Chapter 7

Analysis of the Finite Element Method

7.1 Introduction

Finite element theory is embedded in a very elegant framework that enables accurate \textit{a priori} and \textit{a posteriori} estimates of discretization errors and convergence rates. Unfortunately, a large portion of the theory relies on a knowledge of functional analysis, which has not been assumed in this material. Instead, we present the relevant concepts and key results without proof and cite sources of a more complete treatment. Once again, we focus on the model Galerkin problem: find $u \in H_0^1$ satisfying

$$A(v, u) = (v, f), \quad \forall v \in H_0^1,$$

(7.1.1a)

where

$$(v, f) = \iint_{\Omega} vf \, dx \, dy,$$

(7.1.1b)

$$A(v, u) = \iint_{\Omega} \left[ p(v_x u_x + v_y u_y) + qvu \right] \, dx \, dy,$$

(7.1.1c)

where the two-dimensional domain $\Omega$ has boundary $\partial \Omega = \partial \Omega_E \cup \partial \Omega_N$. For simplicity, we have assumed trivial essential and natural boundary data on $\partial \Omega_E$ and $\partial \Omega_N$, respectively.

Finite element solutions $U \in S_0^N$ of (7.1.1) satisfy

$$A(V, U) = (V, f), \quad \forall V \in S_0^N,$$

(7.1.2)

where $S_0^N$ is a finite-dimensional subspace of $H_0^1$.

As described in Chapter 2, error analysis typically proceeds in two steps:
1. showing that $U$ is optimal in the sense that the error $u - U$ satisfies

$$\|u - U\| = \min_{W \in S^N} \|u - W\|$$

in an appropriate norm, and

2. finding an upper bound for the right-hand side of (7.1.3).

The appropriate norm to use with (7.1.3) for the model problem (7.1.1) is the strain energy norm

$$\|v\|_A = \sqrt{A(v, v)}.$$  \hfill (7.1.4)

The finite element solution might not satisfy (7.1.3) with other norms and/or problems. For example, finite element solutions are not optimal in any norm for non-self-adjoint problems. In these cases, (7.1.3) is replaced by the weaker statement

$$\|u - U\| \leq C \min_{W \in S^N} \|u - W\|,$$

$C > 1$. Thus, the solution is “nearly best” in the sense that it only differs by a constant from the best possible solution in the space.

Upper bounds of the right-hand sides of (7.1.3) or (7.1.5) are obtained by considering the error of an interpolant $W$ of $u$. Using Theorems 2.6.4 and 4.6.5, for example, we could conclude that

$$\|u - W\|_s \leq C h^{p+1-s}\|u\|_{p+1}, \quad s = 0, 1,$$

if $S^N$ consists of complete piecewise polynomials of degree $p$ with respect to a sequence of uniform meshes (cf. Definition 4.6.1) and $u \in H^{p+1}$. The bound (7.1.6) can be combined with either (7.1.3) or (7.1.5) to provide an estimate of the error and convergence rate of a finite element solution.

The Sobolev norm on $H^1$ and the strain energy norm (7.1.4) are equivalent for the model problem (7.1.1) and we shall use this with (7.1.3) and (7.1.6) to construct error estimates. Prior to continuing, you may want to review Sections 2.6, 3.2, and 4.6.

_A priori_ finite element discretization errors, obtained as described, do not account for such “perturbations” as

1. using numerical integration,

2. interpolating Dirichlet boundary conditions by functions in $S^N$, and

3. approximating $\partial \Omega$ by piecewise-polynomial functions.
These effects will have to be appraised. Additionally, the \textit{a priori} error estimates supply information on convergence rates but are difficult to use for quantitative error information. \textit{A posteriori} error estimates, which use the computed solution, provide more practical accuracy appraisals.

\section*{7.2 Convergence and Optimality}

While keeping the model problem (7.1.1) in mind, we will proceed in a slightly more general manner by considering a Galerkin problem of the form (7.1.1a) with a strain energy $A(v, u)$ that is a symmetric bilinear form (cf. Definitions 3.2.2, 3) and is also continuous and coercive.

\textbf{Definition 7.2.1.} A bilinear form $A(v, u)$ is \textit{continuous} in $H^s$ if there exists a constant $\alpha > 0$ such that

\begin{equation}
|A(v, u)| \leq \alpha \|u\|_s \|v\|_s, \quad \forall u, v \in H^s.
\end{equation}

\textbf{Definition 7.2.2.} A bilinear form $A(u, v)$ is \textit{coercive} ($H^s$ - \textit{elliptic} or \textit{positive definite}) in $H^s$ if there exists a constant $\beta > 0$ such that

\begin{equation}
A(u, u) \geq \beta \|u\|_s^2, \quad \forall u \in H^s.
\end{equation}

Continuity and coercivity of $A(v, u)$ can be used to establish the existence and uniqueness of solutions to the Galerkin problem (7.1.1a). These results follow from the Lax-Milgram Theorem. We'll subsequently prove a portion of this result, but more complete treatments appear elsewhere [6, 12, 13, 15]. We'll use examples to provide insight into the meanings of continuity and coercivity.

\textbf{Example 7.2.1.} Consider the variational eigenvalue problem: determine nontrivial $u \in H^1_0$ and $\lambda \in [0, \infty)$ satisfying

$$A(u, v) = \lambda(u, v), \quad \forall v \in H^1_0.$$ 

When $A(v, u)$ is the strain energy for the model problem (7.1.1), smooth solutions of this variational problem also satisfy the differential eigenvalue problem

$$-(pu_x)_x - (pu_y)_y + qu = \lambda u, \quad (x, y) \in \Omega,$$

$$u = 0, \quad (x, y) \in \partial\Omega_E, \quad u_n = 0, \quad (x, y) \in \partial\Omega_N.$$ 

where $n$ is the unit outward normal to $\partial\Omega$. 
Letting $\lambda_r$ and $u^r$, $r \geq 1$, be an eigenvalue-eigenfunction pair and using the variational statement with $v = u = u^r$, we obtain the Rayleigh quotient

$$\lambda_r = \frac{A(u^r, u^r)}{(u^r, u^r)}$$

Since this result holds for all $r$, we have

$$\lambda_1 = \min_{r \geq 1} \frac{A(u^r, u^r)}{(u^r, u^r)}$$

where $\lambda_1$ is the minimum eigenvalue. (As indicated in Problem 1, this result can be extended.)

Using the Rayleigh quotient with (7.2.2), we have

$$\lambda_r \geq \frac{\beta \|u^r\|_s^2}{\|u^r\|_0^2}, \quad r \geq 1.$$ 

Since $\|u^r\|_s \geq \|u^r\|_0$, we have

$$\lambda_r \geq \beta > 0, \quad r \geq 1.$$ 

Thus, $\beta \leq \lambda_r$, $r \geq 1$, and, in particular, $\beta \leq \lambda_1$.

Using (7.2.1) in conjunction with the Rayleigh quotient implies

$$\lambda_r \leq \frac{\alpha \|u^r\|_s^2}{\|u^r\|_0^2}, \quad r \geq 1.$$ 

Combining the two results,

$$\frac{\beta \|u^r\|_s^2}{\|u^r\|_0^2} \leq \lambda_r \leq \frac{\alpha \|u^r\|_s^2}{\|u^r\|_0^2}, \quad r \geq 1.$$ 

Thus, $\beta$ provides a lower bound for the minimum eigenvalue and $\alpha$ provides a bound for the maximum growth rate of the eigenvalues in $H^r$.

**Example 7.2.2.** Solutions of the Dirichlet problem

$$-u_{xx} - u_{yy} = f(x, y), \quad (x, y) \in \Omega, \quad u = 0, \quad (x, y) \in \partial \Omega,$$

satisfy the Galerkin problem (7.1.1) with

$$A(v, u) = \iint_{\Omega} \nabla v \cdot \nabla u \, dx \, dy, \quad \nabla u = [u_x, u_y]^T.$$

An application of Cauchy’s inequality reveals

$$|A(v, u)| = \left| \iint_{\Omega} \nabla v \cdot \nabla u \, dx \, dy \right| \leq \|\nabla v\|_0 \|\nabla u\|_0.$$
where
\[
\|\nabla u\|_0^2 = \iint_{\Omega} (v_x^2 + v_y^2) \, dx \, dy.
\]

Since \( \|\nabla u\|_0 \leq \|u\|_1 \), we have
\[
|A(v, u)| \leq \|v\|_1 \|u\|_1.
\]

Thus, (7.2.1) is satisfied with \( s = 1 \) and \( \alpha = 1 \), and the strain energy is continuous in \( H^1 \).

