\hat{K}} |v|_{2,p,K} + |F_K|_{3,\infty,\hat{K}} |v|_{1,p,K}) \\ &\leq C(\text{meas}(\tilde{K}))^{-1/p} h_K^3 (|v|_{2,p,K} + |v|_{3,p,K}). \end{aligned}$$

Thus inequality (4.3.37) is proved for the values $m = 0, 1$ and 2. The case $m = 3$ is left as a problem (Exercise 4.3.7). □

It is interesting to compare the estimates of the above theorem with the analogous estimates obtained for a regular *affine* family of n-simplices of type (2) (cf. Theorem 3.1.6):

$$\|v - \Pi_K v\|_{m,q,K} \leq C(\text{meas}(\tilde{K}))^{1/q - 1/p} h_K^{3-m} |v|_{3,p,K}.$$

We conclude that the two estimates coincide except for the additional semi-norm $|v|_{2,p,K}$ (which appears when one differentiates a function composed with other than an affine function; cf. the end of the proof of Theorem 4.3.2). Also, the present estimates have been established under the additional assumption that the diameters h_K are sufficiently small, basically to insure the invertibility of the derivatives $DF_K(\hat{x})$, $\hat{x} \in \hat{K}$ (cf. the proof of Theorem 4.3.3).

Remark 4.3.3. (i) Just as in the case of affine families (cf. Remark 3.1.3), the parameter $\text{meas}(\tilde{K})$ can be replaced by h_K^n in inequality (4.3.37), since it satisfies (cf. (4.3.26)) the inequalities

$$\sigma_n \sigma^{-n} h_K^n \leq \text{meas}(\tilde{K}) \leq \sigma_n h_K^n,$$

where σ_n denotes the dx-measure of the unit sphere in $\mathbf{R}^n$.

(ii) If necessary, the expression $(|v|_{2,p,K} + |v|_{3,p,K})$ appearing in the right-hand side of inequality (4.3.37) can be of course replaced by the expression $(|v|_{2,p,K}^p + |v|_{3,p,K}^p)^{1/p}$. □

Similar analyses can be carried out for other types of simplicial finite

elements, such as the isoparametric n-simplex of type (3) (cf. Exercise 4.3.8). For the general theory, which also applies to isoparametric Hermite finite elements, see CIARLET & RAVIART (1972b).

If we turn to quadrilateral finite elements, the situation is different. Of course, we could again consider this case as a perturbation of the affine case. But, as exemplified by Fig. 4.3.4, this would reduce the possible shapes to "nearly parallelograms". Hopefully, a new approach can be developed whereby the admissible shapes correspond to mappings F_K which are perturbations of mappings $\tilde{F}_K$ in the space $(Q_1(\hat{K}))^n$, instead of the space $(P_1(\hat{K}))^n$. Accordingly, a new theory has to be developed, in particular for the quadrilateral of type (1), as indicated in Exercise 4.3.9.

Exercises

4.3.1. Let $(\hat{K}, \hat{P}, \hat{\Sigma})$ be a Hermite finite element where the order of directional derivatives occuring in the definition is one, i.e., the set $\hat{\Sigma}$ is of the form

$$\hat{\Sigma} = \{\hat{\varphi}_i^0, 1 \leq i \leq N_0;\ \hat{\varphi}_{ik}^1, 1 \leq k \leq d_i,\ 1 \leq i \leq N_1\},$$

with degrees of freedom of the following form:

$$\hat{\varphi}_i^0 \colon \hat{p} \to \hat{p}(\hat{a}_i^0),\ \hat{\varphi}_{ik}^1 \colon \hat{p} \to D\hat{p}(\hat{a}_i^1)\hat{\xi}_{ik}^1.$$

(i) Let $F \colon \hat{K} \to \mathbf{R}^n$ be a differentiable one-to-one mapping. Let $a_i^0 = F(\hat{a}_i^0),\ 1 \leq i \leq N_0;\ a_i^1 = F(\hat{a}_i^1)$ and $\xi_{ik}^1 = DF(\hat{a}_i^1)\hat{\xi}_{ik}^1,\ 1 \leq k \leq d_i, 1 \leq i \leq N_1$. Then show that the triple (K, P, Σ) is a Hermite finite element, where

$$\begin{cases} K = F(\hat{K}),\ P = \{p \colon K \to \mathbf{R};\ p = \hat{p} \cdot F^{-1},\ \hat{p} \in \hat{P}\}, \\ \Sigma = \{\varphi_i^0, 1 \leq i \leq N_0;\ \varphi_{ik}^1, 1 \leq k \leq d_i,\ 1 \leq i \leq N_1\}, \\ \varphi_i^0 \colon p \to p(a_i^0),\ \varphi_{ik}^1 \colon p \to Dp(a_i^1)\xi_{ik}^1, \end{cases}$$

and show that $(\Pi v)\hat{} = \hat{\Pi}\hat{v}$.

(ii) If the mapping F belongs to the space $(\hat{P})^n$, one obtains in this fashion an *isoparametric Hermite finite element*. In this case, write the mapping F in terms of the basis functions of the finite element $(\hat{K}, \hat{P}, \hat{\Sigma})$. Deduce that, in practice, the isoparametric finite element is completely determined by the data of the points a_i^0 and of the vectors ξ_{ik}^1. Why are the points a_i^1 not arbitrary? Show that this is again a generalization of affine equivalence.

(iii) Using (ii), construct the *isoparametric Hermite triangle of type* (3) which is thus defined by the data of three "vertices" a_i, $1 \leq i \leq 3$, two directions at each point a_i, $1 \leq i \leq 3$, and a point a_{123}.

(iv) Examine whether the construction of (i) and (ii) could be extended to Hermite finite elements in the definition of which higher order directional derivatives are used.

For a reference about questions (i), (ii) and (iii), see CIARLET & RAVIART (1972b).

4.3.2. Let (K, P, Σ) be an isoparametric finite element derived from a finite element $(\hat{K}, \hat{P}, \hat{\Sigma})$ by the construction of Theorem 4.3.1. Show that if the space $\hat{P}$ contains constant functions, then the space P always contains polynomials of degree 1 in the variables $x_1, x_2, \ldots, x_n$. Isn't there a paradox?

4.3.3. Give a description of the space P_K corresponding to the isoparametric triangle of type (2). In particular show that one has $P_K \neq P_2(K)$ in general (although the inclusion $P_1(K) \subset P_K$ holds; cf. Exercise 4.3.2).

4.3.4. (i) Let a_i and a_j, $i \neq j$, be two "vertices" of an isoparametric triangle of type (2). Show that the curved "side" joining these two points is an arc of parabola uniquely determined by the following conditions: It passes through the points a_i, a_j, a_{ij} and its asymptotic direction is parallel to the vector $a_{ij} - (a_i + a_j)/2$.

(ii) Use the result of (i) to show that the mapping F corresponding to the following data is not invertible:

$$a_1 = (0, 0),\ a_2 = (2, 0),\ a_3 = (0, 2),\ a_{12} = (1, 0),\ a_{13} = (1, 1),$$
$$a_{23} = (0, 1).$$

4.3.5. Given a regular isoparametric family of n-simplices of type (2), do we have diam K = diam $\tilde{K}$ for h_K = diam $\tilde{K}$ sufficiently small?

4.3.6. In Theorem 4.3.3, it was shown that the isoparametric mappings F_K are one-to-one (for h_K small enough) by an argument special to isoparametric n-simplices of type (2). Give a more general proof, which would apply to other isoparametric finite elements.

4.3.7. Complete the proof of inequalities (4.3.37) by considering the case $m = 3$.

4.3.8. (i) Carry out an analysis similar to the one given in the text for the isoparametric n-simplex of type (3) (cf. Fig. 4.3.2 for $n = 2$). Introducing the unique affine mapping $\tilde{F}_K$ which satisfies $\tilde{F}_K(\hat{a}_i) = a_i$, $1 \leq i \leq n + 1$, show that one obtains interpolation error estimates of the

form

$$\|v - \Pi_K v\|_{m,q,K} \leq C(\text{meas}(\hat{K}))^{1/q-1/p} h_K^{4-m} |v|_{4,p,K},$$

i.e., as in the affine case, provided we consider a *regular* isoparametric family for which condition (iii) of (4.3.27) is replaced by the following:

$$(*)\quad \|a_{iij,K} - \tilde{a}_{iij,K}\| = O(h_K^3), \quad 1 \leq i, j \leq n+1, \ i \neq j,$$

$$(**)\quad \|a_{ijk,K} - \tilde{a}_{ijk,K}\| = O(h_K^3), \quad 1 \leq i < j < k \leq n+1,$$

where $\tilde{a}_{iij,K} = \tilde{F}_K(\hat{a}_{iij})$ and $\tilde{a}_{ijk,K} = \tilde{F}(\hat{a}_{ijk})$.

(ii) It is clear however that if the points $a_{iij,K}$ are taken from an actual boundary (as they would be in practice), the above condition (*) cannot be satisfied since one has instead in this situation $\|a_{iij,K} - \tilde{a}_{iij,K}\| = O(h_K^2)$. There is nevertheless one case where this difficulty can be circumvented: Assume that $n = 2$ and that (cf. Fig. 4.3.6 where the indices K have been dropped for convenience)

$$a_{331,K} = \tilde{a}_{331,K}, \quad a_{113,K} = \tilde{a}_{113,K}, \quad a_{332,K} = \tilde{a}_{332,K}, \quad a_{223,K} = \tilde{a}_{223,K}.$$

Then show that the estimates of (i) hold with assumptions (*) and (**)

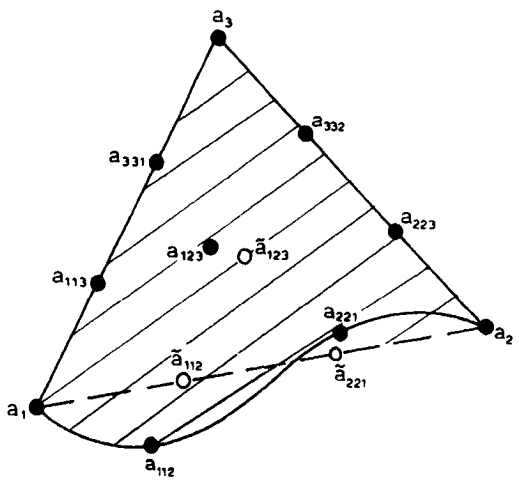


Fig. 4.3.6

replaced by the following:

$$\|a_{112,K} - \tilde{a}_{112,K}\| = O(h_K^2), \quad \|a_{221,K} - \tilde{a}_{221,K}\| = O(h_K^2),$$

$$\|(a_{112,K} - \tilde{a}_{112,K}) - (a_{221,K} - \tilde{a}_{221,K})\| = O(h_K^3),$$

$$\|4(a_{123,K} - \tilde{a}_{123,K}) - (a_{112,K} - \tilde{a}_{112,K}) - (a_{221,K} - \tilde{a}_{221,K})\| = O(h_K^3).$$

Show that the above assumptions are now realistic in the sense that $a_{112,K}$ and $a_{221,K}$ can be actually chosen along a smooth boundary so as to fulfill the above conditions.

For a reference for this problem, see CIARLET & RAVIART (1972b).

4.3.9. (i) Let

$$[v]_{m,p,\Omega} = \left(\int_\Omega \sum_{i=1}^{n} |D^m v(x)(e_i)^m|^p \, dx \right)^{1/p}.$$

Then (cf. Exercise 3.1.1) the semi-norm $[.]_{k,p,\Omega}$ is a norm over the quotient space $W^{k+1,p}(\Omega)/Q_k(\Omega)$, equivalent to the quotient norm. Let there be given two Sobolev spaces $W^{k+1,p}(\Omega)$ and $W^{m,q}(\Omega)$ with $W^{k+1,p}(\Omega) \subset W^{m,p}(\Omega)$ and let $\Pi \in \mathcal{L}(W^{k+1,p}(\Omega); W^{m,q}(\Omega))$ be a mapping which satisfies

$$\forall q \in Q_k(\Omega), \quad \Pi q = q.$$

Show that there exists a constant $C(\Omega, \Pi)$ such that

$$\forall v \in W^{k+1,p}(\Omega), \quad |v - \Pi v|_{m,q,\Omega} \leq C(\Omega, \Pi)[v]_{k+1,p,\Omega}.$$

(ii) For each integer $l \geq 1$, let

$$[\![F]\!]_{l,\infty,\hat{K}} = \max_{1 \leq i \leq n} \sup_{\hat{x} \in \hat{K}} \|D^l F(\hat{x})(e_i)^l\|.$$

With the same assumptions as in Theorem 4.3.2, show that

$$\forall \hat{v} \in W^{2,p}(\hat{\Omega}), \quad [\hat{v}]_{2,p,\hat{\Omega}} \leq C|J_{F^{-1}}|_{0,\infty,\Omega}^{1/p}([\![F]\!]_{1,\infty,\Omega}^2|v|_{2,p,\Omega} + [\![F]\!]_{2,\infty,\Omega}|v|_{1,p,\Omega}),$$

$$\forall \hat{v} \in W^{3,p}(\hat{\Omega}), \quad [\hat{v}]_{3,p,\hat{\Omega}} \leq C|J_{F^{-1}}|_{0,\infty,\Omega}^{1/p}([\![F]\!]_{1,\infty,\Omega}^3|v|_{3,p,\Omega} + [\![F]\!]_{1,\infty,\Omega}[\![F]\!]_{2,\infty,\Omega}|v|_{2,p,\Omega} + [\![F]\!]_{3,\infty,\Omega}|v|_{1,p,\Omega}).$$

(iii) Consider an isoparametric family of quadrilaterals K of type (1)

for $n = 2$ (cf. Fig. 4.3.3). We let for each K,

$$h_K = \text{diam}(K),$$
$$h'_K = \text{smallest length of the sides of } K,$$
$$\gamma_K = \max\{|\cos\{(a_{i+1} - a_i) \cdot (a_{i-1} - a_i)\}|, \quad 1 \leq i \leq 4(\text{mod } 4)\}.$$

Then such a family is said to be *regular* if all the sets K are convex, if there exist constants σ' and γ such that

$$(*)\ \forall K, \frac{h_K}{h'_K} \leq \sigma' \quad \text{and} \quad \gamma_K \leq \gamma < 1,$$

and if the quantity h_K approaches zero. Show that condition (*) implies the usual condition that the ratios h_K/ρ_K be bounded (the converse is clearly false).

Show that, given such a family, the mappings $F_K: \hat{K} = [0, 1]^2 \to K$ are one-to-one and that the following estimates hold:

$$[F_K]_{1,\infty,\hat{K}} \leq Ch_K, \quad [F_K]_{2,\infty,\hat{K}} = 0,$$
$$|F_K^{-1}|_{1,\infty,K} \leq \frac{C}{h_K},$$
$$|J_{F_K}|_{0,\infty,\hat{K}} \leq Ch_K^2, \quad |J_{F_K^{-1}}|_{0,\infty,K} \leq \frac{C}{h_K^2}.$$

Using the above results, derive the following interpolation error estimates (under the assumptions $W^{2,p}(\hat{K}) \hookrightarrow \mathscr{C}^0(\hat{K})$ and $W^{2,p}(\hat{K}) \hookrightarrow W^{m,q}(\hat{K})$): For all $v \in W^{2,p}(K)$,

$$\|v - \Pi_K v\|_{m,q,K} \leq C(h_K^2)^{1/q - 1/p} h_K^{2-m} \|v\|_{2,p,K}, \quad m = 0, 1.$$

(iv) Consider an isoparametric family of quadrilaterals K of type (2) for $n = 2$.

