Chapter 4
Finite Element Methods

Susanne C. Brenner\textsuperscript{1} and Carsten Carstensen\textsuperscript{2}

\textsuperscript{1}University of South Carolina, Columbia, SC, USA
\textsuperscript{2}Humboldt-Universität zu Berlin, Berlin, Germany

1 Introduction 73
2 Ritz–Galerkin Methods for Linear Elliptic Boundary Value Problems 74
3 Finite Element Spaces 77
4 A Priori Error Estimates for Finite Element Methods 82
5 A Posteriori Error Estimates and Analysis 85
6 Local Mesh Refinement 98
7 Other Aspects 104
Acknowledgments 114
References 114

1 INTRODUCTION

The finite element method is one of the most widely used techniques in computational mechanics. The mathematical origin of the method can be traced to a paper by Courant (1943). We refer the readers to the articles by Babuška (1994) and Oden (1991) for the history of the finite element method. In this chapter, we give a concise account of the \( h \)-version of the finite element method for elliptic boundary value problems in the displacement formulation, and refer the readers to the theory of Chapter 5 and Chapter 9 of this Volume.

This chapter is organized as follows. The finite element method for elliptic boundary value problems is based on the Ritz–Galerkin approach, which is discussed in Section 2. The construction of finite element spaces and the a priori error estimates for finite element methods are presented in Sections 3 and 4. The a posteriori error estimates for finite element methods and their applications to adaptive local mesh refinements are discussed in Sections 5 and 6. For the ease of presentation, the contents of Sections 3 and 4 are restricted to symmetric problems on polyhedral domains using conforming finite elements. The extension of these results to more general situations is outlined in Section 7.

For the classical material in Sections 3, 4, and 7, we are content with highlighting the important results and pointing to the key literature. We also concentrate on basic theoretical results and refer the readers to other chapters in this encyclopedia for complications that may arise in applications. For the recent development of a posteriori error estimates and adaptive local mesh refinements in Sections 5 and 6, we try to provide a more comprehensive treatment. Owing to space limitations many significant topics and references are inevitably absent. For in-depth discussions of many of the topics covered in this chapter (and the ones that we do not touch upon), we refer the readers to the following survey articles and books (which are listed in alphabetical order) and the references therein (Ainsworth and Oden, 2000; Apel, 1999; Aziz, 1972; Babuška and Aziz, 1972; Babuška and Strouboulis, 2001; Bangerth and Rannacher, 2003; Bathe, 1996; Becker, Carey and Oden, 1981; Becker and Rannacher, 2001; Braess, 2001; Brenner and Scott, 2002; Ciarlet, 1978, 1991; Eriksson et al., 1995; Hughes, 2000; Oden and Reddy, 1976; Schatz, Thomée and Wendland, 1990; Strang and Fix, 1973; Szabó and Babuška, 1991; Verfürth, 1996; Wahlbin, 1991, 1995; Zienkiewicz and Taylor, 2000).
2 RITZ–GALERKIN METHODS FOR LINEAR ELLIPTIC BOUNDARY VALUE PROBLEMS

In this section, we set up the basic mathematical framework for the analysis of Ritz–Galerkin methods for linear elliptic boundary value problems. We will concentrate on symmetric problems. Nonsymmetric elliptic boundary value problems will be discussed in Section 7.1.

2.1 Weak problems

Let \( \Omega \) be a bounded connected open subset of the Euclidean space \( \mathbb{R}^d \) with a piecewise smooth boundary. For a positive integer \( k \), the Sobolev space \( \mathcal{H}^k(\Omega) \) is the space of square integrable functions whose weak derivatives up to order \( k \) are also square integrable, with the norm

\[
\| v \|_{\mathcal{H}^k(\Omega)} = \left( \sum_{|\alpha| \leq k} \left\| \frac{\partial^{|\alpha|} v}{\partial x^\alpha} \right\|_{L^2(\Omega)}^2 \right)^{1/2}
\]

The seminorm \( \left\| \cdot \right\|_{\mathcal{H}^k(\Omega)} \) will be denoted by \( |v|_{\mathcal{H}^k(\Omega)} \). We refer the readers to Nečas (1967), Adams (1995), Triebel (1978), Grisvard (1985), and Wloka (1987) for the properties of the Sobolev spaces. Here we just point out that \( \| \cdot \|_{\mathcal{H}^k(\Omega)} \) is a norm induced by an inner product and \( \mathcal{H}^k(\Omega) \) is complete under this norm, that is, \( \mathcal{H}^k(\Omega) \) is a Hilbert space. (We assume that the readers are familiar with normed and Hilbert spaces.)

Using the Sobolev spaces we can represent a large class of symmetric elliptic boundary value problems of order \( 2m \) in the following abstract weak form:

Find \( u \in \mathcal{V} \), a closed subspace of a Sobolev space \( \mathcal{H}^m(\Omega) \), such that

\[
\mathcal{A}(u, v) = \mathcal{F}(v) \quad \forall v \in \mathcal{V}
\]

where \( \mathcal{F}: \mathcal{V} \to \mathbb{R} \) is a bounded linear functional on \( \mathcal{V} \) and \( \mathcal{A}(\cdot, \cdot) \) is a symmetric bilinear form that is bounded and \( \mathcal{V} \)-elliptic, that is,

\[
\begin{align*}
|\mathcal{A}(v_1, v_2)| &\leq C_1 \| v_1 \|_{\mathcal{H}^m(\Omega)} \| v_2 \|_{\mathcal{H}^m(\Omega)} \quad \forall v_1, v_2 \in \mathcal{V} \\
\mathcal{A}(v, v) &\geq C_2 \| v \|^2_{\mathcal{H}^m(\Omega)} \quad \forall v \in \mathcal{V}
\end{align*}
\]

Remark 1. We use \( C \), with or without subscript, to represent a generic positive constant that can take different values at different occurrences.

Remark 2. Equation (1) is the Euler–Lagrange equation for the variational problem of finding the minimum of the functional \( v \mapsto \frac{1}{2} \mathcal{A}(v, v) - \mathcal{F}(v) \) on the space \( \mathcal{V} \). In mechanics, this functional often represents an energy and its minimization follows from the Dirichlet principle. Furthermore, the corresponding Euler–Lagrange equations (also called first variation) (1) often represent the principle of virtual work.

It follows from conditions (2) and (3) that \( \mathcal{A}(\cdot, \cdot) \) defines an inner product on \( \mathcal{V} \) which is equivalent to the inner product of the Sobolev space \( \mathcal{H}^m(\Omega) \). Therefore the existence and uniqueness of the solution of (1) follow immediately from (2), (3), and the Riesz Representation Theorem (Yosida, 1995; Reddy, 1986; Oden and Demkowicz, 1996).

The following are typical examples from computational mechanics.

Example 1. Let \( \mathcal{A}(\cdot, \cdot) \) be defined by

\[
\mathcal{A}(v_1, v_2) = \int_\Omega \nabla v_1 \cdot \nabla v_2 \, dx
\]

For \( f \in L^2(\Omega) \), the weak form of the Poisson problem

\[
-\Delta u = f \quad \text{on } \Omega, \quad u = 0 \quad \text{on } \Gamma, \quad \frac{\partial u}{\partial n} = 0 \quad \text{on } \partial \Omega \setminus \Gamma
\]

where \( \Gamma \) is a subset of \( \partial \Omega \) with a positive \( (d-1) \)-dimensional measure, is given by (1) with \( \mathcal{V} = \{ v \in \mathcal{H}^1(\Omega); v|_{\Gamma} = 0 \} \) and

\[
\mathcal{F}(v) = \int_{\Omega} f v \, dx = (f, v)_{L^2(\Omega)}
\]

For the pure Neumann problem where \( \Gamma = \emptyset \), since the gradient vector vanishes for constant functions, an appropriate function space for the weak problem is \( \mathcal{V} = \{ v \in \mathcal{H}^1(\Omega); v|_{\Gamma} = 0 \} \).

Example 2. Let \( \Omega \subset \mathbb{R}^d \) \( (d = 2, 3) \) and \( v \in [\mathcal{H}^1(\Omega)]^d \) be the displacement of an elastic body. The strain tensor \( \varepsilon(v) \) is given by the \( d \times d \) matrix with components

\[
\varepsilon_{ij}(v) = \frac{1}{2} \left( \frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right)
\]
and the stress tensor $\sigma(v)$ is the $d \times d$ matrix defined by

$$\sigma(v) = 2\mu \epsilon(v) + \lambda (\nabla v) \delta$$

(10)

where $\delta$ is the $d \times d$ identity matrix and $\mu > 0$ and $\lambda > 0$ are the Lamé constants.

Let the bilinear form $a(\cdot, \cdot)$ be defined by

$$a(v_1, v_2) = \int_{\Omega} \sum_{i,j=1}^d \sigma_{ij}(v_1) \epsilon_{ij}(v_2) \, dx = \int_{\Omega} \sigma(v_1) : \epsilon(v_2) \, dx$$

(11)

For $f \in L^2(\Omega)^d$, the weak form of the linear elasticity problem (Ciarlet, 1988)

$$\begin{align*}
\text{div} [\sigma(u)] &= f \quad \text{on } \Omega, \quad u = 0 \quad \text{on } \Gamma \\
[\sigma(u)]n &= 0 \quad \text{on } \partial\Omega \setminus \Gamma
\end{align*}$$

(12)

where $\Gamma$ is a subset of $\partial\Omega$ with a positive $(d-1)$-dimensional measure, is given by (1) with $V = \{v \in [H^1(\Omega)]^d : v|_{\Gamma} = 0\}$ and

$$F(v) = \int_{\Omega} f \cdot v \, dx = (f, v)_{L^2(\Omega)}$$

(13)

For the pure traction problem where $\Gamma = \emptyset$, the strain tensor vanishes for all infinitesimal rigid motions, i.e., displacement fields of the form $m = a + \rho x$, where $a \in \mathbb{R}^d$, $\rho$ is a $d \times d$ antisymmetric matrix and $x = (x_1, \ldots, x_d)^T$ is the position vector. In this case an appropriate function space for the weak problem is $V = \{v \in [H^1(\Omega)]^d : v|_{\Gamma} = 0\}$ and $F(v) = \int_{\Omega} f \cdot v \, dx = (f, v)_{L^2(\Omega)}$.

The boundedness of $F$ and $a(\cdot, \cdot)$ is obvious and the coercivity of $a(\cdot, \cdot)$ follows from Korn’s inequalities (Friedrichs, 1947; Duvaut and Lions, 1976; Nitsche, 1981) (see Chapter 2, Volume 2):

$$\|v\|_{H^1(\Omega)} \leq C \left( \|\epsilon(v)\|_{L^2(\Omega)} + \int_{\Gamma} v \, ds \right) \quad \forall v \in [H^1(\Omega)]^d,$$

(14)

$$\|v\|_{H^1(\Omega)} \leq C \left( \|\epsilon(v)\|_{L^2(\Omega)} + \int_{\Omega} \nabla \times v \, dx \right) \quad \forall v \in [H^1(\Omega)]^d$$

(15)

Example 3. Let $\Omega$ be a domain in $\mathbb{R}^2$ and the bilinear form $a(\cdot, \cdot)$ be defined by

$$a(v_1, v_2) = \int_{\Omega} \left[ \Delta v_1 \Delta v_2 + (1 - \sigma) \left( 2 \frac{\partial^2 v_1}{\partial x_1 \partial x_2} \frac{\partial^2 v_2}{\partial x_1 \partial x_2} - \frac{\partial^2 v_1}{\partial x_1^2} \frac{\partial^2 v_2}{\partial x_2^2} - \frac{\partial^2 v_1}{\partial x_1^2} \frac{\partial^2 v_2}{\partial x_1^2} \right) \right] \, dx$$

(16)

where $\sigma \in (0, 1/2)$ is the Poisson ratio.

For $f \in L^2(\Omega)$, the weak form of the clamped plate bending problem (Ciarlet, 1997)

$$\Delta^2 u = f \quad \text{on } \Omega, \quad u = \frac{\partial u}{\partial n} = 0 \quad \text{on } \partial\Omega$$

(17)

is given by (1), where $V = \{v \in H^2(\Omega) : v = \frac{\partial v}{\partial n} = 0 \text{ on } \partial\Omega\} = H^2(\Omega)$ and $F$ is defined by (6). For the simply supported plate bending problem, the function space $V$ is $\{v \in H^2(\Omega) : v = 0 \text{ on } \partial\Omega\} = H^2(\Omega) \cap H^1_0(\Omega)$.

For these problems, the coercivity of $a(\cdot, \cdot)$ is a consequence of the following Poincaré–Friedrichs inequality (Nečas, 1967):

$$\|v\|_{H^1(\Omega)} \leq C |v|_{H^2(\Omega)} \quad \forall v \in H^2(\Omega) \cap H^1_0(\Omega)$$

(18)

Remark 3. The weak formulation of boundary value problems for beams and shells can be found in Chapter 3, Volume 2.

2.2 Ritz–Galerkin methods

In the Ritz–Galerkin approach for (1), a discrete problem is formulated as follows. Find $\tilde{u} \in \tilde{V}$ such that

$$a(\tilde{u}, v) = F(\tilde{v}) \quad \forall \tilde{v} \in \tilde{V}$$

(19)

where $\tilde{V}$, the space of trial/test functions, is a finite-dimensional subspace of $V$.

The orthogonality relation

$$a(u - \tilde{u}, \tilde{v}) = 0 \quad \forall \tilde{v} \in \tilde{V}$$

(20)

follows by subtracting (19) from (1), and hence

$$\|u - \tilde{u}\|_a = \inf \limits_{\tilde{v} \in \tilde{V}} \|u - \tilde{v}\|_a$$

(21)

where $\| \cdot \|_a = (a(\cdot, \cdot))^{1/2}$. Furthermore, (2), (3), and (21) imply that

$$\|u - \tilde{u}\|_{H^m(\Omega)} \leq \left( \frac{C_1}{C_2} \right)^{1/2} \inf \limits_{\tilde{v} \in \tilde{V}} \|u - \tilde{v}\|_{H^m(\Omega)}$$

(22)

that is, the error for the approximate solution $\tilde{u}$ is quasi-optimal in the norm of the Sobolev space underlying the weak problem.

The abstract estimate (22), called Cea’s lemma, reduces the error estimate for the Ritz–Galerkin method to a problem in approximation theory, namely, to the determination of the magnitude of the error of the best approximation of
$u$ by a member of $\tilde{V}$. The solution of this problem depends on the regularity (smoothness) of $u$ and the nature of the space $\tilde{V}$.

One can also measure $u - \tilde{u}$ in other norms. For example, an estimate of $\|u - \tilde{u}\|_{L^2(\Omega)}$ can be obtained by the Aubin–Nitsche duality technique as follows. Let $w \in V$ be the solution of the weak problem

$$a(v, w) = \int_{\Omega} (u - \tilde{u}) v \, dx \quad \forall \, v \in V$$  \hfill (23)

Then we have, from (20), (23), and the Cauchy–Schwarz inequality,

$$\|u - \tilde{u}\|_{L^2(\Omega)}^2 = a(u - \tilde{u}, w) = a(u - \tilde{u}, w - \tilde{v}) \leq C_2 \|u - \tilde{u}\|_{H^m(\Omega)} \|w - \tilde{v}\|_{H^m(\Omega)} \quad \forall \, \tilde{v} \in \tilde{V}$$

which implies that

$$\|u - \tilde{u}\|_{L^2(\Omega)} \leq C_2 \left( \inf_{v \in V} \|w - \tilde{v}\|_{H^m(\Omega)} \right) \|u - \tilde{u}\|_{H^m(\Omega)}$$  \hfill (24)

In general, since $w$ can be approximated by members of $\tilde{V}$ to high accuracy, the term inside the bracket on the right-hand side of (24) is small, which shows that the $L_2$ error is much smaller than the $H^m$ error.

The estimates (22) and (24) provide the basic a priori error estimates for the Ritz–Galerkin method in an abstract setting.

On the other hand, the error of the Ritz–Galerkin method can also be estimated in an a posteriori fashion. Let the computable linear functional (the residual of the approximate solution $\tilde{u}$) $R: V \to \mathbb{R}$ be defined by

$$R(v) = a(u - \tilde{u}, v) = F(v) - a(\tilde{u}, v)$$  \hfill (25)

The global a posteriori error estimate

$$\|u - \tilde{u}\|_{H^m(\Omega)} \leq \frac{1}{C_2} \sup_{v \in V} \|R(v)\|_{H^m(\Omega)}$$  \hfill (26)

then follows from (3) and (25).

Let $D$ be a subdomain of $\Omega$ and $H^m_0(D)$ be the subspace of $V$ whose members vanish identically outside $D$. It follows from (25) and the local version of (2) that we also have a local a posteriori error estimate:

$$\|u - \tilde{u}\|_{H^m(D)} \geq \frac{1}{C_1} \sup_{v \in H^m_0(D)} \|R(v)\|_{H^m(D)}$$  \hfill (27)

The equivalence of the error norm with the dual norm of the residual will be the point of departure in Section 5.1.2 (cf. (70)).

### 2.3 Elliptic regularity

As mentioned above, the magnitude of the error of a Ritz–Galerkin method for an elliptic boundary value problem depends on the regularity of the solution. Here we give a brief description of elliptic regularity for the examples in Section 2.1.

If the boundary $\partial \Omega$ is smooth and the homogeneous boundary conditions are also smooth (i.e. the Dirichlet and Neumann boundary condition in (5) and the displacement and traction boundary conditions in (12) are defined on disjoint components of $\partial \Omega$), then the solution of the elliptic boundary value problems in Section 2.1 obey the classical Shift Theorem (Agmon, 1965; Nečas, 1967; Gilbarg and Trudinger, 1983; Wloka, 1987). In other words, if the right-hand side of the equation belongs to the Sobolev space $H^4(\Omega)$, then the solution of a $2m$-th order elliptic boundary problem belongs to the Sobolev space $H^{2m+\epsilon}(\Omega)$.

The Shift Theorem does not hold for domains with piecewise smooth boundary in general. For example, let $\Omega$ be the $L$-shaped domain depicted in Figure 1 and

$$u(x) = \phi(r) r^{2/3} \sin \left( \frac{2}{3} \left( \theta - \frac{\pi}{2} \right) \right)$$  \hfill (28)

where $r = (x_1^2 + x_2^2)^{1/2}$ and $\theta = \arctan(x_2/x_1)$ are the polar coordinates and $\phi$ is a smooth cut-off function that equals 1 for $0 \leq r < 1/2$ and 0 for $r > 3/4$. It is easy to check that $u \in H^1_0(\Omega)$ and $-\Delta u \in C^\infty(\bar{\Omega})$. Let $D$ be an open neighborhood of the origin in $\Omega$. Then $u \in H^2(\Omega \setminus \bar{D})$ but $u \notin H^2(D)$. In fact $u$ belongs to the Besov space $B^{2,1}(\bar{\Omega})$ (Babuška and Osborn, 1991), which implies that $u \in H^{2/3-\epsilon}(\Omega)$ for any $\epsilon > 0$, but $u \notin H^{2/3}(\Omega)$ (see Triebel (1978) and Grisvard (1985) for a discussion of Besov spaces and fractional order Sobolev spaces). A similar situation occurs when the types of boundary condition change abruptly, such as the Poisson problem with mixed boundary conditions depicted on the circular domain in Figure 1, where the homogeneous Dirichlet boundary condition is assumed on the upper semicircle and the homogeneous Neumann boundary condition is assumed on the lower semicircle.

Therefore (Dauge, 1988), for the second (respectively fourth) order model problems in Section 2.1, the solution in general only belongs to $H^{1+\alpha}(\Omega)$ (respectively $H^{2+\alpha}(\Omega)$) for some $\alpha \in (0, 1]$ even if the right-hand side of the equation belongs to $C^\infty(\bar{\Omega})$. 


For two-dimensional problems, the vertices of $\Omega$ and the points where the boundary condition changes type are the singular points (cf. Figure 1). Away from these singular points, the Shift Theorem is valid. The behavior of the solution near the singular points is also well understood. If the right-hand side function and its derivatives vanish to sufficiently high order at the singular points, then the Shift Theorem holds for certain weighted Sobolev spaces (Nazarov and Plamenevsky, 1994; Kozlov, Maz’ya and Rossman, 1997, 2001). Alternatively, one can represent the solution near a singular point as a sum of a regular part and a singular part (Grisvard, 1985; Dauge, 1988; Nicaise, 1993). For a $2m$-th order problem, the regular part of the solution belongs to the Sobolev space $H^{2m+k}(\Omega)$ if the right-hand side function belongs to $H^l(\Omega)$, and the singular part of the solution is a linear combination of special functions with less regularity, analogous to the function in (28).

The situation in three dimensions is more complicated due to the presence of edge singularities, vertex singularities, and edge-vertex singularities. The theory of three-dimensional singularities remains an active area of research.

### 3 FINITE ELEMENT SPACES

Finite element methods are Ritz–Galerkin methods where the finite-dimensional trial/test function spaces are constructed by piecing together polynomial functions defined on (small) parts of the domain $\Omega$. In this section, we describe the construction and properties of finite element spaces. We will concentrate on conforming finite elements here and leave the discussion of nonconforming finite elements to Section 7.2.

#### 3.1 The concept of a finite element

A $d$-dimensional finite element (Ciarlet, 1978; Brenner and Scott, 2002) is a triple $(K, P_K, N_K)$, where $K$ is a closed bounded subset of $\mathbb{R}^d$ with nonempty interior and a piecewise smooth boundary, $P_K$ is a finite-dimensional vector space of functions defined on $K$ and $N_K$ is a basis of the dual space $P_K^*$. The function space $P_K$ is the space of the shape functions and the elements of $N_K$ are the nodal variables (degrees of freedom).

The following are examples of two-dimensional finite elements.

**Example 4 (Triangular Lagrange Elements)** Let $K$ be a triangle, $P_K$ be the space $P_n$ of polynomials in two variables of degree $\leq n$, and let the set $N_K$ consist of evaluations of shape functions at the nodes with barycentric coordinates $\lambda_1 = i/n$, $\lambda_2 = j/n$ and $\lambda_3 = k/n$, where $i, j, k$ are nonnegative integers and $i + j + k = n$. Then $(K, P_K, N_K)$ is the two-dimensional $P_n$ Lagrange finite element. The nodal variables for the $P_1$, $P_2$, and $P_3$ Lagrange elements are depicted in Figure 2, where $\bullet$ (here and in the following examples) represents pointwise evaluation of shape functions.

**Example 5 (Triangular Hermite Elements)** Let $K$ be a triangle. The cubic Hermite element is the triple $(K, P_3, N_K)$ where $N_K$ consists of evaluations of shape functions and their gradients at the vertices and evaluation of shape functions at the center of $K$. The nodal variables for the cubic Hermite element are depicted in the first figure in Figure 3, where $\circ$ (here and in the following examples) represents pointwise evaluation of gradients of shape functions.

By removing the nodal variable at the center (cf. the second figure in Figure 3) and reducing the space of shape functions to

$$\begin{align*}
\left\{ v \in P_3; 6v(c) - 2 \sum_{i=1}^3 v(p_i) + \sum_{i=1}^3 (\nabla v)(p_i) \cdot (p_i - c) = 0 \right\} \uparrow P_2
\end{align*}$$

where $p_i$ ($i = 1, 2, 3$) and $c$ are the vertices and center of $K$ respectively, we obtain the Zienkiewicz element.

The fifth degree Argyris element is the triple $(K, P_5, N_K)$ where $N_K$ consists of evaluations of the shape functions and their derivatives up to order two at the vertices and evaluations of the normal derivatives at the midpoints of the edges. The nodal variables for the Argyris element are depicted in the third figure in Figure 3, where $\circ$ and $\uparrow$ (here and
in the following examples) represent pointwise evaluation of second order derivatives and the normal derivative of the shape functions, respectively.

By removing the nodal variables at the midpoints of the edges (cf. the fourth figure in Figure 3) and reducing the space of shape functions to \( \{ v \in P_3; (\partial v/\partial n)|_e \in P_3(e) \text{ for each edge } e \} \), we obtain the Bell element.

**Example 6 (Triangular Macro Elements)** Let \( K \) be a triangle that is subdivided into three subtriangles by the center of \( K \), \( P_K \) be the space of piecewise cubic polynomials with respect to this subdivision that belong to \( C^1(K) \), and let the set \( N_K \) consist of evaluations of the shape functions and their first-order derivatives at the vertices of \( K \) and evaluations of the normal derivatives of the shape functions at the midpoints of the edges of \( K \). Then \((K, P_K, N_K)\) is the Hsieh–Clough–Tocher macro element. The nodal variables for this element are depicted in the first figure in Figure 4.

By removing the nodal variables at the midpoints of the edges (cf. the second figure in Figure 4) and reducing the space of shape functions to \( \{ v \in C^1(K); v \) is piecewise cubic and \((\partial v/\partial n)|_e \in P_1(e) \text{ for each edge } e \}\), we obtain the reduced Hsieh–Clough–Tocher macro element.

**Example 7 (Rectangular Tensor Product Elements)** Let \( K \) be the rectangle \([a_1, b_1] \times [a_2, b_2]\), \( P_K \) be the space spanned by the monomials \( x_i^j x_j^l \) for \( 0 \leq i, j \leq n \), and the set \( N_K \) consist of evaluations of shape functions at the nodes with coordinates \((a_1 + i(b_1 - a_1)/n, a_2 + j(b_2 - a_2)/n)\) for \( 0 \leq i, j \leq n \). Then \((K, P_K, N_K)\) is the two-dimensional \( Q_n \) tensor product element. The nodal variables of the \( Q_1, Q_2 \) and \( Q_3 \) elements are depicted in Figure 5.