Establishing that \( A(v, u) \) is coercive in \( H^1 \) is typically done by using Friedrichs’s first inequality which states that there is a constant \( \gamma > 0 \) such that
\[
\|\nabla u\|_0^2 \geq \gamma \|u\|_0^2.
\] (7.2.3)

Now, consider the identity
\[
A(u, u) = \|\nabla u\|_0^2 = (1/2)\|\nabla u\|_0^2 + (1/2)\|\nabla u\|_0^2
\]
and use (7.2.3) to obtain
\[
A(u, u) \geq (1/2)\|\nabla u\|_0^2 + (1/2)\gamma \|u\|_0^2 \geq \beta \|u\|_1^2
\]
where \( \beta = (1/2) \max(1, \gamma) \). Thus, (7.2.2) is satisfied with \( s = 1 \) and \( A(u, v) \) is coercive \( (H^1\text{-elliptic}) \).

Continuity and coercivity of the strain energy reveal the finite element solution \( U \) to be nearly the best approximation in \( S^N \).

**Theorem 7.2.1.** Let \( A(v, u) \) be symmetric, continuous, and coercive. Let \( u \in H^1_0 \) satisfy (7.1.1a) and \( U \in S^N_0 \subset H^1_0 \) satisfy (7.1.2). Then
\[
\|u - U\|_1 \leq \frac{\alpha}{\beta} \|u - V\|_1, \quad \forall V \in S^N_0,
\] (7.2.4a)

with \( \alpha \) and \( \beta \) satisfying (7.2.1) and (7.2.2).

**Remark 1.** Equation (7.2.4a) may also be expressed as
\[
\|u - U\|_1 \leq C \inf_{V \in S^N_0} \|u - V\|_1.
\] (7.2.4b)

Thus, continuity and \( H^1 \)-ellipticity give us a bound of the form (7.1.5).

**Proof.** cf. Problem 2 at the end of this section.

The bound (7.2.4) can be improved when \( A(v, u) \) has the form (7.1.1c).
Theorem 7.2.2. Let \( A(v, u) \) be a symmetric, continuous, and coercive bilinear form; \( u \in H^1_0 \) minimize

\[
I[w] = A(w, w) - 2(w, f), \quad \forall w \in H^1_0;
\] (7.2.5)

and \( S^N_0 \) be a finite-dimensional subspace of \( H^1_0 \). Then

1. The minimum of \( I[W] \) and \( A(u - W, u - W) \), \( \forall W \in S^N_0 \), are achieved by the same function \( U \).

2. The function \( U \) is the orthogonal projection of \( u \) onto \( S^N_0 \) with respect to strain energy, i.e.,

\[
A(V, u - U) = 0, \quad \forall V \in S^N_0. \tag{7.2.6}
\]

3. The minimizing function \( U \in S^N_0 \) satisfies the Galerkin problem

\[
A(V, U) = (V, f), \quad \forall V \in S^N_0. \tag{7.2.7}
\]

In particular, if \( S^N_0 \) is the whole of \( H^1_0 \)

\[
A(v, u) = (v, f), \quad \forall v \in H^1_0. \tag{7.2.8}
\]

Proof. Our proof will omit several technical details, which appear in, e.g., Wait and Mitchell [21], Chapter 6.

Let us begin with (7.2.7). If \( U \) minimizes \( I[W] \) over \( S^N_0 \) then for any \( \epsilon \) and any \( V \in S^N_0 \)

\[
I[U] \leq I[U + \epsilon V].
\]

Using (7.2.5),

\[
I[U] \leq A(U + \epsilon V, U + \epsilon V) - 2(U + \epsilon V, f)
\]

or

\[
I[U] \leq I[U] + 2\epsilon [A(V, U) - (V, f)] + \epsilon^2 A(V, V)
\]

or

\[
0 \leq 2\epsilon [A(V, U) - (V, f)] + \epsilon^2 A(V, V).
\]

This inequality must hold for all possible \( \epsilon \) of either sign; thus, (7.2.7) must be satisfied. Equation (7.2.8) follows by repeating these arguments with \( S^N_0 \) replaced by \( H^1_0 \).

Next, replace \( v \) in (7.2.8) by \( V \in S^N_0 \subset H^1_0 \) and subtract (7.2.7) to obtain (7.2.6).

In order to prove Conclusion 1, consider the identity

\[
\]
Using (7.2.6)
\[ A(u - U, u - U) = A(u - U - V, u - U - V) - A(V, V). \]
Since \( A(V, V) \geq 0 \),
\[ A(u - U, u - U) \leq A(u - U - V, u - U - V), \quad \forall V \in S_0^N. \]
Equality only occurs when \( V = 0 \); therefore, \( U \) is the unique minimizing function. \( \square \)

**Remark 2.** We proved a similar result for one-dimensional problems in Theorems 2.6.1, 2.

**Remark 3.** Continuity and coercivity did not appear in the proof; however, they are needed to establish existence, uniqueness, and completeness. Thus, we never proved that \( \lim_{N \to \infty} U = u \). A complete analysis appears in Wait and Mitchell [21], Chapter 6.

**Remark 4.** The strain energy \( A(v, u) \) not need be symmetric. A proof without this restriction appears in Ciarlet [13].

**Corollary 7.2.1.** With the assumptions of Theorem 7.2.2,
\[ A(u - U, u - U) = A(u, u) - A(U, U). \] (7.2.9)
**Proof.** cf. Problem 3 at the end of this section. \( \square \)

In Section 4.6, we obtained \textit{a priori} estimates of interpolation errors under some mesh uniformity assumptions. Recall (cf. Definition 4.6.1), that we considered a family of finite element meshes \( \Delta_h \) which became finer as \( h \to 0 \). The uniformity condition implied that all vertex angles were bounded away from 0 and \( \pi \) and that all aspect ratios were bounded away from 0 as \( h \to 0 \). Uniformity ensured that transformations from the physical to the computational space were well behaved. Thus, with uniform meshes, we were able to show (cf. Theorem 4.6.5) that the error in interpolating a function \( u \in H^{p+1} \) by a complete polynomial \( W \) of degree \( p \) satisfies
\[ \|u - W\|_s \leq Ch^{p+1-s}\|u\|_{p+1}, \quad s = 0, 1. \] (7.2.10a)

The norm on the right can be replaced by the seminorm
\[ |u|_{p+1}^2 = \sum_{|\kappa|=p+1} \|D^\kappa u\|_0^2 \] (7.2.10b)
to produce a more precise estimate, but this will not be necessary for our present application. If singularities are present so that \( u \in H^{q+1} \) with \( q < p \) then, instead of (7.2.10a), we find
\[ \|u - W\|_1 \leq Ch^q\|u\|_{q+1}. \] (7.2.10c)
With optimality (or near optimality) established and interpolation error estimates available, we can establish convergence of the finite element method.

**Theorem 7.2.3.** Suppose:

1. \( u \in H^1_0 \) and \( U \in S^N_0 \subset H^1_0 \) satisfy (7.2.8) and (7.2.7), respectively.
2. \( A(v, u) \) is a symmetric, continuous, and \( H^1 \)-elliptic bilinear form;
3. \( S^N_0 \) consists of complete piecewise-polynomial functions of degree \( p \) with respect to a uniform family of meshes \( \Delta_h \); and
4. \( u \in H^1_0 \cap H^{p+1} \).

Then

\[
\|u - U\|_1 \leq Ch^p \|u\|_{p+1}
\]  \hspace{1cm} (7.2.11a)

and

\[
A(u - U, u - U) \leq Ch^{2p} \|u\|_{p+1}^2.
\]  \hspace{1cm} (7.2.11b)

**Proof.** From Theorem 7.2.2

\[
A(u - U, u - U) = \inf_{V \in S^N_0} A(u - V, u - V) \leq A(u - W, u - W)
\]

where \( W \) is an interpolant of \( u \). Using (7.2.1) with \( s = 1 \) and \( v \) and \( u \) replaced by \( u - W \) yields

\[
A(u - W, u - W) \leq \alpha \|u - W\|_1^2.
\]

Using the interpolation estimate (7.2.10a) with \( s = 1 \) yields (7.2.11b). In order to prove (7.2.11a), use (7.2.2) with \( s = 1 \) to obtain

\[
\beta \|u - U\|_1^2 \leq A(u - U, u - U).
\]

The use of (7.2.11b) and a division by \( \beta \) yields (7.2.11a).