For each $K = F_K(\hat{K})$, where the mapping $F_K \in (Q_2(\hat{K}))^2$ is uniquely determined by the data of nine points $a_{i,K}$, $1 \leq i \leq 9$, (cf. Fig. 4.3.4), we let $\tilde{F}_K$ denote the mapping uniquely determined by the conditions

$$\tilde{F}_K \in (Q_1(\hat{K}))^2, \quad \tilde{F}_K(\hat{a}_i) = a_{i,K}, \quad 1 \leq i \leq 4.$$

Then we say that the family is *regular* if the family of quadrilaterals $\tilde{K} = \tilde{F}_K(\hat{K})$ is regular in the sense of (iii) and if one has (compare with (4.3.27)):

$$\|a_{i,K} - \tilde{a}_{i,K}\| = O(h_K^2), \quad 5 \leq i \leq 9,$$

where $\tilde{a}_{i,K} = \tilde{F}_K(\hat{a}_i)$, $5 \leq i \leq 9$.

Given such a family, show that the mappings $F_K: \hat{K} = [0, 1]^2 \to K$ are one-to-one for h_K small enough and that the following estimates hold:

$$[F_K]_{1,\infty,\hat{K}} \leq Ch_K, \quad [F_K]_{2,\infty,\hat{K}} \leq Ch_K^2, \quad [F_K]_{3,\infty,\hat{K}} = 0,$$

$$|F_K^{-1}|_{1,\infty,K} \leq \frac{C}{h_K}, \quad |F_K^{-1}|_{2,\infty,K} \leq \frac{C}{h_K},$$

$$|J_{F_K}|_{0,\infty,\hat{K}} \leq Ch_K^2, \quad |J_{F_K^{-1}}|_{0,\infty,K} \leq \frac{C}{h_K^2}.$$

Using the above results, derive the following interpolation error estimates (under the assumptions $W^{3,p}(\hat{K}) \hookrightarrow \mathscr{C}^0(\hat{K})$ and $W^{3,p}(\hat{K}) \hookrightarrow W^{m,q}(\hat{K})$): For all $v \in W^{3,p}(K)$,

$$\|v - \Pi_K v\|_{m,q,K} \leq C(h_K^2)^{1/q-1/p} h_K^{3-m} \|v\|_{3,p,K}, \quad m = 0, 1, 2.$$

A general theory for isoparametric quadrilateral finite elements is given in CIARLET & RAVIART (1972b). Significant improvements have recently been obtained by JAMET (1976b).

4.4 Application to second-order problems over curved domains

Approximation of a curved boundary with isoparametric finite elements

As in Section 4.1, we consider the homogeneous second-order Dirichlet problem which corresponds to the following data:

$$\begin{cases} V = H_0^1(\Omega), \\ a(u, v) = \displaystyle\int_\Omega \sum_{i,j=1}^n a_{ij} \partial_i u \partial_j v \, dx, \\ f(v) = \displaystyle\int_\Omega fv \, dx, \end{cases} \tag{4.4.1}$$

where Ω is a bounded open subset of $\mathbf{R}^n$ with a *curved* boundary Γ (the main novelty) and the functions $a_{ij} \in L^\infty(\Omega)$ and $f \in L^2(\Omega)$ are *everywhere* defined over the set $\bar{\Omega}$. We shall assume that the ellipticity condition holds, i.e.,

$$\exists \beta > 0, \quad \forall x \in \bar{\Omega}, \quad \forall \xi_i, \quad 1 \leq i \leq n, \quad \sum_{i,j=1}^n a_{ij}(x) \xi_i \xi_j \geq \beta \sum_{i=1}^n \xi_i^2. \tag{4.4.2}$$

The elaboration of the discrete problem comprises three steps:

(i) Construction of a *triangulation* of the set $\bar{\Omega}$ using isoparametric finite elements.

(ii) Definition of a discrete problem *without* numerical integration.

(iii) Definition of a discrete problem *with* numerical integration.

We begin by constructing a set $\bar{\Omega}_h$ as a finite union $\bar{\Omega}_h = \bigcup_{K \in \mathcal{T}_h} K$ of isoparametric finite elements (K, P_K, Σ_K), $K \in \mathcal{T}_h$, which we shall assume to be of *Lagrange type*. Following the description given in the previous section, each finite element (K, P_K, Σ_K) is obtained from a reference finite element $(\hat{K}, \hat{P}, \hat{\Sigma})$ through an isoparametric mapping $F_K \in (\hat{P})^n$ which is uniquely determined by the data of the nodes of the finite element K. *These nodes will always be assumed to belong to the set* $\bar{\Omega}$.

In addition, we shall restrict ourselves to finite elements which possess the following property (cf. Remark 2.3.10):

> *Each basis function* $\hat{p}$ *of the reference finite element* $(\hat{K}, \hat{P}, \hat{\Sigma})$ *vanishes along any face of the set* $\hat{K}$ *which does not contain the node associated with* $\hat{p}$. (4.4.3)

As shown by the examples given in the preceding section, this is not a restrictive assumption.

Of course, we shall take advantage of the isoparametric mappings F_K for getting a good approximation of the boundary Γ: By an appropriate choice of nodes along Γ, we construct finite elements with (at least) one curved face which should be very close to Γ, at any rate closer than a straight face would be. Let us assume for definiteness that we are using *simplicial* finite elements. We may then distinguish two cases, depending upon whether the mapping F_K is affine, i.e., $F_K \in (P_1(\hat{K}))^n$, or the mapping F_K is "truly" isoparametric, i.e., $F_K \in (\hat{P})^n$ but $F_K \notin (\hat{P}_1(\hat{K}))^n$. The latter case will in particular apply to "boundary" finite elements, while the former will rather apply to "interior" finite elements. These considerations are illustrated in Fig. 4.4.1, where we consider the case of triangles of type (2).

For computational simplicity, it is clear that we shall try to keep to a minimum the number of curved faces, and this is why, in general, only the "boundary" finite elements will have one curved face. However, all the subsequent analysis applies equally well to all possible cases, including those in which *all* finite elements $K \in \mathcal{T}_h$ are of the isoparametric type.

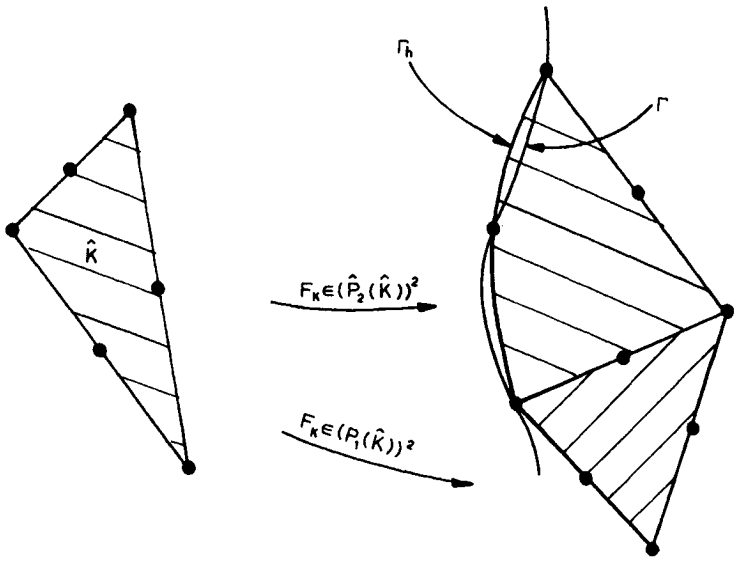


Fig. 4.4.1

In this last case, we must check that the intersection of "adjacent" finite elements is indeed a face for both of them. In other words, there should be no holes and no overlaps. This is true because the finite elements which satisfy (4.4.3) are such that any one of their faces is solely determined by the nodes which are on it (of course, the nodes which define a common face are assumed to be the same for two adjacent finite elements). As an example, we have represented in Fig. 4.4.2 three isoparametric tetrahedra of type (2) "just before assembly": Then the face K' is completely determined by the data of the nodes a_1, a_2, a_3, a_{12}, a_{23}, a_{13}, and the arc $\mathcal{A}$ is completely defined by the data of the nodes a_1, a_2 and a_{12}.

Returning to the general case, we shall assume that *all the nodes which are used in the definition of the faces which approximate the boundary Γ are also the nodes which are on Γ.* Thus, the situation indicated in Fig. 4.4.4 (a) (cf. Exercise 4.4.4) is excluded.

Because each face K' of an isoparametric finite element is necessarily of the form $K' = F_K(\hat{K}')$ with $F_K \in (\hat{P})^n$ and $\hat{K}'$ a face of $\hat{K}$, it is clear that the boundary Γ_h of the set $\bar{\Omega}_h = \bigcup_{K \in \mathcal{T}_h} K$ does not coincide in general with the boundary Γ of the set Ω. Nevertheless, we shall call $\mathcal{T}_h$

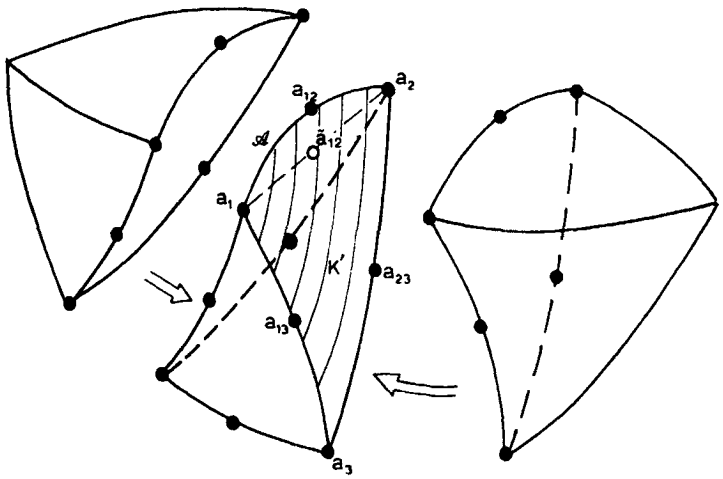


Fig. 4.4.2

a *triangulation of the set* $\bar{\Omega}$ (even though it should more appropriately be called a triangulation of the set $\bar{\Omega}_h$).

Then we let X_h denote the *finite element space* whose functions $v_h: \bar{\Omega}_h \to \mathbf{R}$ are defined as follows:

(i) For each $K \in \mathcal{T}_h$, the restrictions $v_{h|K}$ span the space

$$P_K = \{p: K \to \mathbf{R};\ p = \hat{p} \cdot F_K^{-1},\ \hat{p} \in \hat{P}\}.$$

(ii) Over each $K \in \mathcal{T}_h$, the restrictions $v_{h|K}$ are defined by their values at the nodes of the finite element K.

If the functions of the space $\hat{P}$ are smooth enough, such a space X_h is contained in the space $\mathscr{C}^0(\bar{\Omega}_h)$ (this is again implied by property (4.4.3)), and consequently the inclusion $X_h \subset H^1(\Omega_h)$ holds by Theorem 2.1.1.

We let X_{0h} denote the subspace of X_h whose functions vanish at the *boundary nodes*, i.e., those nodes which are on the boundary Γ. We recall that, by construction, these nodes coincide with those which are on the boundary Γ_h. Therefore another application of property (4.4.3) shows that *the functions in the space* X_{0h} *vanish along the boundary* Γ_h, and thus the inclusion

$$V_h = X_{0h} \subset H_0^1(\Omega_h) \tag{4.4.4}$$

holds (Ω_h denotes the interior of the set $\bar{\Omega}_h$).

Since we expect that the boundaries Γ_h and Γ are close, we shall henceforth assume that *there exists a bounded open set $\tilde{\Omega}$ such that*

$$\Omega \subset \tilde{\Omega} \quad \text{and} \quad \Omega_h \subset \tilde{\Omega} \tag{4.4.5}$$

for all the triangulations $\mathcal{T}_h$ which we shall consider.

Then the most straightforward definition of a *discrete problem* associated with the space V_h consists in finding a function $\tilde{u}_h \in V_h$ such that

$$\forall v_h \in V_h, \quad \int_{\Omega_h} \sum_{i,j=1}^{n} \tilde{a}_{ij} \partial_i \tilde{u}_h \partial_j v_h \, dx = \int_{\Omega_h} \tilde{f} v_h \, dx, \tag{4.4.6}$$

where the functions $\tilde{a}_{ij}$ and $\tilde{f}$ are some *extensions* of the functions a_{ij} and f to the set $\tilde{\Omega}$.

Taking into account isoparametric numerical integration. Description of the resulting discrete problem

In spite of the simplicity and of the natural character of this definition, several questions immediately arise: How should one choose between all possible extensions? How should one construct such extensions in practice? What is the dependence of the discrete solution $\tilde{u}_h$ upon these extensions? Surprisingly, it turns out that these ambiguities will be circumvented by taking into account the effect of *isoparametric numerical integration*:

Just as in Section 4.1, we assume that we have at our disposal a quadrature scheme over the set $\hat{K}$:

$$\int_{\hat{K}} \hat{\varphi}(\hat{x}) \, d\hat{x} \sim \sum_{l=1}^{L} \hat{\omega}_l \hat{\varphi}(\hat{b}_l), \quad \text{with} \quad \hat{\omega}_l \in \mathbf{R}, \hat{b}_l \in \hat{K}, \ 1 \leq l \leq L. \tag{4.4.7}$$

Given two functions $\hat{\varphi}: \hat{K} \to \mathbf{R}$ and $\varphi: K = F_K(\hat{K}) \to \mathbf{R}$ in the usual correspondence (i.e., $\varphi = \hat{\varphi} \cdot F_K^{-1}$), we have

$$\int_K \varphi(x) \, dx = \int_{\hat{K}} \hat{\varphi}(\hat{x}) J_{F_K}(\hat{x}) \, d\hat{x},$$

where the Jacobian J_{F_K} of the mapping F_K may be assumed without loss of generality to be strictly positive over the set $\hat{K}$. Therefore, *the quadrature scheme* (4.4.7) *over the reference element $\hat{K}$ automatically induces a quadrature scheme over the finite element K* (compare with

(4.1.9) and (4.1.10)), namely

$$\int_K \varphi(x)\,dx \sim \sum_{l=1}^{L} \omega_{l,K}\varphi(b_{l,K}), \tag{4.4.8}$$

with *weights* $\omega_{l,K}$ and *nodes* $b_{l,K}$ defined by

$$\omega_{l,K} = \hat{\omega}_l J_{F_K}(\hat{b}_l) \quad \text{and} \quad b_{l,K} = F_K(\hat{b}_l), \quad 1 \leq l \leq L. \tag{4.4.9}$$

Accordingly, we define the *quadrature error functionals*

$$E_K(\varphi) = \int_K \varphi(x)\,dx - \sum_{l=1}^{L} \omega_{l,K}\varphi(b_{l,K}), \tag{4.4.10}$$

$$\hat{E}(\hat{\varphi}) = \int_{\hat{K}} \hat{\varphi}(\hat{x})\,d\hat{x} - \sum_{l=1}^{L} \hat{\omega}_l\hat{\varphi}(\hat{b}_l), \tag{4.4.11}$$

which are related through the equation

$$E_K(\varphi) = \hat{E}(\hat{\varphi}J_{F_K}). \tag{4.4.12}$$

Let us now examine how isoparametric numerical integration affects the definition of the discrete problem (4.4.6). Assuming that the extensions $\tilde{a}_{ij}$ and $\tilde{f}$ are defined everywhere over the set $\bar{\tilde{\Omega}}$, we have to find a discrete solution $u_h \in V_h$ which satisfies (compare with (4.1.24) and (4.1.25)):

$$\forall v_h \in V_h, \sum_{K\in\mathcal{T}_h} \sum_{l=1}^{L} \omega_{l,K} \sum_{i,j=1}^{n} (\tilde{a}_{ij}\partial_i u_h \partial_j v_h)(b_{l,K}) =$$

$$= \sum_{K\in\mathcal{T}_h} \sum_{l=1}^{L} \omega_{l,K}(\tilde{f}v_h)(b_{l,K}). \tag{4.4.13}$$

Then it is clear that *the extensions* $\tilde{a}_{ij}$ *and* $\tilde{f}$ *are not needed in the definition of the above discrete problem if all the quadrature nodes* $b_{l,K}$, $1 \leq l \leq L$, $K \in \mathcal{T}_h$, *belong to the set* $\bar{\Omega}$. To show that this is indeed a common circumstance, let us consider one typical example. Let $n = 2$ and assume that we are using isoparametric triangles of type (2) and that *each node of the quadrature scheme over the set* $\hat{K}$ *either coincides with a node of the triangle* $\hat{K}$ *of type* (2) *or is in the interior* $\mathring{\hat{K}}$ *of the set* $\hat{K}$. *As* shown by the examples given in Section 4.1 (cf. Figs. 4.1.1, 4.1.2 and 4.1.3), this is a realistic situation.