**Example 8 (Quadrilateral \( Q_n \) Elements)** Let \( K \) be a convex quadrilateral; then there exists a bilinear map \((x_1, x_2) \mapsto B(x_1, x_2) = (a_1 + b_1 x_1 + c_1 x_2 + d_1 x_1 x_2, a_2 + b_2 x_1 + c_2 x_2 + d_2 x_1 x_2)\) from the unit square \( S \) with vertices \((\pm 1, \pm 1)\) onto \( K \). The space of shape functions is defined by \( v \in P_K \) if and only if \( v \circ B \in Q_n \) and \( N_K \) consists of pointwise evaluations of the shape functions at the nodes of \( K \) corresponding under the map \( B \) to the nodes of the \( Q_n \) tensor product element on \( S \). The nodal variables of the \( Q_1, Q_2 \) and \( Q_3 \) quadrilateral elements are depicted in Figure 6.

**Example 9 (Other Rectangular Elements)** Let \( K \) be the rectangle \([a_1, b_1] \times [c_1, d_1]\):

\[
P_K = \left\{ v \in Q_4; 4v(e) + \sum_{i=1}^{4} v(p_i) - 2 \sum_{i=1}^{4} v(m_i) = 0 \right\} \text{ (⊃ } P_2)\]

where the \( p_i \)'s are the vertices of \( K \), the \( m_i \)'s are the midpoints of the edges of \( K \) and \( c \) is the center of \( K \); and \( N_K \) consist of evaluations of the shape functions at the vertices and the midpoints (cf. the first figure in Figure 7). Then \((K, P_K, N_K)\) is the 8-node serendipity element.

If we take \( P_K \) to be the space of bicubic polynomials spanned by \( x_i^j x_j^l \) for \( 0 \leq i, j \leq 3 \) and \( N_K \) to be the set consisting of evaluations at the vertices of \( K \) of the shape functions, their first-order derivatives and their second-order...
mixed derivatives, then we have the Bogner–Fox–Schmit element. The nodal variables for this element are depicted in the second figure in Figure 7, where the tilted arrows represent pointwise evaluations of the second-order mixed derivatives of the shape functions.

**Remark 4.** The triangular \( P_n \) elements and the quadrilateral \( Q_n \) elements, which are suitable for second order elliptic boundary value problems, can be generalized to any dimension in a straightforward manner. The Argyris element, the Bell element, the macro elements, and the Bogner–Fox–Schmit element are suitable for fourth-order problems in two space dimensions.

### 3.2 Triangulations and finite element spaces

We restrict \( \Omega \subset \mathbb{R}^d \) \((d = 1, 2, 3)\) to be a polyhedral domain in this and the following sections. The case of curved domains will be discussed in Section 7.4.

A partition of \( \Omega \) is a collection \( \mathcal{P} \) of polyhedral subdomains of \( \Omega \) such that

\[
\overline{\Omega} = \bigcup_{D \in \mathcal{P}} \overline{D} \quad \text{and} \quad D \cap D' = \emptyset \quad \text{if} \quad D, D' \in \mathcal{P}, \; D \neq D'
\]

where we use \( \overline{\Omega} \) and \( \overline{D} \) to represent the closures of \( \Omega \) and \( D \).

A triangulation of \( \Omega \) is a partition where the intersection of the closures of two distinct subdomains is either empty, a common vertex, a common edge or a common face. For \( d = 1 \), every partition is a triangulation. But the two concepts are different when \( d \geq 2 \). A partition that is not a triangulation is depicted in the first figure in Figure 8, where the other three figures represent triangulations. Below we will concentrate on triangulations consisting of triangles or convex quadrilaterals in two dimensions and tetrahedrons or convex hexahedrons in three dimensions.

The shape regularity of a triangle (or tetrahedron) \( D \) can be measured by the parameter

\[
\gamma(D) = \frac{\text{diam } D}{\text{diameter of the largest ball in } \overline{D}} \quad (29)
\]

which will be referred to as the aspect ratio of the triangle (tetrahedron). We say that a family of triangulations of triangles (or tetrahedrons) \( \{T_i; i \in I\} \) is regular (or nondegenerate) if the aspect ratios of all the triangles (tetrahedrons) in the triangulations are bounded, that is, there exists a positive constant \( C \) such that

\[
\gamma(D) \leq C \quad \text{for all} \quad D \in T_i \quad \text{and} \quad i \in I
\]

The shape regularity of a convex quadrilateral (or hexahedron) \( D \) can be measured by the parameter \( \gamma(D) \) defined in (29) and the parameter

\[
\sigma(D) = \max \left\{ \frac{|e_1|}{|e_2|}; e_1 \text{ and } e_2 \text{ are any two edges of } D \right\}
\]

We will refer to the number \( \max(\gamma(D), \sigma(D)) \) as the aspect ratio of the convex quadrilateral (hexahedron). We say that a family of triangulations of convex quadrilaterals (or hexahedrons) \( \{T_i; i \in I\} \) is regular if the aspect ratios of all the quadrilaterals in the triangulations are bounded, that is, there exists a positive constant \( C \) such that

\[
\gamma(D), \sigma(D) \leq C \quad \text{for all} \quad D \in T_i \quad \text{and} \quad i \in I
\]

A family of triangulations is quasi-uniform if it is regular and there exists a positive constant \( C \) such that

\[
h_i \leq C \text{ diam } D \quad \forall D \in T_i, \quad i \in I
\]

where \( h_i \) is the maximum of the diameters of the subdomains in \( T_i \).

**Remark 5.** For a triangle or a tetrahedron \( D \), a lower bound for the angles of \( D \) can lead to an upper bound for \( \gamma(D) \) (and vice versa). Therefore, the regularity of a family of simplicial triangulations (i.e. triangulations consisting of triangles or tetrahedrons) is equivalent to the following minimum angle condition: There exists \( \theta_s > 0 \) such that the angles of the simplex in all the triangulations \( T_i \) are bounded below by \( \theta_s \).

**Remark 6.** A family of triangulations obtained by successive uniform subdivisions of an initial triangulation is quasi-uniform. A family of triangulations generated by a local refinement strategy is usually regular but not quasi-uniform.

Let \( T \) be a triangulation of \( \Omega \), and a finite element \((\overline{D}, \mathcal{P}_i, \mathcal{N}_i)\) be associated with each subdomain \( D \in T \). We define the corresponding finite element space to be

\[
FE_T = \{ v \in L_2(\Omega); v|_D = v|_{\overline{D}} \in \mathcal{P}_i \quad \forall D \in T, \quad \text{and} \quad v|_{\overline{D}}, v|_{\overline{D}} \text{ share the same nodal values on } \overline{D} \cap \overline{D} \}
\]

\[
\gamma(D) \leq C \quad \text{for all} \quad D \in T_i \quad \text{and} \quad i \in I
\]
We say that $FE_T$ is a $C^r$ finite element space if $FE_T \subset C^r(\Omega)$. For example, the finite element spaces constructed from the Lagrange finite elements (Example 4), the tensor product elements (Example 7), the cubic Hermite element (Example 4) and the serendipity element (Example 9) are $C^0$ finite element spaces, and those constructed from the quintic Argyris element (Example 5), the Bell element (Example 5), the macro elements (Example 6) and the Bogner–Fox–Schmit element (Example 9) are $C^1$ finite element spaces.

Note that a $C^r$ finite element space is automatically a subspace of the Sobolev space $H^{r+1}(\Omega)$ and therefore appropriate for elliptic boundary value problems of order $2(r + 1)$.

### 3.3 Element nodal interpolation operators and interpolation error estimates

Let $(K, \mathcal{P}_K, N_K)$ be a finite element. Denote the nodal variables in $N_K$ by $N_1, \ldots, N_n$ ($n = \dim \mathcal{P}_K$) and the dual basis of $\mathcal{P}_K$ by $\phi_1, \ldots, \phi_n$, that is,

$$N_j(\phi_i) = \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Assume that $\xi \mapsto N_j(\xi)$ is well-defined for $\xi \in H^s(K)$ (where $s$ is a sufficiently large positive number), then we can define the element nodal interpolation operator $\Pi_K: H^s(K) \to \mathcal{P}_K$ by

$$\Pi_K \xi = \sum_{j=1}^n N_j(\xi) \phi_j \quad (33)$$

Note that (33) implies

$$\Pi_K v = v \quad \forall v \in \mathcal{P}_K \quad (34)$$

For example, by the Sobolev embedding theorem (Adams, 1995; Nečas, 1967; Wloka, 1987; Gilbarg and Trudinger, 1983), the nodal interpolation operators associated with the Lagrange finite elements (Example 4), the tensor product finite elements (Example 7), and the serendipity element (Example 9) are well-defined on $H^s(K)$ for $s > 1$ if $K \subset \mathbb{R}^2$ and for $s > 3/2$ if $K \subset \mathbb{R}^3$. On the other hand the nodal interpolation operators associated with the Zienkiewicz element (Example 5) and the macro elements (Example 6) are well-defined on $H^s(K)$ for $s > 2$, while the interpolation operators for the quintic Argyris element (Example 5), the Bell element (Example 5) or the Bogner–Fox–Schmit element (Example 9) are well-defined on $H^s(K)$ for $s > 3$.

The error of the element nodal interpolation operator for a triangular (tetrahedral) or convex quadrilateral (hexagonal) element $(K, \mathcal{P}_K, N_K)$ can be controlled in terms of the shape regularity of $K$. Let $\hat{K}$ be the image of $K$ under the scaling map

$$x \mapsto \mathcal{H}(x) = (\text{diam } K)^{-1}x \quad (35)$$

Then $\hat{K}$ is a domain of unit diameter and we can define a finite element $(\hat{K}, \mathcal{P}_{\hat{K}}, N_{\hat{K}})$ as follows: (i) $\hat{v} \in \mathcal{P}_{\hat{K}}$ if and only if $\hat{v} \circ \mathcal{H} \in \mathcal{P}_K$, and (ii) $N \in N_{\hat{K}}$ if and only if the linear functional $v \mapsto N(v \circ \mathcal{H}^{-1})$ on $\mathcal{P}_K$ belongs to $N_K$.

It follows that the dual basis $\hat{\phi}_1, \ldots, \hat{\phi}_n$ of $\mathcal{P}_{\hat{K}}$ is related to the dual basis $\phi_1, \ldots, \phi_n$ of $\mathcal{P}_K$ through the relation $\hat{\phi}_i \circ \mathcal{H} = \phi_i$, and (33) implies that

$$\left(\Pi_K \xi \right) \circ \mathcal{H}^{-1} = \Pi_{\hat{K}} \hat{\xi} \circ \mathcal{H}^{-1} \quad (36)$$

for all sufficiently smooth functions $\xi$ defined on $K$. Moreover, for the dual bases $\hat{\xi}$ and $\xi$ related by $\hat{\xi}(x) = \xi(\mathcal{H}(x))$, we have

$$\| \hat{\xi} \|^2_{H^s(\hat{K})} = (\text{diam } K)^{2s-2d} \| \xi \|^2_{H^s(K)} \quad (37)$$

where $d$ is the spatial dimension.

Assuming that $\mathcal{P}_{\hat{K}} \supset \mathcal{P}_m$ (equivalently $\mathcal{P}_K \supset \mathcal{P}_m$), we have, by (34),

$$\| \hat{\xi} - \Pi_{\hat{K}} \hat{\xi} \|^2_{H^s(\hat{K})} = \| \xi - \Pi_K \xi \|^2_{H^s(K)} \leq 2 \| \Pi_K \|_{m,s} \| \xi - p \|^2_{H^s(K)} \quad \forall p \in \mathcal{P}_m \quad (38)$$

where $\| \Pi_K \|_{m,s}$ is the norm of the operator $\Pi_K: H^s(K) \to H^m(\hat{K})$, and hence

$$\| \hat{\xi} - \Pi_{\hat{K}} \hat{\xi} \|^2_{H^s(\hat{K})} \leq 2 \| \Pi_K \|_{m,s} \inf_{p \in \mathcal{P}_m} \| \hat{\xi} - p \|^2_{H^s(\hat{K})} \quad (39)$$

Since $K$ is convex, the following estimate (Verfürth, 1999) holds provided $m$ is the largest integer strictly less than $s$:

$$\inf_{p \in \mathcal{P}_m} \| \hat{\xi} - p \|^2_{H^s(\hat{K})} \leq C_{s,d} \| \hat{\xi} \|^2_{H^s(\hat{K})} \quad (40)$$

where the positive constant $C_{s,d}$ depends only on $s$ and $d$.

Combining (38) and (39) we find

$$\| \hat{\xi} - \Pi_{\hat{K}} \hat{\xi} \|^2_{H^s(\hat{K})} \leq 2 C_{s,d} \| \Pi_K \|_{m,s} \| \hat{\xi} \|^2_{H^s(\hat{K})} \quad \forall \hat{\xi} \in H^s(\hat{K}) \quad (40)$$

We have therefore reduced the error estimate for the element nodal interpolation operator to an estimate of $\| \Pi_K \|_{m,s}$. Since diam $\hat{K} = 1$, the norm $\| \Pi_K \|_{m,s}$ is a constant depending only on the shape of $K$ (equivalently of $K$), if we considered $s$ and $m$ to be fixed for a given type of element.

For triangular elements, we can use the concept of affine-interpolation-equivalent elements to obtain a more concrete
description of the dependence of \( \| \Pi^{\hat{K}} \|_{m,s} \) on the shape of \( \hat{K} \). A \( d \)-dimensional nondegenerate affine map is a map of the form \( x \mapsto Ax + b \) where \( A \) is a nonsingular \( d \times d \) matrix and \( b \in \mathbb{R}^d \). We say that two finite elements \( (K_1, P_{K_1}, N_{K_1}) \) and \( (K_2, P_{K_2}, N_{K_2}) \) are affine-equivalent if (i) there exists a nondegenerate affine map \( \Phi \) that maps \( K_1 \) onto \( K_2 \), (ii) \( v \in P_{K_1} \) if and only if \( v \circ \Phi \in P_{K_2} \) and (iii) \( (\Pi_{K_1} \xi) \circ \Phi = \Pi_{K_2} (\xi \circ \Phi) \) (41)

for all sufficiently smooth functions \( \xi \) defined on \( K_2 \). For example, any triangular elements in one of the families (except the Bell element and the reduced Hsieh–Clough–Tocher element) described in Section 3.1 are affine-interpolation-equivalent to the corresponding element on the standard simplex \( \hat{S} \) with vertices \((0, 0), (1, 0) \) and \((0, 1) \).

Assuming \( (\hat{K}, P_{\hat{K}}, N_{\hat{K}}) \) (or equivalently \( (K, P_K, N_K) \)) is affine-interpolation-equivalent to the element \( (S, P_S, N_S) \) on the standard simplex, it follows from (41) and the chain rule that

\[
\| \Pi^K \|_{m,s} \leq C \| \Pi_S \|_{m,s}
\] (42)

where the positive constant depends only on the Jacobian matrix of the affine map \( \hat{\Phi}: S \to \hat{K} \) and thus depends only on an upper bound of the parameter \( \gamma(\hat{K}) \) (cf. (29)) which is identical with \( \gamma(K) \).

Combining (36), (37), (40) and (42), we find

\[
\sum_{k=0}^{m} (\text{diam } K)^k |\xi - \Pi^K \xi|_{H^s(K)} \leq C (\text{diam } K)^s |\xi|_{H^s(K)}
\]

\( \forall \xi \in H^s(K) \) (43)

where the positive constant \( C \) depends only on \( s \) and an upper bound of the parameter \( \gamma(K) \) (the aspect ratio of \( K \)), provided that (i) the element nodal interpolation operator is well-defined on \( H^s(K) \), (ii) the triangular element \( (K, P_K, N_K) \) is affine-interpolation-equivalent to a reference element \( (S, P_S, N_S) \) on the standard simplex, (iii) \( P \supseteq P_m \), and (iv) \( m \) is the largest integer \( < s \).

For convex quadrilateral elements, we can similarly obtain a concrete description of the dependence of \( \| \Pi^K \|_{m,s} \) on the shape of \( K \) by assuming that there is a reference element \( (S, P_S, N_S) \) defined on the biunit square \( S \) with vertices \((\pm 1, \pm 1) \) and a bilinear homeomorphism \( \hat{\Phi} \) from \( S \) onto \( \hat{K} \) with the following properties: \( v \in P_K \) if and only if \( v \circ \Phi \in P_S \) and \( (\Pi^K \hat{\xi}) \circ \Phi = \Pi_S (\hat{\xi} \circ \Phi) \) for all sufficiently smooth functions \( \hat{\xi} \) defined on \( \hat{K} \). Note that because of (36) this is equivalent to the existence of a bilinear homeomorphism from \( S \) onto \( K \) such that

\[
v \in P_K \iff v \circ \Phi \in P_S \quad \text{and} \quad (\Pi_K \xi) \circ \Phi = \Pi_S (\xi \circ \Phi)
\] (44)

for all sufficiently smooth functions \( \xi \) defined on \( K \). The estimate (42) holds again by the chain rule, where the positive constant \( C \) depends only on the Jacobian matrix of \( \Phi \) and thus depends only on upper bounds for the parameters \( \gamma(K) \) and \( \sigma(K) \) (cf. (30)), which are identical with \( \gamma(\hat{K}) \) and \( \sigma(\hat{K}) \). We conclude that the estimate (43) also holds for convex quadrilateral elements where the positive constant \( C \) depends on upper bounds of \( \gamma(K) \) and \( \sigma(K) \) (equivalently an upper bound of the aspect ratio of \( K \)) provided condition (ii) is replaced by (44). For example, the estimate (43) is valid for the quadrilateral \( Q_n \) element in Example 8.

**Remark 7.** The general estimate (40) can be refined to yield anisotropic error estimates for certain reference elements. For example, in two dimensions, the following estimates (Apel and Dobrowolski, 1992; Apel, 1999) hold for the \( P_n \) Lagrange elements on the reference simplex \( S \) and the \( Q_n \) tensor product elements on the reference square \( S \):

\[
\left\| \frac{\partial}{\partial x_j} (\xi - \Pi_S \xi) \right\|_{L^2(S)} \leq C \left( \left\| \frac{\partial^2 \xi}{\partial x_i \partial x_j} \right\|_{L^2(S)} + \left\| \frac{\partial^2 \xi}{\partial x_i \partial x_j} \right\|_{L^2(S)} \right)
\] (45)

for \( j = 1, 2 \) and for all \( \xi \in H^2(S) \). We refer the readers to Chapter 3, this Volume, for more details.

**Remark 8.** The analysis of the quadrilateral serendipity elements is more subtle. A detailed discussion can be found in Arnold, Boffi and Falk (2002).

**Remark 9.** The estimate (43) can be generalized naturally to 3-D tetrahedral \( P_n \) elements and hexahedral \( Q_n \) elements.

**Remark 10.** Let \( n \) be a nonnegative integer and \( n < s \leq n + 1 \). The estimate

\[
\inf_{p \in P_s(\Omega)} \| \xi - p \|_{H^s(\Omega)} \leq C_{\Omega,s} |\xi|_{H^s(\Omega)}
\]

\( \forall \xi \in H^s(\Omega) \) (46)

for general \( \Omega \) follows from generalized Poincaré–Friedrichs inequalities (Nečas, 1967). In the case where \( \Omega \) is convex, the constant \( C_{\Omega,s} \) depends only on \( s \) and the dimension of \( \Omega \), but not on the shape of \( \Omega \), as indicated by the estimate (39). For nonconvex domains, the constant \( C_{\Omega,s} \)
does depend on the shape of $\Omega$ (Dupont and Scott, 1980, Verf"{u}rth, 1999).

Let $F$ be a bounded linear functional on $H^s(\Omega)$ with norm $\|F\|$ such that $F(p) = 0$ for all $p \in P_n(\Omega)$. It follows from (46) that

$$|F(\zeta)| \leq \inf_{p \in P_n(\Omega)} |F(\zeta - p)| \leq \|F\| \inf_{p \in P_n(\Omega)} \|\zeta - p\|_{H^s(\Omega)}$$

$$\leq (C_{\Omega, s} \|F\|) \|\zeta\|_{H^s(\Omega)}$$

(47)

for all $\zeta \in H^s(\Omega)$. The estimate (47), known as the Bramble–Hilbert lemma (Bramble and Hilbert, 1970), is useful for deriving various error estimates.

### 3.4 Some discrete estimates

The finite element spaces in Section 3.2 are designed to be subspaces of Sobolev spaces so that they can serve as the trial/test spaces for Ritz–Galerkin methods. On the other hand, since finite element spaces are constructed by piecing together finite-dimensional function spaces, there are discrete estimates valid on the finite element spaces but not the Sobolev spaces.

Let $(K, P_K, N_K)$ be a finite element such that $P_K \subset H^k(K)$ for a nonnegative integer $k$. Since any seminorm on a finite-dimensional space is continuous with respect to a norm, we have, by scaling, the following inverse estimate:

$$|v|_{H^s(K)} \leq C(\text{diam } K)^{k-s} \|v\|_{H^k(K)} \quad \forall v \in P_K, \quad 0 \leq k \leq s$$

(48)

where the positive constant $C$ depends on the domain $K$ (the image of $K$ under the scaling map $\mathcal{H}$ defined by (35)) and the space $P_K$.

For finite elements whose shape functions can be pulled back to a fixed finite-dimensional function space on a reference element, the constant $C$ depends only on the shape regularity of the element domain $K$ and global versions of (48) can be easily derived. For example, for a quasi-uniform family $\{\mathcal{T}_i : i \in I\}$ of simplicial or quadrilateral triangulations of a polygonal domain $\Omega$, we have

$$|v|_{H^s(\Omega)} \leq Ch_i^{-1} \|v\|_{L_2(\Omega)} \quad \forall v \in V_i \quad \text{and} \quad i \in I$$

(49)

where $V_i \subset H^1(\Omega)$ is either the $P_n$ triangular finite element space or the $Q_n$ quadrilateral finite element space associated with $\mathcal{T}_i$. Note that $V_i \subset H^s(\Omega)$ for any $s < 3/2$ and a bit more work shows that the following inverse estimate (Ben Belgacem and Brenner, 2001) also holds:

$$|v|_{H^s(\Omega)} \leq C_i h_i^{-s} \|v\|_{H^s(\Omega)} \quad \forall v \in V_i, \quad i \in I \quad \text{and} \quad 1 \leq s < 3/2$$

(50)

where the positive constant $C_i$ can be uniformly bounded for $s$ in a compact subset of $[1, 3/2)$.

It is well-known that in two dimensions the Sobolev space $H^1(\Omega)$ is not a subspace of $C(\Omega)$. However, the $P_n$ triangular finite element space and the $Q_n$ quadrilateral finite element space do belong to $C(\Omega)$ and it is possible to bound the $L_\infty$ norm of the finite element function by its $H^1$ norm. Indeed, it follows from Fourier transform and extension theorems (Adams, 1995, Wloka, 1987) that, for $\epsilon > 0$,

$$\|v\|_{L_\infty(\Omega)} \leq C\epsilon^{-1/2} \|v\|_{H^{1+\epsilon}(\Omega)} \quad \forall v \in H^{1+\epsilon}(\Omega)$$

(51)

By taking $\epsilon = (1 + \ln h_i)^{-1}$ in (51) and applying (50), we arrive at the following discrete Sobolev inequality:

$$\|v\|_{L_\infty(\Omega)} \leq C(1 + \ln h_i)^{1/2} \|v\|_{H^{1}(\Omega)} \quad \forall v \in V_i \quad \text{and} \quad i \in I$$

(52)

where the positive constant $C$ is independent of $i \in I$.

The discrete Sobolev inequality and the Poincaré–Friedrichs inequality (8) imply immediately the following discrete Poincaré inequality:

$$\|v\|_{L_\infty(\Omega)} \leq \|v - \bar{v}\|_{L_\infty(\Omega)} + \|\bar{v}\|_{L_\infty(\Omega)} \leq 2\|v - \bar{v}\|_{L_\infty(\Omega)}$$

$$\leq C(1 + \ln h_i)^{1/2} \|v\|_{H^{1}(\Omega)}$$

$$\leq C(1 + \ln h_i)^{1/2} \|v\|_{H^{1}(\Omega)}$$

(53)

for all $v \in V_i$ that vanishes at a given point in $\Omega$ and with mean $\bar{v} = \int_\Omega v \, dx / |\Omega|$.

Remark 11. The discrete Sobolev inequality can also be established directly using calculus and inverse estimates (Bramble, Pasciak and Schatz, 1986; Brenner and Scott, 2002), and both (52) and (53) are sharp (Brenner and Sung, 2000).

### 4 A PRIORI ERROR ESTIMATES FOR FINITE ELEMENT METHODS

Let $T$ be a triangulation of $\Omega$ and a finite element $(\overline{D}, P_T, N_T)$ be associated with each subdomain $D \in T$ so that the resulting finite element space $FE_T$ (cf. (32)) is a subspace of $C^{m-1}(\overline{\Omega}) \subset H^m(\Omega)$. By imposing appropriate boundary conditions, we can obtain a subspace $V_T$ of $FE_T$ such that $V_T \subset V$, the subspace of $H^m(\Omega)$, where the weak problem (1) is formulated. The corresponding finite element method for (1) is:

Find $u_T \in V_T$ such that

$$a(u_T, v) = F(v) \quad \forall v \in V_T$$

(54)
In this section, we consider a priori estimates for the discretization error \( u - u_T \). We will discuss the second-order and fourth-order cases separately. We use the letter \( C \) to denote a generic positive constant that can take different values at different appearances.

Let us also point out that the asymptotic error analysis carried out in this section is not sufficient for parameter-dependent problems (e.g., thin structures and nearly incompressible elasticity) that can experience locking (Babuška and Suri, 1992). We refer the readers to other chapters in this encyclopedia that are devoted to such problems for the discussion of the techniques that can overcome locking.