Since the \( H^1 \) norm dominates the \( L^2 \) norm, (7.2.11a) trivially gives us an error estimate in \( L^2 \) as

\[
\|u - U\|_0 \leq Ch^p \|u\|_{p+1}.
\]

This estimate does not have an optimal rate since the interpolation error (7.2.10a) is converging as \( O(h^{p+1}) \). Getting the correct rate for an \( L^2 \) error estimate is more complicated than it is in \( H^1 \). The proof is divided into two parts.
Lemma 7.2.1. (Aubin-Nitsche) Under the assumptions of Theorem 7.2.3, let $\gamma(x,y) \in H^1_0$ be the solution of the “dual problem”

$$A(v, \gamma) = (v, e), \quad \forall v \in H^1_0; \tag{7.2.12a}$$

where

$$e = \frac{u - U}{\|u - U\|_0}. \tag{7.2.12b}$$

Let $\Gamma \in S^N_0$ be an interpolant of $\gamma$, then

$$\|u - U\|_0 \leq \alpha \|u - U\|_1 \|\gamma - \Gamma\|_1. \tag{7.2.12c}$$

Proof. Set $V = \Gamma$ in (7.2.6) to obtain

$$A(\Gamma, u - U) = 0. \tag{7.2.13}$$

Take the $L^2$ inner product of (7.2.12b) with $u - U$ to obtain

$$\|u - U\|_0 = (e, u - U).$$

Setting $v = u - U$ in (7.2.12a) and using the above relation yields

$$\|u - U\|_0 = A(u - U, \gamma).$$

Using (7.2.13)

$$\|u - U\|_0 = A(u - U, \gamma - \Gamma).$$

Now use the continuity of $A(v, u)$ in $H^1$ ((7.2.1) with $s = 1$) to obtain (7.2.12c). \hfill \square

Since we have an estimate for $\|u - U\|_1$, estimating $\|u - U\|_0$ by (7.2.12c) requires an estimate of $\|\gamma - \Gamma\|_1$. This, of course, will be done by interpolation; however, use of (7.2.10a) requires knowledge of the smoothness of $\gamma$. The following lemma provides the necessary a priori bound.

Lemma 7.2.2. Let $A(u, v)$ be a symmetric, $H^1$-elliptic bilinear form and $u$ be the solution of (7.2.8) on a smooth region $\Omega$. Then

$$\|u\|_2 \leq C\|f\|_0. \tag{7.2.14}$$

Remar 5. This result seems plausible since the underlying differential equation is of second order, so the second derivatives should have the same smoothness as the right-hand side $f$. The estimate might involve boundary data; however, we have assumed trivial conditions. Let’s further assume that $\partial \Omega_E$ is not nil to avoid non-uniqueness issues.
Proof. Strang and Fix [18], Chapter 1, establish (7.2.14) in one dimension. Johnson [14], Chapter 4, obtain a similar result.

With preliminaries complete, here is the main result.

**Theorem 7.2.4.** Given the assumptions of Theorem 7.2.3, then

\[ \|u - U\|_0 \leq Ch^{p+1}\|u\|_{p+1}. \]  

(7.2.15)

Proof. Applying (7.2.14) to the dual problem (7.2.12a) yields

\[ \|\gamma\|_2 \leq C\|\varepsilon\|_0 = C, \]

since \( \|\varepsilon\|_0 = 1 \) according to (7.2.12b). With \( \gamma \in H_2 \), we may use (7.2.10c) with \( q = s = 1 \) to obtain

\[ \|\gamma - \Gamma\|_1 \leq Ch\|\gamma\|_2 = Ch. \]

Combining this estimate with (7.2.11a) and (7.2.12c) yields (7.2.15).

Problems

1. Show that the function \( u \) that minimizes

\[ \lambda = \min_{w \in H_0^1, \|w\|_0 \neq 0} \frac{A(w, w)}{(w, w)} \]

is \( u^1 \), the eigenfunction corresponding to the minimum eigenvalue \( \lambda_1 \) of \( A(v, u) = \lambda(v, u) \).

2. Assume that \( A(v, u) \) is a symmetric, continuous, and \( H^1 \)-elliptic bilinear form and, for simplicity, that \( u, v \in H_0^1 \).

2.1. Show that the strain energy and \( H^1 \) norms are equivalent in the sense that

\[ \beta\|u\|_1^2 \leq A(u, u) \leq \alpha\|u\|_1^2, \quad \forall u \in H_0^1, \]

where \( \alpha \) and \( \beta \) satisfy (7.2.1) and (7.2.2).

2.2. Prove Theorem 7.2.1.

3. Prove Corollary 7.2.1 to Theorem 7.2.2.

7.3 Perturbations

In this section, we examine the effects of perturbations due to numerical integration, interpolated boundary conditions, and curved boundaries.
7.3.1 Quadrature Perturbations

With numerical integration, we determine $U^*$ as the solution of

$$A_s(V, U^*) = (V, f)_s, \quad \forall V \in S_0^N,$$

(7.3.1a)

instead of determining $U$ by solving (7.2.8). The approximate strain energy $A_s(V, U)$ or $L^2$ inner product $(V, f)_s$ reflect the numerical integration that has been used. For example, consider the loading

$$(V, f) = \sum_{e=1}^{N_\Delta} (V, f)_e, \quad (V, f)_e = \int_{\Omega_e} V(x, y) f(x, y) dx dy$$

where $\Omega_e$ is the domain occupied by element $e$ in a mesh of $N_\Delta$ elements. Using an $n$-point quadrature rule (cf. (6.1.2a)) on element $e$, we would approximate $(V, f)$ by

$$(V, f)_s = \sum_{e=1}^{N_\Delta} (V, f)_e,\quad (7.3.1b)$$

where

$$(V, f)_e = \sum_{k=1}^n W_k V(x_k, y_k) f(x_k, y_k).\quad (7.3.1c)$$

The effects of transformations to a canonical element have not been shown for simplicity and a similar formula applies for $A_s(V, U)$.

Deriving an estimate for the perturbation introduced by (7.3.1a) is relatively simple if $A(V, U)$ and $A_s(V, U)$ are continuous and coercive.

**Theorem 7.3.1.** Suppose that $A(v, u)$ and $A_s(V, U)$ are bilinear forms with $A$ being continuous and $A_s$ being coercive in $H^1$; thus, there exists constants $\alpha$ and $\beta$ such that

$$|A(u, v)| \leq \alpha ||u||_1 ||v||_1, \quad \forall u, v \in H^1_0;\quad (7.3.2a)$$

and

$$A_s(U, U) \geq \beta ||U||^2_1, \quad \forall U \in S_0^N.\quad (7.3.2b)$$

Then

$$||u - U^*||_1 \leq C \{ ||u - V||_1 + \sup_{W \in S_0^N} \frac{|A(V, W) - A_s(V, W)|}{||W||_1} + \sup_{W \in S_0^N} \frac{|(W, f) - (W, f)_s|}{||W||_1} \}, \quad \forall V \in S_0^N.\quad (7.3.3)$$
Analysis of the Finite Element Method

Proof. Using the triangular inequality
\[ \| u - U^* \|_1 \leq \| u - V + V - U^* \|_1 \leq \| u - V \|_1 + \| W \|_1 \] (7.3.4a)
where
\[ W = U^* - V. \] (7.3.4b)

Using (7.3.2b) and (7.3.4b)
\[ \beta \| W \|_1^2 \leq A_*(U^* - V, W) = A_*(U^*, W) - A_*(V, W). \]

Using (7.3.1a) with \( V \) replaced by \( W \) to eliminate \( A_*(U^*, W) \), we get
\[ \beta \| W \|_1^2 \leq (f, W)_* - A_*(V, W). \]

Adding the exact Galerkin equation (7.2.8) with \( v \) replaced by \( W \)
\[ \beta \| W \|_1^2 \leq (f, W)_* - (f, W) + A(u, W) - A_*(V, W). \]