To prove our assertion, we need of course consider only the case of a "boundary" finite element and, at this point, it becomes necessary to indicate *how the boundary nodes are actually chosen.* With the notations

of Fig. 4.4.3, the point $a_{12,K}$ is chosen at the intersection between the boundary Γ and the line perpendicular to the segment $[a_{1,K}, a_{2,K}]$ which passes through the point $\tilde{a}_{12,K} = (a_{1,K} + a_{2,K})/2$.

This choice has three important consequences:

First, if the boundary Γ is smooth enough, we automatically have

$$\|a_{12,K} - \tilde{a}_{12,K}\| = O(h_K^2), \tag{4.4.14}$$

where h_K is the diameter of the triangle with vertices $a_{i,K}$, $1 \leq i \leq 3$. This estimate will insure that a family made up of such isoparametric triangles of type (2) is regular in the sense understood in Section 4.3. We shall use this property in Theorem 4.4.3.

Secondly, *the image* $b_K = F_K(\hat{b})$ *of any point* $\hat{b} \in \hat{K}$ *belongs to the set* $\tilde{\Omega} \cap K$ *provided* h_K *is small enough.* Intuitively, this seems reasonable from a geometrical point of view, and we leave the complete proof to the reader (Exercise 4.4.1).

Thirdly, it is clear that there exists a bounded open set $\tilde{\Omega}$ such that the inclusions (4.4.5) hold.

Remark 4.4.1. The above construction can be easily extended to an open set with a *piecewise smooth boundary*, i.e., a Lipschitz-continuous boundary which is composed of a finite number of smooth arcs, provided each intersection of adjacent arcs is a "vertex" of at least one isoparametric triangle of type (2). □

Remark 4.4.2. When $n = 3$, a node such as a_{12} (cf. Fig. 4.4.2) may be chosen in such a way that the distance between the points $\tilde{a}_{12}$ and a_{12} is equal to the distance between the point $\tilde{a}_{12}$ and the boundary Γ. □

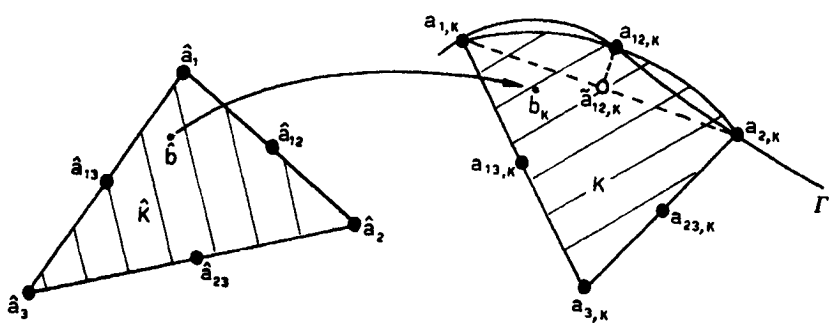


Fig. 4.4.3

Returning to the general case, we are therefore justified in assuming from now on that *the relations*

$$\forall K \in \mathcal{T}_h, \quad b_{l,K} = F_K(\hat{b}_l) \in \bar{\Omega}, \quad 1 \leq l \leq L, \tag{4.4.15}$$

hold for all the triangulations $\mathcal{T}_h$ to be considered.

This being the case, the *discrete problem* (4.4.13) consists in finding a *discrete solution* $u_h \in V_h$ such that

$$\forall v_h \in V_h, \quad a_h(u_h, v_h) = f_h(v_h), \tag{4.4.16}$$

where, for all functions $u_h, v_h \in V_h$, the *approximate bilinear form* $a_h(.,.)$ and the *approximate linear form* $f_h(.)$ are given by

$$a_h(u_h, v_h) = \sum_{K \in \mathcal{T}_h} \sum_{l=1}^{L} \omega_{l,K} \sum_{i,j=1}^{n} (a_{ij} \partial_i u_h \partial_j v_h)(b_{l,K}), \tag{4.4.17}$$

and

$$f_h(v_h) = \sum_{K \in \mathcal{T}_h} \sum_{l=1} \omega_{l,K} (f v_h)(b_{l,K}). \tag{4.4.18}$$

In other words, *thanks to the effect of numerical integration, the discrete problem can be defined without any reference to possible extensions of the functions a_{ij} and f*, and this observation is of course of great practical value (by contrast, extensions explicitly appear in the final error estimate; cf. Theorem 4.4.6).

Remark 4.4.3. Conceivably, *several* quadrature schemes over the reference finite element may be used, depending upon the finite elements $K \in \mathcal{T}_h$. In particular, one would naturally expect that more sophisticated schemes are necessary for dealing with the "truly" isoparametric finite elements. Since our final result (Theorem 4.4.6) shows however that this is not the case, we shall deliberately ignore this possibility (which would require straightforward notational modifications in the writing of (4.4.17) and (4.4.18)). □

Abstract error estimate

Given a family of discrete problems of the form (4.4.16), we shall say that the approximate bilinear forms $a_h(.,.)$ of (4.4.17) are *uniformly*

V_h-elliptic if

$$\exists \tilde{\alpha} > 0, \quad \forall v_h \in V_h, \quad \tilde{\alpha}\|v_h\|^2_{1,\Omega_h} \leq a_h(v_h, v_h), \tag{4.4.19}$$

where the constant $\tilde{\alpha}$ is independent of the subspace V_h.

As usual, we first prove an *abstract error estimate.* The reader should not be surprised by the arbitrariness at this stage in the definition of the functions $\tilde{u}$ and $\tilde{a}_{ij}$ which appear in the next theorem: When this error estimate is actually applied, these will be taken as *extensions* of the functions u and a_{ij} (cf. Theorem 4.4.6).

Theorem 4.4.1. *Given an open set $\tilde{\Omega}$ which contains all the sets Ω_h, and given functions $\tilde{a}_{ij} \in L^\infty(\tilde{\Omega})$, we let*

$$\forall v, w \in H^1(\Omega_h), \quad \tilde{a}_h(v, w) = \int_{\Omega_h} \sum_{i,j=1}^{n} \tilde{a}_{ij}\partial_i v \partial_j w \, dx. \tag{4.4.20}$$

Then if we consider a family of discrete problems of the form (4.4.16), *for which the associated approximate bilinear forms are uniformly V_h-elliptic, there exists a constant C independent of the space V_h such that*

$$\begin{aligned}\|\tilde{u} - u_h\|_{1,\Omega_h} \leq C\Big(& \inf_{v_h \in V_h} \|\tilde{u} - v_h\|_{1,\Omega_h} \\ & + \sup_{w_h \in V_h} \frac{|\tilde{a}_h(v_h, w_h) - a_h(v_h, w_h)|}{\|w_h\|_{1,\Omega_h}} \\ & + \sup_{w_h \in V_h} \frac{|\tilde{a}_h(\tilde{u}, w_h) - f_h(w_h)|}{\|w_h\|_{1,\Omega_h}}\Big),\end{aligned} \tag{4.4.21}$$

where $\tilde{u}$ is any function in the space $H^1(\tilde{\Omega})$, and u_h denotes the solutions of the discrete problems (4.4.16).

Proof. The assumption of uniform V_h-ellipticity insures in particular that each discrete problem has a unique solution u_h. Also, there exists a constant $\tilde{M}$ independent of h such that

$$\forall v, w \in H^1(\Omega_h), \quad |\tilde{a}_h(v, w)| \leq \tilde{M}\|v\|_{1,\Omega_h}\|w\|_{1,\Omega_h}. \tag{4.4.22}$$

Let then v_h denote an arbitrary element in the space V_h. We have

$$\begin{aligned}\tilde{\alpha}\|u_h - v_h\|^2_{1,\Omega_h} &\leq a_h(u_h - v_h, u_h - v_h) \\ &= \tilde{a}_h(\tilde{u} - v_h, u_h - v_h) + \{\tilde{a}_h(v_h, u_h - v_h) \\ &\quad - a_h(v_h, u_h - v_h) + \{f_h(u_h - v_h) - \tilde{a}_h(\tilde{u}, u_h - v_h)\},\end{aligned}$$

so that, using (4.4.22),

$$\begin{aligned}\tilde{\alpha}\|u_h - v_h\|_{1,\Omega_h} &\leq \tilde{M}\|\tilde{u} - v_h\|_{1,\Omega_h} + \frac{|\tilde{a}_h(v_h, u_h - v_h) - a_h(v_h, u_h - v_h)|}{\|u_h - v_h\|_{1,\Omega_h}} \\ &\quad + \frac{|\tilde{a}_h(\tilde{u}, u_h - v_h) - f_h(u_h - v_h)|}{\|u_h - v_h\|_{1,\Omega_h}} \\ &\leq \tilde{M}\|\tilde{u} - v_h\|_{1,\Omega_h} + \sup_{w_h \in V_h} \frac{|\tilde{a}_h(v_h, w_h) - a_h(v_h, w_h)|}{\|w_h\|_{1,\Omega_h}} \\ &\quad + \sup_{w_h \in V_h} \frac{|\tilde{a}_h(\tilde{u}, w_h) - f_h(w_h)|}{\|w_h\|_{1,\Omega_h}}.\end{aligned}$$

Combining the above inequality with the triangular inequality

$$\|\tilde{u} - u_h\|_{1,\Omega_h} \leq \|\tilde{u} - v_h\|_{1,\Omega_h} + \|u_h - v_h\|_{1,\Omega_h},$$

and taking the infimum with respect to $v_h \in V_h$, we obtain inequality (4.4.21). □

Accordingly, the remainder of this section will be devoted to giving sufficient conditions which insure the uniform V_h-ellipticity of the approximate bilinear forms (Theorem 4.4.2) and to estimating the various terms which appear in the right-hand side of inequality (4.4.21). To keep the development within reasonable limits, *we shall however restrict ourselves to finite element spaces made up of isoparametric n-simplices of type* (2).

Finally, we shall make the following assumption:

(H1) *The associated family of triangulations $\mathcal{T}_h$ is regular* in the sense that the family (K), $K \in \bigcup_h \mathcal{T}_h$, is a regular isoparametric family of n-simplices K of type (2) (in the sense understood in Section 4.3; cf. (4.3.26) and (4.3.27)). It is crucial to notice that, in particular, *condition* (4.3.27) *is perfectly compatible with the construction of boundary finite elements* (cf. (4.4.14) and Remark 4.4.2).

Sufficient conditions for uniform V_h-ellipticity

Let us first examine the question of uniform V_h-ellipticity.

Theorem 4.4.2. *Let (V_h) be a family of finite element spaces made up of isoparametric n-simplices of type* (2), *and let there be given a quadrature*

scheme

$$\int_{\hat{K}} \hat{\varphi}(\hat{x})\,d\hat{x} \sim \sum_{l=1}^{L} \hat{\omega}_l \hat{\varphi}(\hat{b}_l) \quad \textit{with} \quad \hat{\omega}_l > 0, \quad 1 \leq l \leq L,$$

such that the union $\bigcup_{l=1}^{L} \{\hat{b}_l\}$ *contains a* $P_1(\hat{K})$*-unisolvent subset and/or the quadrature scheme is exact for the space* $P_2(\hat{K})$.

Then, if hypothesis (H1) *holds, the associated approximate bilinear forms are uniformly* V_h*-elliptic, i.e.,*

$$\exists \tilde{\alpha} > 0, \quad \forall V_h, \quad \forall v_h \in V_h, \quad \tilde{\alpha}\|v_h\|^2_{1,\Omega_h} \leq a_h(v_h, v_h). \tag{4.4.23}$$

Proof. (i) Arguing as in part (i) of the proof of Theorem 4.1.2, we find that there exists a constant $\hat{C} > 0$ such that

$$\forall \hat{p} \in \hat{P} = P_2(\hat{K}), \quad \hat{C}|\hat{p}|^2_{1,\hat{K}} \leq \sum_{l=1}^{L} \hat{\omega}_l \sum_{i=1}^{n} (\partial_i \hat{p}(\hat{b}_l))^2. \tag{4.4.24}$$

(ii) Given a finite element $K \in \mathcal{T}_h$ and a function $v_h \in V_h$, let $p = v_{h|K}$. With the ellipticity condition (4.4.2), we obtain

$$\sum_{l=1}^{L} \omega_{l,K} \sum_{i,j=1}^{n} (a_{ij}\partial_i v_h \partial_j v_h)(b_{l,K}) \geq \beta \sum_{l=1}^{L} \omega_{l,K} \sum_{i=1}^{n} (\partial_i p(b_{l,K}))^2. \tag{4.4.25}$$

We recognize in the expression $\sum_{i=1}^{n} (\partial_i p(b_{l,K}))^2$ the square of the Euclidean norm $\|.\|$ of the vector $Dp(b_{l,K})$. For all points $x = F_K(\hat{x})$, $\hat{x} \in \hat{K}$, we have (by Theorem 4.3.3, F_K is invertible for h small enough)

$$Dp(x) = D\hat{p}(\hat{x})DF_K^{-1}(x),$$

where $Dp(x)$ and $D\hat{p}(\hat{x})$ may be identified with the row vectors $(\partial_1 p(x), \ldots, \partial_n p(x))$ and $(\partial_1 \hat{p}(\hat{x}), \ldots, \partial_n \hat{p}(\hat{x}))$ respectively, and where $DF_K^{-1}(x)$ may be identified with the Jacobian matrix of the mapping F_K^{-1} at x. Using the inequality $\xi A A^T \xi^T \geq (1/\|A^{-1}\|^2)\xi\xi^T$ valid for any invertible matrix A and any row vector ξ (the subscript T denotes transposition), we obtain

$$\forall x = F_K(\hat{x}) \in \hat{K}, \quad \sum_{i=1}^{n} (\partial_i p(x))^2 = Dp(x)Dp(x)^T \geq$$

$$\geq \frac{1}{\|DF_K(\hat{x})\|^2} \sum_{i=1}^{n} (\partial_i \hat{p}(\hat{x}))^2. \tag{4.4.26}$$