### 4.1 Second-order problems

We will devote most of our discussion to the case where \( \Omega \subset \mathbb{R}^2 \) and only comment briefly on the 3-D case. For preciseness, we also assume the right-hand side of the elliptic boundary value problem to be square integrable. We first consider the case where \( V \subset H^1(\Omega) \) is defined by homogeneous Dirichlet boundary conditions (cf. Section 2.1) on \( \Gamma \subset \partial \Omega \). Such problems can be discretized by triangular \( P_n \) elements (Example 4) or quadrilateral \( Q_n \) elements (Example 8).

Let \( T \) be a triangulation of \( \Omega \) by triangles (convex quadrilaterals) and each triangle (quadrilateral) in \( T \) be equipped with the \( P_n \) \((n \geq 1)\) Lagrange element (\( Q_n \) quadrilateral element). The resulting finite element space \( FE_T \) is a subspace of \( C^0(\Omega) \subset H^1(\Omega) \). We assume that \( \Gamma \) is the union of the edges of the triangles (quadrilaterals) in \( T \) and take \( V_T = V \cap FE_T \), the subspace defined by the homogeneous Dirichlet boundary condition on \( \Gamma \).

We know from the discussion in Section 2.3 that \( u \in H^{1+\alpha}(D) \) for each \( D \in T \), where the number \( \alpha(D) \in (0, 1] \) and \( \alpha(D) = 1 \) for \( D \) far away from the singular points. Hence, the element nodal interpolation operator \( \Pi_T \) is well-defined on \( u \) for all \( D \in T \). We can therefore piece together a global nodal interpolant \( \Pi_T u \in V_T \) by the formula

\[
(\Pi_T u)|_D = \Pi_T(u)|_D
\]  

(55)

From the discussion in Section 3.3, we know that (43) is valid for both the triangular \( P_n \) element and the quadrilateral \( Q_n \) element. We deduce from (43) and (55) that

\[
\|u - \Pi_T u\|_{H^1(\Omega)}^2 \leq C \sum_{D \in T} (\text{diam } D)^{2\alpha(D)} \|u\|_{H^{1+\alpha}(D)}^2
\]  

(56)

where \( C \) depends only on the maximum of the aspect ratios of the element domains in \( T \). Combining (22) and (56) we have the a priori discretization error estimate

\[
\|u - u_T\|_{H^1(\Omega)} \leq C \left( \sum_{D \in T} (\text{diam } D)^{2\alpha(D)} \|u\|_{H^{1+\alpha}(D)}^2 \right)^{1/2}
\]  

(57)

where \( C \) depends only on the constants in (2) and (3) and the maximum of the aspect ratios of the element domains in \( T \).

Hence, if \( \{T_i; i \in I\} \) is a regular family of triangulations, and the solution \( u \) of (1) belongs to the Sobolev space \( H^{1+\alpha}(\Omega) \) for some \( \alpha \in (0, 1] \), then we can deduce from (57) that

\[
\|u - u_T\|_{H^1(\Omega)} \leq C h_i \|u\|_{H^{1+\alpha}(\Omega)}
\]  

(58)

where \( h_i = \max_{D \in T_i} \text{diam } D \) is the mesh size of \( T_i \) and \( C \) is independent of \( i \in I \). Note that the solution \( w \) of (23) with \( \bar{u} \) replaced by \( u_T \) also belongs to \( H^{1+\alpha}(\Omega) \) and satisfies the elliptic regularity estimate

\[
\|w\|_{H^{1+\alpha}(\Omega)} \leq C \|u - u_T\|_{L^2(\Omega)}
\]

Therefore, we have

\[
\inf_{v \in V_T} \|w - v\|_{H^1(\Omega)} \leq \|w - \Pi_T^N w\|_{H^1(\Omega)} \leq C h_i \|w\|_{H^{1+\alpha}(\Omega)} \leq \|u - u_T\|_{L^2(\Omega)}
\]  

(59)

The abstract estimate (24) with \( \bar{u} \) replaced by \( u_T \) and (59) yield the following \( L^2 \) estimate:

\[
\|u - u_T\|_{L^2(\Omega)} \leq C h_i \|u\|_{H^{1+\alpha}(\Omega)}
\]  

(60)

where \( C \) is also independent of \( i \in I \).

**Remark 12.** In the case where \( \alpha = 1 \) (for example, when \( \Gamma = \partial \Omega \) in Example 1 and \( \Omega \) is convex), the estimate (58) is optimal and it is appropriate to use a quasi-uniform family of triangulations. In the case where \( \alpha(D) < 1 \) for \( D \) near to singular points, the estimate (57) allows the possibility of improvement by graded meshes (cf. Section 6).

**Remark 13.** In the derivations of (58) and (60) above for the triangular \( P_n \) elements, we have used the minimum angle condition (cf. Remark 5). In view of the anisotropic estimates (45), these estimates also hold for triangular \( P_n \) elements under the maximum angle condition (Babuška and Aziz, 1976; Jamet, 1976; Ženíšek, 1995; Apel, 1999): there exists \( \theta_\alpha \in \pi \) such that all the angles in the family of triangulations are \( \leq \theta_\alpha \). The estimates (58) and (60) are also valid for \( Q_n \) elements on parallelograms satisfying the maximum angle condition. They can also be established for certain thin quadrilateral elements (Apel, 1999).
The 2-D results above also hold for 3-D tetrahedral $P_n$ elements and 3-D hexagonal $Q_n$ elements if the solution $u$ of (1) belongs to $H^{1+\alpha}(\Omega)$ where $1/2 < \alpha \leq 1$, since the nodal interpolation operator are then well-defined by the Sobolev embedding theorem. This is the case, for example, if $\Gamma = \partial \Omega$ in Example 1. However, new interpolation operators that require less regularity are needed if $0 < \alpha \leq 3/2$. Below, we construct an interpolation operator $\Pi_T^2: H^1(\Omega) \rightarrow V_T$ using the local averaging technique of Scott and Zhang (1990).

For simplicity, we take $V_T$ to be a tetrahedral $P_1$ finite element space. Therefore, we only need to specify the value of $\Pi_T^2 \xi$ at the vertices of $T$ for a given function $\xi \in H^1(\Omega)$. Let $p$ be a vertex. We choose a face (or edge in 2-D) $F$ of a subdomain in $T$ such that $p \in F$. The choice of $F$ is of course not unique. But we always choose $F \subseteq \partial \Omega$ if $p \in \partial \Omega$ so that the resulting interpolant will satisfy the appropriate Dirichlet boundary condition. Let $\{\psi_j\}_{j=1}^{62} \subset P_1(F)$ be biorthogonal to the nodal basis $\{\phi_j\}_{j=1}^{62} \subset P_1(F)$ with respect to the $L_2(F)$ inner product. In other words $\phi_j$ equals 1 at the $j$th vertex of $F$ and vanishes at the other vertices, and

$$
\int_F \psi_j \phi_j \, dx = \delta_{ij} \quad (61)
$$

Suppose $p$ corresponds to the $j$th vertex of $F$. We then define

$$
(\Pi_T^2 \xi)(p) = \int_F \psi_j \xi \, dx \quad (62)
$$

where the integral is well-defined because of the trace theorem.

It is clear in view of (61) and (62) that $\Pi_T^2 v = v$ for all $v \in FE_T$ and $\Pi_T^2 \xi = 0$ on $\Gamma$ if $\xi = 0$ on $\Gamma$. Note also that $\Pi_T^2$ is a local operator, i.e., $(\Pi_T^2 \xi)|_D$ is in general determined by $\xi|_{S(D)}$, where $S(D)$ is the polyhedral domain formed by the subdomains in $T$ sharing (at least) a vertex with $D$ (cf. Figure 9 for a 2-D example). It follows that the interpolation error estimate for $\tilde{\Pi}_T$ takes the following form:

$$
\|\xi - \Pi_T^2 \xi\|_{L^2(D)} + (\text{diam } D)^2 \|\xi - \Pi_T^2 c\|_{H^1(D)}^2 \\
\leq C (\text{diam } D)^{2(1+\alpha(S(D)))} \|\xi\|_{H^{1+\alpha(S(D))}(S(D))}^2
$$

(63)

where $C$ depends on the shape regularity of $T$, provided that $\xi \in H^{1+\alpha(S(D))}(S(D))$ for some $\alpha(S(D)) \in (0, 1]$. The estimates (58) and (60) for tetrahedral $P_1$ elements can be derived for general $\alpha \in (0, 1]$ and regular triangulations using the estimate (63).

Next, we consider the case where $V$ is a closed subspace of $H^1(\Omega)$ with finite codimension $n < \infty$, as in the case of the Poisson problem with pure homogeneous Neumann boundary condition (where $n = 1$) or the elasticity problem with pure homogeneous traction boundary condition (where $n = 1$ when $d = 1$, and $n = 3(d - 1)$ when $d = 2$ or 3). The key assumption here is that there exists a bounded linear projection $Q$ from $H^1(\Omega)$ onto an $n$ dimensional subspace of $FE_T$ such that $\xi \in H^1(\Omega)$ belongs to $V$ if and only if $Q \xi = 0$. We can then define an interpolation operator $\tilde{\Pi}_T$ from appropriate Sobolev spaces onto $V_T$ by

$$
\tilde{\Pi}_T = (I - Q) \Pi_T
$$

where $\Pi_T$ is either the nodal interpolation operator $\Pi_T^N$ or the Scott–Zhang averaging interpolation operator $\Pi_T^A$ introduced earlier. Observe that, since the weak solution $u$ belongs to $V$,

$$
u - \tilde{\Pi}_T u = u - Qu - (I - Q) \Pi_T u = (I - Q)(u - \Pi_T u)
$$

and the interpolation error of $\tilde{\Pi}_T$ can be estimated in terms of the norm of $Q: H^1(\Omega) \rightarrow H^1(\Omega)$ and the interpolation error of $\Pi_T$. Therefore, the a priori discretization error estimates for Dirichlet or Dirichlet/Neumann boundary value problems also hold for this second type of (pure Neumann) boundary value problems.

For the Poisson problem with homogeneous Neumann boundary condition, we can take

$$
Q \xi = \frac{1}{|\Omega|} \int_\Omega \xi \, dx
$$

the mean of $\xi$ over $\Omega$. For the elasticity problem with pure homogeneous traction boundary condition, the operator $Q$ from $[H^1(\Omega)]^d$ onto the space of infinitesimal rigid motions.
is defined by
\[ \int_{\Omega} Q \xi \, dx = \int_{\Omega} \xi \, dx \quad \text{and} \quad \int_{\Omega} \nabla \times Q \, dx = \int_{\Omega} \nabla \times \xi \, dx \quad \forall \xi \in [H^1(\Omega)]^d \]
In both cases, the norm of \( Q \) is bounded by a constant \( C_{\Omega} \).

Remark 15. In the case where \( f \in H^k(\Omega) \) for \( k > 0 \), the solution \( u \) belongs to \( H^{2+k}(\Omega) \) away from the geometric or boundary data singularities and, in particular, away from \( \partial \Omega \). Therefore, it is advantageous to use higher-order elements in certain parts of \( \Omega \), or even globally (with curved elements near \( \partial \Omega \)) if singularities are not present. In the case where \( f \in L^2(\Omega) \), the error estimate (57) indicates that the order of the discretization error for the triangular \( P_n \) element or the quadrilateral \( Q_n \) element is independent of \( n \geq 1 \). However, the convergence of the finite element solutions to a particular solution as \( h \downarrow 0 \) can be improved by using higher-order elements because of the existence of \emph{nonuniform error estimates} (Babuška and Kellogg, 1975).

\section{4.2 Fourth-order problems}

We restrict the discussion of fourth-order problems to the two-dimensional plate bending problem of Example 3.

Let \( T \) be a triangulation of \( \Omega \) by triangles and each triangle in \( T \) be equipped with the Hsieh–Clough–Tocher macro element (cf. Example 6). The finite element space \( FE_T \) defined by (32) is a subspace of \( C^1(\Omega) \subset H^2(\Omega) \). We take \( V_T \) to be \( V \cap FE_T \), where \( V = H^1_0(\Omega) \) for the clamped plate and \( V = H^1_0(\Omega) \cap H^2(\Omega) \) for the simply supported plate.

The solution \( u \) of the plate-bending problem belongs to \( H^{2+\alpha(D)}(D) \) for each \( D \in T \), where \( \alpha(D) \in (0, 2] \) and \( \alpha(D) = 2 \) for \( D \) away from the corners of \( \Omega \). The elemental nodal interpolation operator \( \Pi^N_D \) is well-defined on \( u \) for all \( D \in T \). We can therefore define a global nodal interpolation operator \( \Pi^N_T \) by the formula (55). Since the Hsieh–Clough–Tocher macro element is affine-interpolation-equivalent to the reference element on the standard simplex, we deduce from (55) and (43) that
\[ \| u - \Pi^N_T u \|_{H^2(\Omega)} \leq C \sum_{D \in T} (\text{diam } D)^{2\alpha(D)} |u|_{H^{2+\alpha(D)}(D)}^2 \quad (64) \]
where \( C \) depends only on the maximum of the aspect ratios of the triangles in \( T \) (or equivalently the minimum angle of \( T \)). From (22) and (64), we have
\[ \| u - u_T \|_{H^2(\Omega)} \leq C \left( \sum_{D \in T} (\text{diam } D)^{2\alpha(D)} |u|_{H^{2+\alpha(D)}(D)}^2 \right)^{1/2} \quad (65) \]
where \( C \) depends only on the constants in (2) and (3) and the minimum angle of \( T \).

Hence, if \( \{ T_i \mid i \in I \} \) is a regular family of triangulations, and the solution \( u \) of the plate bending problem belongs to the Sobolev space \( H^{2+\alpha}(\Omega) \) for some \( \alpha \in (0, 2] \), we can deduce from (65) that
\[ \| u - u_{T_i} \|_{H^2(\Omega)} \leq C h_i^\alpha |u|_{H^{2+\alpha}(\Omega)} \quad (66) \]
where \( h_i = \max_{D \in T_i} \text{diam } D \) is the mesh size of \( T_i \) and \( C \) is independent of \( i \in I \). Since the solution \( w \) of (23) also belongs to \( H^{2+\alpha}(\Omega) \), the abstract estimate (24) combined with an error estimate for \( w \) in the \( H^2 \)-norm analogous to (66) yields the following \( L_2 \) estimate:
\[ \| u - u_{T_i} \|_{L_2(\Omega)} \leq C h_i^{2\alpha} |u|_{H^{2+\alpha}(\Omega)} \quad (67) \]
where \( C \) is also independent of \( i \in I \).

Remark 16. The analysis of general triangular and quadrilateral \( C^1 \) macro elements can be found in Douglas et al. (1979).

The plate-bending problem can also be discretized by the Argyris element (cf. Example 5). If \( \alpha(D) > 1 \) for all \( D \in D \), then the nodal interpolation operator \( \Pi^N_T \) is well-defined for the Argyris finite element space. If \( \alpha(D) \leq 1 \) for some \( D \in T \), then the nodal interpolation operator \( \Pi^N_T \) must be replaced by an interpolation operator constructed by the technique of local averaging. In either case, the estimates (65)–(67) remain valid for the Argyris finite element solution.

\section{5 A POSTERIORI ERROR ESTIMATES AND ANALYSIS}

In this section, we review explicit and implicit estimators as well as averaging and multilevel estimators for a posteriori finite element error control.

Throughout this section, we adopt the notation of Sections 2.1 and 2.2 and recall that \( u \) denotes the (unknown) exact solution of (1) while \( \tilde{u} \in \tilde{V} \) denotes the discrete and given solution of (19). It is the aim of Section 5.1–5.6 to estimate the error \( e := u - \tilde{u} \) in the energy norm \( \| \cdot \|_E = (\langle \cdot, \cdot \rangle)^{1/2} \) in terms of computable quantities while Section 5.7 concerns other error norms or goal functionals.
Throughout this section, we assume $0 < \|e\|_a$ to exclude the exceptional situation $u = \tilde{u}$.

5.1 Aims and concepts in a posteriori finite element error control

The following five sections introduce the notation, the concepts of efficiency and reliability, the definitions of residual and error, a posteriori error control and adaptive algorithms, and comment on some relevant literature.

5.1.1 Error estimators, efficiency, reliability, asymptotic exactness

Regarded as an approximation to the (unknown) error norm $\|e\|_a$, a (computable) quantity $\eta$ is called a posteriori error estimator, or estimator for brevity, if it is a function of the known domain $\Omega$ and its boundary $\Gamma$, the quantities of the right-hand side $F$, cf. (6) and (13), as well as of the (given) discrete solution $\tilde{u}$, or the underlying triangulation.

An estimator $\eta$ is called reliable if

$$\|e\|_a \leq C_{\text{rel}} \eta + \text{h.o.t. rel} \quad (68)$$

An estimator $\eta$ is called efficient if

$$\eta \leq C_{\text{eff}} \|e\|_a + \text{h.o.t. eff} \quad (69)$$

An estimator is called asymptotically exact if it is reliable and efficient in the sense of (68)–(69) with $C_{\text{rel}} = C_{\text{eff}}^{-1}$.

Here, $C_{\text{rel}}$ and $C_{\text{eff}}$ are multiplicative constants that do not depend on the mesh size of an underlying finite element mesh $T$ for the computation of $\tilde{u}$ and h.o.t. denotes higher-order terms. The latter are generically much smaller than $\eta$ or $\|e\|_a$, but usually, this depends on the (unknown) smoothness of the exact solution or the (known) smoothness of the given data. The readers are warned that, in general, h.o.t. may not be neglected; in case of high oscillations they may even dominate (68) or (69).

5.1.2 Error and residual

Abstract examples for estimators are (26) and (27), which involve dual norms of the residual (25). Notice carefully that $R := F - a(\tilde{u}, \cdot)$ is a bounded linear functional in $V$, written $R \in V^*$, and hence the dual norm

$$\|R\|_{\cdot, V^*} := \sup_{v \in V \backslash \{0\}} \frac{R(v)}{\|v\|_a} = \sup_{v \in V \backslash \{0\}} \frac{a(e, v)}{\|v\|_a} = \|e\|_a < \infty \quad (70)$$

The second equality immediately follows from (25). A Cauchy inequality in (70) with respect to the scalar product $a$ results in $\|R\|_{\cdot, V^*} \leq \|e\|_a$ while $v = e$ in (70) yields finally the equality $\|R\|_{\cdot, V^*} = \|e\|_a$.

That is, the error (estimation) in the energy norm is equivalent to the (computation of the) dual norm of the given residual. Furthermore, it is even of comparable computational effort to compute an optimal $v = e$ in (70) or to compute $e$. The proof of (70) yields even a stability estimate: The relative error of $R(v)$ as an approximation to $\|e\|_a$ equals

$$\frac{\|e\|_a - R(v)}{\|e\|_a} = \frac{1}{2} \left| v - \frac{e}{\|e\|_a} \right|^2$$

for all $v \in V$ with $\|v\|_a = 1 \quad (71)$

In fact, given any $v \in V$ with $\|v\|_a = 1$, the identity (71) follows from

$$1 - a\left( \frac{e}{\|e\|_a}, v \right) = \frac{1}{2} a\left( \frac{e}{\|e\|_a}, \frac{e}{\|e\|_a} \right) - a\left( \frac{e}{\|e\|_a}, v \right) + \frac{1}{2} a(v, v) = \frac{1}{2} \left| v - \frac{e}{\|e\|_a} \right|^2$$

The error estimate (71) implies that the maximizing $v$ in (70) (i.e. $v \in V$ with maximal $R(v)$ subject to $\|v\|_a \leq 1$) is unique and equals $e/\|e\|_a$. As a consequence, the computation of the maximizing $v$ in (70) is equivalent to and indeed equally expensive as the computation of the unknown $e/\|e\|_a$ and so (since $\tilde{u}$ is known) of the exact solution $u$. Therefore, a posteriori error analysis aims to compute lower and upper bounds of $\|R\|_{\cdot, V^*}$, rather than its exact value.

5.1.3 Error estimators and error control

For an idealized termination procedure, one is given a tolerance $Tol > 0$ and interested in a stopping criterion (of successively adapted mesh refinements)

$$\|e\|_a \leq Tol \quad (72)$$

Since the error $\|e\|_a$ is unknown, it is replaced by its upper bound (68) and then leads to

$$C_{\text{rel}} \eta + \text{h.o.t. rel} \leq Tol \quad (72)$$

For a verification of (72), in practice, one requires not only $\eta$ but also $C_{\text{rel}}$ and h.o.t. rel. The later quantity cannot be dropped; it is not sufficient to know that h.o.t. rel is (possibly) negligible for sufficient small mesh-sizes.

Section 5.6 presents numerical examples and further discussions of this aspect.
5.1.4 Adaptive mesh-refining algorithms

Error estimators are used in adaptive mesh-refining algorithms to motivate a refinement rule, which determines whether an element or edge and so on shall be refined or coarsened. This will be discussed in Section 6 below.

At this stage two remarks are in order. First, one should be precise in the language and distinguish between error estimators, which are usually global and fully involve constants and higher-order terms and (local) refinement indicators used in refinement rules. Second, constants and higher-order terms might be seen as less important and are often omitted in the usage as refinement indicators for the step MARKING in Section 6.2.

5.1.5 Literature

Amongst the most influential pioneering publications on a posteriori error control are Babuška and Rheinboldt (1978), Ladeveze and Leguillon (1983), Bank and Weiser (1985), Babuška and Miller (1987), Eriksson and Johnson (1991), followed by many others. The readers may find it enlightening to motivate a refinement rule by considering the works of Verfürth (1996), Ainsworth and Oden (2000), Babuška and Strouboulis (2001), Bangerth and Rannacher (2003) and the books of Becker and Rannacher (2001) and the books of Becker and Rannacher (2001) for a first insight and further references.

5.2 Explicit residual-based error estimators

The most frequently considered and possibly easiest class of error estimators consists of local norms of explicitly given volume and jump residuals multiplied by mesh-depending weights.

To derive them for a general class of abstract problems from Section 2.1, let \( u \in V \) be an exact solution of the problem (1) and let \( \tilde{u} \in \tilde{V} \) be its Galerkin approximation from (19) with residual \( R(v) \) from (25). Moreover, as in Example 1 or 2, it is supposed throughout this chapter that the strong form of the equilibration associated with the weak form (19) is of the form

\[
-\text{div} p = f \quad \text{for some flux or stress} \quad p \in L^2(\Omega; \mathbb{R}^{m \times n})
\]

The discrete analog \( \tilde{p} \) is piecewise smooth but, in general, discontinuous; at several places below, it is a \( T \) piecewise constant \( m \times n \) matrix as it is proportional to the gradient of some (piecewise) \( P_1 \) FE function \( \tilde{u} \). The description of the residuals is based on the weak form of \( f + \text{div} p = 0 \).

5.2.1 Residual representation formula

It is the aim of this section to recast the residual in the form

\[
R(v) = \sum_{T \in \mathcal{T}} \int_T r_T \cdot v \, dx - \sum_{E \in \mathcal{E}} \int_E r_E \cdot v \, ds \quad (73)
\]

of a sum of integrals over all element domains \( T \in \mathcal{T} \) plus a sum of integrals over all edges or faces \( E \in \mathcal{E} \) and to identify the explicit volume residual \( r_T \) and the jump residual \( r_E \).

The boundary \( \partial T \) of each finite element domain \( T \in \mathcal{T} \) is a union of edges or faces, which form the set \( \mathcal{E}(T) \), written \( \partial T = \partial \mathcal{E}(T) \). Each edge or face \( E \in \mathcal{E} \) is in the set of all possible edges or faces \( \mathcal{E} = \cup \{ \mathcal{E}(T) : T \in \mathcal{T} \} \) is associated with a unit normal vector \( \nu_E \), which is unique up to an orientation \( \pm \nu_E \), which is globally fixed. By convention, the unit normal \( \nu \) on the domain \( \Omega \) or on an element \( T \) points outwards.