Adding and subtracting \( A(V, W) \) and taking an absolute value
\[ \beta \| W \|_1^2 \leq |(f, W)_* - (f, W)| + |A(u - V, W)| + |A(V, W) - A_*(V, W)|. \]

Now, using the continuity condition (7.3.2a) with \( u \) replaced by \( u - V \) and \( v \) replaced by \( W \), we obtain
\[ \beta \| W \|_1^2 \leq |(f, W)_* - (f, W)| + \alpha \| u - V \|_1 \| W \|_1 + |A(V, W) - A_*(V, W)|. \]

Dividing by \( \beta \| W \|_1 \)
\[ \| W \|_1 \leq \frac{1}{\beta} \left\{ \alpha \| u - V \|_1 + \frac{|(f, W)_* - (f, W)|}{\| W \|_1} + \frac{|A(V, W) - A_*(V, W)|}{\| W \|_1} \right\}. \]

Combining the above inequality with (7.3.4a), maximizing the inner product ratios over \( W \), and choosing \( C \) as the larger of \( 1 + \alpha/\beta \) or \( 1/\beta \) yields (7.3.3). \qed

Remark 1. Since the error estimate (7.3.3) is valid for all \( V \in S_0^N \) it can be written in the form
\[ \| u - U^* \|_1 \leq C \inf_{V \in S_0^N} \left\{ \| u - V \|_1 + \sup_{W \in S_0^N} \frac{|A(V, W) - A_*(V, W)|}{\| W \|_1} \right\} + \sup_{W \in S_0^N} \frac{|(W, f) - (W, f)_*|}{\| W \|_1}. \] (7.3.5)

To bound (7.3.3) or (7.3.5) in terms of a mesh parameter \( h \), we use standard interpolation error estimates \((cf. \text{Sections 2.6 and 4.6})\) for the first term and numerical integration error estimates \((cf. \text{Chapter 6})\) for the latter two terms. Estimating quadrature errors is relatively easy and the following typical result includes the effects of transforming to a canonical element.
Theorem 7.3.2. Let $\mathbf{J}(\xi, \eta)$ be the Jacobian of a transformation from a computational $(\xi, \eta)$-plane to a physical $(x, y)$-plane and let $W \in S_0^N$. Relative to a uniform family of meshes $\Delta_h$, suppose that $\det(\mathbf{J}(\xi, \eta))W_x(\xi, \eta)$ and $\det(\mathbf{J}(\xi, \eta))W_y(\xi, \eta)$ are piecewise polynomials of degree at most $r_1$ and $\det(\mathbf{J}(\xi, \eta))W(\xi, \eta)$ is a piecewise polynomial of degree at most $r_0$. Then:

1. If a quadrature rule is exact (in the computational plane) for all polynomials of degree at most $r_1 + r$,
   \[
   \frac{|A(V, W) - A_s(V, W)|}{\|W\|_1} \leq C h^{r+1} \|V\|_{r+2}, \quad \forall V, W \in S_0^N, \tag{7.3.6a}
   \]
   2. If a quadrature rule is exact for all polynomials of degree at most $r_0 + r - 1$,
   \[
   \frac{|(f, W) - (f, W)_s|}{\|W\|_1} \leq C h^{r+1} \|f\|_{r+1}, \quad \forall W \in S_0^N. \tag{7.3.6b}
   
   Proof. cf. Wait and Mitchell [21], Chapter 6, or Strang and Fix [18], Chapter 4.

Example 7.3.1. Suppose that the coordinate transformation is linear so that $\det(\mathbf{J}(\xi, \eta))$ is constant and that $S_0^N$ consists of piecewise polynomials of degree at most $p$. In this case, $r_1 = p - 1$ and $r_0 = p$. The interpolation error in $H^1$ is
   \[
   \|u - V\|_1 = O(h^p).
   
   Suppose that the quadrature rule is exact for polynomials of degree $\rho$ or less. Thus, $\rho = r_1 + r$ or $\rho = r_0 + r - 1$ and (7.3.6a) implies that
   \[
   \frac{|A(V, W) - A_s(V, W)|}{\|W\|_1} \leq C h^{\rho-p+2} \|V\|_{\rho-p+3}, \quad \forall V, W \in S_0^N.
   \]

With $\rho = r_0 + r - 1$ and $r_0 = p$, we again find $\rho = r_0 = p + 1$ and, using (7.3.6b),
   \[
   \frac{|(f, W) - (f, W)_s|}{\|W\|_1} \leq C h^{\rho-p+2} \|f\|_{\rho-p+2}, \quad \forall W \in S_0^N.
   
   • If $\rho = 2(p-1)$ so that $r = p-1$ then the above perturbation errors are $O(h^p)$. Hence, all terms in (7.3.3) or (7.3.5) have the same order of accuracy and we conclude that
   \[
   \|u - U^*\|_1 = O(h^p).
   
   This situation is regarded as optimal. If the coefficients of the differential equation are constant and, as is the case here, the Jacobian is constant, this result is equivalent to integrating the differentiated terms in the strain energy exactly (cf., e.g., (7.1.1c)).
Analysis of the Finite Element Method

- If $\rho > 2(p-1)$ so that $r > p-1$ then the error in integration is higher order than the $O(h^p)$ interpolation error; however, the interpolation error dominates and
  
  $$\|u - U^*\|_1 = O(h^p).$$

  The extra effort in performing the numerical integration more accurately is not justified.

- If $\rho < 2(p-1)$ so that $r < p-1$ then the integration error dominates the interpolation error and determines the order of accuracy as
  
  $$\|u - U^*\|_1 = O(h^{\rho - p+2}).$$

  In particular, convergence does not occur if $\rho \leq p-2$.

Let us conclude this example by examining convergence rates for piecewise-linear (or bilinear) approximations ($p = 1$). In this case, $r_1 = 0$, $r_0 = 1$, and $r = \rho$. Interpolation errors converge as $O(h)$. The optimal order of accuracy of the quadrature rule is $\rho = 0$, i.e., only constant functions need be integrated exactly. Performing the integration more accurately yields no improvement in the convergence rate.

**Example 7.3.2.** Problems with variable Jacobians are more complicated. Consider the term

$$\det(J(\xi, \eta))W_x(\xi, \eta) = J(W_\xi \xi_x + W_\eta \eta_x)$$

where $J = \det(J(\xi, \eta))$. The metrics $\xi_x$ and $\eta_x$ are obtained from the inverse Jacobian

$$J^{-1} = \begin{bmatrix} \xi_x & \xi_y \\ \eta_x & \eta_y \end{bmatrix} = \frac{1}{J} \begin{bmatrix} y_\eta & -x_\eta \\ -y_\xi & x_\xi \end{bmatrix}.$$ 

In particular, $\xi_x = y_\eta / J$ and $\eta_x = -y_\xi / J$ and

$$\det(J)W_x = W_\xi y_\eta - W_\eta y_\xi.$$ 

Consider an isoparametric transformation of degree $p$. Such triangles or quadrilaterals in the computational plane have curved sides of piecewise polynomials of degree $p$ in the physical plane. If $W$ is a polynomial of degree $p$ then $W_x$ has degree $p - 1$. Likewise, $x$ and $y$ are polynomials of degree $p$ in $\xi$ and $\eta$. Thus, $y_\eta$ and $y_\xi$ also have degrees $p - 1$. Therefore, $JW_x$ and, similarly, $JW_\eta$ have degrees $r_1 = 2(p-1)$. With $J$ being a polynomial of degree $2(p-1)$, we find $JW$ to be of degree $r_0 = 3p - 2$.

For the quadrature errors (7.3.6) to have the same $O(h^p)$ rate as the interpolation error, we must have $r = p - 1$ in (7.3.6a,b). Thus, according to Theorem 7.3.2, the order $\rho$ of the quadrature rules in the $(\xi, \eta)$-plane should be

$$\rho = r_1 + r = 2(p-1) + (p-1) = 3(p-1).$$
for (7.3.6a) and
\[
\rho = r_0 + r - 1 = (3p - 2) + (p - 1) - 1 = 4(p - 1)
\]
for (7.3.6b). These results are to be compared with the order of \(2(p - 1)\) that was needed with the piecewise polynomials of degree \(p\) and linear transformations considered in Example 7.3.1. For quadratic transformations and approximations \((p = 2)\), we need third- and fourth-order quadrature rules for \(O(h^2)\) accuracy.