Since $\omega_{l,K} = \hat{\omega}_l J_{F_K}(\hat{b})$ (cf. (4.4.9)) and since the weights $\hat{\omega}_l$ are positive,

we deduce from (4.4.26) and (4.4.24):

$$\sum_{l=1}^{L} \omega_{l,K} \sum_{i=1}^{n} (\partial_i p(b_{l,K}))^2$$

$$\geq \inf_{\hat{x}\in\hat{K}} J_{F_K}(\hat{x}) \inf_{\hat{x}\in\hat{K}} \left(\frac{1}{\|DF_K(\hat{x})\|^2}\right) \sum_{l=1}^{L} \hat{\omega}_l \sum_{i=1}^{n} (\partial_i \hat{p}(\hat{b}_l))^2$$

$$\geq \hat{C} \frac{1}{|J_{F_K^{-1}}|_{0,\infty,K} |F_K'|_{1,\infty,\hat{K}}^2} |\hat{p}|_{1,\hat{K}}^2, \tag{4.4.27}$$

where, here and subsequently, we use the notations introduced in (4.3.15). Using Theorem 4.3.2, we know that there exists a constant C such that

$$\forall p = \hat{p} \cdot F_K^{-1}, \quad \hat{p} \in \hat{P}, \ |\hat{p}|_{1,\hat{K}} \geq C \frac{1}{|J_{F_K}|_{0,\infty,\hat{K}}^{1/2} |F_K^{-1}|_{1,\infty,K}} |p|_{1,K}. \tag{4.4.28}$$

Hence, upon combining inequalities (4.4.25), (4.4.27) and (4.4.28), we obtain

$$\sum_{l=1}^{L} \omega_{l,K} \sum_{i,j=1}^{n} (a_{ij} \partial_i v_h \partial_j v_h)(b_{l,K}) \geq$$

$$\geq \beta \hat{C} C^2 \frac{|v_h|_{1,K}^2}{|J_{F_K^{-1}}|_{0,\infty,K} |J_{F_K}|_{0,\infty,\hat{K}} (|F_K|_{1,\infty,\hat{K}} |F_K^{-1}|_{1,\infty,K})^2}. \tag{4.4.29}$$

If we next make use of the estimates established in Theorem 4.3.3 (which we may apply in view of the assumption of regularity), we find that the denominators appearing in the right-hand side of inequality (4.4.29) are uniformly bounded for all $K \in \mathcal{T}_h$, $v_h \in V_h$ and all V_h. Therefore we have shown that

$$\exists \tilde{\alpha}' > 0, \quad \forall v_h \in V_h, \quad \forall K \in \mathcal{T}_h, \quad \forall h,$$

$$\sum_{l=1}^{L} \omega_{l,K} \sum_{i,j=1}^{n} (a_{ij} \partial_i v_h \partial_j v_h)(b_{l,K}) \geq \tilde{\alpha}' |v_h|_{1,K}^2. \tag{4.4.30}$$

(iii) With inequality (4.4.30), we obtain

$$\forall v_h \in V_h, \quad a_h(v_h, v_h) = \sum_{K\in\mathcal{T}_h} \sum_{l=1}^{L} \omega_{l,K} \sum_{i,j=1}^{n} (a_{ij} \partial_i v_h \partial_j v_h)(b_{l,K})$$

$$\geq \tilde{\alpha}' \sum_{K\in\mathcal{T}_h} |v_h|_{1,K}^2 = \tilde{\alpha}' |v_h|_{1,\Omega_h}^2. \tag{4.4.31}$$

Since all the sets Ω_h are contained in a bounded open set $\tilde{\Omega}$ (cf. (4.4.5)),

there exists a constant C independent of h such that

$$\forall v \in H_0^1(\Omega_h), \quad \|v\|_{1,\Omega_h} \leq C|v|_{1,\Omega_h}. \tag{4.4.32}$$

To see this, it suffices, for each function $v \in H_0^1(\Omega_h)$, to apply the Poincaré–Friedrichs inequality over the set $\tilde{\Omega}$ to the function $\tilde{v} \in H_0^1(\tilde{\Omega})$ which equals v on Ω_h and vanishes on $\tilde{\Omega} - \Omega_h$.

Inequality (4.4.23) is then a consequence of inequalities (4.4.31) and (4.4.32) and the proof is complete. □

Interpolation error and consistency error estimates

In what follows, we shall consider the *X_h-interpolation operator* Π_h, whose definition is the natural extension of the definition given in Section 2.3 in the case of straight finite elements: Given a function $v \in \text{dom}\, \Pi_h = \mathscr{C}^0(\bar{\Omega}_h)$, the *$X_h$-interpolant* $\Pi_h v$ is the unique function which satisfies

$$\begin{cases} \Pi_h v \in X_h, \\ \forall K \in \mathscr{T}_h, \quad \Pi_h v(a_{i,K}) = v(a_{i,K}), \quad 1 \leq i \leq n+1, \\ \Pi_h v(a_{ij,K}) = v(a_{ij,K}), \quad 1 \leq i < j \leq n+1, \end{cases} \tag{4.4.33}$$

so that it is clear that the relations

$$\forall K \in \mathscr{T}_h, \quad \Pi_h v|_K = \Pi_K v \tag{4.4.34}$$

hold.

We now estimate the difference $(v - \Pi_h v)$ in various norms. In particular, these estimates will subsequently allow us to obtain (for a specific choice of function $\tilde{u}$) an estimate of the term $\inf_{v_h \in V_h} \|\tilde{u} - v_h\|_{1,\Omega_h}$ which appears in inequality (4.4.21). As usual, the same letter C stands for various constants independent of h and of the various functions involved.

Theorem 4.4.3. *Let (X_h) be a family of finite element spaces made up of isoparametric n-simplices of type* (2), *and assume that $n \leq 5$.*

Then if hypothesis (H1) *holds, there exists a constant C independent of h such that*

$$\forall v \in H^3(\tilde{\Omega}), \quad \|v - \Pi_h v\|_{m,\Omega_h} \leq Ch^{3-m}\|v\|_{3,\Omega_h}, \quad m = 0, 1, \tag{4.4.35}$$

$$\left(\sum_{K \in \mathscr{T}_h} \|v - \Pi_K v\|_{m,K}^2\right)^{1/2} \leq Ch^{3-m}\|v\|_{3,\Omega_h}, \quad m = 2, 3, \tag{4.4.36}$$

where

$$h = \max_{K \in \mathcal{T}_h} h_K, \tag{4.4.37}$$

and $\tilde{\Omega}$ is any open set such that the inclusions (4.4.5) *hold. We also have the implication*

$$v \in H^3(\tilde{\Omega}) \quad \textit{and} \quad v = 0 \quad \textit{on} \quad \Gamma \Rightarrow \Pi_h v \in X_{0h}. \tag{4.4.38}$$

Proof. Since $n \leq 5$, the inclusion $H^3(\hat{K}) \hookrightarrow \mathscr{C}^0(\hat{K})$ holds and thus we may apply Theorem 4.3.4: For all functions $v \in H^3(K)$, we have

$$\|v - \Pi_K v\|_{m,K} \leq Ch_K^{3-m}(|v|_{2,K} + |v|_{3,K}) \leq Ch_K^{3-m}\|v\|_{3,K}, \quad 0 \leq m \leq 3,$$

and inequalities (4.4.35) and (4.4.36) follow from the above inequalities and relations (4.4.34).

If a function vanishes on Γ, its X_h-interpolant vanishes at all the nodes situated on Γ_h (by construction) and therefore it vanishes on the boundary Γ_h of the set $\bar{\Omega}_h = \bigcup_{K \in \mathcal{T}_h} K$. Thus implication (4.4.38) is proved. □

Just as in Section 4.1, the *consistency errors*

$$\sup_{w_h \in V_h} \frac{|\tilde{a}_h(v_h, w_h) - a_h(v_h, w_h)|}{\|w_h\|_{1,\Omega_h}} \quad \text{and} \quad \sup_{w_h \in V_h} \frac{|\tilde{a}_h(\tilde{u}, w_h) - f_h(w_h)|}{\|w_h\|_{1,\Omega_h}}$$

(cf. inequality (4.4.21)) will be estimated as a consequence of careful analyses of similar "local" terms. These are the object of the next two theorems (compare with Theorems 4.1.4 and 4.1.5). The quadrature error functionals $E_K(.)$ and $\hat{E}(.)$ have been defined in (4.4.10) and (4.4.11).

Theorem 4.4.4. *Let there be given a regular isoparametric family of n-simplices K of type* (2), *and let the quadrature scheme over the reference finite element be exact for the space $P_2(\hat{K})$, i.e.,*

$$\forall \hat{\varphi} \in P_2(\hat{K}), \quad \hat{E}(\hat{\varphi}) = 0. \tag{4.4.39}$$

Then there exists a constant C independent of K such that

$$\forall a \in W^{2,\infty}(K), \quad \forall p \in P_K, \quad \forall p' \in P_K, \tag{4.4.40}$$

$$|E_K(a\partial_i p' \partial_j p)| \leq Ch_K^2 \|a\|_{2,\infty,K} \|p'\|_{2,K} |p|_{1,K}.$$

Proof. For notational convenience, the indices K will be dropped throughout the proof.

(i) To begin with, we shall record some consequences of Theorem 4.3.3. First, inequalities (4.3.28) imply that

$$|\partial_i F_k|_{0,\infty,\hat{K}} \leq Ch, \quad 1 \leq i, k \leq n, \tag{4.4.41}$$

$$|\partial_{ij} F_k|_{0,\infty,\hat{K}} \leq Ch^2, \quad 1 \leq i, j, k \leq n, \tag{4.4.42}$$

and inequalities (4.3.30) imply that

$$|J_F|_{0,\infty,\hat{K}} \leq Ch^n, \quad |J_{F^{-1}}|_{0,\infty,K} = |(J_F)^{-1}|_{0,\infty,\hat{K}} \leq \frac{C}{h^n}. \tag{4.4.43}$$

Next, we show that

$$|\partial_i J_F|_{0,\infty,\hat{K}} \leq Ch^{n+1}, \quad 1 \leq i \leq n, \quad |\partial_{ij} J_F|_{0,\infty,\hat{K}} \leq Ch^{n+2}, \quad 1 \leq i, j \leq n, \tag{4.4.44}$$

Let us denote by $\partial_i F(\hat{x})$ and $\partial_{ij} F(\hat{x})$ the column vectors with components $\partial_i F_k(\hat{x})$, $1 \leq k \leq n$, and $\partial_{ij} F_k(\hat{x})$, $1 \leq k \leq n$, respectively. Then to prove the first inequalities of (4.4.44), we observe that, for any $\hat{x} \in \hat{K}$, we have

$$\partial_i J_F(\hat{x}) = \sum_{j=1}^{n} \det(\partial_1 F(\hat{x}), \ldots, \partial_{j-1} F(\hat{x}), \partial_{ij} F(\hat{x}), \partial_{j+1} F(\hat{x}), \ldots, \partial_n F(\hat{x})),$$

and it suffices to use inequalities (4.4.41) and (4.4.42). The second inequalities of (4.4.44) are proved in a similar fashion (since $F \in (P_2(\hat{K}))^n$, the partial derivatives $\partial_{ijk} F$ are identically zero).

(ii) The expression to be estimated can be written as

$$E_K(a \partial_i p' \partial_j p) = \hat{E}(\hat{a}(\partial_i p')\hat{}\,(\partial_j p)\hat{}\, J_F). \tag{4.4.45}$$

Then it is clear that, by contrast with the affine case, *the functions* $(\partial_i p')\hat{}$ *and* $(\partial_j p)\hat{}$ *no longer belong to the space* $P_1(\hat{K})$ in general. This is why our first task is to determine the nature of these functions: Denoting by e_j the j-th basis vector of $\mathbf{R}^n$, we have

$$(\partial_j p)\hat{}\,(\hat{x}) = \partial_j p(x) = Dp(x) e_j = D\hat{p}(\hat{x}) DF^{-1}(x) e_j = D\hat{p}(\hat{x})(DF(\hat{x}))^{-1} e_j.$$

Expressing the fact that the vector $f_j = (DF(\hat{x}))^{-1} e_j$ is the solution of the linear system $DF(\hat{x}) f_j = e_j$, we find that

$$(\partial_j p)\hat{}\,(\hat{x}) = (J_F(\hat{x}))^{-1} \sum_{k=1}^{n} \partial_k \hat{p}(\hat{x}) \times \det(\partial_1 F(\hat{x}), \ldots, \partial_{k-1} F(\hat{x}), e_j, \partial_{k+1} F(\hat{x}), \ldots, \partial_n F(\hat{x})). \tag{4.4.46}$$

Consequently, the expression $(\partial_j p)\hat{}(\hat{x})J_F(\hat{x})$ is a finite sum of terms of the form $\pm \partial_k \hat{p}(\hat{x}) \prod_{l \neq k} \partial_l F_{j_l}(\hat{x})$, and likewise the quantity $(\partial_i p')\hat{}(\hat{x})$ is a finite sum of terms of the form $\pm (J_F(\hat{x}))^{-1} \partial_r (p')\hat{}(\hat{x}) \prod_{s \neq r} \partial_s F_{j_s}(\hat{x})$. Using (4.4.45), we have obtained a sum of the form

$$E_K(a\partial_i p' \partial_j p) = \sum_{\substack{k;j_l,l\neq k \\ r;j_s,s\neq r}}{}' \pm \hat{E}\Big(J_F^{-1}\hat{a} \prod_{s\neq r} \partial_s F_{j_s} \prod_{l\neq k} \partial_l F_{j_l} \partial_r \hat{p}' \partial_k \hat{p}\Big), \tag{4.4.47}$$

where, by the symbol Σ', we simply mean that the indices j_l and j_s do not take all possible values $1, 2, \ldots, n$.