For the ease of exploration, suppose that the underlying boundary value problem allows the bilinear form \( a(\tilde{u}, v) \) to equal the sum over all \( \int_T \tilde{p}_{jk} D_j v_k \, dx \) with given fluxes or stresses \( \tilde{p}_{jk} \). Moreover, Neumann data are excluded from the description in this section and hence only interior edges contribute with a jump residual. An integration by parts on \( T \) with outer unit normal \( \nu_T \) yields

\[
\int_T \tilde{p}_{jk} D_j v_k \, dx = \int_{\partial T} \tilde{p}_{jk} v_k \nu_{T,j} \, ds - \int_T v_k D_j \tilde{p}_{jk} \, dx
\]

which, with the divergence operator \( \text{div} \) and proper evaluation of \( \tilde{p} \nu \), reads

\[
a(\tilde{u}, v) + \sum_{T \in \mathcal{T}} \int_T v \cdot \text{div} \tilde{p} \, dx = \sum_{T \in \mathcal{T}} \int_{\partial T} (\tilde{p} \nu) \cdot v \, ds
\]

Each boundary \( \partial T \) is rewritten as a sum of edges or faces. Each such edge or face \( E \) belongs either to the boundary \( \partial \Omega \), written \( E \in \mathcal{E}_{\partial \Omega} \), or is an interior edge, written \( E \in \mathcal{E}_{\Omega} \). For \( E \in \mathcal{E}_{\partial \Omega} \) there exists exactly one element \( T \) with \( E \in \mathcal{E}(T) \) and one defines \( T_+ = T \), \( T_- = E \subset \partial \Omega \), \( \omega_E = \text{int}(T) \) and \( \nu_E := \nu_T = \nu_{\Omega} \). Any \( E \in \mathcal{E}_{\Omega} \) is the intersection of exactly two elements, which we name \( T_+ \) and \( T_- \) and which essentially determine the patch \( \omega_E := \text{int}(T_+ \cup T_-) \) of \( E \). This description of \( T_\pm \) is unique up to the order that is fixed in the sequel by the convention that \( \nu_E = \nu_{T_+} \) is exterior to \( T_+ \). Then,

\[
\sum_{T \in \mathcal{T}} \int_{\partial T} (\tilde{p} \nu) \cdot v \, ds = \sum_{E \in \mathcal{E}_{\partial \Omega}} \int_E [\tilde{p} \nu_E] \cdot v \, ds
\]

where \( [\tilde{p} \nu_E] := (\tilde{p}|_{T_+} - \tilde{p}|_{T_-}) \nu_E \) for \( E = \partial T_+ \cap \partial T_- \in \mathcal{E}_{\partial \Omega} \) and \( [\tilde{p} \nu_E] := 0 \) for \( E \in \mathcal{E}(T) \cap \mathcal{E}_{\partial \Omega} \). Altogether, one obtains
the error residual formula (73) with the

\[
\begin{align*}
\text{volume residuals} & \quad r_T := f + \text{div} \, \tilde{\rho} \quad \text{in } T \in \mathcal{T} \\
\text{jump residuals} & \quad r_E := [\tilde{\rho} v_E] \quad \text{along } E \in \mathcal{E}_{\Omega}
\end{align*}
\]

5.2.2 Weak approximation operators

In terms of the residual \( R \), the orthogonality condition (20)
is rewritten as \( R(\tilde{v}) = 0 \) for all \( \tilde{v} \in \tilde{V} \). Hence, given any
\( v \in V \) with norm \( \|v\|_a = 1 \), there holds \( R(v) = R(v - \tilde{v}) \).

Explicit error estimators rely on the design of \( \tilde{v} := \Pi_1^2(v) \) as a function of \( v \), \( \Pi_1^2 \) is called approximation operator as in (61)–(63) and discussed further in Section 4. See also (Carstensen, 1999; Carstensen and Funken, 2006; Nochetto and Wahlbin, 2002). For the understanding of this section, it suffices to know that there are several choices of \( \tilde{v} \in \tilde{V} \) that satisfy first-order approximation and stability properties in the sense of

\[
\sum_{T \in \mathcal{T}} \|h_T^{-1}(v - \tilde{v})\|^2_{L^2(T)} + \sum_{E \in \mathcal{E}_{\Omega}} \|h_E^{-1/2}(v - \tilde{v})\|^2_{L^2(E)} + \|v - \tilde{v}\|^2_{H^1(\Omega)} \leq C \|v\|^2_{H^1(\Omega)} \tag{74}
\]

Here, \( h_T \) and \( h_E \) denotes the diameter of an element \( T \in \mathcal{T} \) and an edge \( E \in \mathcal{E} \), respectively. The multiplicative constant \( C \) is independent of the mesh-sizes \( h_T \) or \( h_E \), but depends on the shape of the element domains through their minimal angle condition (for simplices) or aspect ratio (for tensor product elements).

5.2.3 Reliability

Given the explicit volume and jump residuals \( r_T \) and \( r_E \) in (73), one defines the explicit residual-based estimator \( \eta_{R,R} \):

\[
\eta_{R,R}^2 := \sum_{T \in \mathcal{T}} h_T^2 \|r_T\|^2_{L^2(T)} + \sum_{E \in \mathcal{E}_{\Omega}} h_E \|r_E\|^2_{L^2(E)} \tag{75}
\]

which is reliable, that is

\[
\|e\|_a \leq C \eta_{R,R} \tag{76}
\]

The proof of (76) follows from (73)–(75) and Cauchy inequalities:

\[
\begin{align*}
R(v) &= R(v - \tilde{v}) = \sum_{T \in \mathcal{T}} \int_T r_T \cdot (v - \tilde{v}) \, dx \\
&\quad - \sum_{E \in \mathcal{E}_{\Omega}} \int_E r_E \cdot (v - \tilde{v}) \, ds \\
&\leq \sum_{T \in \mathcal{T}} (h_T \|r_T\|_{L^2(T)}) (h_T^{-1} \|v - \tilde{v}\|_{L^2(T)}) \\
&\quad + \sum_{E \in \mathcal{E}_{\Omega}} (h_E \|r_E\|_{L^2(E)}) (h_E^{-1/2} \|v - \tilde{v}\|_{L^2(E)}) \\
&\leq \left( \sum_{T \in \mathcal{T}} h_T^2 \|r_T\|^2_{L^2(T)} \right)^{1/2} \left( \sum_{E \in \mathcal{E}_{\Omega}} h_E \|r_E\|^2_{L^2(E)} \right)^{1/2} \\
&\leq C \eta_{R,R} \|v\|_{H^1(\Omega)}
\end{align*}
\]

For first-order finite element methods in the situation of Example 1 or 2, the volume term \( r_T = f \) can be substituted by the higher-order term of oscillations, that is

\[
\|e\|_a^2 \leq C \left( \text{osc}(f)^2 + \sum_{E \in \mathcal{E}_{\Omega}} h_E \|r_E\|^2_{L^2(E)} \right) \tag{77}
\]

For each node \( z \in \mathcal{N} \) with nodal basis function \( z \), and patch \( \omega_z := \{ x \in \Omega : z(x) \neq 0 \} \) of diameter \( h_z \) and the source term \( f \in L^2(\Omega)^m \) with integral mean \( f_z := |\omega_z|^{-1} \int_{\omega_z} f(x) \, dx \in \mathbb{R}^m \), the oscillations of \( f \) are defined by

\[
\text{osc}(f) := \left( \sum_{z \in \mathcal{N}} h_z^2 \|f - f_z\|^2_{L^2(\omega_z)} \right)^{1/2}
\]

Notice for \( f \in H^1(\Omega)^m \) and the mesh size \( h_T \in P_0(T) \) there holds

\[
\text{osc}(f) \leq C \|h_T^2 Df\|_{L^2(\Omega)}
\]

and so \( \text{osc}(f) \) is of quadratic and hence of higher-order. We refer to Carstensen and Verfürth (1999), Nochetto (1993), Becker and Rannacher (1996), and Rodriguez (1994b) for further details on and proofs of (77).

5.2.4 Efficiency

Following a technique with inverse estimates due to Verfürth (1996), this section investigates the proof of efficiency of \( \eta_{R,R} \) in a local form, namely,

\[
\begin{align*}
&h_T \|r_T\|_{L^2(T)} \leq C \left( \|e\|_{H^1(T)} + \text{osc}(f, T) \right) \tag{78} \\
&h_E^{1/2} \|r_E\|_{L^2(E)} \leq C \left( \|e\|_{H^1(\omega_E)} + \text{osc}(f, \omega_E) \right) \tag{79}
\end{align*}
\]

where \( \tilde{f} \) denotes an elementwise polynomial (best-) approximation of \( f \) and

\[
\text{osc}(f, T) := h_T \|f - \tilde{f}\|_{L^2(T)}
\]

and

\[
\text{osc}(f, \omega_E) := h_E \|f - \tilde{f}\|_{L^2(\omega_E)}
\]
The main tools in the proof of (79) and (78) are bubble functions \( b_E \) and \( b_T \) based on an edge or face \( E \in \mathcal{E} \) and an element \( T \in \mathcal{T} \) with nodes \( N(E) \) and \( N(T) \), respectively. Given a nodal basis \( \{ \varphi_z : z \in N \} \) of a first-order finite element method with respect to \( T \) define, for any \( T \in \mathcal{T} \) and \( E \in \mathcal{E}_T \), the element- and edge-bubble functions:

\[
b_T := \prod_{z \in N(T)} \varphi_z \quad \text{and} \quad b_E := \prod_{z \in N(E)} \varphi_z \quad \text{(80)}
\]

\( b_E \) and \( b_T \) are nonnegative and continuous piecewise polynomials \( \leq 1 \) with support \( \text{supp} b_E = \overline{\omega E} = \overline{E} \) and \( \text{supp} b_T = T \). Utilizing the bubble functions (80), the proof of (78)–(79) essentially consists in design of test functions \( \tilde{w}_E \) and \( \tilde{w}_T \);

\[\tilde{w}_E \cdot \tilde{r}_E = 0 \quad \text{and} \quad \tilde{w}_T \cdot \tilde{r}_T = 0 \quad \text{in the last step). A Friedrichs inequality with respect to the scalar product \( \alpha \) prove (78)–(79).}

An integration by parts concludes the proof of (82) for

\[w_T := h_T^2 b_T \tilde{r}_T \quad \text{(84)}\]

the proof of (81) for this \( w_T \) is immediate.

Given an interior edge \( E = T_+ \cap T_- \in \mathcal{E}_T \) with its neighboring elements \( T_+ \) and \( T_- \), simultaneously addressed as \( T \in \mathcal{T} \), extend the edge residual \( r_E \) from the edge \( E \) to its patch \( \omega_E = \text{int}(T_+ \cup T_-) \) such that

\[\|b_E r_E\|_{L^2(\omega_E)} + h_E \|b_T r_T\|_{H^1(\omega_E)} \leq C_1 h_T^{1/2} \|r_E\|_{L^2(E)} \leq C_2 h_T^{1/2} \|b_E r_E\|_{L^2(E)} \quad \text{(85)}\]

with an inverse inequality at the end. The choice of the two real constants

\[\alpha_{\pm} = \frac{\int_{T_+} h_E b_E \tilde{r}_T \cdot r_E \, dx}{\int_{T_-} w_{T_-} \cdot \tilde{r}_{T_-} \, dx} \]

in the definition

\[w_E := \alpha_+ w_{T_+} + \alpha_- w_{T_-} - h_E b_E r_E \quad \text{(86)}\]

yields \( \int_{T_+} w_E \cdot \tilde{r}_{T_+} \, dx = 0 \). Since \( \int_{T_-} w_{T_-} \cdot \tilde{r}_{T_-} \, dx \) one eventually deduces \( |\alpha_{\pm}| \|w_{T_{\pm}}\|_{H^1(T_{\pm})} \leq C h_T^{1/2} \|r_E\|_{L^2(E)} \) and then concludes (81). An integration by parts shows

\[C^{-2} h_T \|r_E\|_{L^2(E)}^2 \leq h_T \|b_E^{1/2} r_E\|_{L^2(E)}^2 \]

\[= - \int_{\omega_E} w_E \cdot r_E \, dx \quad \int_{\omega_E} w_E \cdot [(p - \tilde{p}) \cdot v_E] \, dx \]

\[= \int_{\omega_E} (p - \tilde{p}) \cdot Dw_E \, dx + \int_{\omega_E} w_E \cdot \text{div}_T (p - \tilde{p}) \, dx \]

\[= R(w_E) - \int_{\omega_E} w_E \cdot (f + \text{div}_T \tilde{p}) \, dx \]

\[= R(w_E) - \int_{\omega_E} w_E \cdot (f - \tilde{f}) \, dx \]

(with \( \int_{T_-} w_E \cdot \tilde{r}_{T_-} \, dx = 0 \) in the last step). A Friedrichs inequality \( \|w_E\|_{L^2(\omega_E)} \leq C h_T \|w_E\|_{H^1(\omega_E)} \) and (81) then conclude the proof of (83).

### 5.3 Implicit error estimators

Implicit error estimators are based on a local norm of a solution of a localized problem of a similar type with the residual terms on the right-hand side. This section
introduces two different versions based on a partition of unity and based on an equilibration technique.

5.3.1 Localization by partition of unity

Given a nodal basis \((\varphi_z : z \in \mathcal{N})\) of a first-order finite element method with respect to \(T\), there holds the partition of unity property

\[
\sum_{z \in \mathcal{N}} \varphi_z = 1 \quad \text{in} \quad \Omega
\]

Given the residual \(R = F - a(\bar{u}, \cdot) \in V^*\), we observe that \(R_z(v) := R(\varphi_z, v)\) defines a bounded linear functional \(R_z\) on a localized space called \(V_z\) and specified below. The bilinear form \(a\) is an integral over \(\omega\) on some integrand. The latter may be weighted with \(\varphi_z\) to define some (localized) bilinear form \(a_z : V_z \times V_z \to \mathbb{R}\). Supposing that \(a_z\) is \(V\)-elliptic one defines the norm \(\| \cdot \|_{a_z}\) on \(V_z\) and considers

\[
\eta_z := \sup_{v \in V_z \setminus \{0\}} \frac{R_z(v)}{\|v\|_{a_z}}
\]

The dual norm is as in (70)–(71) and hence equivalent to the computation of the norm \(\|e\|_{a_z}\) of a local solution

\[
e_z \in V_z \quad \text{with} \quad a_z(e_z, \cdot) = R_z \in V_z^* \quad (88)
\]

(The proof of \(\|e_z\|_{a_z} = \eta_z\) follows the arguments of Section 5.1.2 and hence is omitted.)

**Example 10.** Adopt notation from Example 1 and let \((\varphi_z : z \in \mathcal{N})\) be the first-order finite element nodal basis functions. Then define \(R_z\) and \(a_z\) by

\[
R_z(v) := \int_\Omega \varphi_z f v \, dx - \int_\Omega \nabla \bar{u} \cdot \nabla (\varphi_z v) \, dx \quad \forall v \in V
\]

\[
a_z(v_1, v_2) := \int_\Omega \varphi_z \nabla v_1 \cdot \nabla v_2 \, dx \quad \forall v_1, v_2 \in V
\]

Let \(V_z\) denote the completion of \(V\) under the norm given by the scalar product \(a_z\), when \(\varphi_z \neq 0\) on \(\Gamma\) or otherwise its quotient space with \(\mathbb{R}\), i.e.

\[
V_z = \begin{cases} 
\{v \in H^1_0(\omega_z) : a_z(v, v) < \infty, \varphi_z v = 0 \quad \text{on} \quad \Gamma \cap \partial \omega_z \} & \text{if} \quad \varphi_z \neq 0 \quad \text{on} \quad \Gamma \\
\{v \in H^1_0(\omega_z) : a_z(v, v) = \infty, \int_\Omega \varphi_z v \, dx = 0 \} & \text{if} \quad \varphi_z \equiv 0 \quad \text{on} \quad \Gamma
\end{cases}
\]

Notice that \(R_z(1) = 0\) for a free node \(z\) such that (88) has a unique solution and hence \(\eta_z < \infty\).

In practical applications, the solution \(e_z\) of (88) has to be approximated by some finite element approximation \(\tilde{e}_z\) on a discrete space \(\tilde{V}_z\) based on a finer mesh or of higher order. (Arguing as in the stability estimate (71), leads to an error estimate for an approximation \(\|\tilde{e}_z\|_{a_z}\) of \(\eta_z\).

Suppose that \(\eta_z\) is known exactly (or computed with high and controlled accuracy) and that the bilinear form \(a\) is localized through the partition of unity such that (e.g. in Example 10)

\[
a(u, v) = \sum_{z \in \mathcal{N}} a_z(u, v) \quad \forall u, v \in V \quad (89)
\]

Then the implicit error estimator \(\eta_L\) is reliable with \(C_{rel} = 1\) and \(h.o.t. = 0\),

\[
\|e\|_a \leq \eta_L := \left( \sum_{z \in \mathcal{N}} \eta_z^2 \right)^{1/2} \quad (90)
\]

The proof of (90) follows from the definition of \(R_z, \eta_z\), and \(e\) and Cauchy inequalities:

\[
\|e\|_a^2 = R(e) = \sum_{z \in \mathcal{N}} R_z(e) \leq \sum_{z \in \mathcal{N}} \eta_z \|e\|_{a_z}
\]

\[
\leq \left( \sum_{z \in \mathcal{N}} \eta_z^2 \right)^{1/2} \left( \sum_{z \in \mathcal{N}} \|e\|_{a_z}^2 \right)^{1/2} = \eta_L \|e\|_a
\]

Notice that \(\|\tilde{e}_z\|_{a_z} = \tilde{\eta}_z \leq \eta_z\) for any approximated local solution

\[
\tilde{e}_z \in \tilde{V}_z \quad \text{with} \quad a_z(\tilde{e}_z, \cdot) = R_z \in \tilde{V}_z^* \quad (91)
\]

and all of them are efficient estimators. The proof of efficiency is based on a weighted Poincaré or Friedrichs inequality which reads

\[
\|\varphi_z v\| \leq C \|v\|_{a_z} \quad \forall v \in V_z \quad (92)
\]

In fact, in Example 1, 2, and 3, one obtains efficiency in a more local form than indicated in

\[
\eta_z \leq C \|e\|_{a_z} \quad \text{with} \quad h.o.t. = 0 \quad (93)
\]

(This follows immediately from (92):

\[
R_z(v) = R(\varphi_z v) = a(e, \varphi_z v) \leq \|e\|_{a_z} \|\varphi_z v\|_{a_z}
\]

\[
\leq C \|e\|_{a_z} \|v\|_{a_z}
\]

In the situation of Example 10, the estimator \(\eta_L\) dates back to Babuška and Miller (1987); the use of weights was established in Carstensen and Funken (1999/00). A reliable computable estimator \(\tilde{\eta}_L\) is introduced in Morin, Nochetto
and Siebert (2003a) based on a proper finite-dimensional space $\bar{V}_\chi$ of some piecewise quadratic polynomials on $\omega_\chi$.

### 5.3.2 Equilibration estimators

The nonoverlapping domain decomposition schemes employ artificial unknowns $g_T \in L_2(\partial T)^m$ for each $T \in T$ at the interfaces, which allow a representation of the form

$$ R(v) = \sum_{T \in \mathcal{T}} R_T(v) $$

where

$$ R_T(v) := \int_T f \cdot v \, dx - \int_T \tilde{p} : Du \, dx + \int_{\partial T} g_T \cdot v \, ds $$  (94)

Adopting the notation from Section 5.2.1, the new quantities $g_T$ satisfy

$$ g_{T_+} + g_{T_-} = 0 \quad \text{along } E = \partial T_+ \cap \partial T_- \in \partial \Omega $$

(where $T_\pm$ and $T$ denote neighboring element domains) to guarantee (94). (There are non-displayed modifications on any Neumann boundary edge $E \subset \partial \Omega$.) Moreover, the bilinear form $a$ is expanded in an elementwise form

$$ a(u, v) = \sum_{T \in \mathcal{T}} a_T(u, v) \quad \forall u, v \in V $$  (95)

Under the equilibration condition $R_T(c) = 0$ for all kernel functions $c$ (namely, the constant functions for the Laplace model problem), the resulting local problem reads

$$ e_T \in V_T \quad \text{with } a_T(e_T, \cdot) = R_T \in V_T^* $$  (96)

and is equivalent to the computation of

$$ \eta_T := \sup_{v \in V_T \setminus \{0\}} \frac{R_T(v)}{\|v\|_{ar}} = \|e_T\|_{ar} $$  (97)

The sum of all local contributions defines the reliable equilibration error estimator $\eta_{EQ}$,

$$ \|e\|_a \leq \eta_{EQ} := \left( \sum_{T \in \mathcal{T}} \eta_T^2 \right)^{1/2} $$  (98)

(The immediate proof of (98) is analogous to that of (90) and hence is omitted.)

**Example 11.** In the situation of Example 10 there holds $\eta_T < \infty$ if and only if either $\Gamma \cap \partial T$ has positive surface measure (with $V_T = \{v \in H^1(T) : v = 0 \text{ on } \Gamma \cap \partial T\}$) or otherwise $R_T(1) = 0$ (with $V_T = \{v \in H^1(T) : f_T \cdot v \, dx = 0\}$). Ladeveze and Leguillon (1983) suggested a certain choice of the interface corrections to guarantee this and even higher-order equilibrations are established. Details on the implementation are given in Ainsworth and Oden (2000); a detailed error analysis with higher-order equilibrations and the perturbation by a finite element simulation of the local problems with corrections can be found in Ainsworth and Oden (2000) and Babuška and Strouboulis (2001).

The error estimator $\eta = \eta_{EQ}$ is efficient in the sense of (69) with higher-order terms $\text{h.o.t.}(T)$ on $T$ that depend on the given data provided

$$ h_T^{1/2} \|g_T - \tilde{p}v_T\|_{L_2(E)} \leq C \|e\|_{ar} + h_{\text{h.o.t.}}(T) $$

for all $E \in \mathcal{E}(T)$  (99)

(Recall that $\mathcal{E}(T)$ denotes the set of edges or faces of $T$.) This stability property depends on the design of $g_T$; a positive example is given in Theorem 6.2 of Ainsworth and Oden (2000) for Example 1. Given Inequality (99), the efficiency of $\eta_T$ follows with standard arguments, for example, an integration by parts, a trace and Poincaré or Friedrichs inequality $h_T^{-1} \|v\|_{L_2(T)} + h_T^{-1/2} \|v\|_{L_2(\partial T)} \leq C \|v\|_{ar}$, for $v \in V_T$:

$$ R_T(v) = \int_T r_T \cdot v \, dx + \int_{\partial T} (g_T - \tilde{p}v) \cdot v \, ds $$

$$ \leq C \left( h_T \|r_T\|_{L_2(T)} + h_T^{1/2} \|g_T - \tilde{p}v\|_{L_2(\partial T)} \right) \|v\|_{ar} $$

followed by (79) and (99).

### 5.4 Multilevel error estimators

While the preceding estimators evaluate or estimate the residual of one finite element solution $u_H$, multilevel estimators concern at least two meshes $\mathcal{T}_H$ and $\mathcal{T}_h$ with associated discrete spaces $V_H \subset V_h \subset V$ and two discrete solutions $u_H = \tilde{u}$ and $u_h$. The interpretation is that $p_h$ is computed on a finer mesh (e.g., $\mathcal{T}_h$ is a refinement of $\mathcal{T}_H$) or that $p_h$ is computed with higher polynomial order than $p_H = \tilde{p}$.

#### 5.4.1 Error-reduction property and multilevel error estimator

Let $V_H \subset V_h \subset V$ denote two nested finite element spaces in $V$ with coarse and fine finite element solution $u_H = \tilde{u} \in V_H = \tilde{V}$ and $u_h \in V_h$ of the discrete problem (19), respectively, and with the exact solution $u$. Let $p_H = \tilde{p}$, $p_h$, and $p$ denote the respective fluxes and let $\|\cdot\|$ be a norm associated to the energy norm, for example,
a norm with \( p - \tilde{p} = \| u - \tilde{u} \|_a \) and \( p - p_h = \| u - u_h \|_a \). Then, the multilevel error estimator

\[
\eta_{ML} := \| p_h - p_H \| = \| u_h - u_H \|_a \tag{100}
\]

simply is the norm of the difference of the two discrete solutions. The interpretation is that the error \( \| p - p_h \| \) of the finer discrete solution is systematically smaller than the error \( \| e \|_a = \| p - p_H \| \) of the coarser discrete solution in the sense of an error-reduction property: For some constant \( \varrho < 1 \), there holds

\[
\| p - p_h \| \leq \varrho \| p - p_H \| \tag{101}
\]

Notice the bound \( \varrho \leq 1 \) for Galerkin errors in the energy norm (because of the best-approximation property). The point is that \( \varrho < 1 \) in (101) is bounded away from one. Then, the error-reduction property (101) immediately implies reliability and efficiency of \( \eta_{ML} \):

\[
(1 - \varrho) \| p - p_H \| \leq \eta_{ML} = \| p_h - p_H \| \leq (1 + \varrho) \| p - p_H \| \tag{102}
\]

(The immediate proof of (102) utilizes (101) and a simple triangle inequality.)

Four remarks on the error-reduction conclude this section: Efficiency of \( \eta_{ML} \) in the sense of (69) is robust in \( \varrho \to 1 \), but reliability is not: The reliability constant \( C_{\text{rel}} = (1 - \varrho)^{-1} \) in (68) tends to infinity as \( \varrho \) approaches 1.

Higher-order terms are invisible in (102): h.o.t. rel = 0 = h.o.t. eff. This is unexpected when compared to all the other error estimators and hence indicates that (101) should fail to hold for heavily oscillating right-hand sides.

The error-reduction property (101) is often observed in practice for fine meshes and can be monitored during the calculation. For coarse meshes and in the preasymptotic range, (101) may fail to hold.

The error-reduction property (101) is often called saturation assumption in the literature and frequently has a status of an unproven hypothesis.

### 5.4.2 Counterexample for error-reduction

The error-reduction property (101) may fail to hold even if \( f \) shows no oscillations: Figure 10 displays two triangulations, \( T_H \) and its refinement \( T_h \), with one and five free nodes, respectively. If the right-hand side is constant and if the problem has homogeneous Dirichlet conditions for the Poisson problem

\[
1 + \Delta u = 0 \quad \text{in } \Omega := (0, 1)^2 \quad \text{and} \quad u = 0 \quad \text{on } \partial \Omega
\]

then the corresponding \( P_1 \) finite element solutions coincide: \( u_{H} = u_{h} \). A direct proof is based on the nodal basis function

**Figure 10.** Two triangulations \( T_H \) and \( T_h \) with equal discrete \( P_1 \) finite element solutions \( u_H = u_h \) for a Poisson problem with right-hand side \( f = 1 \). The refinement \( T_h \) of \( T_H \) is generated by two (newest-vertex) bisections per element (initially with the interior node in \( T_H \) as newest vertex).

The error-reduction property (101) may fail to hold even if \( \varrho \) is reliability and efficiency of \( \eta_{ML} \):

\[
(1 - \varrho) \| p - p_H \| \leq \eta_{ML} = \| p_h - p_H \| \leq (1 + \varrho) \| p - p_H \| \tag{102}
\]

Thus \( u_H \) is the finite element solution in \( W_h := \text{span}\{T_1, \Phi_2\} \subset V_h \). Since, by symmetry, the finite element solution \( u_h \) in \( V_h \) belongs to \( W_h \), there holds \( u_H = u_h \).