### 7.3.2 Interpolated Boundary Conditions

Assume that integration is exact and the boundary \(\partial \Omega\) is modeled exactly, but Dirichlet boundary data is approximated by a piecewise polynomial in \(S^N\), *i.e.*, by a polynomial having the same degree \(p\) as the trial and test functions. Under these conditions, Wait and Mitchell [21], Chapter 6, show that the error in the solution \(U\) of a Galerkin problem with interpolated boundary conditions satisfies

\[
\|u - U\|_1 \leq C \{h^p \|u\|_{p+1} + h^{p+1/2} \|u\|_{p+1}\}. \tag{7.3.7}
\]

The first term on the right is the standard interpolation error estimate. The second term corresponds to the perturbation due to approximating the boundary condition. As usual, computation is done on a uniform family of meshes \(\Delta_h\) and \(u\) is smooth enough to be in \(H^{p+1}\). Brenner and Scott [12], Chapter 8, obtain similar results under similar conditions when interpolation is performed at the Lobatto points on the boundary of an element. The Lobatto polynomial of degree \(p\) is defined on \([-1, 1]\) as

\[
L_p(\xi) = \frac{d^{p-2}}{d\xi^{p-2}}(1 - \xi^2)^{p-1}, \quad \xi \in [-1, 1], \quad p \geq 2.
\]

These results are encouraging since the perturbation in the boundary data is of slightly higher order than the interpolation error. Unfortunately, if the domain \(\Omega\) is not smooth and, *e.g.*, contains corners solutions will not be elements of \(H^{p+1}\). Less is known in these cases.

### 7.3.3 Perturbed Boundaries

Suppose that the domain \(\Omega\) is replaced by a polygonal domain \(\tilde{\Omega}\) as shown in Figure 7.3.1. Strang and Fix [18], analyze second-order problems with homogeneous Dirichlet data of the form: determine \(u \in H^1_0\) satisfying

\[
A(v, u) = (v, f), \quad \forall v \in H^1_0, \tag{7.3.8a}
\]
where functions in $H^1_0$ satisfy $u(x, y) = 0$, $(x, y) \in \partial \Omega$. The finite element solution $U \in \tilde{S}^N_0$ satisfies

$$A(V, U) = (V, f), \quad \forall V \in \tilde{S}^N_0,$$

where functions in $\tilde{S}^N_0$ vanish on $\partial \Omega$. (Thus, $\tilde{S}^N_0$ is not a subspace of $H^1_0$.)

For piecewise linear polynomial approximations on triangles they show that $\|u - U\|_1 = O(h)$ and for piecewise quadratic approximations $\|u - U\|_1 = O(h^3/2)$. The poor accuracy with quadratic polynomials is due to large errors in a narrow “boundary layer” near $\partial \Omega$. Large errors are confined to the boundary layer and results are acceptable elsewhere. Wait and Mitchell [21], Chapter 6, quote other results which prove that $\|u - U\|_1 = O(h^p)$ for $p$th degree piecewise polynomial approximations when the distance between $\partial \Omega$ and $\partial \Omega$ is $O(h^{p+1})$. Such is the case when $\partial \Omega$ is approximated by $p$th degree piecewise-polynomial interpolation.

### 7.4 A Posteriori Error Estimation

In previous sections of this chapter, we considered *a priori error estimates*. Thus, we can, without computation, infer that finite element solutions converge at a certain rate depending on the exact solution’s smoothness. Error bounds are expressed in terms of unknown constants which are difficult, if not impossible, to estimate. Having computed a finite element solution, it is possible to obtain *a posteriori error estimates* which give more quantitative information about the accuracy of the solution. Many error estimation techniques are available and before discussing any, let’s list some properties that a good *a posteriori* error estimation procedure should possess.

- The error estimate should give an accurate measure of the discretization error for a wide range of mesh spacings and polynomial degrees.
• The procedure should be inexpensive relative to the cost of obtaining the finite element solution. This usually means that error estimates should be calculated using only local computations, which typically require an effort comparable to the cost of generating the stiffness matrix.

• A technique that provides estimates of pointwise errors which can subsequently be used to calculate error measures in several norms is preferable to one that only works in a specific norm. Pointwise error estimates and error estimates in local (elemental) norms may also provide an indications as to where solution accuracy is insufficient and where refinement is needed.

* * *  

A *posteriori* error estimates can roughly be divided into four categories.

1. *Residual error estimates.* Local finite element problems are created on either an element or a subdomain and solved for the error estimate. The data depends on the residual of the finite element solution.

2. *Flux-projection error estimates.* A new flux is calculated by post processing the finite element solution. This flux is smoother than the original finite element flux and an error estimate is obtained from the difference of the two fluxes.

3. *Extrapolation error estimates.* Two finite element solutions having different orders or different meshes are compared and their differences used to provide an error estimate.

4. *Interpolation error estimates.* Interpolation error bounds are used with estimates of the unknown constants.

The four techniques are not independent but have many similarities. Surveys of error estimation procedures [7, 20] describe many of their properties, similarities, and differences. Let us set the stage by briefly describing two simple extrapolation techniques. Consider a one-dimensional problem for simplicity and suppose that an approximate solution $U^p_h(x)$ has been computed using a polynomial approximation of degree $p$ on a mesh of spacing $h$ (Figure 7.4.1). Suppose that we have an *a priori* interpolation error estimate of the form

$$u(x) - U^p_h(x) = C_{p+1}h^{p+1} + O(h^{p+2}).$$

We have assumed that the exact solution $u(x)$ is smooth enough for the error to be expanded in $h$ to $O(h^{p+2})$. The leading error constant $C_{p+1}$ generally depends on (unknown) derivatives of $u$. Now, compute a second solution with spacing $h/2$ (Figure 7.4.1) to obtain

$$u(x) - U^p_{h/2}(x) = C_{p+1}h^{p+1}/2 + O(h^{p+2}).$$
Figure 7.4.1: Solutions $U_h^1$ and $U_{h/2}^1$ computed on meshes having spacing $h$ and $h/2$ with piecewise linear polynomials ($p = 1$) and a third solution $U_h^2$ computed on a mesh of spacing $h$ with a piecewise quadratic polynomial ($p = 2$).

Subtracting the two solutions we eliminate the unknown exact solution and obtain

$$U_{h/2}^p(x) - U_h^p(x) = C_{p+1}h^{p+1} \left(1 - \frac{1}{2^p + 1}\right) + O(h^{p+2}).$$

Neglecting the higher-order terms, we obtain an approximation of the discretization error as

$$C_{p+1}h^{p+1} \approx \frac{U_{h/2}^p(x) - U_h^p(x)}{1 - 1/2^{p+1}}.$$

Thus, we have an estimate of the discretization error of the coarse-mesh solution as

$$u(x) - U_h^p(x) \approx \frac{U_{h/2}^p(x) - U_h^p(x)}{1 - 1/2^{p+1}}.$$

The technique is called Richardson’s extrapolation or $h$-extrapolation. It can also be used to obtain error estimates of the fine-mesh solution. The cost of obtaining the error estimate is approximately twice the cost of obtaining the solution. In two and three dimensions the cost factors rise to, respectively, four and eight times the solution cost. Most would consider this to be excessive. The only way of justifying the procedure is to consider the fine-mesh solution as being the result and the coarse-mesh solution as furnishing the error estimate. This strategy only furnishes an error estimate on the coarse mesh.

Another strategy for obtaining an error estimate by extrapolation is to compute a second solution using a higher-order method (Figure 7.4.1), e.g.,

$$u(x) - U_h^{p+1} = C_{p+2}h^{p+2} + O(h^{p+3}).$$

Now, use the identity

$$u(x) - U_h^p(x) = [u(x) - U_h^{p+1}(x)] + [U_h^{p+1}(x) - U_h^p].$$
The first term on the right is the $O(h^{p+1})$ error of the higher-order solution and, hence, can be neglected relative to the second term. Thus, we obtain the approximation

$$u(x) - U_h^p(x) \approx U_h^{p+1}(x) - U_h^p(x).$$

The difference between the lower- and higher-order solutions furnish an estimate of the error of the lower-order solution. The technique is called order embedding or p-extrapolation. There is no error estimate for the higher-order solution, but some use it without an error estimate. This strategy, called local extrapolation, can be dangerous near singularities. Unless there are special properties of the scheme that can be exploited, the work involved in obtaining the error estimate is comparable to the work of obtaining the solution. With a hierarchical embedding, computations needed for the lower-order method are also needed for the higher-order method and, hence, need not be repeated.