(iii) We shall now take crucial advantage of the fact that the functions $(\partial_i p')\hat{}$ and $(\partial_j p)\hat{}$ can be expressed in terms of the functions $\partial_k \hat{p}$, $1 \leq k \leq n$, which *do* belong to the space $P_1(\hat{K})$: Consider one of the terms occuring in the sum (4.4.47). It can be written as

$$\hat{E}\Big(J_F^{-1}\hat{a} \prod_{s\neq r} \partial_s F_{j_s} \prod_{l\neq k} \partial_l F_{j_l} \partial_r \hat{p}' \partial_k \hat{p}\Big) = \hat{E}(\hat{b}\hat{v}\hat{w}), \tag{4.4.48}$$

with

$$\begin{cases} \hat{b} = J_F^{-1}\hat{a} \prod_{s\neq r} \partial_s F_{j_s} \prod_{l\neq k} \partial_l F_{j_l} \in W^{2,\infty}(\hat{K}), \\ \hat{v} = \partial_r \hat{p}' \in P_1(\hat{K}), \\ \hat{w} = \partial_k \hat{p} \in P_1(\hat{K}), \end{cases} \tag{4.4.49}$$

and consequently, we may apply inequality (4.1.47) with the value $k = 2$. We find in this fashion that

$$\begin{aligned} |\hat{E}(\hat{b}\hat{v}\hat{w})| &\leq C(|\hat{b}|_{2,\infty,\hat{K}}|\hat{v}|_{0,\hat{K}} + |\hat{b}|_{1,\infty,\hat{K}}|\hat{v}|_{1,\hat{K}})|\hat{w}|_{0,\hat{K}} \\ &\leq C(|\hat{b}|_{2,\infty,\hat{K}}|\hat{p}'|_{1,\hat{K}} + |\hat{b}|_{1,\infty,\hat{K}}|\hat{p}'|_{2,\hat{K}})|\hat{p}|_{1,\hat{K}}, \end{aligned} \tag{4.4.50}$$

and it remains to express the various semi-norms occuring in the above inequality in terms of appropriate norms over the set K. Using Theorems 4.3.2 and 4.3.3, we obtain:

$$\begin{cases} |\hat{p}|_{l,\hat{K}} \leq Ch^{-n/2}h^l|p|_{l,K}, \quad l = 0, 1, \\ |\hat{p}'|_{l,\hat{K}} \leq Ch^{-n/2}h^l\|p'\|_{l,K}, \quad l = 1, 2. \end{cases} \tag{4.4.51}$$

Next, we have (cf. (4.1.42))

$$|\hat{b}|_{1,\infty,\hat{K}} = \left| J_F^{-1}\hat{a} \prod_{s\neq r} \partial_s F_{j_s} \prod_{l\neq k} \partial_l F_{j_l} \right|_{1,\infty,\hat{K}}$$
$$\leq C\Big(|J_F^{-1}|_{0,\infty,\hat{K}} |\hat{a}|_{1,\infty,\hat{K}} \left| \prod_{s\neq r} \partial_s F_{j_s} \prod_{l\neq k} \partial_l F_{j_l} \right|_{0,\infty,\hat{K}} +$$
$$+ |J_F^{-1}|_{0,\infty,\hat{K}} |\hat{a}|_{0,\infty,\hat{K}} \left| \prod_{s\neq r} \partial_s F_{j_s} \prod_{l\neq k} \partial_l F_{j_l} \right|_{1,\infty,\hat{K}} +$$
$$+ |J_F^{-1}|_{1,\infty,\hat{K}} |\hat{a}|_{0,\infty,\hat{K}} \left| \prod_{s\neq r} \partial_s F_{j_s} \prod_{l\neq k} \partial_l F_{j_l} \right|_{0,\infty,\hat{K}} \Big), \tag{4.4.52}$$

and we could likewise write an analogous inequality for the semi-norm $|\hat{b}|_{2,\infty,\hat{K}}$. Using inequalities (4.4.41) and (4.4.42), we obtain

$$\left| \prod_{s\neq r} \partial_s F_{j_s} \prod_{l\neq k} \partial_l F_{j_l} \right|_{\lambda,\infty,\hat{K}} \leq Ch^{2n-2+\lambda}, \quad \lambda = 0, 1, 2, \tag{4.4.53}$$

and, using inequalities (4.4.43) and (4.4.44), we obtain

$$|J_F^{-1}|_{\mu,\infty,\hat{K}} \leq Ch^{\mu-n}, \quad \mu = 0, 1, 2, \tag{4.4.54}$$

so that, upon combining inequalities (4.4.52), (4.4.53), (4.4.54) with the inequalities (cf. Theorems 4.3.2 and 4.3.3)

$$|\hat{a}|_{\nu,\infty,\hat{K}} \leq Ch^{\nu}\|a\|_{\nu,\infty,K}, \quad \nu = 0, 1, 2, \tag{4.4.55}$$

we eventually find that

$$|\hat{b}|_{1,\infty,\hat{K}} \leq Ch^{n-1}\|a\|_{1,\infty,K}. \tag{4.4.56}$$

By a similar analysis, we would find that

$$|\hat{b}|_{2,\infty,\hat{K}} \leq Ch^{n}\|a\|_{2,\infty,K}. \tag{4.4.57}$$

Then the conjunction of inequalities (4.4.50), (4.4.51), (4.4.56) and (4.4.57) with equation (4.4.48) shows that

$$\left| \hat{E}\Big(J_F^{-1}\hat{a} \prod_{s\neq r} \partial_s F_{j_s} \prod_{l\neq k} \partial_l F_{j_l} \partial_n \hat{p}' \partial_k \hat{p} \Big) \right| \leq$$
$$\leq Ch^2\|a\|_{2,\infty,K}\|p'\|_{2,K}|p|_{1,K}. \tag{4.4.58}$$

By adding up inequalities (4.4.58), we find that the expression $E_K(a\partial_i p'\partial_j p)$ (cf. (4.4.47)) satisfies an inequality similar to (4.4.58), and the proof is complete. □

Theorem 4.4.5. *Let there be given a regular isoparametric family of n-simplices K of type* (2), *let the quadrature scheme over the reference*

finite element be such that

$$\forall \hat{\varphi} \in P_2(\hat{K}), \quad \hat{E}(\hat{\varphi}) = 0, \tag{4.4.59}$$

and finally, let $q \in [1, \infty]$ *be any number which satisfies the inequality*

$$2 - \frac{n}{q} > 0. \tag{4.4.60}$$

Then there exists a constant C *independent of* K *such that*

$$\forall f \in W^{2,q}(K), \quad \forall p \in P_K,$$

$$|E_K(fp)| \leq Ch_K^2(\text{meas}(\tilde{K}))^{1/2-1/q}\|f\|_{2,q,K}\|p\|_{1,K}, \tag{4.4.61}$$

where, for each K, $\tilde{K}$ *denotes the* n*-simplex with the same vertices as those of* K.

Proof. We have, for all $f \in W^{2,q}(K)$ and all $p \in P_K$,

$$E_K(fp) = \hat{E}(\hat{f}\hat{p}J_F). \tag{4.4.62}$$

It follows from the proof of Theorem 4.1.5 (cf. (4.1.54) and (4.1.55)) that there exists a constant C such that

$$\forall \hat{g} \in W^{2,q}(\hat{K}), \quad \forall \hat{p} \in P_2(\hat{K}),$$

$$|\hat{E}(\hat{g}\hat{p})| \leq C((|\hat{g}|_{1,q,\hat{K}} + |\hat{g}|_{2,q,\hat{K}})|\hat{p}|_{1,\hat{K}} + |\hat{g}|_{2,q,\hat{K}}|\hat{p}|_{0,\hat{K}}). \tag{4.4.63}$$

By letting

$$\hat{g} = \hat{f}J_F \tag{4.4.64}$$

in the above inequality and by making use of inequalities (4.1.42), (4.4.43), (4.4.44) and

$$|\hat{f}|_{\mu,q,\hat{K}} \leq C(\text{meas}(\tilde{K}))^{-1/q}h^{\mu}\|f\|_{\mu,q,K}, \quad \mu = 0, 1, 2$$

(cf. Theorems 4.3.2 and 4.3.3), we obtain

$$|\hat{f}J_F|_{l,q,\hat{K}} \leq C\left(\sum_{j=0}^{l} |J_F|_{j,\infty,\hat{K}}|\hat{f}|_{l-j,q,\hat{K}}\right)$$

$$\leq C(\text{meas}(\tilde{K}))^{-1/q}h^{n+l}\|f\|_{l,q,K}, \; l = 1, 2.$$

These last inequalities, coupled with relations (4.4.62), (4.4.63), (4.4.64) and the first inequalities of (4.4.51) with $l = 0, 1$, yield inequality (4.4.61). □

Estimate of the error $\|\tilde{u} - u_h\|_{1,\Omega_h}$

Combining the previous theorems, we are in a position to prove the main result of this section, which the reader would profit from comparing with Theorem 4.1.6. We recall that u is the solution of the variational problem corresponding to the data (4.4.1). For references concerning the existence of extensions such as $\tilde{u}$ and $\tilde{a}_{ij}$ below, see LIONS (1962, chapter 2), NEČAS (1967, chapter 2).

Theorem 4.4.6. *Let $n \leq 5$, let (V_h) be a family of finite element spaces made up of isoparametric n-simplices of type* (2)*, and let there be given a quadrature scheme on the reference finite element such that*

$$\forall \hat{\varphi} \in P_2(\hat{K}), \quad \hat{E}(\hat{\varphi}) = 0. \tag{4.4.65}$$

Let $\tilde{\Omega}$ be an open set such that the inclusions

$$\Omega \subset \tilde{\Omega}, \quad \text{and} \quad \Omega_h \subset \tilde{\Omega} \text{ for all } h, \tag{4.4.66}$$

hold, and such that the functions u and a_{ij}, $1 \leq i, j \leq n$, possess extensions $\tilde{u}$ and $\tilde{a}_{ij}$, $1 \leq i, j \leq n$, which satisfy

$$\tilde{u} \in H^3(\tilde{\Omega}), \quad \tilde{a}_{ij} \in W^{2,\infty}(\tilde{\Omega}), \quad 1 \leq i, j \leq n, \tag{4.4.67}$$

$$\tilde{f} = \sum_{i,j=1}^{n} \partial_j(\tilde{a}_{ij}\partial_i\tilde{u}) \in W^{2,q}(\tilde{\Omega}) \quad \text{for some} \quad q \geq 2 \text{ with } 2 > \frac{n}{q}. \tag{4.4.68}$$

Then, if hypothesis (H1) *holds, there exists a constant C independent of h such that*

$$\|\tilde{u} - u_h\|_{1,\Omega_h} \leq Ch^2\left(\|\tilde{u}\|_{3,\tilde{\Omega}} + \sum_{i,j=1}^{n} \|\tilde{a}_{ij}\|_{2,\infty,\tilde{\Omega}}\|\tilde{u}\|_{3,\tilde{\Omega}} + \|\tilde{f}\|_{2,q,\tilde{\Omega}}\right), \tag{4.4.69}$$

where $h = \max_{K \in \mathcal{T}_h} h_K$.

Proof. By Theorem 4.4.2, the approximate bilinear forms are uniformly V_h-elliptic and therefore, we can use the abstract error estimate (4.4.21) of Theorem 4.4.1.

(i) Since $n \leq 5$, the inclusion $H^3(\tilde{\Omega}) \hookrightarrow \mathscr{C}^0(\bar{\tilde{\Omega}})$ holds, and by Theorem 4.4.3, the function $\Pi_h\tilde{u}$ belongs to the space X_{0h} since on the boundary Γ, we have $u = \tilde{u} = 0$. Thus we may let $v_h = \Pi_h\tilde{u}$ in the term $\inf_{v_h \in V_h}\{\ldots\}$

which appears in the abstract error estimate. In this fashion we obtain

$$\|\tilde{u} - u_h\|_{1,\Omega_h} \leq C\Big(\|\tilde{u} - \Pi_h\tilde{u}\|_{1,\Omega_h} + \sup_{w_h \in V_h} \frac{|\tilde{a}_h(\Pi_h\tilde{u}, w_h) - a_h(\Pi_h\tilde{u}, w_h)|}{\|w_h\|_{1,\Omega_h}} + \sup_{w_h \in V_h} \frac{|\tilde{a}_h(\tilde{u}, w_h) - f_h(w_h)|}{\|w_h\|_{1,\Omega_h}}\Big). \tag{4.4.70}$$

By Theorem 4.4.3, we know that

$$\|\tilde{u} - \Pi_h\tilde{u}\|_{1,\Omega_h} \leq Ch^2\|\tilde{u}\|_{3,\Omega_h} \leq Ch^2\|\tilde{u}\|_{3,\tilde{\Omega}}. \tag{4.4.71}$$

(ii) To evaluate the two consistency errors, a specific choice must be made for the functions $\tilde{a}_{ij}$ which appear in the bilinear form $\tilde{a}_h(.,.)$: *We shall choose precisely the functions given in* (4.4.67). Notice that, since the inclusion $W^{2,\infty}(\tilde{\Omega}) \hookrightarrow \mathscr{C}^1(\bar{\tilde{\Omega}})$ holds, the functions $\tilde{a}_{ij}$ are in particular defined everywhere on the set $\bar{\tilde{\Omega}}$. Then we have, for all $w_h \in V_h$,

$$\tilde{a}_h(\Pi_h\tilde{u}, w_h) - a_h(\Pi_h\tilde{u}, w_h) = \int_{\Omega_h} \sum_{i,j=1}^{n} \tilde{a}_{ij}\partial_i\Pi_h\tilde{u}\partial_j w_h \, dx - \sum_{K \in \mathscr{T}_h} \sum_{l=1}^{L} \omega_{l,K} \sum_{i,j=1}^{n} (a_{ij}\partial_i\Pi_h\tilde{u}\partial_j w_h)(b_{l,K}),$$

and, since all the quadrature nodes $b_{l,K}$ belong to the set $\bar{\Omega}$, we have $a_{ij}(b_{l,K}) = \tilde{a}_{ij}(b_{l,K})$. Consequently, we can rewrite the above expression as

$$\tilde{a}_h(\Pi_h\tilde{u}, w_h) - a_h(\Pi_h\tilde{u}, w_h) = \sum_{K \in \mathscr{T}_h} \sum_{i,j=1}^{n} E_K(\tilde{a}_{ij}\partial_i\Pi_h\tilde{u}\partial_j w_h).$$

Using the estimates of Theorem 4.4.4 and the Cauchy–Schwarz inequality, we obtain

$$|\tilde{a}_h(\Pi_h\tilde{u}, w_h) - a_h(\Pi_h\tilde{u}, w_h)| \leq C \sum_{K \in \mathscr{T}_h} h_K^2 \sum_{i,j=1}^{n} \|\tilde{a}_{ij}\|_{2,\infty,K}\|\Pi_K\tilde{u}\|_{2,K}|w_h|_{1,K}$$

$$\leq Ch^2\Big(\sum_{i,j=1}^{n} \|\tilde{a}_{ij}\|_{2,\infty,\tilde{\Omega}}\Big)\Big(\sum_{K \in \mathscr{T}_h} \|\Pi_K\tilde{u}\|_{2,K}^2\Big)^{1/2} |w_h|_{1,\Omega_h}.$$

By another application of Theorem 4.4.3,

$$\left(\sum_{K\in\mathcal{T}_h}\|\Pi_K\tilde{u}\|_{2,K}^2\right)^{1/2}\leq\|\tilde{u}\|_{2,\Omega_h}+\left(\sum_{K\in\mathcal{T}_h}\|\tilde{u}-\Pi_K\tilde{u}\|_{2,K}^2\right)^{1/2}$$
$$\leq\|\tilde{u}\|_{2,\Omega_h}+Ch\|\tilde{u}\|_{3,\Omega_h}\leq C\|\tilde{u}\|_{3,\hat{\Omega}},$$

and thus, we have shown that

$$\sup_{w_h\in V_h}\frac{|\tilde{a}_h(\Pi_h\tilde{u},w_h)-a_h(\Pi_h\tilde{u},w_h)|}{\|w_h\|_{1,\Omega_h}}\leq Ch^2\sum_{i,j=1}^{n}\|\tilde{a}_{ij}\|_{2,\infty,\hat{\Omega}}\|\tilde{u}\|_{3,\hat{\Omega}}. \tag{4.4.72}$$

(iii) Let us next examine the expression which appears in the numerator of the second consistency error. First it is easily verified that assumptions (4.4.67) imply in particular that the functions $(a_{ij}\partial_i\tilde{u})$ belong to the space $H^1(\Omega)$.