#### 5.4.3 Affirmative example for error-reduction

Adopt notation from Section 5.2.4 with a coarse discrete space \( V_H = \tilde{V} \) and consider the fine space \( V_h := V_H \oplus W_h \) for

\[
W_h := \text{span}\{r_T b_T : T \in T\} \oplus \text{span}\{r_E b_E : E \in E_{\Omega}\} \subset V \tag{103}
\]

Then there holds the error-reduction property up to higher-order terms

\[
\text{osc}(f) := \left( \sum_{T \in \mathcal{T}} h_T^2 \| f - \tilde{f} \|^2_{L_2(T)} \right)^{1/2}
\]

namely

\[
\| p - p_h \|^2 \leq \varrho \| p - p_H \|^2 + \text{osc}(f)^2 \tag{104}
\]

The constant \( \varrho \) in (104) is uniformly smaller than one, independent of the mesh size, and depends on the shape of the elements and the type of ansatz functions through constants in (81)–(83).
The proof of (104) is based on the test functions \( w_T \) and \( w_E \) in (84) and (86) of Section 5.2.4 and equation: 

\[
w_h := \sum_{T \in \mathcal{T}} w_T + \sum_{E \in \mathcal{E}_h} w_E \in W_h \subset V
\]

Utilizing (81)–(83) one can prove the equality:

\[
\|w_h\|_a^2 \leq C \left( \sum_{T \in \mathcal{T}} h_T^2 \|\tilde{r}_T\|_{L^2(T)}^2 + \sum_{E \in \mathcal{E}_h} h_E^2 \|r_E\|_{L^2(E)}^2 \right)
\]

\[
\leq C \left( \sum_{T \in \mathcal{T}} R(w_T) + \sum_{E \in \mathcal{E}_h} R(w_E) + \text{osc}(f)^2 \right)
\]

\[
= C (R(w_h) + \text{osc}(f)^2)
\]

Since \( w_h \) belongs to \( V_h \) and \( u_h \) is the finite element solution with respect to \( V_h \) there holds

\[
R(w_h) = a(u_h - u_H, w_h) \leq \|u_h - u_H\|_a \|w_h\|_a
\]

The combination of the preceding inequalities yields the key inequality

\[
\eta_{R,R}^2 = \sum_{T \in \mathcal{T}} h_T^2 \|\tilde{r}_T\|_{L^2(T)}^2 + \sum_{E \in \mathcal{E}_h} h_E^2 \|r_E\|_{L^2(E)}^2
\]

\[
\leq C \|u_h - u_H\|_a^2 + \text{osc}(f)^2
\]

This, the Galerkin orthogonality \( \|u - u_H\|_a^2 = \|u - u_h\|_a^2 + \|u_h - u_H\|_a^2 \), and the reliability of \( \eta_R \) show

\[
\|u - u_H\|_a^2 \leq C \text{C}_{\text{rel}}^2 (\|u - u_H\|_a^2 + \|u - u_h\|_a^2 + \text{osc}(f)^2)
\]

and so imply (104) with \( q = 1 - C^{-1} \text{C}_{\text{rel}}^2 < 1 \).

Example 12. In the Poisson problem with \( P_1 \) finite elements and discrete space \( V_H \), let \( W_h \) consist of the quadratic and cubic bubble functions (80). Then (104) holds; cf. also Example 14 below.

Example 13. Other affirmative examples for the Poisson problem consist of the \( P_1 \) and \( P_2 \) finite element spaces \( V_H \) and \( V_h \) over one regular triangulation \( \mathcal{T} \) or of the \( P_1 \) finite element spaces with respect to a regular triangulation \( \mathcal{T}_H \) and its red-refinement \( \mathcal{T}_h \). The observation that the element-bubble functions are in fact redundant is due to Dörfler and Nochetto (2002).

5.4.4 Hierarchical error estimator

Given a smooth right-hand side \( f \) and based on the example of the previous section, the multilevel estimator (100) is reliable and efficient up to higher-order terms. The costly calculation of \( u_h \), however, exclusively allows for an accurate error control of \( u_H \) (and no reasonable error control for \( u_h \)). Instead of (100), cheaper versions are favored where \( u_h \) is replaced by some quantity computed by a localized problem. One reliable and efficient version is the hierarchical error estimator

\[
\eta_H := \left( \sum_{T \in \mathcal{T}} \eta_T^2 + \sum_{E \in \mathcal{E}} \eta_E^2 \right)^{1/2}
\]

where, for each \( T \in \mathcal{T} \) and \( E \in \mathcal{E} \) and their test functions (84) and (86),

\[
\eta_T := \frac{R(w_T)}{\|w_T\|_a} \quad \text{and} \quad \eta_E := \frac{R(w_E)}{\|w_E\|_a}
\]

(The proof of reliability and efficiency follows from (81)–(83) by the arguments from Section 5.2.4.)

Example 14. In the Poisson problem with \( P_1 \) finite elements and discrete space \( V_H \), let \( W_h \) consist of the quadratic and cubic bubble functions (80). Then,

\[
\eta_H := \left( \sum_{T \in \mathcal{T}} \frac{R(b_T)^2}{\|b_T\|_a^2} + \sum_{E \in \mathcal{E}_h} \frac{R(b_E)^2}{\|b_E\|_a^2} \right)^{1/2}
\]

is a reliable and efficient hierarchical error estimator. With the error-reduction property of \( P_1 \) and \( P_2 \) finite elements due to Dörfler and Nochetto (2002),

\[
\eta_H := \left( \sum_{E \in \mathcal{E}_h} \frac{R(b_E)^2}{\|b_E\|_a^2} \right)^{1/2}
\]

is reliable and efficient as well. The same is true if, for each edge \( E \in \mathcal{E} \), \( b_E \) defines a hat function of the midpoint of \( E \) with respect to a red-refinement \( \mathcal{T}_h \) of \( \mathcal{T}_H \) (that is, each edge is halved and each triangle is divided into four congruent subtriangles; cf. Figure 15, left).

5.5 Averaging error estimators

Averaging techniques, also called (gradient) recovery estimators, focus on one mesh and one known low-order flux approximation \( \tilde{p} \) and the difference to a piecewise polynomial \( \hat{q} \) in a finite-dimensional subspace \( \mathcal{Q} \subset L^2(\Omega; \mathbb{R}^{m \times n}) \) of higher polynomial degrees and more restrictive continuity conditions than those generally satisfied by \( \tilde{p} \). Averaging techniques are universal in the sense that there is no need for any residual or partial differential equation in order to apply them.
5.5.1 Definition of averaging error estimators

The procedure consists of taking a piecewise smooth \( \tilde{p} \) and approximating it by some globally continuous piecewise polynomials (denoted by \( \bar{Q} \)) of higher degree \( A \tilde{p} \). A simple example, frequently named after Zienkiewicz and Zhu and sometimes even called the ZZ estimator, reads as follows: For each node \( z \in \mathcal{N} \) and its patch \( \omega_z \), let

\[
(A \tilde{p})(z) = \int_{\omega_z} \tilde{p} \, dx \quad \in \mathbb{R}^{m \times n}
\]

be the integral mean of \( \tilde{p} \) over \( \omega_z \). Then, define \( A \tilde{p} \) by interpolation with (conforming, i.e. globally continuous) hat functions \( \varphi_z \), for \( z \in \mathcal{N} \),

\[
A \tilde{p} = \sum_{z \in \mathcal{N}} (A \tilde{p})(z) \varphi_z \in \bar{Q}
\]

Let \( \bar{Q} = \text{span}\{\varphi_z : z \in \mathcal{N}\} \) denote the (conforming) first-order finite element space and let \( \| \cdot \| \) be the norm for the fluxes. Then the averaging estimator is defined by

\[
\eta_A := \| \tilde{p} - A \tilde{p} \|
\]

Notice that there is a minimal version

\[
\eta_M := \min_{\tilde{q} \in \bar{Q}} \| \tilde{p} - \tilde{q} \| \leq \eta_A
\]

The efficiency of \( \eta_M \) follows from a triangle inequality, namely

\[
\eta_M \leq \| p - \tilde{p} \| + \| p - \tilde{q} \| \quad \text{for all} \quad \tilde{q} \in \bar{Q}
\]

and the fact that \( \| p - \tilde{p} \| = \mathcal{O}(h) \) while (in all the examples of this chapter)

\[
\min_{\tilde{q} \in \bar{Q}} \| p - \tilde{q} \| = \text{h.o.t.}(p) =: \text{h.o.t.}_{\text{eff}}
\]

This is of higher order for smooth \( p \) and efficiency follows for \( \eta = \eta_M \) and \( C_{\text{eff}} = 1 \).

It turns out that \( \eta_A \) and \( \eta_M \) are very close and accurate estimators in many numerical examples; cf. Section 5.6.4 below. This and the fact that the calculation of \( \eta_A \) is an easy postprocessing made \( \eta_A \) extremely popular.

For proper treatment of Neumann boundary conditions, we refer to Carstensen and Bartels (2002) and for applications in solid and fluid mechanics to Alberty and Carstensen (2003) and Carstensen and Funken (2001a,b) and for higher-order FEM to Bartels and Carstensen (2002).

Multigrid smoothing steps may be successfully employed as averaging procedures as proposed in Bank and Xu (2003).

5.5.2 All averaging error estimators are reliable

The first proof of reliability dates back to Rodriguez (1994a,b) and we refer to Carstensen (2004) for an overview. A simplified reliability proof for \( \eta_M \) and hence for all averaging techniques (Carstensen, Bartels and Klose, 2001) is outlined in the sequel.

First let \( \Pi \) be the \( L_2 \) projection onto the first-order finite element space \( \bar{V} \) and let \( \tilde{q} \) be arbitrary in \( \bar{Q} \), that is, each of the \( m \times n \) components of \( L \tilde{q} \) is a first-order finite element function in \( FE_T \). The Galerkin orthogonality shows for the error \( e := u - \bar{u} \) and \( \tilde{p} := \nabla \bar{u} \) in the situation of Example 1 that

\[
\|e\|_2^2 = \int_{\Omega} (\nabla u - \tilde{q}) \cdot (\nabla (e - \Pi e)) \, dx
\]

\[
+ \int_{\Omega} (\tilde{q} - \tilde{p}) \cdot (\nabla (e - \Pi e)) \, dx
\]

A Cauchy inequality in the latter term is combined with the \( H^1 \)-stability of \( \Pi \), namely,

\[
\| \nabla (e - \Pi e) \|_{L_2(\Omega)} \leq C_{\text{stab}} \| \nabla e \|_{L_2(\Omega)}
\]

(For sufficient conditions for this we refer to Crouzeix and Thomée (1987), Bramble, Pasciak and Steinbach (2002), Carstensen (2002, 2003b), and Carstensen and Verfürth (1999).) The \( H^1 \)-stability in the second term and an integration by parts in the first term on the right-hand side show

\[
\|e\|_2^2 \leq \int_{\Omega} f \cdot (e - \Pi e) \, dx + \int_{\Omega} (e - \Pi e) \cdot \div \tilde{q} \, dx
\]

\[
+ C_{\text{stab}} \| \nabla e \|_{L_2(\Omega)} \| \tilde{p} - \tilde{q} \|_{L_2(\Omega)}
\]

Since \( e - \Pi e \) is \( L_2 \)-orthogonal onto \( f_h := \Pi f \in \bar{V} \),

\[
\int_{\Omega} f \cdot (e - \Pi e) \, dx
\]

\[
= \int_{\Omega} (f - f_h) \cdot (e - \Pi e) \, dx
\]

\[
\leq \|h_T^{-1}(e - \Pi e)\|_{L_2(\Omega)} \|h_T(f - \Pi f)\|_{L_2(\Omega)}
\]

Notice that, despite possible boundary layers, \( \|h_T(f - \Pi f)\|_{L_2(\Omega)} = \text{h.o.t.} \) is of higher order. The first-order approximation property of the \( L_2 \) projection,

\[
\|h_T^{-1}(e - \Pi e)\|_{L_2(\Omega)} \leq C_{\text{approx}} \| \nabla e \|_{L_2(\Omega)}
\]
follows from the \( H^1 \)-stability (cf. e.g. Carstensen and Verfürth (1999) for a proof). Similar arguments for the remaining term \( \text{div} \, \tilde{q} \) and the \( \mathcal{T} \)-piecewise divergence operator \( \text{div}_\mathcal{T} \) with \( \text{div}_\mathcal{T} \tilde{q} = \text{div}_\mathcal{T} \tilde{q} - \tilde{p} \) (recall that \( \tilde{u} \) is of first-order and hence \( \Delta \tilde{u} = 0 \) on each element) lead to

\[
\int_\Omega (e - \Pi e) \cdot \text{div} \, \tilde{q} \, dx \leq C_{\text{approx}} \| \nabla e \|_{L^2(\Omega)} \| \mathcal{H}_\mathcal{T} \text{div}_\mathcal{T} (\tilde{q} - \tilde{p}) \|_{L^2(\Omega)}
\]

An inverse inequality \( \mathcal{H}_\mathcal{T} \text{div}_\mathcal{T} (\tilde{q} - \tilde{p}) \|_{L^2(\mathcal{T})} \leq C_{\text{inv}} \| \tilde{q} - \tilde{p} \|_{L^2(\mathcal{T})} \) (cf. Section 3.4) shows

\[
\int_\Omega (e - \Pi e) \cdot \text{div} \, \tilde{q} \, dx \leq C_{\text{approx}} C_{\text{inv}} \| \nabla e \|_{L^2(\Omega)} \| \tilde{q} - \tilde{p} \|_{L^2(\Omega)}
\]

The combination of all established estimates plus a division by \( \| e \|_a = \| \nabla e \|_{L^2(\Omega)} \) yield the announced reliability result

\[
\| e \|_a \leq (C_{\text{stab}} + C_{\text{approx}} C_{\text{inv}}) \| \tilde{p} - \tilde{q} \|_{L^2(\Omega)} + \text{h.o.t.}
\]

In the second step, one designs a more local approximation operator \( J \) to substitute \( \Pi \) as in Carstensen and Bartels (2002); the essential properties are the \( H^1 \)-stability, the first-order approximation property, and a local form of the orthogonality condition; we omit the details.

### 5.5.3 Averaging error estimators and edge contributions

There is a local equivalence of the estimators \( \eta_A \) from a local averaging process (109) and the edge estimator

\[
\eta_E := \left( \sum_{E \in \mathcal{E}_1} h_E \| \tilde{p} \nabla E \|_{L^2(E)}^2 \right)^{1/2}
\]

The observation that, with some mesh-size-independent constant \( C \),

\[
C^{-1} \eta_E \leq \eta_A \leq C \eta_E
\]

dates back to Rodriguez (1994a) and can be found in Verfürth (1996). The proof of (113) for piecewise linears in \( \mathcal{V} \) is based on the equivalence of the two seminorms

\[
\eta_1(\tilde{q}) := \min_{r \in \mathbb{R}^{m \times n}} \| \tilde{q} - r \|_{L^2(\omega)}
\]

and

\[
\eta_2(\tilde{q}) := \left( \sum_{E \in \mathcal{E}_z} h_E \| \tilde{q} \nabla E \|_{L^2(E)}^2 \right)^{1/2}
\]

for piecewise constant vector-valued functions \( \tilde{q} \) in \( P_1 \), the set of possible fluxes \( \tilde{q} \) restricted on the patch \( \omega_z \) of a node \( z \), and with the set of edges \( \mathcal{E}_z := \{ E \in \mathcal{E} : z \in E \} \). The main observation is that \( \eta_1 \) and \( \eta_2 \) vanish exactly for constants functions \( \mathbb{R}^{m \times n} \) and hence they are norms on the quotient space \( P_1 / \mathbb{R}^{m \times n} \). By the equivalence of norms on any finite-dimensional space \( P_1 / \mathbb{R}^{m \times n} \), there holds

\[
C^{-1} \eta_1(\tilde{q}) \leq \eta_2(\tilde{q}) \leq C \eta_1(\tilde{q}) \quad \forall \tilde{q} \in P_1
\]

This is a local version of (113), which eventually implies (113) by localization and composition; we omit the details.

For triangles and tetrahedra and piecewise linear finite element functions, it is proved in Carstensen (2004) that

\[
\eta_M \leq \eta_A \leq C_d \eta_M
\]

with universal constants \( C_2 = \sqrt{10} \) and \( C_3 = \sqrt{15} \) for 2-D and 3-D, respectively. This equivalence holds for a larger class of elements and (first-order) averaging operators and then proves efficiency for \( \eta_A \) whereas efficiency of \( \eta_M \) follows from a triangle inequality in (112).

### 5.6 Comparison of error bounds in benchmark example

This section is devoted to a numerical comparison of energy errors and its a posteriori error estimators for an elliptic model problem.

#### 5.6.1 Benchmark example

The numerical comparisons are computed for the Poisson problem

\[
1 + \Delta u = 0 \text{ in } \Omega \quad \text{and} \quad u = 0 \text{ on } \partial \Omega
\]

on the \( L \)-shaped domain \( \Omega = (-1, +1)^2 \setminus ([0, 1] \times [-1, 0]) \) and its boundary \( \partial \Omega \). The first mesh \( \mathcal{T}_1 \) consists of 17 free nodes and 48 elements and is obtained by, first, a decomposition of \( \Omega \) in 12 congruent squares of size 1/2 and, second, a decomposition of each of the boxes along its two diagonals into 4 congruent triangles. The subsequent meshes are successively red-refined (i.e. each triangle is partitioned into 4 congruent subtriangles of Figure 15, left). This defines the (conforming) \( P_1 \) finite element spaces \( V_1 \subset V_2 \subset V_3 \subset \cdots \subset V := H^1_0(\Omega) \). The error \( e := u - u_j \) of the finite element solution \( u_j = \bar{u} \in V_j \subset \mathcal{V} \) of dimension \( N = \text{dim}(V_j) \) is measured in the energy norm (the Sobolev seminorm in \( H^1(\Omega) \))

\[
|e|_{1,2} := |e|_{H^1(\Omega)} := \left( \int_\Omega |D e|^2 \, dx \right)^{1/2} = a(u - \bar{u}, u + \bar{u})^{1/2} = \left( |u|^2_{H^1(\Omega)} - |\bar{u}|^2_{H^1(\Omega)} \right)^{1/2}
\]
Figure 11. Experimental results for the benchmark problem (115) with meshes $T_1, \ldots, T_6$. The relative error $|e|_{1,2}/|u|_{1,2}$ and various estimators $\eta/|u|_{1,2}$ are plotted as functions of the number of degrees of freedom $N$. Both axes are in a logarithmic scaling such an algebraic curve $N^{-a}$ is visible as a straight line with slope $-a$. A triangle with slope $-0.33$ is displayed for comparison. A color version of this image is available at http://www.mrw.interscience.wiley.com/ecm

(by the Galerkin orthogonality) computed with the approximation $|u|^2_{H^1(\Omega)} \approx 0.21407315683398$.

Figure 11 displays the computed values of $|e|_{1,2}$ for a sequence of uniformly refined meshes $T_1, T_2, \ldots, T_6$ as a function of the number of degrees of freedom $N = 17, 81, 353, 1473, 6017, 24321, 97793$. It is this curve that will be estimated by computable upper and lower bounds explained in the subsequent sections where we use the fact that each element is a right isosceles triangle.

Notice that the two axes in Figure 11 scale logarithmically such that any algebraic curve of growth $a$ is mapped into a straight line of slope $-a$. The experimental convergence rate is $2/3$ in agreement with the (generic) singularity of the domain and resulting theoretical predictions. More details can be found in Carstensen, Bartels and Klose (2001).

5.6.2 Explicit error estimators

For the benchmark problem of Section 5.6.1, the error estimator (75) can be written in the form

$$\eta_{R,R} := \left( \sum_{T \in T} h_T^2 \|1\|_{L^2(T)}^2 \right)^{1/2} + \left( \sum_{E \in E_0} h_E \int_E \left( \frac{\partial u}{\partial v_E} \right)^2 \, dx \right)^{1/2}$$

(116)

and is reliable with $C_{rel} = 1$ and h.o.t.$_{rel} = 0$ (Carstensen and Funken, 2001a). Figure 11 displays $\eta_{R,R}$ as a function of the number of unknowns $N$ and illustrates $|e|_{1,2} \leq \eta_{R,R}$.

Guaranteed lower error bounds, i.e. with the efficiency constant $C_{eff}$, are less established and the higher-order terms h.o.t.$_{eff}$ usually involve $\|f - f_T\|_{L^2(T)}$ for the elementwise integral mean $f_T$ of the right-hand side. Here, $f \equiv 1$ and so h.o.t.$_{eff} = 0$. Following a derivation in Carstensen, Bartels and Klose (2001), Figure 11 also displays an efficient variant $\eta_{R,E}$ as a lower error bound: $\eta_{R,E} \leq |e|_{1,2}$.

The guaranteed lower and upper bounds with explicit error estimators leave a very large region for the true error. Our interpretation is that various geometries (the shape of the patches) lead to different constants and $C_{rel} = 1$ reflects the worst possible situation for every patch in the current mesh.

A more efficient reliable explicit error estimator $\eta_{R,C}$ from Carstensen and Funken (2001a) displayed in Figure 11 requires the computation of local (patchwise) analytical eigenvalues and hence is very expensive. However, the explicit estimators $\eta_{R,C}$ and $\eta_{R,E}$ still overestimate and underestimate the true error by a huge factor (up to 10 and even more) in the simple situation of the benchmark. One conclusion is that the involved constants estimate a worst-case scenario with respect to every right-hand side or every exact solution.

This experimental evidence supports the design of more elaborate estimators: The stopping criterion (72) with the reliable explicit estimators may appear very cheap and easy. But the decision (72) may have too costly consequences.

5.6.3 Implicit estimators

For comparison, the two implicit estimators $\eta_L$ and $\eta_EQ$ are displayed in Figure 11 as functions of $N$. It is stressed that both estimators are efficient and reliable (Carstensen and Funken, 1999/00)

$$|e|_{1,2} \leq \eta_L \leq 2.37 |e|_{1,2}$$

The practical performance of $\eta_L$ and $\eta_EQ$ in Figure 11 is comparable and in fact is much sharper than that of $\eta_{R,E}$ and $\eta_{R,R}$.

5.6.4 Averaging estimator

The averaging estimators $\eta_A$ and $\eta_M$ are as well displayed in Figure 11 as a function of $N$. Here, $\eta_M$ is efficient up to higher-order terms (since the exact solution $u \in H^{5/3-\varepsilon}(\Omega)$ is singular, this is not really guaranteed) while its reliability is open, i.e. the corresponding constants have not been computed. Nevertheless, the behavior of $\eta_A$ and $\eta_M$ is exclusively seen here from an experimental point of view.
The striking numerical result is an amazing high accuracy of $\eta_M \approx \eta_A$ as an empirical guess of $|e|_{1,2}$. If we took $C_{\text{rel}}$ into account, this effect would be destroyed: The high accuracy is an empirical observation in this example (and possibly many others) but does not yield an accurate guaranteed error bound.

5.6.5 Adapted meshes

The benchmark in Figure 11 is based on a sequence of uniform meshes and hence results in an experimental convergence rate 2/3 according to the corner singularity of this example. Adaptive mesh-refining algorithms, described below in more detail, are empirically studied also in Carstensen, Bartels and Klose (2001). The observations can be summarized as follows: The quality of the estimators and their relative accuracy is similar to what is displayed in Figure 11 even though the convergence rates are optimally improved to one.

5.7 Goal-oriented error estimators

This section provides a brief introduction to goal-oriented error control.

5.7.1 Goal functionals

Given the Sobolev space $V$, for example, $V_0^2$, we can define a linear form $\eta: V \to \mathbb{R}$ as an empirical guess of $|e|_{1,2}$.

5.7.2 Duality technique

To bound or approximate $\eta$ one considers the dual problem

$$a(v, z) = J(v) \quad \forall v \in V$$

with exact solution $z \in V$ (guaranteed by the Lax–Milgram lemma) and the discrete solution $\tilde{z} \in \tilde{V}$ of

$$a(\tilde{v}, \tilde{z}) = J(\tilde{v}) \quad \forall \tilde{v} \in \tilde{V}$$

Set $f := z - \tilde{z}$. On the basis of the Galerkin orthogonality $a(e, \tilde{z}) = 0$ one infers

$$J(e) = a(e, z) = a(e, z - \tilde{z}) = a(e, f)$$

As a result of (119) and the boundedness of $a$ one obtains the a posteriori estimate

$$|J(e)| \leq \|a\| \|e\|_V \|f\|_V \leq 2 \eta_u \eta_z$$

Indeed, utilizing the primal and dual residual $R_u$ and $R_z$ in $V^*$, defined by

$$R_u := F - a(\tilde{u}, \cdot) \quad \text{and} \quad R_z := J - a(\cdot, \tilde{z})$$

computable upper error bounds for $\|e\|_V \leq \eta_u$ and $\|f\|_V \leq \eta_z$ can be found by the arguments of the energy error estimators of the previous sections. This yields a computable upper bound $\|a\| \eta_u \eta_z$ for $|J(e)|$ which is global, that is, the interaction of $e$ and $f$ is not reflected. One might therefore speculate that the upper bound is often too coarse and inappropriate for goal-oriented adaptive mesh refinement.