The extrapolation techniques just described are typically too expensive for use as error estimates. We’ll develop a residual-based error estimation procedure that follows Bank (cf. [8], Chapter 7) and uses many of the ideas found in order embedding. We’ll follow our usual course of presenting results for the model problem

\begin{align*}
-\nabla \cdot p \nabla u + qu &= -(pu_x)_x - (pu_y)_y + qu = f(x,y), \quad (x,y) \in \Omega, \quad (7.4.1a) \\
u(x,y) &= \alpha, \quad (x,y) \in \partial \Omega_E, \quad pu_n(x,y) = \beta, \quad (x,y) \in \partial \Omega_N; \quad (7.4.1b)
\end{align*}

however, results apply more generally. Of course, the Galerkin form of (7.4.1) is: determine $u \in H^1_E$ such that

\begin{align*}
A(v,u) &= (v,f) + \langle v, \beta \rangle, \quad \forall v \in H^1_0, \quad (7.4.2a) \\

\text{where}
\begin{align*}
(v,f) &= \iint_{\Omega} vf \, dx \, dy, \quad (7.4.2b) \\
A(v,u) &= \iint_{\Omega} [p \nabla v \cdot \nabla u + qvu] \, dx \, dy, \quad (7.4.2c) \\
\langle v, u \rangle &= \int_{\partial \Omega_N} vuds, \quad (7.4.2d)
\end{align*}
\end{align*}

and

\begin{align*}
\langle v, u \rangle &= \int_{\partial \Omega_N} vuds, \quad (7.4.2d)
\end{align*}

Similarly, the finite element solution $U \in S^N_E \subset H^1_E$ satisfies

\begin{align*}
A(V,U) &= (V,f) + \langle V, \beta \rangle, \quad \forall V \in S^N_0. \quad (7.4.3)
\end{align*}
We seek an error estimation technique that only requires local (element level) mesh computations, so let's construct a local Galerkin problem on element $e$ by integrating (7.4.1a) over $\Omega_e$ and applying the divergence theorem to obtain: determine $u \in H^1(\Omega_e)$ such that

$$A_e(v, u) = (v, f)_e + <v, pu_n>_{e}, \quad \forall v \in H^1(\Omega_e),$$

(7.4.4a)

where

$$(v, f)_e = \iint_{\Omega_e} vf \, dx \, dy,$$

(7.4.4b)

$$A_e(v, u) = \iint_{\Omega_e} [p \nabla v \cdot \nabla u + qvu] \, dx \, dy,$$

(7.4.4c)

and

$$<v, u>_{e} = \int_{\partial \Omega_e} vu \, ds.$$  (7.4.4d)

As usual, $\Omega_e$ is the domain of element $e$, $s$ is a coordinate along $\partial \Omega_e$, and $n$ is a unit outward normal to $\partial \Omega_e$.

Let

$$u = U + e,$$

(7.4.5)

where $e(x, y)$ is the discretization error of the finite element solution, and substitute (7.4.5) into (7.4.4a) to obtain

$$A_e(v, e) = (v, f)_e - A_e(v, U) + <v, pu_n>_{e}, \quad \forall v \in H^1(\Omega_e).$$

(7.4.6)

Equation (7.4.6), of course, cannot be solved because (i) $v$, $u$, and $e$ are elements of an infinite-dimensional space and (ii) the flux $pu_n$ is unknown on $\partial \Omega_e$. We could obtain a finite element solution of (7.4.6) by approximating $e$ and $v$ by $E$ and $V$ in a finite-dimensional subspace $\tilde{S}^N(\Omega_e)$ of $H^1(\Omega_e)$. Thus,

$$A_e(V, E) = (V, f)_e - A_e(V, U) + <V, pu_n>_{e}, \quad \forall V \in \tilde{S}^N(\Omega_e).$$

(7.4.7)

We will discuss selection of $\tilde{S}^N$ momentarily. Let us first prescribe the flux $pu_n$ appearing in the last term of (7.4.7). The simplest possibility is to use an average flux obtained from $pU_n$ across the element boundary, i.e.,

$$A_e(V, E) = (V, f)_e - A_e(V, U) + <V, \frac{1}{2}(pU_n^+ + pU_n^-)>_{e}, \quad \forall V \in \tilde{S}^N(\Omega_e),$$

(7.4.8)
where superscripts $+$ and $-$, respectively, denote values of $pU_n$ on the exterior and interior of $\partial \Omega_e$.

Equation (7.4.8) is a local Neumann problem for determining the error approximation $E$ on each element. No assembly and global solution is involved. Some investigators prefer to apply the divergence theorem to the second term on the right to obtain

$$A_e(V, E) = (V, r)_e - < V, (pU_n)^- >_e + < V, \frac{(pU_n)^+ + (pU_n)^-}{2} >_e$$

or

$$A_e(V, E) = (V, r)_e + < V, \frac{(pU_n)^+ - (pU_n)^-}{2} >_e$$

(7.4.9a)

where

$$r(x, y) = f + \nabla \cdot p \nabla U - qU$$

(7.4.9b)

is the residual. This form involves jumps in the flux across element boundaries.

Now let us select the error approximation space $\tilde{S}^N$. Choosing $\tilde{S}^N = S^N$ does not work since there are no errors in the solution subspace. Bank [10] chose $\tilde{S}^N$ as a space of discontinuous polynomials of the same degree $p$ used for the solution space $S^N_E$; however, the algebraic system for $E$ resulting from (7.4.8) or (7.4.9) could be ill-conditioned when the basis is nearly continuous. A better alternative is to select $\tilde{S}^N$ as a space of piecewise $p + 1$ st-degree polynomials when $S^N_E$ is a space of $p$ th degree polynomials. Hierarchical bases (cf. Sections 2.5 and 4.4) are the most efficient to use in this regard. Let us illustrate the procedure by constructing error estimates for a piecewise bilinear solution on a mesh of quadrilateral elements. The bilinear shape functions for a canonical $2 \times 2$ square element are

$$N_{i,j}^1(\xi, \eta) = \tilde{N}_i(\xi) \tilde{N}_j(\eta), \quad i, j = 1, 2,$$

(7.4.10a)

where

$$\tilde{N}_1(\xi) = \frac{1 - \xi}{2}, \quad \tilde{N}_2(\xi) = \frac{1 + \xi}{2}.$$  

(7.4.10b)

The four second-order hierarchical shape functions are

$$N_{3,j}^2(\xi, \eta) = \tilde{N}_j(\eta) \tilde{N}_3(\xi), \quad j = 1, 2,$$

(7.4.11a)

$$N_{i,3}^2(\xi, \eta) = \tilde{N}_i(\xi) \tilde{N}_3(\eta), \quad i = 1, 2,$$

(7.4.11b)

where

$$\tilde{N}_3^2(\xi) = \frac{3(\xi^2 - 1)}{2\sqrt{6}}.$$  

(7.4.11c)
Figure 7.4.2: Nodal placement for bilinear and hierarchical biquadratic shape functions on a canonical $2 \times 2$ square element.

Node indexing is given in Figure 7.4.2.

The restriction of a piecewise bilinear finite element solution $U$ to the square canonical element is

$$U(\xi, \eta) = \sum_{i=1}^{2} \sum_{j=1}^{2} c_{ij}^1 N_{ij}^1(\xi, \eta).$$  \hfill (7.4.12)

Using either (7.4.8) or (7.4.9), the restriction of the error approximation $E$ to the canonical element is the second-order hierarchical function

$$E(\xi, \eta) = \sum_{i=1}^{2} \sum_{j=1}^{2} c_{ij}^2 N_{ij}^1(\xi, \eta) + \sum_{i=1}^{2} d_{ij}^2 N_{ij}^2(\xi, \eta) + \sum_{j=1}^{2} d_{ij}^3 N_{ij}^2(\xi, \eta).$$  \hfill (7.4.13)

The local problems (7.4.8) or (7.4.9) are transformed to the canonical element and solved for the eight unknowns, $c_{ij}^2, i, j = 1, 2, d_{ij}^2, i = 1, 2, d_{ij}^3, j = 1, 2$, using the test functions $V = N_{ij}^k, i, j = 1, 2, 3, k = 1, 2$.