Therefore Green's formula yields

$$\forall w_h\in V_h\subset H_0^1(\Omega_h),\quad \tilde{a}_h(\tilde{u},w_h)=\int_{\Omega_h}\sum_{i,j=1}^{n}\tilde{a}_{ij}\partial_i\tilde{u}\partial_j w_h\,dx$$
$$=-\int_{\Omega_h}\sum_{i,j=1}^{n}\partial_j(\tilde{a}_{ij}\partial_i\tilde{u})w_h\,dx$$
$$=\int_{\Omega_h}\tilde{f}w_h\,dx.$$

Since we have

$$-\sum_{i,j=1}^{n}\partial_j(\tilde{a}_{ij}\partial_i\tilde{u})=-\sum_{i,j=1}^{n}\partial_j(a_{ij}\partial_i u)=f\quad\text{on}\quad\Omega,$$

the function $\tilde{f}$ given in (4.4.68) *is an extension of the function f.* Besides, using once again the fact that all integration nodes $b_{l,K}$ belong to the set $\bar{\Omega}$, we obtain $f(b_{l,K})=\tilde{f}(b_{l,K})$ and consequently, we can write

$$\tilde{a}_h(\tilde{u},w_h)-f_h(w_h)=\int_{\Omega_h}\tilde{f}w_h\,dx-\sum_{K\in\mathcal{T}_h}\sum_{l=1}^{L}\omega_{l,K}(fw_h)(b_{l,K})$$
$$=\sum_{K\in\mathcal{T}_h}E_K(\tilde{f}w_h).$$

Using the estimates of Theorem 4.4.5, we get

$$|\tilde{a}_h(\tilde{u},w_h)-f_h(w_h)|\leq C\sum_{K\in\mathcal{T}_h}h_K^2(\text{meas}(\hat{K}))^{1/2-1/q}\|\tilde{f}\|_{2,q,K}\|w_h\|_{1,K}$$
$$\leq Ch^2\left(\sum_{K\in\mathcal{T}_h}\text{meas}(\hat{K})\right)^{1/2-1/q}\|\tilde{f}\|_{2,q,\Omega_h}\|w_h\|_{1,\Omega_h}.$$

By construction, the interiors of the n-simplices do not overlap and therefore the quantity $\Sigma_{K\in\mathcal{T}_h}\operatorname{meas}(\tilde{K}) = \operatorname{meas}(\bigcup_{K\in\mathcal{T}_h}\tilde{K})$ is clearly bounded independently of h. Thus, we have shown that

$$\sup_{w_h\in V_h}\frac{|\tilde{a}_h(\tilde{u}, w_h) - f_h(w_h)|}{\|w_h\|_{1,\Omega_h}} \leq Ch^2\|\tilde{f}\|_{2,q,\tilde{\Omega}}, \tag{4.4.73}$$

and inequality (4.4.69) follows from inequalities (4.4.70), (4.4.71), (4.4.72) and (4.4.73). □

We have therefore reached a *remarkable conclusion*: *In order to retain the same order of convergence as in the case of polygonal domains (when only straight finite elements are used), the same quadrature scheme should be used, whether it be for straight or for isoparametric finite elements.* Thus, if $n = 2$ for instance, we can use the quadrature scheme of (4.1.17), which is exact for polynomials of degree ≤ 2.

Remark 4.4.4. (i) As one would expect, it is of course true that, in the absence of numerical integration, the order of convergence is the same, i.e., one has $\|\tilde{u} - \tilde{u}_h\|_{1,\Omega_h} = O(h^2)$, where $\tilde{u}$ is now the solution of the discrete problem (4.4.6). To show this is the object of Exercise 4.4.3.

(ii) To make the analysis even more complete, it would remain to show that for a given domain with a curved boundary (irrespectively of whether or not numerical integration is used), isoparametric n-simplices of type (2) yield better estimates than their straight counterparts! Indeed, STRANG & BERGER (1971) and THOMÉE (1973b) have shown that one gets in the latter case an $O(h^{3/2})$ convergence. In this direction, see Exercise 4.4.4. □

Remark 4.4.5. By contrast with the case of straight finite elements (cf. Remark 4.1.8), *the integrals* $\int_K a_{ij}\partial_i u_h \partial_j v_h \, dx$ *are no longer computed exactly when the coefficients* a_{ij} *are constant functions.* If K is an isoparametric n-simplex of type (2), we have

$$\forall p', p \in P_K, \quad \int_K \partial_i p' \partial_j p \, dx = \int_{\hat{K}} J_F(\partial_i p')\hat{}\,(\partial_j p)\hat{} \, d\hat{x},$$

and (cf. (4.4.46)),

$$J_F(\hat{x})(\partial_i p')\hat{}(\hat{x})$$
$$= \sum_{k=1}^{n} \partial_k \hat{p}'(\hat{x}) \det(\partial_1 F(\hat{x}), \ldots,$$
$$\times \partial_{k-1} F(\hat{x}), e_i, \partial_{k+1} F(\hat{x}), \ldots, \partial_n F(\hat{x}))$$
$$= \{\text{polynomial of degree} \leq n \text{ in } \hat{x}\},$$
$$(\partial_i p)\hat{}(x) = (J_F(\hat{x}))^{-1} \times \{\text{polynomial of degree} \leq n \text{ in } \hat{x}\}.$$

Since

$$J_F(\hat{x}) = \det(\partial_1 F(\hat{x}), \ldots, \partial_n F(\hat{x}))$$
$$= \{\text{polynomial of degree} \leq n \text{ in } \hat{x}\},$$

we eventually find that

$$\int_K \partial_i p' \partial_j p \, dx = \int_{\hat{K}} \frac{\{\text{polynomial of degree} \leq 2n \text{ in } \hat{x}\}}{\{\text{polynomial of degree} \leq n \text{ in } \hat{x}\}} d\hat{x}.$$

Therefore the exact computation of such integrals would require a quadrature scheme which is exact for rational functions of the form N/D with $N \in P_{2n}(\hat{K})$, $D \in P_n(\hat{K})$. □

Exercises

4.4.1. With the notations of Fig. 4.4.3, show that the image $b_K = F_K(\hat{b})$ of any point $\hat{b} \in \hat{K}$ belongs to the set $\bar{\Omega} \cap K$ provided h_K is small enough.

4.4.2. With the same assumptions as in Theorem 4.4.4, show that the estimates

$$|E_K(a\partial_i p' \partial_j p)| \leq C h_K \|a\|_{2,\infty,K} \|p'\|_{1,K} |p|_{1,K}$$

hold. Deduce from these another proof of the uniform V_h-ellipticity of the approximate bilinear forms (this type of argument is used by ZLÁMAL (1974)).

4.4.3. Analyze the case where isoparametric n-simplices of type (2) are used *without* numerical integration, i.e., the discrete problem is defined as in (4.4.6).

[Hint: After defining appropriate extensions of the functions a_{ij} so that the discrete bilinear forms are uniformly V_h-elliptic, use the abstract error estimates of Theorem 4.4.1. This type of analysis is carried out in SCOTT (1973a).]

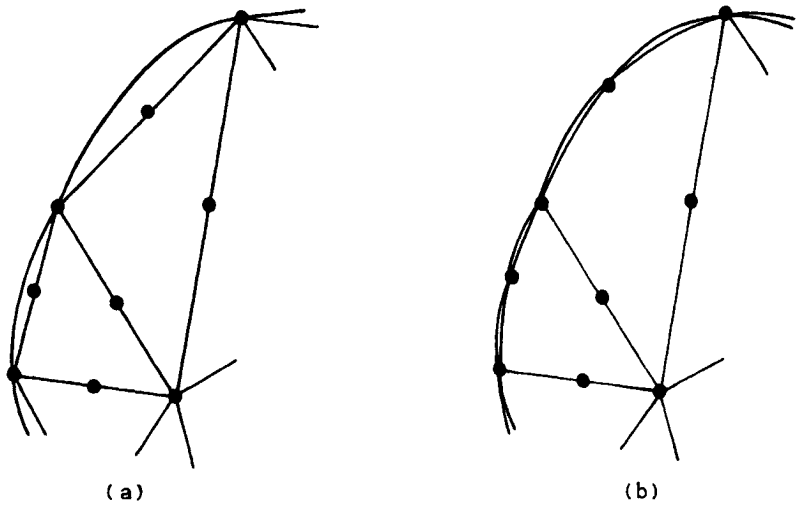


Fig. 4.4.4

4.4.4. Assume that the set Ω is a bounded convex domain in $\mathbf{R}^2$. Given a triangulation $\mathcal{T}_h$ made up of triangles *with straight sides only*, let X_h denote the finite element space whose generic finite element is the triangle of type (2), and let $V_h = \{v_h \in X_h;\ v_h = 0 \text{ on } \Gamma_h\}$, where Γ_h is the boundary of the set $\bigcup_{K \in \mathcal{T}_h} K$ (cf. Fig. 4.4.4(a)).

Show that (cf. STRANG & BERGER (1971), THOMÉE (1973b); see also STRANG & FIX (1973, Chapter 4))

$$\|u - u_h\|_{1,\Omega_h} = O(h^{3/2}),$$

where $u_h \in V_h$ is the solution of the equations

$$\forall v_h \in V_h, \quad \int_{\Omega} \sum_{i,j=1}^{n} a_{ij} \partial_i u_h \partial_j v_h \, dx = \int_{\Omega} f v_h \, dx$$

(one should notice that in this case, the X_h-interpolant of the solution u does *not* belong to the space V_h). In other words, triangulations of type (b) are asymptotically better than triangulations of type (a) (cf. Fig. 4.4.4).

4.4.5. Analyze the case where n-simplices of type (1) are used, with or without numerical integration, over a curved domain Ω. It is assumed that all the vertices which are on the boundary of the set $\bar{\Omega}_h = \bigcup_{K \in \mathcal{T}_h} K$ are also on the boundary Γ.

Bibliography and comments

4.1. The content of this section is essentially based on CIARLET & RAVIART (1972c, 1975) and RAVIART (1972).

The abstract error estimate of Theorem 4.1.1 is based on STRANG (1972b). The proof of the uniform V_h-ellipticity given in Theorem 4.1.2 is based on, and generalizes, an idea of G. Strang (STRANG & FIX (1973, Section 4.3)).

Theorem 4.1.3 is due to BRAMBLE & HILBERT (1970). It is recognized as an important tool in getting error estimates in numerical integration and interpolation theory (although we did not use it in Section 3.1).

In CIARLET & RAVIART (1975), the content of this section is given a general treatment so as to comprise as special cases the inclusions $P_k(\hat{K}) \subset \hat{P} \subset P_{k'}(\hat{K})$ (cf. Exercise 4.1.6), the case of quadrilateral elements (cf. Exercise 4.1.7), etc... As regards in particular the error estimate in the norm $|.|_{0,\Omega}$ (cf. the abstract error estimate of Exercise 4.1.3), the following is proved: Assuming that the adjoint problem is regular and that $\hat{P} = P_k(\hat{K})$, one has $|u - u_h|_{0,\Omega} = O(h^{k+1})$ if the quadrature scheme is exact for the space $P_{2k-2}(\hat{K})$ if $k \geq 2$, or if the quadrature scheme is exact for the space $P_1(\hat{K})$ if $k = 1$.

For other references concerning the effect of numerical integration, see BABUŠKA & AZIZ (1972, Ch. 9), FIX (1972a, 1972b), HERBOLD (1968) where this problem was studied for the first time, HERBOLD, SCHULTZ & VARGA (1969), HERBOLD & VARGA (1972), ODEN & REDDY (1976a, Section 8.8), SCHULTZ (1972), STRANG & FIX (1973, Section 4.3).

Comparisons between finite element methods (with or without numerical integration) and finite-difference methods are found in BIRKHOFF & GULATI (1974), TOMLIN (1972), WALSH (1971).

Examples of numerical quadrature schemes used in actual computations are found in the book of ZIENKIEWICZ (1971, Section 8.10).

For general introductions to the subject of numerical integration (also known as: *numerical quadrature, approximate integration, approximate quadrature*), see the survey of HABER (1970), and the books of DAVIS & RABINOWITZ (1974), STROUD (1971).

For studies of numerical integration along the lines developed here, see also MANSFIELD (1971, 1972a). In ARCANGELI & GOUT (1976) and MEINGUET (1975), the constants appearing in the quadrature error estimates are evaluated.

4.2. The abstract error estimate of Theorem 4.2.2 is due to STRANG

(1972b). The description of Wilson's brick is given in WILSON & TAYLOR (1971).

In analyzing the consistency error, we have followed the method set up in CIARLET (1974a) for studying nonconforming methods, the main idea being to obtain two polynomial invariances in the functions $D_K(.,.)$ so as to apply the bilinear lemma. For the specific application of this method to Wilson's brick, we have extended to the three-dimensional case the analysis which LESAINT (1976) has made for Wilson's rectangle. P. Lesaint has considered the use of this element for approximating the system of plane elasticity, for which he was able to show the uniform ellipticity of the corresponding approximate bilinear forms. In this fashion, P. Lesaint obtains an $O(h)$ convergence in the norm $\|.\|_h$ and an $O(h^2)$ convergence in the norm $|.|_{0,\Omega}$ (the corresponding technique is indicated in Exercise 4.2.3). Also, the idea of introducing the degrees of freedom $\int_K \partial_{ij} p \, dx$ is due to P. Lesaint.

In his pioneering work on the mathematical analysis of nonconforming methods, G. Strang (cf. STRANG (1972b), and also STRANG & FIX (1973, Section 4.2) where the study of Wilson's brick is sketched) has shown in particular the importance of the patch test of B. Irons (cf. IRONS & RAZZAQUE (1972a)). For more recent developments on the connection with the patch test, see OLIVEIRA (1976).

There are other ways of generating nonconforming finite element methods. See for example RACHFORD & WHEELER (1974). In NITSCHE (1974), several types of such methods are analyzed in a systematic way. See also CÉA (1976).

References more specifically concerned with nonconforming methods for fourth-order problems are postponed till Section 6.2.

4.3. This section is based on CIARLET & RAVIART (1972b), where an attempt was made to establish an interpolation theory for general isoparametric finite elements (in this direction see Exercises 4.3.1, 4.3.8 and 4.3.9). A survey is given in CIARLET (1973).

To see that our description indeed coincides with the one used by the Engineers, let us consider for example the isoparametric triangle of type (2) as described by FELIPPA & CLOUGH (1970, p. 224): Given six points $a_i = (a_{1i}, a_{2i})$ $1 \leq i \leq 6$, in the plane (the points a_4, a_5 and a_6 play momentarily the role of the points which we usually call a_{12}, a_{23} and a_{13}, respectively), a "natural" coordinate system is defined, whereby the following relation (written in matrix form) should hold between the Cartesian coordinates x_1 and x_2 describing the finite element and the

"new" coordinates λ_1, λ_2 and λ_3:

$$\begin{pmatrix} x_1 \\ x_2 \\ 1 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{16} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & a_{26} \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1(2\lambda_1 - 1) \\ \lambda_2(2\lambda_2 - 1) \\ \lambda_3(2\lambda_3 - 1) \\ 4\lambda_1\lambda_2 \\ 4\lambda_2\lambda_3 \\ 4\lambda_3\lambda_1 \end{pmatrix}$$

Then we observe that the first two lines of the above relation precisely represent relation (4.3.7), with $F(\hat{x}) = (F_1(\hat{x}), F_2(\hat{x}))$ now denoted (x_1, x_2). The last line of the above matrix equation implies either $\lambda_1 + \lambda_2 + \lambda_3 = 1$ or $\lambda_1 + \lambda_2 + \lambda_3 = -\frac{1}{2}$, so that the solution $\lambda_1 + \lambda_2 + \lambda_3 = 1$ is the only one which is acceptable if we impose the restriction that $\lambda_i \geq 0$, $1 \leq i \leq 3$.

Therefore, the "natural" coordinates λ_1, λ_2 and λ_3 are nothing but the barycentric coordinates with respect to a fixed triangle $\hat{K}$, and the isoparametric finite element associated with the points a_i, $1 \leq i \leq 6$, is in this formulation the set of those points (x_1, x_2) given by the first two lines of the above matrix equation when the "natural" coordinates λ_i (also known as "*curvilinear*" *coordinates*) satisfy the inequalities $0 \leq \lambda_i \leq 1$, $1 \leq i \leq 3$, and the equality $\sum_{i=1}^{3} \lambda_i = 1$.

A general description of isoparametric finite elements along these lines is also found in ZIENKIEWICZ (1971, chapter 8). The first references where such finite elements are found are ARGYRIS & FRIED (1968) and ERGATOUDIS, IRONS & ZIENKIEWICZ (1968).

In case of isoparametric quadrilateral elements, JAMET (1976b) has significantly contributed to the interpolation error analysis, by relaxing some assumptions of CIARLET & RAVIART (1972b).