5.7.3 Upper and lower bounds of $J(e)$

In many cases, the main interest is on a point value and then $J(e)$ is given by a mollification $\rho$ of a singular measure in order to guarantee the boundedness of $J : V \to \mathbb{R}$.

$$2J(e) = 2a(e, f) = \|e + f\|_a^2 - \|e\|_a^2 - \|f\|_a^2$$

This right-hand side can be written in terms of residuals, in the spirit of (70), namely, $\|e\|_a = \|\text{Res}_u\|_{V^*}$, $\|f\|_a = \|\text{Res}_z\|_{V^*}$, and

$$\|e + f\|_a = \|\text{Res}_{u+z}\|_{V^*}$$

for $\text{Res}_{u+z} := F + J - a(\tilde{u} + \tilde{z}, \cdot)$

$$= \text{Res}_u + \text{Res}_z \in V^*$$
Therefore, the estimation of \( J(e) \) is reduced to the computation of lower and upper error bounds for the three residuals \( \text{Res}_u, \text{Res}_z, \) and \( \text{Res}_{u+z} \) with respect to the energy norm. This illustrates that the energy error estimation techniques of the previous sections may be employed for goal-oriented error control.

For more details and examples of a refined estimation see Ainsworth and Oden (2000) and Babuška and Strouboulis (2001).

### 5.7.4 Computing an approximation to \( J(e) \)

An immediate consequence of (119) is

\[
J(e) = R(z)
\]

and hence \( J(e) \) is easily computed once the dual solution \( z \) of (118) is known or at least approximated to sufficient accuracy. An upper error bound for \( |J(e)| = |R(z)| \) is obtained following the methodology of Becker and Rannacher (1996, 2001) and Bangerth and Rannacher (2003).

To outline this methodology, consider the residual representation formula (73) following the notation of Section 5.2.1. Suppose that \( z \in H^2(\Omega) \) (e.g., for a \( H^2 \) regular dual problem) and let \( I_z \) denote the nodal interpolant in the lowest-order finite element space \( \hat{V} \). With some interpolation constant \( C_I > 0 \), there holds, for any element \( T \in T \),

\[
h_T^{-2} \| z - I_z \|_{L^2(T)} + h_T^{-3/2} \| z - I_z \|_{L^2(\partial T)} \leq C_I |z|_{H^2(T)}
\]

The combination of this with (73) shows

\[
J(e) = \text{Res}(z - I_z) = \sum_{T \in T} \int_T r_T \cdot (z - I_z) \, dx - \sum_{E \in \mathcal{E}} \int_E r_E \cdot (z - I_z) \, ds \\
\leq \sum_{T \in T} \left( \| r_T \|_{L_2(T)} \| z - I_z \|_{L_2(T)} + \| r_E \|_{L_2(\partial T)} \| z - I_z \|_{L_2(\partial T)} \right) \\
\leq \sum_{T \in T} C_I \left( h_T^{-2} \| r_T \|_{L_2(T)} + h_T^{-3/2} \| r_E \|_{L_2(\partial T)} \right) |z|_{H^2(T)}
\]

The influence of the goal functional in this upper bound is through the unknown \( H^2 \)-seminorm \( |z|_{H^2(T)} \), which is to be replaced by some discrete analog based on a computed approximation \( \tilde{z}_h \). The justification of some substitute \( |D_2^2 \tilde{z}_h|_{L^2(T)} \) (postprocessed with some averaging technique) for \( |z|_{H^2(T)} \) is through striking numerical evidence; we refer to Becker and Rannacher (2001) and Bangerth and Rannacher (2003) for details and numerical experiments.

### 6 LOCAL MESH REFINEMENT

This section is devoted to the mesh-design task in the finite element method based on a priori and a posteriori information. Examples of the former type are graded meshes or geometric meshes with an a priori choice of refinement toward corner singularities briefly mentioned in Section 6.1. Examples of the latter type are adaptive algorithms for automatic mesh refinement (or mesh coarsening) strategies with a successive call of the steps

\[
\text{SOLVE} \Rightarrow \text{ESTIMATE} \Rightarrow \text{MARK} \Rightarrow \text{REFINE/COARSESEN}
\]

Given the current triangulation, one has to compute the finite element solution in step SOLVE; cf. Section 7.8 for a MATLAB realization of that. The accuracy of this finite element approximation is checked in the step ESTIMATE. On the basis of the refinement indicators of Section 6.2 the step MARK identifies the elements, edges or patches in the current mesh in need of refinement (or coarsening). The new data structure is generated in step REFINE/COARSESEN where a partition is given and a closure algorithm computes a triangulation described in Section 6.3. The convergence and optimality of the adaptive algorithm is discussed in Section 6.4. Brief remarks on coarsening strategies in Section 6.5 conclude the section.

### 6.1 A priori mesh design

The singularities of the exact solution of the Laplace equation on domains with corners (cf. Figure 1) are reasonably well understood and motivate the (possibly anisotropic) mesh refinement toward vertices or edges. This section aims a short introduction for two-dimensional \( P_1 \) finite elements – Chapter 3, this Volume, will report on three-dimensional examples.

Given a polygonal domain with a coarse triangulation into triangles (which specify the geometry), macro elements can be used to fill the domain with graded meshes. Figure 12(a) displays a macro element described in the sequel while Figure 12(b) illustrates the resulting fine mesh for an \( L \)-shaped domain.

The description is restricted to the geometry on the reference element \( T_{\text{ref}} \) with vertices \((0,0), (1,0), \) and \((0,1)\) of Figure 12(a). The general situation is then obtained by an affine transformation illustrated in Figure 12(b). The macro element \( T_{\text{ref}} \) is generated as follows: Given a grading parameter \( \beta > 0 \) for a grading function \( g(t) = t^\beta \), and given a natural number \( N \), set \( \xi_j := g(j/N) \) and draw line segments aligned to the antidiagonal through \((0, \xi_j)\) and \((\xi_j, 0)\) for \( j = 0, 1, \ldots, N \). Each of these segments
The geometric mesh depicted in Figure 13 yields a finer refinement toward some corners of the polygonal domain. The nodes $ξ_0 := 0$ and $ξ_j := β^{N−j}$ for $j = 1, \ldots, N$ define antidiagonals through $(ξ_j, 0)$ and $(0, ξ_j)$, which are in turn bisected. For such a triangulation, the polynomial degrees $p_T$ on each triangle $T$ are distributed as follows: $p_T = 1$ for the two triangles $T$ with vertex $(0,0)$ and $p_T = j + 2$ for the four elements in the convex quadrilateral with the vertices $(ξ_j, 0)$, $(0, ξ_j)$, $(ξ_{j+1}, 0)$, and $(0, ξ_{j+1})$ for $j = 0, \ldots, N − 1$. Figure 14 compares experimental convergence rates of the error in $H^{−1}$-seminorm $|e|_{H}$ for various graded meshes for the $P_1$ finite element method, the $p$- and $hp$-finite element method, and for the adaptive algorithm of Section 6.4. The $P_1$ finite element method on graded meshes with $β = 3/2$, $β = 2$ and $h$-adaptivity recover optimality in the convergence rate as opposite to the uniform refinement ($β = 1$), leading only to a sub-optimal convergence due to the corner singularity. The $hp$-finite element method performs better convergence rate compared to the $p$-finite element method.

Tensor product meshes are more appropriate for smaller values of $β$; the one-dimensional model analysis of Babuška and Guo (1986) suggests $β = (2)^{1/2} − 1 ≈ 0.171573$.

### 6.2 Adaptive mesh-refining algorithms

The automatic mesh refinement for regular triangulations called MARK and REFINE/COARSEN frequently consists of three stages:

(i) the marking of elements or edges for refinement (or coarsening);
(ii) the closure algorithm to ensure that the resulting triangulation is (or remains) regular;
(iii) the refinement itself, i.e. the change of the underlying data structures.
This section will focus on the marking strategies (i) while the subsequent section addresses the closure algorithm (ii) and the refinement procedure (iii).

In a model situation with a sum over all elements \( T \in \mathcal{T} \) (or over all edges, faces, or nodes), the a posteriori error estimators of the previous section give rise to a lower or upper error bound \( \eta \) in the form

\[
\eta = \left( \sum_{T \in \mathcal{T}} \eta_T^2 \right)^{1/2}
\]

Then, the marking strategy is an algorithm, which selects a subset \( \mathcal{M} \) of \( \mathcal{T} \) called the marked elements; these are marked with the intention of being refined during the later refinement algorithm.

A typical algorithm computes a threshold \( L \), a positive real number, and then utilizes the refinement rule or marking criterion

\[
\text{mark } T \in \mathcal{T} \quad \text{if} \quad L \leq \eta_T
\]

Therein, \( \eta_T \) is referred to as the refinement indicator whereas \( L \) is the threshold; that is,

\[
\mathcal{M} := \{ T \in \mathcal{T}; L \leq \eta_T \}
\]

Typical examples for the computation of a threshold \( L \) are the maximum criterion

\[
L := \Theta \max \{ \eta_T; T \in \mathcal{T} \}
\]

or the bulk criterion where \( L \) is the largest value such that

\[
(1 - \Theta)^2 \sum_{T \in \mathcal{T}} \eta_T^2 \leq \sum_{T \in \mathcal{M}} \eta_T^2
\]

The parameter \( \Theta \) is chosen with \( 0 \leq \Theta \leq 1 \); \( \Theta = 0 \) corresponds to an almost uniform refinement and \( \Theta = 1 \) to a raw refinement of just a few elements (no refinements in the bulk criterion).

The justification of the refinement criteria is essentially based on the heuristic of an equi-distribution of the refinement indicator; see Babuška and Rheinboldt (1978, 1979) and Babuška and Vogelius (1984) for results in 1-D. A rigorous justification for some class of problems started with Dörfler (1996); it is summarized in Morin, Nochetto and Siebert (2003b) and will be addressed in Section 6.4.

A different strategy is possible if the error estimator gives rise to a quantitative bound of a new meshsize. For instance, the explicit error estimator can be rewritten as \( \eta_R := \| h_T R \|_{L^2(\Omega)} \) with a given function \( R \in L^2(\Omega) \) and the local mesh size \( h_T \) (when edge contributions are recast as volume contributions). Then, given a tolerance \( Tol \) and using the heuristic that \( R \) would not change (at least not dramatically increase) during a refinement, the new local mesh size \( h_{\text{new}} \) can be calculated from the condition

\[
\| h_{\text{new}} R \|_{L^2(\Omega)} = Tol
\]

upon the equi-distribution hypothesis \( h_{\text{new}} \propto Tol/R \). Another approach that leads to a requested mesh-size distribution is based on sharp a priori error bounds, such as \( \| h_T D^2 u \|_{L^2(\Omega)} \) where \( D^2 u \) denotes the matrix of all second derivatives of the exact solution \( u \). Since \( D^2 u \) is unknown, it has to be approximated by postprocessing a finite element solution with some averaging technique.

The aforementioned refinement rules for the step MARK (intended for conforming FEM) ignore further decisions such as the particular type of anisotropic refinement or the increase of polynomial degrees versus the mesh refinements for \( hp \)-FEM. One way to handle such subtle decisions will be addressed under the heading coarsening strategy in Section 6.5.

### 6.3 Mesh refining of regular triangulations

Given a marked set of objects such as nodes, edges, faces, or elements, the refinement of the element (triangles or tetrahedra) plus further refinements (closure algorithm) for the design of regular triangulations are considered in this section.
6.3.1 Refinement of a triangle

Triangular elements in two dimensions are refined into two, three, or four subtriangles as indicated in Figure 15. All these divisions are based on hidden information on some reference edge. Rivara (1984) assumed the longest edge in the triangle as base of the refinement strategies while the one below is based on the explicit marking of a reference edge. In Figure 15, the bottom edge of the original triangle acts as the reference edge and, in any refinement, is halved. The four divisions displayed correspond to the bisection of one (called green-refinement), of two (the two versions of blue-refinement), or of all three edges (for red-refinement) as long as the bottom edge is refined. In Figure 15, the reference edges in the generated subtriangles are drawn with a bold line.

6.3.2 Refinement of a tetrahedron

The three-dimensional situation is geometrically more complicated. Therefore, this section is focused on the bisection algorithm and the readers are referred to Bey (1995) for 3-D red-refinements.

Figure 16 displays five types of tetrahedra which shows the bottom triangle with vertices of local label 1, 2, and

Figure 16. Bisection rules in 3-D after Arnold, Mukherjee and Pouly (2000) which essentially dates back to Bänsch (1991). It involves five types of tetrahedra (indicated as $P_u$, $A$, $M$, $O$, and $P_f$ on the left) and their bisections with an initialization of $M \rightarrow 2 P_u$ and $O \rightarrow 2 P_u$ followed by a successive loop of the form $A \rightarrow 2 P_u \rightarrow 4 P_f \rightarrow 8 A$ and so on. Notice that the forms of $P_u$ and $P_f$ are identical, but play a different role in the refinement loop.
3 while the remaining three faces with the vertex label 4 are drawn in the same plane as the bottom face; hence, the same number 4 for the top vertex is visible at the end-points of the three outer subtriangles.

Each triangular face has a reference edge and (at least) one edge is a reference edge of the two neighboring faces and that determines the vertices with local numbers 1 and 2. In the language of Arnold, Mukherjee and Pouly (2000), the five appearing types are called $P_0$, $A$, $M$, $O$, and $P_f$. The bisection of the edges between the vertices number 1 and 2 with the new vertex number 5 are indicated in Figure 16.

It is an important property of the inherited reference edges that the edges of each element $K$ in the original regular triangulation $T$ are refined in a cyclic way such that four consecutive bisections inside $K$ are necessary before a new vertex is generated, which is not the midpoint of some edge of $K$.

### 6.3.3 Closure algorithms

The bisection of some set of marked elements does not always lead to a regular triangulation – the new vertices may be hanging nodes for the neighboring elements. Further refinements are necessary to make those nodes regular. The default procedure within the class of bisection algorithms is to work on a set of marked elements $M(k)$.

**Closure Algorithm for Bisection.** Input a regular triangulation $T(0) := T$ and an initial subset $M(0) := M \subset T(0)$ of marked elements, set $Z = \emptyset$ and $k := 0$.

While $M(k) \neq \emptyset$ repeat (i)–(iv):

(i) choose some element $K$ in $M(k)$ with reference edge $E$ and initiate its midpoint $z_K$ as new vertex, set $Z := Z \cup \{z_K\}$;

(ii) bisect $E$ and divide $K$ into $K_+$ and $K_-$ and set $T(k+1) := \{K_+, K_-\} \cup (T(k) \setminus \{K\})$;

(iii) find all elements $T_1, \ldots, T_{m_K} \in T(k+1)$ with hanging node $z \in Z$ (if any) and set $M(k+1) := \{T_1, \ldots, T_{m_K}\} \cup (M(k) \setminus \{K\})$;

(iv) update $k := k + 1$ and go to (i).

Output a refined triangulation $T(k)$.

According to step (iii) of the closure algorithm, any termination leads to a regular triangulation $T(k)$. The remaining essential detail is to guarantee that there will always occur a termination via $M(k) = \emptyset$ for some $k$. In the newest-vertex bisection, for instance, the reference edges are inherited in such a way that any element $K \in T$, the initial regular triangulation, is refined only by subdivisions which, at most, halve each edge of $K$. Since the closure algorithm only halves some edge in $T$ and prohibits any further refinements, any intermediate (irregular) $T(k)$ remains coarser than or equal to some regular uniform refinement $\hat{T}$ of $T$. This is the main argument to prove that the closure algorithm cannot refine forever and stops after a finite number of steps.

Figure 13(b) shows an example where, given the initial mesh of Figure 13(a), only one edge, namely the second on the diagonal, is marked for refinement and the remaining refinement is induced by the closure algorithm. Nevertheless, the number of new elements can be bounded in terms of the initial triangulation and the number of marked elements (Binev, Dahmen and DeVore, 2004).

The closure algorithm for the red–green–blue refinement in 2-D is simpler when the focus is on marking of edges. One main ingredient is that each triangle $K$ is assigned a reference edge $E(K)$. If we are given a set of marked elements, let $M$ denote the set of corresponding assigned reference edges.

**Closure Algorithm for Red–Green–Blue Refinement.** Input a regular triangulation $T$ with a set of edges $E$ and an initial subset $M \subset E$ of marked edges, set $k := 0$ and $M(0) := N(0) := M$.

While $M(k) \neq \emptyset$ repeat (i)–(iv):

(i) choose some edge $E$ in $N(k)$ and let $T_{\pm} \subset T$ denote the (at most) two triangles that share the edge $E \subset \partial T_{\pm}$;

(ii) set $M(k+1) := M(k) \cup \{E_+, E_-\}$ with the reference edge $E_{\pm} := E(T_{\pm})$ of $T_{\pm}$;

(iii) if $M(k+1) = M(k)$ set $N(k+1) := N(k) \setminus \{E\}$ else set $N(k+1) := N(k) \cup \{E_+, E_-\} \setminus \{E\}$;

(iv) update $k := k + 1$ and go to (i).

Bisect the marked edges $\{E \in M(k) : E \subset \partial T\}$ of each element $T \subset T$ and refine $T$ by one of the red–green–blue refinement rules to generate elementwise a partition $\hat{T}$.

Output the regular triangulation $\hat{T}$.

The closure algorithm for red–green–blue refinement terminates as $N(k)$ is decreasing and $M(k)$ is increasing and outputs a set $\hat{M} := M(k)$ of marked edges with the following closure property: Any element $T \subset T$ with an edge in $\hat{M}$ satisfies $E(T) \in \hat{M}$, i.e. if $T$ is marked, then at least its reference edge will be halved. This property allows the application of one properly chosen refinement of Figure 15 and leads to a regular triangulation.

The reference edge $E(K)$ in the closure algorithm is assigned to each element $K$ of the initial triangulation and then is inherited according to the rules of Figure 15. For newest-vertex bisection, each triangle with vertices of global numbers $j, k$, and $\ell$ has the reference edge opposite to the vertex number $\max\{j, k, \ell\}$. 

**Finite Element Methods**
On the basis of refinement rules that inherit a reference edge to the generated elements, one can prove that a finite number of affine-equivalent elements domains occur.

### 6.4 Convergence of adaptive algorithms

This section discusses the convergence of a class of adaptive $P_1$ finite element methods based on newest-vertex bisection or red–green–blue refinement of Section 6.3.1. For the ease of this presentation, let us adopt the notation of the residual representation formula (73) of Section 5.2.1 in 2-D and perform a refinement with four bisections of each triangle with a marked edge as illustrated in Figure 17.

**Adaptive Algorithm.** Input a regular triangulation $\mathcal{T}^{(0)} := T$ with a set of reference edges $\{E(K) : K \in \mathcal{T}^{(0)}\}$ and a parameter $0 < \Theta \leq 1$, set $k := 0$.

Repeat (i)–(v) until termination:

(i) compute discrete solution $u_k \in V_k$ based on the triangulation $\mathcal{T}^{(k)}$ with set of interior edges $\mathcal{E}^{(k)}$;

(ii) compute $\eta^2_E := h_E \| r_E \|^2_{L_2(E)} + \sum_{T \in \mathcal{T}(E)} h_T^2 \| r_T \|^2_{L_2(T)}$ for any $E \in \mathcal{E}^{(k)}$ with the set $\mathcal{T}(E)$ of its at most two neighboring elements;

(iii) mark edges in $\mathcal{M}^{(k)} \subset \mathcal{E}^{(k)}$ by the bulk criterion such that

$$(1 - \Theta)^2 \sum_{E \in \mathcal{E}^{(k)}} \eta^2_E \leq \sum_{E \in \mathcal{M}^{(k)}} \eta^2_E$$

(iv) for any $E \in \mathcal{M}^{(k)}$, bisect each element $T \in \mathcal{T}(E)$ (at least) four times (as in Figure 17(a)) such that (at least) each midpoint $z_E$ of any edge $E \in \mathcal{E}(T)$ and the midpoint of $T$ become new nodes (cf. Figure 17(b));

(v) call closure algorithm to generate a refined regular triangulation $\mathcal{T}^{(k+1)}$, update $k := k + 1$ and go to (i).

Output a sequence of finite element meshes $\mathcal{T}^{(k)}$ with associated finite element spaces $V_k$ and discrete solutions $u_k$.

Recall that $u$ is the exact solution. Then there holds the error-reduction property

$$\| u - u_{k+1} \|^2_a \leq \Theta \| u - u_k \|^2_a + \text{osc}(f)^2$$

for any $k$ and with a constant $0 \leq \Theta < 1$, which depends on $\mathcal{T}^{(0)}$ and $\Theta > 0$. The oscillations in $\text{osc}(f)$ may be seen as higher-order terms when compared with the volume term $\| h_T f \|_{L_2(\Omega)}$ of the error estimator.

An outline of the proof of the error-reduction property concludes this section to illustrate that $\Theta < 1$ is independent of any mesh size, the level $k$, and number of elements $N_k := \text{card} (T^{(k)})$. The proof follows the arguments of Section 5.4.3 for $V_h = \mathcal{V} := V_k$ and $V_h := V_h \oplus W_h \subseteq V_{k+1}$. The space $W_h$ is spanned by all $b_T$ for all $T \in \mathcal{T}(E)$ for $E \in \mathcal{M}^{(k)} \subset \mathcal{E}^{(k)}$, which, here, are given by a nodal basis function in the new triangulation of the new interior node of $T \in \mathcal{T}(E)$ and of the midpoint of $E \in \mathcal{M}^{(k)}$. These basis functions substitute the bubble functions (80) in Section 5.2.4 and have in fact the properties (81)–(83). One then argues as in Section 5.4.3 with some

$$w_h := \sum_{E \in \mathcal{M}^{(k)}} (w_E + \sum_{T \in \mathcal{T}(E)} w_T) \in W_h$$

to show $\| u_k \|^2_a \leq C(R(w_h) + \text{osc}(f))$ and eventually

$$\sum_{E \in \mathcal{M}^{(k)}} \eta^2_E \leq C(\| u_h - u_H \|^2_a + \text{osc}(f)^2)$$

for $u_h := u_{k+1}$ and $u_H := u_k$. This and the bulk criterion lead to

$$\| u - u_h \|^2_a \leq C_{\text{rel}} \left( \sum_{E \in \mathcal{M}^{(k)}} \eta^2_E + \text{osc}(f)^2 \right)$$

$$\leq (1 - \Theta)^{-2}C_{\text{rel}} \left( \sum_{E \in \mathcal{M}^{(k)}} \eta^2_E + \text{osc}(f)^2 \right)$$

$$\leq C (1 - \Theta)^{-2}C_{\text{rel}} (\| u_h - u_H \|^2_a + \text{osc}(f)^2)$$

Utilizing the Galerkin orthogonality

$$\| u_h - u_H \|^2_a = \| u - u_H \|^2_a - \| u - u_h \|^2_a$$

one concludes the proof.

It is interesting to notice that the conditions on the refinement can be weakened. For instance, it suffices to refine solely one element in $\mathcal{T}(E)$ such that there is an interior node as in Figure 17(c). (But then the oscillations are coarser and involve terms like $h_E \| f - f_E \|_{L_2(\Omega_E)}$ with an integral mean $f_E$ over an edge-patch $\omega_E$.) However, the counterexample of Figure 10 clearly indicates that some specific conditions are necessary to ensure the error-reduction property.

![Figure 17](image-url)
6.5 Coarsening

The constructive task in approximation theory is the design of effective approximations of a given function \( u \). Replacing the unknown \( u \) by a very accurate known approximation (e.g., obtained by some overkill computation) of it, one may employ all the approximation techniques available and find an efficient representation in terms of an adapted finite element space, i.e. in terms of an underlying adapted mesh and/or an adapted distribution of polynomial degrees.

The input of this coarsening step consists, for instance, of a given triangulation \( T_k \) and a given finite element solution \( u_k \) and some much more accurate approximation \( \hat{u}_k \). Then, the difference

\[
\eta_k := \| u_k - \hat{u}_k \|_a
\]

is an error estimator as accurate as \( \| u - u_k \|_a \) (reliability and efficiency is proved by a triangle inequality). Moreover, one might use \( \eta_k = \| u_k - \hat{u}_k \|_{H(T)} \) as a local refinement indicator in the setting of the benchmark example.

Moreover, many decisions of the mesh design (e.g. choice of anisotropic elements, choice of polynomial degree, etc.) can be based on the explicit availability of \( \hat{u}_k \).

One single coarsening step is certainly pointless, for the reference solution \( \hat{u}_k \) is known and regarded as very accurate (and so there is no demand for further work). On the other hand, a cascade of coarsening steps requires in each step, a high accuracy relative to the precision of that level and hence a restricted accuracy. A schematic description of this procedure reads as follows:

**Coarsening Algorithm for Adaptive Mesh Design.** Input a finite element space \( V_0 \) with a finite element solution \( u_0 \), set \( k := 0 \).

- Repeat (i)–(iv) until termination:
  1. design a super space \( \hat{V}_k \) of \( V_k \) by uniform refinements plus uniform increase of all the polynomial degrees;
  2. compute accurate finite element solution \( \hat{u}_k \) with respect to \( \hat{V}_k \);
  3. coarsen \( \hat{V}_k \) based on given \( \hat{u}_k \) and design a new finite element space \( V_{k+1} \);
  4. update \( k := k + 1 \) and go to (i).

The paper of Binev, Dahmen and DeVore (2004) presents a realization of this algorithm with an adaptive algorithm on stage (i) and then proves that the convergence rate of the overall procedure is optimal with respect to the energy error as a function of the number of degrees of freedom.

An automatic version of the coarsening strategy is called the *hp-adaptive finite element method* in Demkowicz (2003) for higher-order polynomials adaptivity. This involves algorithms for the determination of edges that will be refined and of their polynomial degrees.