Several simplifications and variations are possible. One of these may be called vertex superconvergence which implies that the solution at vertices converges more rapidly than it does globally. Vertex superconvergence has been rigorously established in certain circumstances (e.g., for uniform meshes of square elements), but it seems to hold more widely than current theory would suggest. In the present context, vertex superconvergence implies that the bilinear vertex solution $c_{ij}^1, i, j = 1, 2$, converges at a higher rate than the solution elsewhere on Element $e$. Thus, the error at the vertices $c_{ij}^2, i, j = 1, 2$, may be neglected relative to $d_{ij}^2, i = 1, 2$, and $d_{ij}^3, j = 1, 2$. With this simplification,
(7.4.13) becomes

\[ E(\xi, \eta) = \sum_{i=1}^{2} d_{i1}^2 N_{i3}^2(\xi, \eta) + \sum_{j=1}^{2} d_{3j}^2 N_{3j}^2(\xi, \eta). \]  

(7.4.14)

Thus, there are four unknowns \(d_{13}^2, d_{23}^2, d_{31}^2\), and \(d_{32}^2\) per element. This technique may be carried to higher orders. Thus, if \(S_E^N\) contains complete polynomials of degree \(p\), \(S_E^N\) only contains the hierarchical correction of order \(p+1\). All lower-order terms are neglected in the error estimation space.

The performance of an error estimate is typically appraised in a given norm by computing an \textit{effectivity index} as

\[ \Theta = \frac{\|E(x, y)\|}{\|e(x, y)\|}. \]

(7.4.15)

Ideally, the effectivity index should not differ greatly from unity for a wide range of mesh spacings and polynomial degrees. Bank and Weiser [11] and Oden et al. [17] studied the error estimation procedure (7.4.8) with the simplifying assumption (7.4.14) and were able to establish upper bounds of the form \(\Theta \leq C\) in the strain energy norm

\[ \|e\|_A = \sqrt{A(e, e)}. \]

They could not, however, show that the estimation procedure was \textit{asymptotically correct} in the sense that \(\Theta \to 1\) under mesh refinement or order enrichment.

\textit{Example 7.4.1.} Strouboulis and Haque [19] study the properties of several different error estimation procedures. We report results for the residual error estimation procedure (7.4.8, 7.4.14) on the “Gaussian Hill” problem. This problem involves a Dirichlet problem for Poisson’s equation on an equilateral triangle having the exact solution

\[ u(x, y) = 100e^{-1.5[(x-4.5)^2+(y-2.6)^2]}. \]

Errors are shown in Figure 7.4.3 for uniform \(p\)-refinement on a mesh of uniform triangular elements having an edge length of 0.25 and for uniform \(h\)-refinement with \(p = 2\). “Extrapolation” refers to the \(p\)-refinement procedure described earlier in this section. This order embedding technique appears to produce accurate error estimates for all polynomial degrees and mesh spacings. The “residual” error estimation procedure is (7.4.8) with errors at vertices neglected and the hierarchical corrections of order \(p+1\) forming \(S_E^N\) (7.4.14). The procedure does well for even-degree approximations, but less well for odd-degree approximations.

From (7.4.8), we see that the error estimate \(E\) is obtained by solving a Neumann problem. Such problems are only solvable when the edge loading (the flux average across
Figure 7.4.3: Effectivity indices for several error estimation procedures using uniform $h$-refinement (left) and $p$-refinement (right) for the Gaussian Hill Problem \[\text{(1)}\] of Example 7.4.1.

element edges) is equilibrated. The flux averaging used in (7.4.8) is, apparently, not sufficient to ensure this when $p$ is odd. We’ll pursue some remedies to this problem later in this section, but, first, let us look at another application.

Example 7.4.2. Aiffa [4] considers the nonlinear parabolic problem

$$u_t + qu^2(u - 1) = \frac{u_{xx} + u_{yy}}{2}, \quad (x, y) \in (0, 1) \times (0, 1), \quad t > 0,$$

with the initial and Dirichlet boundary conditions specified so that the exact solution is

$$u(x, y, t) = \frac{1}{1 + e^{\sqrt{q/2}(y + t\sqrt{q/2})}}.$$

He estimates the spatial discretization error using the residual estimate (7.4.8) neglecting the error at vertices. The error estimation space $\tilde{S}^N$ consists of the hierarchical corrections of degree $p + 1$; however, some lower-degree hierarchical terms are used in some cases. This is to provide a better equilibration of boundary terms and improve results. although this is a time-dependent problem, which we haven’t studied yet, Aiffa [4] keeps the temporal errors small to concentrate on spatial error estimation. With $q = 500$, Aiffa’s [4] effectivity indices in $H^1$ at $t = 0.06$ are presented in Table 7.4.1 for computations performed on uniform meshes of $N_\Delta$ triangles with polynomial degrees $p$ ranging from 1 to 4.

The results with $\tilde{S}^N$ consisting only of hierarchical corrections of degree $p + 1$ are reasonable. Effectivity indices are in excess of 0.9 for the lower-degree polynomials $p =$
1, 2, but degrade with increasing polynomial degree. The addition of a lower (third) degree polynomial correction has improved the error estimates with \( p = 4 \); however, a similar tactic provided little improvement with \( p = 3 \). These results and those of Strouboulis and Haque [19] show that the performance of a posteriori error estimates is still dependent on the problem being solved and on the mesh used to solve it.

Another way of simplifying the error estimation procedure (7.4.8) and of understanding the differences between error estimates for odd- and even-order finite element solutions involves a profound, but little known, result of Babuška (cf. [1, 2, 3, 9, 22, 23]). Concentrating on linear second-order elliptic problems on rectangular meshes, Babuška indicates that asymptotically (as mesh spacing tends to zero) errors of odd-degree finite element solutions occur near element edges while errors of even-degree solutions occur in element interiors. These findings suggest that error estimates may be obtained by neglecting errors in element interiors for odd-degree polynomials and neglecting errors on element boundaries for even-degree polynomials.

Thus, for piecewise odd-degree approximations, we could neglect the area integrals on the right-hand sides of (7.4.8) or (7.4.9a) and calculate an error estimate by solving

\[
A_e(V, E) = \langle V, \frac{(pU_n)^+ + (pU_n)^-}{2} \rangle_e, \quad \forall V \in \tilde{S}^N. \tag{7.4.16a}
\]

or

\[
A_e(V, E) = \langle V, \frac{(pU_n)^+ - (pU_n)^-}{2} \rangle_e, \quad \forall V \in \tilde{S}^N. \tag{7.4.16b}
\]

For piecewise even-degree approximations, the boundary terms in (7.4.8) or (7.4.9a) can be neglected to yield

\[
A_e(V, E) = (V, f)_e - A_e(V, U), \quad \forall V \in \tilde{S}^N. \tag{7.4.17a}
\]
or
\[ A_e(V, E) = (V, r)_e, \quad \forall V \in \mathcal{S}^N. \quad (7.4.17b) \]

Yu [22, 23] used these arguments to prove asymptotic convergence of error estimates to true errors for elliptic problems. Adjerid et al. [2, 3] obtained similar results for transient parabolic systems. Proofs, in both cases, apply to a square region with square elements of spacing \( h = 1/\sqrt{N_{\Delta}} \). A typical result follows.

**Theorem 7.4.1.** Let \( u \in H^1_E \cap H^{p+2} \) and \( U \in S^N_E \) be solutions of (7.4.2) using complete piecewise-bi-polynomial functions of order \( p \).

1. If \( p \) is an odd positive integer then

\[
\| e(\cdot, \cdot) \|^2_1 = \| E(\cdot, \cdot) \|^2_1 + O(h^{2p+1})
\]

where

\[
\| E \|^2_1 = \frac{h^2}{16(2p+1)} \sum_{e=1}^{N_{\Delta}} \sum_{i=1}^2 \sum_{k=1}^4 \left[ U_{x_i}(P_{k,e}) \right]^2,
\]

\( P_{k,e}, k = 1, 2, 3, 4, \) are the coordinates of the vertices of \( \Omega_e \), and \( [f(P)]_i \) denotes the jump in \( f(x) \) in the direction \( x_i, i = 1, 2 \), at the point \( P \).