Curved finite elements of other than isoparametric type have also been considered, notably by ZLÁMAL (1970, 1973a, 1973b, 1974) and SCOTT (1973a). Both authors begin by constructing a curved face K' by approximating a smooth surface through an $(n-1)$-dimensional interpolation process. This interpolation serves to define a mapping F_K which in turn allows to define a finite element with K' as a curved face. Then the corresponding interpolation theory follows basically the same pattern as here. In particular, R. Scott constructs in this fashion a curved finite element which resembles the isoparametric triangle of type (3) and

for which an interpolation theory can be developed which requires weaker assumptions than those indicated in Exercise 4.3.8.

In ARCANGÉLI & GOUT (1976), a polynomial interpolation process over a curved domain is analyzed. For curved finite elements based on the so-called *blending function* interpolation process, see notably CAVENDISH, GORDON & HALL (1976), GORDON & HALL (1973), the paper of BARNHILL (1976a) and the references therein. WACHSPRESS (1971, 1973, 1975) uses rational functions for constructing general polygonal finite elements in the plane with straight or curved sides. For additional references, see LEAF, KAPER & LINDEMAN (1976), LUKÁŠ (1974), MCLEOD & MITCHELL (1972, 1976), MITCHELL (1976), MITCHELL & MARSHALL (1975).

4.4. The error analysis developed in this section follows the general approach set up in CIARLET & RAVIART (1972c) (however it was thought at that time that more accurate quadrature schemes were needed for isoparametric elements), where an estimate of the error in the norm $|.|_{0,\Omega}$ was also obtained.

An analogous study is made in ZLÁMAL (1974), where it is shown that, for two-dimensional curved elements for which $\hat{P} = P_k(\hat{K})$, k even, it is sufficient to use quadrature schemes exact for polynomials of degree $\leq 2k-2$, in order to retain the $O(h^k)$ convergence in the norm $\|.\|_{1,\Omega_h}$. ZLÁMAL (1973b) has also evaluated the error in the absence of numerical integration. For complementary results, see VEIDINGER (1975). Likewise, SCOTT (1973a) has shown that quadrature schemes of higher order of accuracy are not needed when curved finite elements are used. However, the finite elements considered by M. Zlámal and R. Scott are not of the isoparametric type as understood here. For such elements, a general theory is yet to be developed, in particular for quadrilateral finite elements.

In spite of the absence of a uniform V_h-ellipticity condition, GIRAULT (1976a) has successfully studied the use of quadrilaterals of type (1) in conjunction with a one-point quadrature scheme.

Alternate ways of handling Dirichlet problems posed over domains with curved boundaries have been proposed, which rely on various alterations of the bilinear form of the given problem. In this direction, we notably mention

(i) penalty methods, as advocated by AUBIN (1969) and BABUŠKA (1973b), and later improved by KING (1974),

(ii) methods where the boundary condition is considered as a con-

straint and as such is treated via techniques from duality theory, as in BABUŠKA (1973a),

(iii) least square methods as proposed and studied in BRAMBLE & SCHATZ (1970, 1971), BRAMBLE & NITSCHE (1973), BAKER (1973),

(iv) methods where the domain is approximated by a polygonal domain, as in BRAMBLE, DUPONT & THOMÉE (1972),

(v) various methods proposed by NITSCHE (1971, 1972b).

For additional references for the finite element approximation of boundary value problems over curved boundaries, see BABUŠKA (1971b), BERGER (1973), BERGER, SCOTT & STRANG (1972), BLAIR (1976), BRAMBLE (1975), NITSCHE (1972b), SCOTT (1975), SHAH (1970), STRANG & BERGER (1971), STRANG & FIX (1973, Chapter 4), THOMÉE (1973a, 1973b). See also Chapter 6 for fourth-order problems.

We finally mention that, following the terminology of STRANG (1972b), we have perpetrated in this chapter three *variational crimes*: numerical integration, nonconforming methods, approximation of curved boundaries.

Additional bibliography and comments

Problems on unbounded domains

Let us consider one physical example: Given an electric conductor which occupies a bounded volume $\bar{\Omega}$ in $\mathbf{R}^3$, and assuming that the electric potential u_0 is known along the boundary Γ of the set Ω, the *electric conductor problem* consists in finding the space distribution of the electric potential u. This potential u is the solution of

$$\begin{cases} \Delta u = 0 & \text{in} \quad \Omega, \\ \Delta u = 0 & \text{in} \quad \Omega' = \mathbf{C}\bar{\Omega}, \\ u = u_0 & \text{on} \quad \Gamma. \end{cases}$$

Thus, in addition to a standard problem on the set $\bar{\Omega}$, we have to solve a boundary value problem on the *unbounded* set $\bar{\Omega}'$. Classically, this problem is solved in the following fashion: Denoting by $\partial_\nu u$ the normal derivative of $u|_{\bar{\Omega}}$ across Γ and by $(\partial_\nu u)'$ the normal derivative of $u|_{\bar{\Omega}'}$ across Γ (both normals being oriented in the same direction), let

$$q = \partial_\nu u - (\partial_\nu u)'.$$

Then if the function q is known on Γ, the solution u is obtained in $\mathbf{R}^3$ as a *single layer potential* through the formula

$$\forall x \in \mathbf{R}^3, \quad u(x) = \frac{1}{4\pi} \int_\Gamma \frac{q(y)}{\|x - y\|} \, d\gamma(y).$$

By specializing the points x to belong to the boundary Γ in the above formula, we are therefore led to solve the *integral equation*:

$$\forall x \in \Gamma, \quad u_0(x) = \frac{1}{4\pi} \int_\Gamma \frac{q(y)}{\|x - y\|} \, d\gamma(y)$$

in the unknown q. For details about this classical approach, see for instance PETROVSKY (1954).

Interestingly, this integral equation can be given a variational formulation which, among other things, make it amenable to finite element approximations, as shown by NÉDÉLEC & PLANCHARD (1973). First we need a new *Sobolev space*, the space

$$H^{1/2}(\Gamma) = \{r \in L^2(\Gamma); \exists v \in H^1(\Omega); \text{ tr } v = r \quad \text{on} \quad \Gamma\},$$

which is dense in the space $L^2(\Gamma)$. It is a Hilbert space when it is equipped with the quotient norm

$$r \in H^{1/2}(\Gamma) \to \|r\|_{H^{1/2}(\Gamma)} = \inf\{\|v\|_{1,\Omega}; v \in H^1(\Omega), \text{ tr } v = r \quad \text{on} \quad \Gamma\}.$$

We shall denote by $H^{-1/2}(\Gamma)$ its dual space, and by $\|\cdot\|_{H^{-1/2}(\Gamma)}$ the dual norm. Denoting by $\langle \cdot, \cdot \rangle_\Gamma$ the duality pairing between the spaces $H^{-1/2}(\Gamma)$ and $H^{1/2}(\Gamma)$, we note that

$$\forall r \in L^2(\Gamma) \subset H^{-1/2}(\Gamma), \quad \forall s \in H^{1/2}(\Gamma), \quad \langle r, s \rangle_\Gamma = \int_\Gamma rs \, d\gamma.$$

For details about these spaces (and more generally about the spaces $H^t(\Gamma)$, $t \in \mathbf{R}$), see LIONS & MAGENES (1968).

The bilinear form

$$(q, r) \to \frac{1}{4\pi} \int_\Gamma \int_\Gamma \frac{q(x)r(y)}{\|x - y\|} \, d\gamma(x) \, d\gamma(y)$$

is well-defined over the space $\mathcal{D}(\Gamma) \times \mathcal{D}(\Gamma)$, and it is continuous when the space $\mathcal{D}(\Gamma)$ is equipped with the norm $\|\cdot\|_{H^{-1/2}(\Gamma)}$. Consequently, it has a unique extension over the space $H^{-1/2}(\Gamma) \times H^{-1/2}(\Gamma)$, which shall be denoted by $a(\cdot, \cdot)$, and one can show that this bilinear form is $H^{-1/2}(\Gamma)$-

elliptic. Therefore, the natural variational formulation of the problem posed above as an integral equation consists in finding the unique function q which satisfies

$$q \in H^{-1/2}(\Gamma) \quad \text{and} \quad \forall r \in H^{-1/2}(\Gamma), \quad a(q, r) = \langle u_0, r \rangle_\Gamma,$$

assuming the data u_0 belongs to the space $H^{1/2}(\Gamma)$.

Once the function q is found in this fashion, the solution u of the original problem is obtained as follows. Define the space $W_0^1(\mathbf{R}^3)$ as being the completion of the space $\mathcal{D}(\mathbf{R}^3)$ with respect to the norm $|\cdot|_{1,\mathbf{R}^3}$. This space (which does *not* coincide with the space $H^1(\mathbf{R}^3)$, i.e., the completion of the space $\mathcal{D}(\mathbf{R}^3)$ with respect to the norm $\|\cdot\|_{1,\mathbf{R}^3}$) can be equally characterized by (cf. BARROS, NETO (1965), DENY & LIONS (1953–1954))

$$\begin{aligned} W_0^1(\mathbf{R}^3) &= \{v \in L^6(\mathbf{R}^3);\ \partial_i v \in L^2(\mathbf{R}^3);\quad 1 \leq i \leq 3\} \\ &= \left\{ v \in \mathcal{D}'(\mathbf{R}^3);\ \frac{v}{(1+\|x\|^2)^{1/2}} \in L^2(\mathbf{R}^3);\ \partial_i v \in L^2(\mathbf{R}^3), \right. \\ &\qquad \left. 1 \leq i \leq 3 \right\}. \end{aligned}$$

Then for each function $q \in \mathcal{D}(\Gamma)$, the function

$$u: x \in \mathbf{R}^3 \to u(x) = \frac{1}{4\pi} \int_\Gamma \frac{q(y)}{\|x - y\|} \, d\gamma(y)$$

belongs to the space $W_0^1(\mathbf{R}^3)$ and, besides, the mapping $q \in \mathcal{D}(\Gamma) \to u \in W_0^1(\mathbf{R}^3)$ defined in this fashion is continuous when the space $\mathcal{D}(\Gamma)$ is equipped with the norm $\|\cdot\|_{H^{-1/2}(\Gamma)}$. Therefore, it has a unique extension over the space $H^{-1/2}(\Gamma)$. In other words, we have solved the original problem via the mappings $u_0 \in H^{1/2}(\Gamma) \to q \in H^{-1/2}(\Gamma) \to u \in W_0^1(\mathbf{R}^3)$ (as indicated in NÉDÉLEC & PLANCHARD (1973), one can also solve directly the problem $u_0 \in H^{1/2}(\Gamma) \to u \in W_0^1(\mathbf{R}^3)$). We mention that the related *perfect dielectric problem* can be also handled in an analogous manner (the boundary condition $u = u_0$ on Γ is then replaced by $\partial_\nu u - c(\partial_\nu u)' = u_1$ on Γ, $c > 0$).

J.C. Nédélec and J. Planchard then construct a general finite element approximation of the above problem. Given a subspace of the space $H^{-1/2}(\Gamma)$, they first derive an abstract error estimate: Let $u_{0h} \in V_h$ be an approximation of the function u_0 and let the discrete solution q_h be such that

$$\forall r_h \in V_h, \quad a(q_h, r_h) = \langle u_{0h}, r_h \rangle_\Gamma.$$

Then there exists a constant C independent of the subspace V_h such that

$$\|q - q_h\|_{H^{-1/2}(\Gamma)} \leq C\Big(\inf_{r_h \in V_h} \|q - r_h\|_{H^{-1/2}(\Gamma)} + \|u_0 - u_{0h}\|_{H^{1/2}(\Gamma)}\Big).$$

To apply this error estimate the authors assume that the boundary Γ is polygonal so that it may be triangulated in an obvious fashion, i.e., the set Γ is written as a union $\bigcup_{K \in \mathcal{T}_h} K$ of triangles K. Then they look for a discrete solution in either space

$$V_{0h} = \{v_h : \Gamma \to \mathbf{R};\ \forall K \in \mathcal{T}_h,\ v_h|_K \in P_0(K)\} \subset L^2(\Gamma) \subset H^{-1/2}(\Gamma),$$

$$V_{1h} = \{v_h \in \mathcal{C}^0(\Gamma);\ \forall K \in \mathcal{T}_h,\ v_h|_K \in P_1(K)\} \subset H^1(\Gamma) \subset H^{-1/2}(\Gamma)$$

(note that the functions in the space V_{0h} are discontinuous across adjacent triangles) and they show that

$$\|q - q_h\|_{H^{1/2}(\Gamma)} \leq \begin{cases} C(q)h^{3/2} + C\|u_0 - u_{0h}\|_{H^{1/2}(\Gamma)} & \text{in } V_{0h}, \\ C(q)h^{5/2} + C\|u_0 - u_{0h}\|_{H^{1/2}(\Gamma)} & \text{in } V_{1h}, \end{cases}$$

assuming the function q is smooth enough. To conclude their analysis, they compute the function

$$u_h : x \in \mathbf{R}^3 \to \frac{1}{4\pi} \int_\Gamma \frac{q_h(y)}{\|x - y\|}\, d\gamma(y),$$

and they obtain in both cases

$$\|u - u_h\|_{W_0^1(\mathbf{R}^3)} \leq C(q)h^{3/2} + \|u - u_{0h}\|_{H^{1/2}(\Gamma)}.$$

Of course, the major computational difficulty in this approach is the evaluation of the coefficients of the resulting linear system. For a review of the numerical aspects of such integral equation techniques for solving problems on unbounded domains arising in the study of 2- or 3-dimensional incompressible potential flows around obstacles, see HESS (1975a, 1975b).

NÉDÉLEC (1976) next considers the case of a curved surface Γ which needs therefore to be approximated by another surface Γ_h made up of finite elements of isoparametric type (such a construction is related to – and is of interest for – the surface approximation found in the shell problem; cf. Section 8.2). Again error estimates for the differences $(q - q_h)$ and $(u - u_h)$ are obtained in appropriate Hilbert spaces.

LE ROUX (1974, 1977) considers the finite element approximation of the analogous problem in dimension two. In this case, the kernel in the

integral transform is $\ln\|x-y\|$ instead of $1/\|x-y\|$. A similar analysis is found in HSIAO & WENDLAND (1976).

There are other ways of handling problems on unbounded domains. In particular, there are methods where the unbounded domain is triangulated and then the triangulation is "truncated" in some fashion. In this spirit, BABUŠKA (1972c) considers the model problem: Find $u \in H^1(\mathbf{R}^n)$ such that $-\Delta u + u = f$ in $\mathbf{R}^n$, $f \in L^2(\mathbf{R}^n)$. Using an "abstract" variational approximation (cf. BABUŠKA (1970, 1971a)), he obtains orders of convergence on compact subsets of $\mathbf{R}^n$ which are arbitrarily close to the orders of convergence obtained in the case of bounded domains. By contrast with the method of FIX & STRANG (1969), the discrete solution is obtained via the solution of a linear system with a finite number of unknowns. In SILVESTER & HSIEH (1971), a bounded subdomain is triangulated in the usual way while the remaining unbounded part is represented by a single "finite element" of a special type.

As we shall mention in the section "Bibliography and Comments" of Section 5.1, problems on unbounded domains which typically arise in the study of 2-dimensional compressible flows may be reduced to variational inequalities, as in CIAVALDINI & TOURNEMINE (1977) and ROUX (1976).