7 OTHER ASPECTS

In this section, we discuss briefly several topics in finite element methodology. Some of the discussions involve the Sobolev space \( W^k_p(\Omega) \) (1 \( \leq p \leq \infty \)), which is the space of functions in \( L_p(\Omega) \) whose weak derivatives up to order \( k \) also belong to \( L_p(\Omega) \), with the norm

\[
\| v \|_{W^k_p(\Omega)} = \left( \sum_{|\alpha| \leq k} \left\| \frac{\partial^\alpha v}{\partial x^\alpha} \right\|_{L_p(\Omega)}^p \right)^{1/p}
\]

for \( 1 \leq p < \infty \) and

\[
\| v \|_{W^k_\infty(\Omega)} = \max_{|\alpha| \leq k} \left( \left\| \frac{\partial^\alpha v}{\partial x^\alpha} \right\|_{L_\infty(\Omega)} \right)
\]

For \( 1 \leq p < \infty \), the seminorm \( \left( \sum_{|\alpha| = k} \| \partial^\alpha v/\partial x^\alpha \|_{L_p(\Omega)} \right)^{1/p} \) will be denoted by \( |v|_{W^k_p(\Omega)} \), and the seminorm \( \max_{|\alpha| = k} \| (\partial^\alpha v/\partial x^\alpha) \|_{L_\infty(\Omega)} \) will be denoted by \( |v|_{W^k_\infty(\Omega)} \).

7.1 Nonsymmetric/indefinite problems

The results in Section 4 can be extended to the case where the bilinear form \( a(\cdot, \cdot) \) in the weak problem (1) is nonsymmetric and/or indefinite due to lower order terms in the partial differential equation. We assume that \( a(\cdot, \cdot) \) is bounded (cf. (2)) on the closed subspace \( V \) of the Sobolev space \( H^m(\Omega) \) and replace (3) by the condition that

\[
a(v, v) + L \| v \|_{L^2(\Omega)}^2 \geq C_3 \| v \|_{H^m(\Omega)}^2 \quad \forall v \in V \quad (120)
\]

where \( L \) is a positive constant.

**Example 15.** Let \( a(\cdot, \cdot) \) be defined by

\[
a(v_1, v_2) = \int_\Omega \nabla v_1 \cdot \nabla v_2 \, dx + \sum_{j=1}^d \int_\Omega b_j(x) \frac{\partial v_1}{\partial x_j} v_2 \, dx
\]

\[
+ \int_\Omega c(x) v_1 v_2 \, dx \quad (121)
\]

for all \( v_1, v_2 \in H^1(\Omega) \), where \( b_j(x) \) (1 \( \leq j \leq d \)), \( c(x) \in L_\infty(\Omega) \). If we take \( V = \{ v \in H^1(\Omega) : v |_{\Gamma} = 0 \} \) and \( F \) is
defined by (6), then (1) is the weak form of the nonsymmetric boundary value problem

\[-\Delta u + \sum_{j=1}^{d} b_j \frac{\partial u}{\partial x_j} + cu = f, \quad u = 0 \quad \text{on} \quad \Gamma,\]

\[\frac{\partial u}{\partial n} = 0 \quad \text{on} \quad \partial \Omega \setminus \Gamma \quad (122)\]

and the coercivity condition (120) follows from the well-known Gårding’s inequality (Agmon, 1965).

Unlike the symmetric positive definite case, we need to assume that the weak problem (1) has a unique solution, and that the adjoint problem is also uniquely solvable, that is, given any \( G \in V^* \) there is a unique \( w \in V \) such that

\[a(v, w) = G(v) \quad \forall \; v \in V \quad (123)\]

Furthermore, we assume that the solution \( w \) of (123) enjoys some elliptic regularity when

\[w \in H^m(\Omega) \quad \text{for some} \quad m > 0 \quad \text{and} \quad a(v, w) \leq C \| g \|_{L^2(\Omega)} \quad (124)\]

Let \( T \) be a triangulation of \( \Omega \) with mesh size \( h_T = \max_{T \in \mathcal{T}} \text{diam} \; T \) and \( V_T \subset V \) be a finite element space associated with \( T \) such that the following approximation property is satisfied:

\[\inf_{w \in V_T} \| w - v \|_{H^m(\Omega)} \leq \epsilon_T \| w \|_{H^m(\Omega)} \quad \forall \; w \in H^m(\Omega) \quad (125)\]

where

\[\epsilon_T \downarrow 0 \quad \text{as} \quad h_T \downarrow 0 \quad (126)\]

The discrete problem is then given by (54).

Following Schatz (1974) the well-posedness of the discrete problem and the error estimate for the finite element approximate solution can be addressed simultaneously. Assume for the moment that \( u_T \in V_T \) is a solution of (54). Then we have

\[a(u - u_T, v) = 0 \quad \forall \; v \in V_T \quad (127)\]

We use (127) and a duality argument to estimate \( \| u - u_T \|_{L^2(\Omega)} \) in terms of \( \| u - u_T \|_{H^m(\Omega)} \). Let \( w \in V \) satisfy

\[a(v, w) = (u - u_T, v)_{L^2(\Omega)} \quad \forall \; v \in V_T \quad (128)\]

We obtain, from (2), (127), and (128), the following analog of (24):

\[\| u - u_T \|_{L^2(\Omega)}^2 = a(u - u_T, w) \leq C \left( \inf_{v \in V_T} \| w - v \|_{H^m(\Omega)} \right) \times \| u - u_T \|_{H^m(\Omega)} \quad (129)\]

and hence, by (124) and (125),

\[\| u - u_T \|_{L^2(\Omega)} \leq \epsilon_T \| u - u_T \|_{H^m(\Omega)} \quad (130)\]

It follows from (120) and (130) that

\[\| u - u_T \|_{H^m(\Omega)}^2 \leq a(u - u_T, u - u_T) + C \| u - u_T \|_{H^m(\Omega)}^2 \quad (131)\]

which together with (126) implies, for \( h_T \) sufficiently small,

\[\| u - u_T \|_{H^m(\Omega)}^2 \leq a(u - u_T, u - u_T) \quad (131)\]

For the special case where \( F = 0 \) and \( u = 0 \), any solution \( u_T \) of the homogeneous discrete problem

\[a(u_T, v) = 0 \quad \forall \; v \in V_T \quad (127)\]

must satisfy, by (131),

\[\| u_T \|_{H^m(\Omega)}^2 \leq 0 \quad (129)\]

We conclude that any solution of the homogeneous discrete problem must be trivial and hence the discrete problem (19) is uniquely solvable provided \( h_T \) is sufficiently small. Under this condition, we also obtain immediately from (2), (127), and (131), the following analog of (22):

\[\| u - u_T \|_{H^m(\Omega)} \leq C \inf_{v \in V_T} \| u - v \|_{H^m(\Omega)} \quad (132)\]

Concrete error estimates now follow from (132), (129), and the results in Section 3.3.

### 7.2 Nonconforming finite elements

When the finite element space \( FE_T \) defined by (32) does not belong to the Sobolev space \( H^m(\Omega) \) where the weak problem (1) of the boundary value problem is posed, it is referred to as a nonconforming finite element space. Nonconforming finite element spaces are more flexible and are useful for problems with constraints where conforming finite element spaces are more difficult to construct.

**Example 16 (Triangular Nonconforming Elements)** Let \( K \) be a triangle. If the set \( N_K \) consists of evaluations of the shape functions at the midpoints of the edges of \( K \) (Figure 18a), then \( (K, P_1, N_K) \) is the nonconforming \( P_1 \) element of Crouzeix and Raviart. It is the simplest triangular element that can be used to solve the incompressible Stokes equation.
If the set \( N_K \) consists of evaluations of the shape functions at the vertices of \( K \) and the evaluations of the normal derivatives of the shape functions at the midpoints of the edges of \( K \) (Figure 18b), then \((K, P_2, N_K)\) is the Morley element. It is the simplest triangular element that can be used to solve the plate bending problem.

**Example 17 (Rectangular Nonconforming Elements)**

Let \( K \) be a rectangle. If \( P_k \) is the space spanned by the functions \( 1, x_1, x_2 \) and \( x_1^2 - x_2^2 \) and the set \( N_K \) consists of the mean values of the shape functions on the edges of \( K \), then \((K, P_2, N_K)\) is the rotated \( Q_1 \) element of Rannacher and Turek (Figure 19a), where the thick lines represent mean values over the edges). It is the simplest rectangular element that can be used to solve the incompressible Stokes equation.

If \( P_k \) is the space spanned by the functions \( 1, x_1, x_2, x_1^2, x_2^2, x_1^2 x_2 \) and \( x_1 x_2^2 \) and the set \( N_K \) consists of evaluations of the shape functions at the vertices of \( K \) and evaluations of the normal derivatives at the midpoints of the edges (Figure 19b), then \((K, P_2, N_K)\) is the incomplete \( Q_2 \) element. It is the simplest rectangular element that can be used to solve the plate bending problem.

Consider the weak problem (1) for a symmetric positive definite boundary value problem, where \( F \) is defined by (6) for a function \( f \in L_2(\Omega) \). Let \( V_T \) be a nonconforming finite element space associated with the triangulation \( T \). We assume that there is a (mesh-dependent) symmetric bilinear form \( a_T(\cdot, \cdot) \) defined on \( V + V_T \) such that (i) \( a_T(v, v) = a(v, v) \) for \( v \in V \), (ii) \( a_T(\cdot, \cdot) \) is positive definite on \( V_T \).

The discrete problem for the nonconforming finite element method reads: Find \( u_T \in V_T \) such that

\[
a_T(u_T, v) = \int_\Omega f v \, dx \quad \forall v \in V_T \tag{133}
\]

**Example 18.** The Poisson problem in Example 1 can be solved by the nonconforming \( P_1 \) finite element method in which the finite element space is \( V_T = \{ v \in L_2(\Omega) : v|_T \in P_1(T) \text{ for every triangle } T \in \mathcal{T} \} \), \( v \) is continuous at the midpoints of the edges of \( \mathcal{T} \) and \( v \) vanishes at the midpoints of the edges of \( \mathcal{T} \) along \( \Gamma \}).

The nonconforming Ritz–Galerkin method (133) can be analyzed as follows. Let \( \tilde{u}_T \in V_T \) be defined by

\[
a_T(\tilde{u}_T, v) = a_T(u, v) \quad \forall v \in V_T
\]

Then we have

\[
\|u - \tilde{u}_T\|_{a_T} = \inf_{v \in V_T} \|u - v\|_{a_T}
\]

where \( \|w\|_{a_T} = (a_T(w, w))^{1/2} \) is the nonconforming energy norm defined on \( V + V_T \), and we arrive at the following generalization (Berger, Scott and Strang, 1972) of (21):

\[
\|u - u_T\|_{a_T} \leq \|u - \tilde{u}_T\|_{a_T} + \|\tilde{u}_T - u_T\|_{a_T}
\]

\[
= \inf_{v \in V_T} \|u - v\|_{a_T} + \sup_{v \in V_T} a_T(\tilde{u}_T - u_T, v)\|v\|_{a_T}
\]

\[
= \inf_{v \in V_T} \|u - v\|_{a_T} + \sup_{v \in V_T} a_T(u - u_T, v)\|v\|_{a_T} \tag{135}
\]

**Remark 17.** The second term on the right-hand side of (135), which vanishes in the case of conforming Ritz–Galerkin methods, measures the consistency errors of nonconforming methods.

As an example, we analyze the nonconforming \( P_1 \) method for the Poisson problem in Example 18. For each \( T \in \mathcal{T} \), we define an interpolation operator \( \Pi_T : H^1(T) \rightarrow P_1(T) \) by

\[
(\Pi_T \zeta)(m) = \frac{1}{|e|} \int_e \zeta \, ds
\]

![Figure 18. Triangular nonconforming finite elements.](image)

![Figure 19. Rectangular nonconforming finite elements.](image)
where \( m_e \) is the midpoint for the edge \( e \) of \( T \). The interpolation operator \( \Pi_T \) satisfies the estimate (43) for \( m = 1 \) and they can be pieced together to form an interpolation operator \( \Pi : H^1(\Omega) \to V_T \). Since the solution \( u \) of (5) belongs to \( H^{1+\alpha(T)}(T) \), where \( 0 < \alpha(T) \leq 1 \), the first term on the right-hand side of (135) satisfies the estimate

\[
\inf_{v \in V_T} \| u - v \|_{a_T} \leq \| u - \Pi u \|_{a_T} \leq C \left( \sum_{T \in \mathcal{T}} (\text{diam } T)^{2\alpha(T)} |u|^2_{H^{1+\alpha(T)}(T)} \right)^{1/2}
\]  

(136)

where the constant \( C \) depends only on the minimum angle in \( \mathcal{T} \).

To analyze the second term on the right-hand side of (135), we write, using (5), (133), and (134),

\[
a_T(u - u_T, v) = - \sum_{e \in \mathcal{E}_T} \int_e \frac{\partial u}{\partial n_e} [v]_e \, ds
\]  

(137)

where \( \mathcal{E}_T \) is the set of edges in \( \mathcal{T} \) that are not on \( \Gamma \), \( n_e \) is a unit vector normal to \( e \), and \([v]_e = v_+ - v_-\) is the jump of \( v \) across \( e \) (\( n_e \) is pointing from the minus side to the plus side and \( v = 0 \) outside \( \Omega \)). Since \([v]_e \) vanishes at the midpoint of \( e \in \mathcal{E}_T \), we have

\[
\int_e \frac{\partial u}{\partial n_e} [v]_e \, ds = \int_e (\partial (u - p) - \partial n_e) [v]_e \, ds \quad \forall p \in P_1
\]  

(138)

Let \( T_e = \{ T \in \mathcal{T} : e \subset \partial T \} \). It follows from (138), the trace theorem and the Bramble–Hilbert lemma (cf. Remark 10) that

\[
\left| \int_e \frac{\partial u}{\partial n_e} [v]_e \, ds \right| \leq C \inf_{p \in P_1} \left[ \sum_{T \in \mathcal{T}_e} |u - p|_{H^1(T)}^2 \right]^{1/2} \left[ \sum_{T \in \mathcal{T}_e} |v|_{H^1(T)}^2 \right]^{1/2} + (\text{diam } T)^{2\alpha(T)} |u|^2_{H^{1+\alpha(T)}(T)} \right]^{1/2} \left[ \sum_{T \in \mathcal{T}_e} |v|_{H^1(T)}^2 \right]^{1/2}
\]  

\[
\leq C \left( \sum_{T \in \mathcal{T}_e} (\text{diam } T)^{2\alpha(T)} |u|^2_{H^{1+\alpha(T)}(T)} \right)^{1/2} \left( \sum_{T \in \mathcal{T}_e} |v|^2_{H^1(T)} \right)^{1/2}
\]  

(139)

We conclude from (137) and (139) that

\[
\sup_{v \in V_T} \frac{a_T(u - u_T, v)}{\|v\|_{a_T}} \leq C \left( \sum_{T \in \mathcal{T}} (\text{diam } T)^{2\alpha(T)} |u|^2_{H^{1+\alpha(T)}(T)} \right)^{1/2}
\]  

(140)

where the constant \( C \) depends only on the minimum angle in \( \mathcal{T} \).

Combining (135), (136), and (140) we have the following analog of (57)

\[
\sum_{T \in \mathcal{T}} |u - u_T|^2_{H^1(T)} \leq C \sum_{T \in \mathcal{T}} (\text{diam } T)^{2\alpha(T)} |u|^2_{H^{1+\alpha(T)}(T)}
\]  

(141)

Remark 18. Estimates of \( \|u - u_T\|_{L^2(\Omega)} \) can also be obtained for nonconforming finite element methods (Crouzeix and Raviart, 1973). There is also a close connection between certain nonconforming methods and mixed methods (Arnold and Brezzi, 1982).

### 7.3 Effects of numerical integration

The explicit form of the finite element equation (54) involves the evaluations of integrals which, in general, cannot be computed exactly. Thus, the effects of numerical integration must be taken into account in the error analysis. We will illustrate the ideas in terms of finite element methods for simplicial triangulations. The readers are referred to Davis and Rabinowitz (1984) for a comprehensive treatment of numerical integration.

Consider the second order elliptic boundary value problem (5) in Example 1 on a bounded polyhedral domain \( \Omega \subset \mathbb{R}^d \) (\( d = 2 \) or 3). Let \( \mathcal{T} \) be a simplicial triangulation of \( \Omega \) such that \( \Gamma \) is a union of the edges (faces) of \( \mathcal{T} \) and \( V_T \subset V \) be the corresponding \( P_n \) Lagrange finite element space.

In Section 4, the finite element approximate solution \( u_T \in V_T \) is defined by (54). But in practice, the integral \( F(v) = \int_\Omega f v \, dx \) is evaluated by a quadrature scheme and the approximate solution \( u_T \in V_T \) is actually defined by

\[
a_T(u_T, v) = F_T(v) \quad \forall v \in V_T
\]  

(142)

where \( F_T(v) \) is the result of applying the quadrature scheme to the integral \( F(v) \).

More precisely, let \( D \in \mathcal{T} \) be arbitrary and \( \Phi_D : S \to D \) be an affine homeomorphism from the standard (closed) simplex \( S \) onto \( D \). It follows from a change of variables that

\[
\int_D f v \, dx = \int_S (\det J_{\Phi_D})(f \circ \Phi_D)(v \circ \Phi_D) \, d\hat{x}
\]  

where without loss of generality \( \det J_{\Phi_D} \) (the determinant of the Jacobian matrix of \( \Phi \)) is assumed to be a positive number. The integral on \( S \) is evaluated by a quadrature scheme \( I_S \) and the right-hand side of (142) is then given by

\[
F_h(v) = \sum_{D \in \mathcal{T}} I_S((\det J_{\Phi_D})(f \circ \Phi_D)(v \circ \Phi_D))
\]  

(143)
The error \( u - u_T \) can be estimated by the following analog of (135):

\[
\|u - u_T\|_a \leq \inf_{v \in V_T} \|u - v\|_a + \sup_{v \in \mathcal{V} \setminus \{0\}} \frac{a(u - u_T, v)}{\|v\|_a}
\]

(144)

The first term on the right-hand side of (144) can be estimated by \( \|u - \Pi_T u\|_a \) as in Section 4.1. The second term on the right-hand side of (144) measures the effect of numerical quadrature. Below we give conditions on the quadrature scheme \( w \mapsto I_S(w) \) and the function \( f \) so that the magnitude of the quadrature error is identical with that of the optimal interpolation error for the finite element space.

We assume that the quadrature scheme \( w \mapsto I_S(w) \) has the following properties:

\[
|I_S(w)| \leq C_S \max_{x \in S} |w(\hat{x})| \quad \forall w \in C^0(S),
\]

(145)

\[
I_S(w) = \int_S w \, d\hat{x} \quad \forall w \in P_{2n-2}
\]

(146)

We also assume that \( f \in W^q_0(\Omega) \) such that \( q \geq 2 \) and \( n > d/q \), which implies, in particular, by the Sobolev embedding theorem that \( f \in C^0(\Omega) \) so that (143) makes sense.

Under these conditions it can be shown (Ciarlet, 1978) by using the Bramble–Hilbert lemma on \( S \) that

\[
\left| \int_D f \, v \, dx - I_S((\det J_{\Phi_D})(f \circ \Phi_D)(v \circ \Phi_D)) \right| \leq C (\text{diam } D)^q |D|^{(1/2) - (1/q)} \|f\|_{W^q_0(\Omega)} \|v\|_{H^1(D)} \quad \forall v \in V_T
\]

(147)

where the positive constant \( C \) depends only on the shape regularity of \( D \). It then follows from (1), (143), (147) and Hölder’s inequality that

\[
|a(u - u_T, v)| = \left| \int_\Omega f \, v \, dx - F_h(v) \right| \leq Ch^p_d \|f\|_{W^p_0(\Omega)} \|v\|_{H^1(\Omega)} \quad \forall v \in V_T
\]

(148)

We conclude from (144) and (148) that

\[
\|u - u_T\|_a \leq \|u - \Pi_T u\|_a + Ch^p_d \|f\|_{W^p_0(\Omega)}
\]

(149)

We see by comparing (43) and (149) that the overall accuracy of the finite element method is not affected by the numerical integration.

We now consider a general symmetric positive definite elliptic boundary problem whose variational form is defined by

\[
a(w, v) = \sum_{i,j=1}^d \int_\Omega a_{ij}(x) \frac{\partial w}{\partial x_i} \frac{\partial v}{\partial x_j} \, dx + \int_\Omega b(x) w v \, dx
\]

(150)

where \( a_{ij}, b \in W^2_0(\Omega), \ b \geq 0 \) on \( \Omega \) and there exists a positive constant \( c \) such that

\[
\sum_{i,j=1}^d a_{ij}(x) \xi_i \xi_j \geq c |\xi|^2 \quad \forall x \in \Omega \quad \text{and} \quad \xi \in \mathbb{R}^d
\]

(151)

In this case, the bilinear form (150) must also be evaluated by numerical integration and the approximation solution \( u_T \in V_T \) to (1) is defined by

\[
a_T(u_T, v) = F_T(v) \quad \forall v \in V_T
\]

(152)

where \( a_T(u_T, v) \) is the result of applying the quadrature scheme \( I_S \) to the pull-back of

\[
\sum_{i,j=1}^d \int_D a_{ij}(x) \frac{\partial w}{\partial x_i} \frac{\partial v}{\partial x_j} \, dx + \int_D b(x) w v \, dx
\]

on the standard simplex \( S \) under the affine homeomorphism \( \Phi_D \) over all \( D \in T \).

The error \( u - u_T \) can be estimated under (145), (146) and the additional condition that

\[
g \geq 0 \text{ on } \hat{S} \implies I_S(g) \geq 0
\]

(153)

Indeed (146), (151), (153) and the sign of \( b \) imply that

\[
a_T(v, v) \geq c |v|_{H^1(\Omega)}^2 \quad \forall v \in V_T
\]

(154)

and we have the following analog of (144)

\[
\|u - u_T\|_{H^1(\Omega)} \leq \inf_{v \in V_T} \|u - v\|_{H^1(\Omega)} \]

\[
+ \frac{1}{c} \left\{ \sup_{v \in V_T \setminus \{0\}} \frac{a_T(u, v) - a_T(u_T, v)}{|v|_{H^1(\Omega)}} \right\} + \sup_{v \in V_T \setminus \{0\}} \frac{a(u, v) - a_T(u_T, v)}{|v|_{H^1(\Omega)}}
\]

(155)

The first term on the right-hand side of (155) is dominated by \( \|u - \Pi_T u\|_{H^1(\Omega)} \). Since \( a(u, v) - a_T(u_T, v) = \int_\Omega f \, v \, dx - F_T(v) \), the third term is controlled by the estimate (148). The second term, which measures the effect of
numerical quadrature on the bilinear form $a(\cdot, \cdot)$, is controlled by the estimate
\[
|a_T(u, v) - a(u, v)| \leq C \sum_{D \in T} (\text{diam } D)^{\alpha}(D) \times |u|_{H^{1+\alpha}(D)} |v|_{H^{1}(D)},
\]
provided the solution $u$ belongs to $H^{1+\alpha}(D)$ for each $D \in T$ and $1/2 < \alpha(D) \leq 1$. The estimate (156) follows from (145), (146) and the Bramble–Hilbert lemma, and the positive constant $C$ in (156) depends only on the $W^2_\infty$ norms of $a_{ij}$ and $b$ and the shape regularity of $T$. Again we see by comparing (56), (149), and (156) that the overall accuracy of the finite element method is not affected by the numerical integration.

**Remark 19.** For problems that exhibit the phenomenon of locking, the choice of a lower order quadrature scheme in the evaluation of the stiffness matrix may help alleviate the effect of locking (Malkus and Hughes, 1978).

### 7.4 Curved domains

So far, we have restricted the discussion to polygonal (polyhedral) domains. In this section, we consider the second-order elliptic boundary value problem (5) on a domain $\Omega \subset \mathbb{R}^2$ with a curved boundary. For simplicity, we assume that $\Gamma = \partial \Omega$.

First, we consider the $P_1$ Lagrange finite element method for a domain $\Omega$ with a $C^2$ boundary. We approximate $\Omega$ by a polygonal domain $\Omega_h$ on which a simplicial triangulation $T_h$ of mesh size $h$ is imposed. We assume that the vertices of $T_h$ belong to the closure of $\Omega$. A typical triangle in $T_h$ near $\partial \Omega$ is depicted in Figure 20(a).

Let $V_h \subset H^1_0(\Omega_h)$ be the $P_1$ finite element space associated with $T_h$. The approximate solution $u_h \in V_h$ for (5) is then defined by
\[
a_h(u_h, v) = F_h(v) \quad \forall v \in V_h. \tag{157}
\]

![Figure 20. Triangulations for curved domains.](image)

where
\[
a_h(w, v) = \int_{\Omega_h} \nabla w \cdot \nabla v \, dx \quad \forall w, v \in H^1(\Omega_h) \tag{158}
\]
and $F_h(v)$ represented the result of applying a numerical quadrature scheme to the integral $\int_{\Omega_h} \tilde{f} \, v \, dx$ (cf. (143) with $f$ and $T$ replaced by $\tilde{f}$ and $T_h$). Here $\tilde{f}$ is an extension of $f$ to $\mathbb{R}^2$. We assume that the numerical scheme uses only the values of $\tilde{f}$ at the nodes of $T_h$ and hence the discrete problem (157) is independent of the extension $\tilde{f}$.

We assume that $f \in W^1_q(\Omega)$ for $2 < q < \infty$ and hence $u \in W^1_q(\Omega)$ by elliptic regularity. Let $\tilde{u} \in W^q_3(\mathbb{R}^2)$ be an extension of $u$ such that
\[
\|\tilde{u}\|_{W^q_3(\mathbb{R}^2)} \leq C_\Omega \|u\|_{W^q_1(\Omega)} \leq C_\Omega \|f\|_{W^1_q(\Omega)} \tag{159}
\]
We can then take $\tilde{f} = -\Delta \tilde{u} \in W^1_q(\mathbb{R}^2)$ to be the extension appearing in the definition of $F_h$.