2. If \( p \) is a positive even integer then (7.4.18a) is satisfied with

\[
A_e(V_i, E) = (V, f)_e - A_e(V_i, U),
\]

where

\[
E(x_1, x_2) = b_{1,e} \Phi^{p+1}_e(x_1) + b_{2,e} \Phi^{p+1}_e(x_2),
\]

\[
V_i(x_1, x_2) = x_i \frac{\Phi^{p+1}_e(x_1) \Phi^{p+1}_e(x_2)}{x_1 x_2}, \quad i = 1, 2,
\]

and \( \Phi^{m}_e(x) \) is the mapping of the hierarchical basis function

\[
\tilde{N}_3^m(\xi) = \sqrt{\frac{2m-1}{2}} \int_{-1}^{\xi} P_{m-1}(\zeta) d\zeta
\]

from \([-1, 1]\) to the appropriate edge of \( \Omega_e \).

**Proof.** cf. Adjerid et al. [2, 3] and Yu [22, 23]. Coordinates are written as \( x = [x_1, x_2]^T \) instead of \((x, y)\) to simplify notation within summations. The hierarchical basis element (7.4.18f) is consistent with prior usage. Thus, the subscript 3 refers to a midside node as indicated in Figure 7.4.2. \( \square \)
Remark 1. The error estimate for even-degree approximations has different trial and test spaces. The functions $V_i(x_1, x_2)$ vanish on $\partial \Omega_e$. Each function is the product of a “bubble function” $\Phi_e^{p+1}(x_1)\Phi_e^{p+1}(x_2)$ biased by a variation in either the $x_1$ or the $x_2$ direction. As an example, consider the test functions on the canonical element with $p = 2$. Restricting $(7.4.18e)$ to the canonical element $-1 \leq \xi_1, \xi_2 \leq 1$, we have

$$V_i(\xi_1, \xi_2) = \xi_i \frac{\tilde{N}_3^3(\xi_1) \tilde{N}_3^3(\xi_2)}{\xi_2}, \quad i = 1, 2.$$  

Using $(7.4.18f)$ with $m = 3$ or $(2.5.8)$,

$$\tilde{N}_3^3(\xi) = \frac{5}{2\sqrt{10}}\xi(\xi^2 - 1).$$

Thus,

$$V_i(\xi_1, \xi_2) = \frac{5\xi_i}{8}(\xi_1^2 - 1)(\xi_2^2 - 1), \quad i = 1, 2.$$

Remark 2. Theorem 7.4.1 applies to tensor-product bi-polynomial bases. Adjerid et al. [1] show how this theorem can be modified for use with hierarchical bases.

Example 7.4.3. Adjerid et al. [2] solve the nonlinear parabolic problem of Example 7.4.2 with $q = 20$ on uniform square meshes with $p$ ranging from 1 to 4 using the error estimates $(7.4.18a,b)$ and $(7.4.18a,c-f)$. Temporal errors were controlled to be negligible relative to spatial errors; thus, we need not be concerned that this is a parabolic and not an elliptic problem. The exact $H^1$ errors and effectivity indices at $t = 0.5$ are presented in Table 7.4.2. Approximate errors are within ten percent of actual for all but one mesh and appear to be converging at the same rate as the actual errors under mesh refinement.

<table>
<thead>
<tr>
<th>$p$</th>
<th>$N_\Delta = 100$</th>
<th>400</th>
<th>900</th>
<th>1600</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$|e|_1/|u|_1$</td>
<td>$\Theta$</td>
<td>$|e|_1/|u|_1$</td>
<td>$\Theta$</td>
</tr>
<tr>
<td>1</td>
<td>0.262(-1)</td>
<td>0.949</td>
<td>0.129(-1)</td>
<td>0.977</td>
</tr>
<tr>
<td>2</td>
<td>0.872(-3)</td>
<td>0.995</td>
<td>0.218(-3)</td>
<td>0.999</td>
</tr>
<tr>
<td>3</td>
<td>0.278(-4)</td>
<td>0.920</td>
<td>0.348(-5)</td>
<td>0.966</td>
</tr>
<tr>
<td>4</td>
<td>0.848(-6)</td>
<td>0.999</td>
<td>0.530(-7)</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 7.4.2: Errors and effectivity indices in $H^1$ for Example 7.4.3 on $N_\Delta$-element uniform meshes with piecewise bi-$p$ polynomial bases. Numbers in parentheses indicate a power of ten.

The error estimation procedures $(7.4.8)$ and $(7.4.9)$ use average flux values on $\partial \Omega_e$. As noted, data for such (local) Neumann problems cannot be prescribed arbitrarily. Let us examine this further by concentrating on $(7.4.9)$ which we write as

$$A_e(V, E) = (V, r)_e + < V, R >_e$$  

(7.4.19a)
where the elemental residual $r$ was defined by $(7.4.9b)$ and the boundary residual is

$$R = \kappa[(pU_n)^+ - (pU_n)^-]. \quad (7.4.19b)$$

The function $\kappa$ on $\partial \Omega_e$ was taken as $1/2$ to obtain $(7.4.9a)$; however, this may not have been a good idea for reasons suggested in Example 7.4.1.

Recall (cf. Section 3.1) that smooth solutions of the weak problem $(7.4.19)$ satisfy the Neumann problem

$$-\nabla \cdot p\nabla E + q E = r, \quad (x, y) \in \Omega_e, \quad (7.4.20a)$$

$$pE_n = R, \quad (x, y) \in \partial \Omega_e. \quad (7.4.20b)$$

Solutions of $(7.4.20)$ only exist when the data $R$ and $r$ satisfy the equilibrium condition

$$\int_{\Omega_e} r(x,y)dxdy + \int_{\partial \Omega_e} R(s)ds = 0. \quad (7.4.20c)$$

This condition will most likely not be satisfied by the choice of $\kappa = 1/2$. Ainsworth and Oden [5] describe a relatively simple procedure that requires the solution of the Poisson problem

$$-\Delta \omega_e = r, \quad (x, y) \in \Omega_e, \quad (7.4.21a)$$

$$\frac{\partial \omega_e}{\partial n} = R, \quad (x, y) \in \partial \Omega_e - \partial \Omega_E, \quad (7.4.21b)$$

$$\omega_e = 0, \quad (x, y) \in \partial \Omega_E. \quad (7.4.21c)$$

The error estimate is

$$\|E\|_A^2 = \sum_{e=1}^{N_\Delta} A_e(\omega_e, \omega_e). \quad (7.4.21d)$$

The function $\kappa$ is approximated by a piecewise-linear polynomial in a coordinate $s$ on $\partial \Omega_e$ and may be determined explicitly prior to solving $(7.4.21)$. Let us illustrate the effect of this equilibrated error estimate.

**Example 7.4.4.** Oden [16] considers a “cracked panel” as shown in Figure 7.4.4 and determines $u$ as the solution of

$$A(v, u) = \iint_{\Omega} (v_x u_x + v_y u_y) dxdy = 0.$$
The essential boundary condition

\[ u(r, \theta) = r^{1/2} \cos \theta/2 \]

is prescribed on all boundaries except \( x > 0, y = 0 \). Thus, the solution of the Galerkin problem will satisfy the natural boundary condition \( u_y = 0 \) there. These conditions have been chosen so that the exact solution is the specified essential boundary condition. This solution is singular since \( u_r \sim r^{-1/2} \) near the origin \((r = 0)\).

Results for the effectivity indices in strain energy for the entire region and for the two elements, \( \Omega_L \) and \( \Omega_R \), adjacent to the singularity are shown in Table 7.4.3. Computations were performed on a square grid with uniform spacing \( h \) in each coordinate direction (Figure 7.4.4). Piecewise linear and quadratic polynomials were used as finite element bases.

Local effectivity indices on \( \Omega_L \) and \( \Omega_R \) are not close to unity and don’t appear to be converging as either the mesh spacing is refined or \( p \) is increased. Global effectivity indices are near unity. Convergence to unity is difficult to appraise with the limited data.
At this time, the field of *a posteriori* error estimation is still emerging. Error estimates for problems with singularities are not generally available. The performance of error estimates is dependent on both the problem, the mesh, and the basis. Error estimates for realistic nonlinear and transient problems are just emerging. Verfürth [20] provides an excellent survey of methods and results.
Bibliography


Analysis of the Finite Element Method