The Stokes problem

Classically, the *Stokes problem* for an incompressible viscous fluid in a domain $\bar{\Omega} \subset \mathbf{R}^n$, $n = 2$ or 3, consists in finding functions $\boldsymbol{u} = (u_i)_{i=1}^n$ and p defined over the set $\bar{\Omega}$, which satisfy $(\Delta \boldsymbol{u} = (\Delta u_i)_{i=1}^n)$

$$\begin{cases} -\nu\Delta \boldsymbol{u} + \nabla p = \boldsymbol{f} & \text{in } \Omega, \\ \operatorname{div} \boldsymbol{u} = 0 & \text{in } \Omega, \\ \boldsymbol{u} = \boldsymbol{0} & \text{on } \Gamma. \end{cases}$$

The vector function $\boldsymbol{u}$ represents the velocity distribution, the scalar function p is the pressure, and the given vector function $\boldsymbol{f} = (f_i)_{i=1}^n \in (L^2(\Omega))^n$ represents the volumic forces per unit mass. The constant $\nu > 0$ is the *dynamic viscosity*, a constant which is inversely proportional to the *Reynolds number.*

In order to derive the variational formulation of this problem, we introduce the space

$$V = \{\boldsymbol{v} \in (H_0^1(\Omega))^n;\ \operatorname{div} \boldsymbol{v} = 0\}$$

provided with the norm

$$|\cdot|_{1,\Omega}: v = (v_i)_{i=1}^n \to |v|_{1,\Omega} = \left(\sum_{i=1}^n |v_i|_{1,\Omega}^2\right)^{1/2},$$

and we introduce the bilinear form

$$u, v \in (H^1(\Omega))^n \to a(u, v) = \sum_{i,j=1}^n \int_\Omega \partial_j u_i \partial_j v_i \, dx,$$

which is clearly V-elliptic. We shall use the notation

$$(u, v) = \int_\Omega uv \, dx, \quad (u, v) = \int_\Omega u \cdot v \, dx.$$

Since one has $(\nabla p, v) = -(p, \operatorname{div} v)$ for all $v \in \mathcal{D}(\Omega)$ for smooth functions p, *the natural variational formulation of this problem consists in finding a pair* (u, p) *such that*

$$(u, p) \in V \times (L^2(\Omega)/P_0(\Omega)), \quad \text{and}$$
$$\forall v \in (H_0^1(\Omega))^n, \quad \nu a(u, v) - (p, \operatorname{div} v) = (f, v)$$

(notice that the definition of the space $L^2(\Omega)/P_0(\Omega)$ reflects the fact that the unknown p can be determined up to additive constants only). Then we observe that the relations

$$\forall v \in V, \quad \nu a(u, v) = (f, v)$$

determine uniquely the function u (a word of caution: Since the space $(\mathcal{D}(\Omega))^n$ is *not* contained in the space V, the above variational problem *cannot* be interpreted in the usual way as a boundary value problem). Once the function u is known, it remains to find a function $p \in L^2(\Omega)/P_0(\Omega)$ such that

$$\forall v \in (H_0^1(\Omega))^n, \quad (p, \operatorname{div} v) = g(v),$$

where the linear form

$$g: v \in (H_0^1(\Omega))^n \to g(v) = \nu a(u, v) - (f, v)$$

is continuous over the space $(H_0^1(\Omega))^n$ and vanishes over its subspace V, by definition of the function u.

It then follows that there exists a function $p \in L^2(\Omega)$, unique up to an additive constant factor, such that the linear form can also be written as

$$\forall v \in (H_0^1(\Omega))^n, \quad g(v) = \int_\Omega p \operatorname{div} v \, dx.$$

This is a nontrivial fact whose proof may be found in de RHAM (1955) (the converse is clear).

Notice that if $n = 2$, the Stokes problem can be reduced to a familiar problem: Since $\operatorname{div} \boldsymbol{u} = 0$, there exists a *stream function* ψ such that $u_1 = \partial_2\psi$ and $u_2 = -\partial_1\psi$. Then a simple computation shows that $\nu\Delta^2\psi = f$ with $f = \partial_1 f_2 - \partial_2 f_1$. When the set Ω is simply connected, we may impose the boundary condition $\psi = 0$ on Γ, so that we also have $\partial_\nu\psi = 0$ on Γ as a consequence of the boundary condition $\boldsymbol{u} = \boldsymbol{0}$ on Γ. Therefore *the solution of the Stokes problem is reduced in this case to the solution of a biharmonic problem* (cf. Section 1.2). Observe that the *vorticity* $-\Delta\psi$ is then nothing but the value of the rotational of the velocity $\boldsymbol{u}$. Finite element methods for this problem will be described in Section 6.1 and Chapter 7.

As regards the finite element approximation of the general Stokes problem, it is realized that a major difficulty consists in taking properly into account the *incompressibility condition* $\operatorname{div} \boldsymbol{u} = 0$. A first approach is to use standard finite element spaces $\boldsymbol{V}_h$ in which the condition $\operatorname{div} \boldsymbol{v}_h = 0$ is exactly imposed. However, this process often results in sophisticated elements. Methods of this type have been extensively studied by FORTIN (1972a, 1972b).

In a second approach, whose applicability seems wider, the *incompressibility condition is approximated.* This is the method advocated by CROUZEIX & RAVIART (1973), who seek the discrete solution in a space of the form

$$\boldsymbol{V}_h = \left\{ \boldsymbol{v}_h \in \boldsymbol{X}_{0h};\ \forall \phi_h \in \Phi_h,\ \sum_{K \in \mathcal{T}_h} \int_K \phi_h \operatorname{div} \boldsymbol{v}_h \, dx = 0 \right\},$$

where $\boldsymbol{X}_{0h}$ is a product of standard finite element spaces and Φ_h appears as an appropriate space of "Lagrange multipliers", following the terminology of duality theory (cf. the section "Additional Bibliography and Comments" in Chapter 7). For instance if the generic finite element in the space X_{0h} is the triangle – or tetrahedron – of type (k), the space Φ_h is the product $\Pi_{K \in \mathcal{T}_h} P_{k-1}(K)$. In their remarkable paper, M. Crouzeix and P.-A. Raviart construct both conforming and nonconforming finite elements of special type (cf. Exercise 2.3.9) and they obtain estimates for the error $(\Sigma_{K \in \mathcal{T}_h} |u - u_h|^2_{1,K})^{1/2}$, and for the error $(\Sigma_{i=1}^n |u_i - u_{ih}|^2_{0,\Omega})^{1/2}$ through an extension of the Aubin–Nitsche lemma. They also compute an approximation p_h of the pressure p and they evaluate the norm $\|p - p_h\|_{L^2(\Omega)/P_0(\Omega)}$. Finally, they briefly consider the case of the in-

homogeneous boundary condition $\boldsymbol{u} = \boldsymbol{u}_0$ on Γ. As usual the error estimates depend upon the smoothness of the solution, a question studied in KELLOGG & OSBORN (1976), OSBORN (1976b), and TÉMAM (1973). It seems however that the most promising finite element approximations of the Stokes problem are of the so-called *mixed* type. For such methods, the reader is referred to the section "Additional Bibliography and Comments" at the end of Chapter 7.

Further references concerning the finite element approximation of the Stokes problem are FALK (1976a, 1976c), FALK & KING (1976), and the thorough treatment given by TÉMAM (1977). We also mention that CROUZEIX & LE ROUX (1976) have proposed and analyzed a finite element method for two-dimensional irrotational fluid flows, in which the unknown $\boldsymbol{u} = (u_1, u_2)$ satisfies

$$\begin{cases} rot\, \boldsymbol{u} = \boldsymbol{0} & \text{in} \quad \Omega, \\ \operatorname{div} \boldsymbol{u} = f & \text{in} \quad \Omega, \\ \boldsymbol{u} \cdot \boldsymbol{\nu} = g & \text{on} \quad \Gamma. \end{cases}$$

Eigenvalue problems

Given an elliptic operator $\mathcal{L}$ defined on a bounded open subset Ω of $\mathbf{R}^n$ and given the boundary condition $u = 0$ on Γ, the associated *eigenvalue problem* classically consists in finding real numbers λ and functions $u \neq 0$ such that

$$\begin{cases} \mathcal{L}u = \lambda u & \text{in} \quad \Omega, \\ u = 0 & \text{in} \quad \Gamma. \end{cases}$$

Indeed, eigenvalue problems may be associated with any other homogeneous boundary conditions but, for simplicity, we shall consider only the Dirichlet condition.

Such problems typically arise when one looks for *periodic* (in time) solutions of evolution problems of the form $\partial_{00}u + \mathcal{L}u = 0$, $u = 0$ on Γ, where ∂_{00} denotes the second partial derivative with respect to the time variable t. Such particular solutions being of the form $u(x)e^{i\mu t}$, $\mu \in \mathbf{R}$, the pair (λ, u), $\lambda = \mu^2$, is therefore obtained through the solution of an eigenvalue problem. This is why such problems are of fundamental importance, in the analysis of vibrations of structures for instance.

We shall in fact consider the *variational formulation of this eigenvalue problem*, which consists in finding pairs (λ, u), $\lambda \in \mathbf{R}$, $u \in V - \{0\}$, such that

$$\forall v \in V \quad a(u, v) = \lambda(u, v),$$

where $V = H_0^1(\Omega)$ or $H_0^2(\Omega)$, depending upon whether $\mathcal{L}$ is a second-order or fourth-order operator, $a(\cdot, \cdot)$ is the associated bilinear form (i.e., which satisfies $a(u, v) = (\mathcal{L}u, v)$ for all $v \in \mathcal{D}(\Omega)$), and $(\cdot, \cdot)$ is the inner-product in the space $L^2(\Omega)$. If (λ, u) is a solution, then u is called an *eigenfunction* associated with the *eigenvalue* λ.

Let us make the usual assumptions that the bilinear form is continuous and V-elliptic, so that for each $f \in L^2(\Omega)$, there exists a unique function $u \in V$ which satisfies $a(u, v) = (f, v)$ for all $v \in V$ (if we identify the function f with an element of V', we have $u = A^{-1}f$ with the notations of Theorem 1.1.3). In this fashion, we define a mapping

$$G: f \in L^2(\Omega) \to u = Gf \in V,$$

which is continuous (cf. Remark 1.1.3), and consequently, *the mapping*

$$G: V \to V$$

is compact, by Rellich theorem. Since $(u, v) = a(Gu, v)$ for all $u, v \in V$ by definition of the mapping G, *the eigenvalue problem amounts to finding the inverses of the eigenvalues of the mapping* $G: V \to V$ (clearly, zero cannot be an eigenvalue of the mapping G nor of the original problem).

If we finally add the assumption of *symmetry* of the bilinear form, then the problem is reduced to that of finding the eigenvalues and eigenfunctions of a compact symmetric operator in the Hilbert space V, considered as equipped with the inner-product $a(\cdot, \cdot)$ (the symmetry is a consequence of the equalities $a(Gu, v) = (u, v) = (v, u) = a(Gv, u) = a(u, Gv)$). Consequently, an application of the spectral theory of such operators (cf. e.g. RIESZ & NAGY (1952)) yields the following result concerning the existence and characterizations of the solutions of the eigenvalue problem: *There exists an increasing sequence of strictly positive eigenvalues*:

$$0 < \lambda_1 \leqslant \lambda_2 \leqslant \cdots \leqslant \lambda_k \leqslant \lambda_{k+1} \leqslant \cdots \text{ with } \lim_{k \to \infty} \lambda_k = \infty,$$

associated with eigenfunctions u_k, $k \geqslant 1$, *which can be orthonormalized*

in the sense that

$$a(u_k, u_l) = \lambda_k \delta_{kl}, \quad (u_k, u_l) = \delta_{kl}, \quad k,l \geqslant 1,$$

and which form a complete system in both the Hilbert spaces V and $L^2(\Omega)$.

Moreover, if we introduce the *Rayleigh quotient*

$$R: v \in V - \{0\} \to R(v) = \frac{a(v, v)}{(v, v)},$$

the eigenvalues are characterized by the relations

$$\begin{cases} \lambda_1 = \inf\{R(v);\ v \in V\} = R(u_1), \\ \lambda_k = \inf\{R(v);\ v \in V,\ (v, u_l) = 0,\ 0 \leqslant l \leqslant k-1\} = R(u_k),\ k \geqslant 2. \end{cases}$$

Other characterizations (quite useful for the analysis of such problems and of their approximations) as well as further developments may be found in COURANT & HILBERT (1953, Chapter V).

The simplest discretization of such problems is called the *Rayleigh–Ritz method* and it is defined as follows: Given a subspace V_h, of dimension M, of the space V, find pairs $(\lambda_h, u_h) \in \mathbf{R} \times V$ such that

$$\forall v_h \in V_h, \quad a(u_h, v_h) = \lambda_h (u_h, v_h).$$

Equivalently, if we let w_k, $1 \leqslant k \leqslant M$, denote a basis in the space V_h, the problem consists in finding the solutions of the generalized matrix eigenvalue problem in $\mathbf{R}^M$:

$$\mathcal{A}_h u = \lambda_h \mathcal{B}_h u,$$

where the coefficients of the symmetric and positive definite matrices $\mathcal{A}_h$ and $\mathcal{B}_h$ have respectively for expressions $a(w_h, w_l)$ and (w_k, w_l). In this fashion we obtain M strictly positive *approximate eigenvalues*

$$0 < \lambda_{1h} \leqslant \lambda_{2h} \leqslant \cdots \leqslant \lambda_{Mh}$$

and M *approximate eigenfunctions* u_{kh}, $1 \leqslant k \leqslant M$, which can be orthonormalized in the sense that

$$a(u_{kh}, u_{lh}) = \lambda_{kh} \delta_{kl}, \quad (u_{kh}, u_{lh}) = \delta_{kl}, \quad 1 \leqslant k, l \leqslant M.$$

One can then show that for any fixed integer l, one has $\lim_{h \to 0} \lambda_{kh} = \lambda_k$ (from above), $1 \leqslant k \leqslant l$, provided $\lim_{h \to 0} \inf_{v_h \in V_h} \|u_k - v_h\| = 0$, $1 \leqslant k \leqslant l$ (compare with Theorem 2.4.1). More precisely, one obtains inequalities

of the form

$$\lambda_k \leqslant \lambda_{kh} \leqslant \lambda_k + C(l) \sum_{r=1}^{k} \inf_{v_h \in V_h} \|u_r - v_h\|^2, \quad 1 \leqslant k \leqslant l,$$

which show that the order of the error $|\lambda_{kh} - \lambda_k| = (\lambda_{kh} - \lambda_k)$ is the *square* of the order of the interpolation error (provided as usual the eigenfunctions are smooth enough). For the eigenfunctions one can show (again with appropriate smoothness assumptions) that the orders of the errors $\|u_{kh} - u_k\|$ are the same as those of the interpolation error (with additional difficulties in case of multiple eigenvalues, as expected). Such error estimates are found in BIRKHOFF, de BOOR, SWARTZ & WENDROFF (1966), CIARLET, SCHULTZ & VARGA (1968b), FIX (1969), PIERCE & VARGA (1972a, 1972b), CHATELIN & LEMORDANT (1975).

For treatments more directly connected with the finite element method, see BABUŠKA & ARIZ (1972, Chapter 10), FIX (1972a) (where in particular the effect of numerical integration is studied), FIX (1973), GRÉGOIRE, NÉDÉLEC & PLANCHARD (1976), STRANG & FIX (1973, Chapter 6).

Extensions to the nonsymmetric case have been obtained by BRAMBLE & OSBORN (1972, 1973). See also BRAMBLE (1972), OSBORN (1974). For an extension to a noncompact operator, see RAPPAZ (1976, 1977).

For the practical implementation of such methods, see for example BATHE & WILSON (1973), LINDBERG & OLSON (1970).