The error $\tilde{u} - u_h$ over $\Omega_h$ can be estimated by the following analog of (135):
\[
\|\tilde{u} - u_h\|_{\alpha_h} \leq \inf_{v \in V_h} \|\tilde{u} - v\|_{\alpha_h} + \sup_{v \in V_h} \frac{a_h(\tilde{u} - u_h, v)}{\|v\|_{\alpha_h}} \tag{160}
\]
The first term on the right-hand side is dominated by $\|\tilde{u} - \Pi_h \tilde{u}\|_{\alpha_h}$ where $\Pi_h$ is the nodal interpolation operator. The second term is controlled by
\[
|a_h(\tilde{u} - u_h, v)| = \left| \int_{\Omega_h} \tilde{f} \, v \, dx - F_h(v) \right| \leq C_h \|\tilde{f}\|_{W^1_q(\Omega_h)} \|v\|_{H^1(\Omega_h)} \tag{161}
\]
which is a special case of (148), provided that the conditions (145) and (146) (with $n = 1$) on the numerical quadrature scheme are satisfied.

Combining (43) and (159)–(161) we see that
\[
\|\tilde{u} - u_h\|_{H^1(\Omega_h)} = \|\tilde{u} - u_h\|_{\alpha_h} \leq C_h \|f\|_{W^1_q(\Omega)} \tag{162}
\]
that is, the $P_1$ finite element method for the curved domain retains the optimal $O(h)$ accuracy.

The approximation of $\Omega_h$ to $\Omega$ can be improved if we replace straight-edge triangles by triangles with a curved edge (Figure 20(b)). This can be achieved by the *isoparametric* finite element methods. We will illustrate the idea using the $P_2$ Lagrange element.

Let $\Omega_{h2}$, an approximation of $\Omega$, be the union of straight-edge triangles (in the interior of $\Omega_{h2}$) and triangles with one curved edge (at the boundary of $\Omega_{h2}$), which form a triangulation $T_{h2}$ of $\Omega_{h2}$. The finite element associated with the interior triangles is the standard $P_2$ Lagrange element.
For a triangle $D$ at the boundary, we assume that there is a homeomorphism $\Phi_D$ from the standard simplex $S$ onto $D$ such that $\Phi_D(\hat{x}) = (\Phi_{D,1}(\hat{x}), \Phi_{D,2}(\hat{x}))$ where $\Phi_{D,1}(\hat{x})$ and $\Phi_{D,2}(\hat{x})$ are quadratic polynomials in $\hat{x} = (\hat{x}_1, \hat{x}_2)$. The space of shape functions $\mathcal{P}_D$ is then defined by

$$\mathcal{P}_D = \{ v \in C^\infty(D) : v \circ \Phi_D \in P_2(S) \} \quad (163)$$

that is, the functions in $\mathcal{P}_D$ are quadratic polynomials in the curvilinear coordinates on $D$ induced by $\Phi_D^{-1}$. The set $N_D$ of nodal variables consist of pointwise evaluations of the shape functions at the nodes corresponding to the nodes of the $P_2$ element on $S$ (cf. Figures 2 and 20) under the map $\Phi_D$. We assume that all such nodes belong to $\hat{S}$ and the nodes on the curved edge of $D$ belong to $\partial\Omega$ (cf. Figure 20).

In other words, the finite element $(\hat{D}, \mathcal{P}_D, N_D)$ is pulled back to the $P_2$ Lagrange finite element on $\hat{S}$ under $\Phi_D$. It is called an isoparametric element because the components of the parameterization map $\Phi_D$ belong to the shape functions of the $P_2$ element on $\hat{S}$. The corresponding finite element space defined by (32) (with $T$ replaced by $\mathcal{T}_h$) is a subspace of $H^1(\Omega_h)$ that contains all the continuous piecewise linear polynomials with respect to $\mathcal{T}_h$. By setting the nodal values on $\partial\Omega_h$ to be zero, we have a finite element space $V_h \subset H^1_0(\Omega_h)$. The discrete problem for $u_h \in V_h$ is then defined by

$$\tilde{a}_h(u_h, v) = F_h(v) \quad (164)$$

where the numerical quadrature scheme in the definition of $F_h$ involves only the nodes of the finite element space so that the discrete problem is independent of the choice of the extension of $f$ and the variational form $\tilde{a}_h(\cdot, \cdot)$ is obtained from $a_h(\cdot, \cdot)$ by the numerical quadrature scheme.

We assume that $f \in W^2_q(\Omega)$ for $1 < q < \infty$ and hence $u \in W^2_q(\Omega)$ by elliptic regularity (assuming that $\Omega$ has a $C^3$ boundary). Let $\tilde{u} \in W^2_q(\mathbb{R}^2)$ be an extension of $u$ such that

$$\|u\|_{W^2_q(\Omega)} \leq C_{\Omega} f = C_{\Omega} \|f\|_{W^2_q(\Omega)} \quad (165)$$

Under the condition (153), we have

$$\tilde{a}_h(v, v) \geq c|v|^2_{H^1(\Omega_h)} \quad \forall v \in V_h \quad (166)$$

and the error of $\tilde{u} - u_h$ over $\Omega_h$ can be estimated by the following analog of (155)

$$\|\tilde{u} - u_h\|_{H^1(\Omega_h)} \leq \inf_{v \in V_h} |\tilde{u} - v|_{H^1(\Omega_h)}$$

$$+ \frac{1}{c} \sup_{v \in V_h} |\tilde{a}_h(\tilde{u}, v) - a_h(\tilde{u}, v)|$$

$$\leq 1 + \frac{1}{c} \sup_{v \in V_h} \|a_h(\tilde{u}, v) - a_h(\tilde{u}, v)\|_{H^1(\Omega_h)}$$

$$\approx 1 + \frac{1}{c} \sup_{v \in V_h} \|a_h(\tilde{u}, v) - a_h(\tilde{u}, v)\|_{H^1(\Omega_h)}$$

$$\approx 1 + \frac{1}{c} \sup_{v \in V_h} \|a_h(\tilde{u}, v) - a_h(\tilde{u}, v)\|_{H^1(\Omega_h)}$$

$$+ \frac{1}{c} \sup_{v \in V_h} |\tilde{a}_h(\tilde{u}, v) - a_h(\tilde{u}, v)|$$

$$\leq 1 + \frac{1}{c} \sup_{v \in V_h} \|a_h(\tilde{u}, v) - a_h(\tilde{u}, v)\|_{H^1(\Omega_h)}$$

The analysis of the terms on the right-hand side of (167) involves the shape regularity of a curved triangle, which can be defined as follows. Let $\mathcal{A}_D$ be the affine map that agrees with $\Phi_D$ at the vertices of the standard simplex $S$, and $\tilde{D}$ be the image of $S$ under $\mathcal{A}_D$. (The triangle is in Figure 20(a), while $D$ is the curved triangle in (b).) The shape regularity of the curved triangle $D$ is measured by the aspect ratio $\gamma(D)$ (cf. (29)) of the straight-edged triangle $\tilde{D}$ and the parameter $\kappa(D)$ defined by

$$\kappa(D) = \max \{ h^{-1}|\mathcal{A}_D \cdot \mathcal{A}_D^{-1}|_{W^1_2(D)}, h^{-2}|\mathcal{A}_D \cdot \mathcal{A}_D^{-1}|_{W^1_2(S)} \} \quad (168)$$

and we can take the aspect ratio $\gamma(D)$ to be the maximum of $\gamma(\tilde{D})$ and $\kappa(D)$. Note that in the case where $D = \tilde{D}$ the parameter $\kappa(D) = 0$ and $\gamma(D) = \gamma(D)$.

The first term on the right-hand side of (167) is dominated by $\|\tilde{u} - \Pi_h \tilde{u}\|_{a_h}$, where $\Pi_h$ is the nodal interpolation operator. Note that, by using the Bramble–Hilbert lemma on $S$ and scaling, we have the following generalization of (43):

$$\|\tilde{u} - \Pi_h \tilde{u}\|_{H^1(D)} \leq C(h^2) \|	ilde{u}\|_{H^1(\Omega_h)} \quad (169)$$

where $\Pi_h$ is the element nodal interpolation operator and the constant $C$ depends only on an upper bound of $\gamma(D)$, and hence

$$\|\tilde{u} - \Pi_h \tilde{u}\|_{H^1(\Omega_h)} \leq C h^2 \|\tilde{u}\|_{H^1(\Omega_h)} \quad (170)$$

In order to analyze the third term on the right-hand side of (167), we take $\tilde{f} = -\Delta \tilde{u}$ and impose the conditions (145) and (146) (with $n = 2$) on the numerical quadrature scheme. We then have the following special case of (148):

$$|a_h(\tilde{u}, v) - a_h(\tilde{u}, v)| \leq \left| \int_{\Omega_h} \tilde{f} v dx - F_h(v) \right|$$

$$\leq C h^2 \|	ilde{f}\|_{W^1_2(\Omega_h)} \|	ilde{u}\|_{H^1(\Omega_h)} \quad (171)$$

Similarly, the second term on the right-hand side of (167), which measures the effect of numerical integration on the variational form $a_h(\cdot, \cdot)$, is controlled by the estimate

$$|\tilde{a}_h(\tilde{u}, v) - a_h(\tilde{u}, v)| \leq C h^2 \|	ilde{u}\|_{H^1(\Omega_h)} \|v\|_{H^1(\Omega_h)} \quad \forall v \in V_h$$

Combining (167) and (170)–(172) we have

$$\|\tilde{u} - u_h\| \leq C h^2 \|f\|_{W^2_2(\Omega)} \quad (173)$$
where $C$ depends only on an upper bound of $\{\gamma(D) : D \in T_h\}$ and the constants in (165). Therefore, the $P_2$ isoparametric finite element method retains the optimal $O(h^2)$ accuracy. On the other hand, if only straight-edged triangles are used in the construction of $\Omega_h$, then the accuracy of the $P_2$ Lagrange finite element method is only of order $O(h^{3/2})$ (Strang and Berger, 1971).

The discussion above can be generalized to higher-order isoparametric finite element methods, higher dimensions, and elliptic problems with variable coefficients (Ciarlet, 1978).

**Remark 20.** Estimates such as (160) and (162) are useful only when a sequence of domains $\Omega_h$ with corresponding triangulations $T_h$ can be constructed so that $h_i \downarrow 0$ and the aspect ratios of all the triangles (straight or curved) in the triangulations remain bounded. We refer the readers to Scott (1973) for 2-D constructions and to Lenoir (1986) for the 3-D case.

**Remark 21.** Other finite element methods for curved domains can be found in Zlámal (1973, 1974), Scott (1975), and Bernardi (1989).

**Remark 22.** Let $\Omega_i$ be a sequence of convex polygons approaching the unit disc. The displacement of the simply supported plate on $\Omega_i$ with unit loading does not converge to the displacement of the simply supported plate on the unit disc (also with unit loading) as $i \to \infty$. This is known as Babuška’s plate paradox (Babuška and Pitkäranta, 1990). It shows that numerical solutions obtained by approximating a curved domain with polygonal domains, in general, do not converge to the solution of a fourth-order problem defined on the curved domain. We refer the readers to Mansfield (1978) for the construction of finite element spaces that are subspaces of $H^2(\Omega)$.

### 7.5 Pointwise estimates

Besides the estimates in $L_2$-based Sobolev spaces discussed in Section 4, there also exist a priori error estimates for finite element methods in $L_p$-based Sobolev spaces with $p \neq 2$. In particular, error estimates in the $L_{\infty}$-based Sobolev spaces can provide pointwise error estimates. Below we describe some results for second-order elliptic boundary value problems with homogeneous Dirichlet boundary conditions.

In the one-dimensional case (Wheeler, 1973) where $\Omega$ is an interval, the finite element solution $u_T$ for a given triangulation $T$ with mesh size $h_f$ satisfies

$$
\|u - u_T\|_{L_\infty(\Omega)} \leq C h_f^2 |u|_{W^2_2(\Omega)}
$$

(174)

provided the solution $u$ of (5) belongs to $W^2_\infty(\Omega)$ and the finite element space contains all the piecewise polynomial functions of degree $\leq n - 1$. The estimate (174) also holds in higher dimensions (Douglas, Dupont and Wheeler, 1974) in the case where $\Omega$ is a product of intervals and $u_T$ is the solution in the $Q_{n-1}$ finite element space of Example 7.

For a two-dimensional convex polygonal domain (Natterer, 1975; Scott, 1976, Nitsche, 1977), the estimate (174) holds in the case where $n \geq 3$ and $u_T$ is the $P_{n-1}$ triangular finite element solution for a general triangulation $T$. In the case where $u_T$ is the $P_1$ finite element solution, (174) is replaced by

$$
\|u - u_T\|_{L_\infty(\Omega)} \leq C h_f^2 \ln h_f |u|_{W^2_2(\Omega)}
$$

(175)

$L_{\infty}$ estimates for general triangulations on polygonal domains with reentrant corners and higher-dimensional domains can be found in Schatz and Wahlbin (1978, 1979, 1982) and Schatz (1998). The estimate (175) was also established in Gastaldi and Nochetto (1987) for the Crouzeix-Raviart nonconforming $P_1$ element of Example 16.

It is also known (Rannacher and Scott, 1982; Brenner and Scott, 2002) that

$$
|u - u_T|_{W^1_\infty(\Omega)} \leq C \inf_{v \in V_T} C |u - v|_{W^1_\infty(\Omega)}
$$

(176)

where $\Omega$ is a convex polygonal domain in $\mathbb{R}^2$ and $u_T$ is the $P_n$ ($n \geq 1$) triangular finite element solution obtained from a general triangulation $T$ of $\Omega$. Optimal order estimates for $|u - u_T|_{W^1_\infty(\Omega)}$ can be derived immediately from (176). Extension of (176) to higher dimensions can be found in Schatz and Wahlbin (1995).

### 7.6 Interior estimates and pollution effects

Let $\Omega$ be the $L$-shaped polygon in Figure 1. The solution $u$ of the Poisson problem (5) on $\Omega$ with homogeneous Dirichlet boundary condition is singular near the reentrant corner and $u \notin H^2(\Omega)$. Consequently, the error estimate $|u - u_T|_{H^1(\Omega)} \leq C h_f \|f\|_{L_2(\Omega)}$ does not hold for the $P_1$ triangular finite element solution $u_T$ associated with a quasi-uniform triangulation $T$ of mesh size $h_T$.

However, $u$ does belong to $H^2(\Omega)$ where $\Omega$ is the subset of the points of $\Omega$ whose distances to the reentrant
corner are strictly greater than the positive number \( \delta \). Therefore, it is possible that

\[
|u - u_T|_{H^1(\Omega_3)} \leq C h_T \|f\|_{L^2(\Omega)} \tag{177}
\]

That the estimate (177) indeed holds is a consequence of the following interior estimate (Nitsche and Schatz, 1974):

\[
|u - u_T|_{H^1(\Omega)} \leq C \left( \inf_{v \in V_T} |u - v|_{H^1(\Omega_3)} + \|u - u_T\|_{L^2(\Omega_3)} \right) \tag{178}
\]

where \( V_T \subset H^1_0(\Omega) \) is the \( P_1 \) triangular finite element space. Interior estimates in various Sobolev norms can be established for subdomains of general \( \Omega \) in \( \mathbb{R}^d \) and general finite elements. We refer the readers to Wahlbin (1991) for a survey of such results and to Schatz (2000) for some recent developments.

On the other hand, since \( u_T \) is obtained by solving a global system that involves the nodal values near the reentrant corner of the \( L \)-shaped domain, the effect of the singularity at the reentrant corner can propagate into other parts of \( \Omega \). This is known as the pollution effect and is reflected, for example, by the following estimate (Wahlbin, 1984):

\[
\|u - u_T\|_{L^2(\Omega)} \geq C h_T \tag{179}
\]

where \( \beta = \pi/(3\pi/2) = 2/3 \). Similar estimates can also be established for other Sobolev norms.

### 7.7 Superconvergence

Let \( u_T \) be the finite element solution of a second-order elliptic boundary value problem. Suppose that the space of shape functions on each element contains all the polynomials of degree \( \leq n \) but not all the polynomials of degree \( n + 1 \). Then the \( L_\infty \) norm of the error \( u - u_T \) is at most of order \( h^{n+1} \), even if the solution \( u \) is smooth. However, the absolute value of \( u - u_T \) at certain points can be of order \( h^{n+1+\sigma} \) for some \( \sigma > 0 \). This is known as the phenomenon of superconvergence and such points are the superconvergence points for \( u_T \). Similarly, a point where the absolute value of a derivative of \( u - u_T \) is of order \( h^{n+\sigma} \) is a superconvergence point for the derivative of \( u_T \).

The division points of a partition \( T \) for a two point boundary value problem with smooth coefficients provides

![Figure 21](image-url)

**Figure 21.** Picture of a triangulation \( T \) = \( \text{conv}((-1/2, -1), (-1, -1/2), (-1, -1), \text{conv}((-1, -1/2), (-1/2, -1), (-1/2, -1/2), \ldots, \text{conv}((1/2, 1), (1, 1/2), (1, 1)) \) with \( m = 24 \) triangles and \( n = 21 \) nodes (a). The picture indicates an enumeration of nodes (numbers in circles) and elements (numbers in boxes) given in the matrices coordinates (b) and elements (c). The Dirichlet boundary conditions on the exterior nodes are included in the vector \( \text{dirichlet} = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16) \) of the labels in a counterclockwise enumeration. The data coordinates, elements, and dirichlet are the input of the finite element program to compute a displacement vector \( x \) as its output.
the simplest example of superconvergence points. Let \( u_T \) be the finite element solution from the \( P_n \) Lagrange finite element space. Since the Green’s function \( G_p \) associated with a division point \( p \) is continuous in the interval \( \Omega \) and smooth on the two subintervals divided by \( p \), we have (Douglas and Dupont, 1974)

\[
| (u - u_T)(p) | \leq | a(u - u_T, G_p) | \\
= | a(u - u_T, G_p - \Pi_T G_p) | \\
\leq C \| u - u_T \|_{H^1(\Omega)} \| G_p - \Pi_T G_p \|_{H^1(\Omega)} \leq Ch^n
\]

provided that \( u \) is sufficiently smooth. Therefore, \( p \) is a superconvergence point for \( u_T \) if \( n \geq 2 \).

For general superconvergence results in various dimensions, we refer the readers to Krížek and Neittaanmäki (1987), Chen and Huang (1995), Wahlbin (1987), Chen and Huang (1995), Wahlbin (1996), Chen and Huang (1999), and Babuška and Strouboulis (2001).

### 7.8 Finite element program in 15 lines of MATLAB

It is the purpose of this section to introduce a short (two-dimensional) \( P_1 \) finite element program.

The data for a given triangulation \( T = \{ T_1, \ldots, T_m \} \) into triangles with a set of nodes \( \mathcal{N} = \{ z_1, \ldots, z_n \} \) are described in user-specified matrices called coordinates and elements. Figure 21 displays a triangulation with \( m \) triangles and \( n \) nodes as well as a fixed enumeration and the corresponding data. The coordinates of the nodes \( z_k = (x_k, y_k) \) are real components in general) are stored in the \( k \)th row of the two-dimensional matrix coordinates. Each element \( T_j = \text{conv}(z_k, z_\ell, z_m) \) is represented by the labels of its vertices \( (k, \ell, m) \) stored in the \( j \)th row of the two-dimensional matrix elements. The chosen permutation of \( (k, \ell, m) \) describes the element in a counterclockwise orientation. Homogeneous Dirichlet conditions are prescribed on the boundary specified by an input vector dirichlet of all fixed nodes at the outer boundary; cf. Figure 21.

Given the aforementioned data in the model Dirichlet problem with right-hand side \( f = 1 \), the \( P_1 \) finite element space \( \mathcal{V} := \text{span}(\varphi_j : z_j \in \mathcal{K}) \) is formed by the nodal basis functions \( \varphi_j \) of each free node \( z_k \); the set \( \mathcal{K} \) of free nodes, the interior nodes, is represented in the \( N \) vector freenodes, the vector of labels in \( 1 : n \) without dirichlet.

The resulting discrete equation is the \( N \times N \) linear system of equations \( Ax = b \) with the positive definite symmetric stiffness matrix \( A \) and right-hand side \( b \). Their components are defined (as a subset of)

\[
A_{jk} := \int_\Omega \nabla \varphi_j \cdot \nabla \varphi_k \, dx
\]

and

\[
b_j := \int_\Omega f \varphi_j \, dx \quad \text{for } j, k = 1, \ldots, n
\]

The computation of the entries \( A_{jk} \) and \( b_j \) is performed elementwise for the additivity of the integral and since \( T \) is a partition of the domain \( \Omega \). Given the triangle \( T_j \) number \( j \), the MATLAB command \( \text{coordinates}(\text{elements}(j,:), :) \) returns the \( 3 \times 2 \) matrix \( (P_1, P_2, P_3)^T \) of its vertices. Then, the local stiffness matrix reads

\[
\text{STIMA}(T_j)_{\alpha\beta} := \int_{T_j} \nabla \varphi_\alpha \cdot \nabla \varphi_\beta \, dx \quad \text{for } \alpha, \beta = 1, 2, 3
\]

for those numbers \( k \) and \( \ell \) of two vertices \( z_k = P_a \) and \( z_\ell = P_\beta \) of \( T_j \). The correspondence of global and local indices, i.e. the numbers of vertices in \((z_k, z_\ell, z_m) = (P_1, P_2, P_3)\), of \( T_j \) can be formalized by

\[
I(T_j) = \{(\alpha, k) \in \{1, 2, 3\} \times \{1, \ldots, n\} : P_\alpha = z_k \in \mathcal{N}\}
\]

The local stiffness matrix is in fact

\[
\text{STIMA}(T_j) = \det \frac{P}{2} (QQ^T) \quad \text{with } P := \begin{bmatrix} 1 & 1 & 1 \\ P_1 & P_2 & P_3 \end{bmatrix}
\]

and

\[
Q := P^{-1} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}
\]

This formula allows a compact programming in MATLAB as shown (for any dimension \( d \))

\[
\text{function stima=stima(vertices)} \\
P=\text{[ones(1,size(vertices,2)+1);vertices']}; \\
Q=P\backslash\{\text{zeros(1,size(vertices,2))}\}; \\
\text{eye(size(vertices,2))}; \\
\text{stima=det(P)*Q*Q'/prod(1:size(vertices,2));}
\]

Utilizing the index sets \( I \), the assembling of all local stiffness matrices reads

\[
\text{STIMA} = \sum_{T_j \in T} \sum_{\alpha, k \in I(T_j)} \sum_{\beta, \ell \in I(T_j)} \text{STIMA}(T_j)_{\alpha\beta} e_k \otimes e_\ell
\]

(\( e_k \) is the \( k \)th canonical unit vector with the \( \ell \)th component equal to the Kronecker delta \( \delta_{k\ell} \) and \( \otimes \) is the dyadic product.) The implementation of each summation is realized by adding \( \text{STIMA}(T_j) \) to the \( 3 \times 3 \) submatrix of the rows and columns corresponding to \( k, \ell, m \); see the MATLAB program below.
function x=FEM(coordinates,elements,dirichlet)

A=sparse(size(coordinates,1),size(coordinates,1));
b=sparse(size(coordinates,1),1);x=zeros(size(coordinates,1),1);
for j=1:size(elements,1)
    A(elements(j,:),elements(j,:))=A(elements(j,:),elements(j,:)) ...
        +stima(coordinates(elements(j,:),:));
    b(elements(j,:))=b(elements(j,:))+ones(3,1) ...
        *det([1,1,1;coordinates(elements(j,:),:)'])/6;
end

freeNodes=setdiff(1:size(coordinates,1),dirichlet);
x(freeNodes)=A(freeNodes,freeNodes) \ b(freeNodes);

Given the output vector $x$, a plot of the discrete solution

$$\tilde{u} = \sum_{j=1}^{n} x_j \phi_n$$

is generated by the command `trisurf(elements,coordinates(:,1),coordinates(:,2),x)` and displayed in Figure 22.

For alternative programs with numerical examples and full documentation, the interested readers are referred to Alberty, Carstensen and Funken (1999) and Alberty et al. (2002). The closest more commercial finite element package might be FEMLAB. The internet provides over 200 000 entries under the search for ‘Finite Element Method Program’. Amongst public domain software are the programs ALBERT (Freiburg), DEAL (Heidelberg), UG (Heidelberg), and so on.

![Figure 22. Discrete solution of $-\Delta u = 1$ with homogeneous Dirichlet boundary data based on the triangulation of Figure 21. A color version of this image is available at http://www.mrw.interscience.wiley.com/ecm](http://www.mrw.interscience.wiley.com/ecm)

ACKNOWLEDGMENTS

The work of Susanne C. Brenner was partially supported by the National Science Foundation under Grant Numbers DMS-00-74246 and DMS-03-11790. The work of Carsten Carstensen has been initiated while he was visiting the Isaac-Newton Institute of Mathematical Sciences, Cambridge, England; the support of the EPSRC under grant N09176/01 is thankfully acknowledged. The authors thank Dipl.-Math. Jan Bolte for his assistance with the numerical examples of this chapter.

REFERENCES


Kiříček M, Neittaanmäki P and Stenberg R (eds.). Finite Element Methods: Superconvergence, Post-processing and A Posteriori


Nicai S. *Polygonal Interface Problems.* Peter D Lang Verlag, Frankfurt am Main, 1993.


