SIAM J. NUMER. ANAL.
Vol. 15, No. 4, August 1978


0036-1429/78/1504-0007 $01.00/0



# ERROR ESTIMATES FOR ADAPTIVE FINITE ELEMENT COMPUTATIONS*

I. BABUŠKA† AND W. C. RHEINBOLDT‡

**Abstract.** A mathematical theory is developed for a class of a-posteriori error estimates of finite element solutions. It is based on a general formulation of the finite element method in terms of certain bilinear forms on suitable Hilbert spaces. The main theorem gives an error estimate in terms of localized quantities which can be computed approximately. The estimate is optimal in the sense that, up to multiplicative constants which are independent of the mesh and solution, the upper and lower error bounds are the same. The theoretical results also lead to a heuristic characterization of optimal meshes, which in turn suggests a strategy for adaptive mesh refinement. Some numerical examples show the approach to be very effective.

**1. Introduction.** In the application of the finite element method one of the most critical decisions is the design of the mesh and the selection of the elements. In practice, the user has to base his choices on some—necessarily incomplete and often conflicting—experience with earlier, similar computations (see, e.g., [1]). Moreover, the reliability of the final results and thereby the adequacy of the original decisions are generally judged on a corresponding experimental basis. Without question, there is much need for techniques to compute reliable, a posteriori error estimates of finite element solutions at reasonable cost.

Error estimates of this type are not only important for an assessment of the reliability of the results, but provide also a means for adaptive optimization of the finite element mesh. Optimal mesh design has been considered by several authors (see, e.g., [2]–[9]). The approaches vary considerably. For example, in [6] heuristic techniques are derived from energy considerations that are natural to finite element analysis. On the other hand, [7] uses heuristic methods analogous to those in finite-difference calculations. In [8] and [9] some mathematical results are evolved and proved on the basis of some combination of finite difference and finite element analysis.

In this paper we present a mathematical theory of a class of a-posteriori error estimates for finite element solutions. A general formulation is employed using bilinear forms on pairs of suitable spaces. The main theorem shows that localized computations provide for an error estimate which is optimal in the sense that, up to multiplicative constants, the upper and lower bounds of the error are the same. The constants are independent of the mesh and the specific solution and, moreover, in practice they are not large. The results are in a sense similar to those encountered in connection with error control in the solution of initial value problems for initial value problems of ordinary differential equations. The concepts leading to our estimates may also be applied to the estimation of the formulation error of the problem itself in comparison to a "higher" problem (see also [10]).

The theoretical results lead to a heuristic characterization of optimal meshes which in turn translates itself into a strategy for adaptive mesh refinement. Finally, some numerical examples show the practical usefulness of the results.

* Received by the editors May 31, 1977, and in revised form September 30, 1977.

† Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742. The work of this author was supported in part by the U.S. Energy Research and Development Administration under Contract E(40-1)3443.

‡ Computer Science Center, University of Maryland, College Park, Maryland 20742. The work of this author was supported by the National Science Foundation under Grant MCS 72-03721A06 (formerly GJ-35568X).

## 2. Preliminaries.

**2.1. Basic notations.** Throughout this article $\Omega \subset R^n$ shall be a given, bounded domain with Lipschitzian boundary $\partial\Omega$ in the $n$ dimensional, real Euclidean space $R^n$ of vectors $x = (x_1, \cdots, x_n)^T$. While this excludes domains with slits, it can be shown that the results also extend to that case.

We denote by $\mathscr{E}(\bar{\Omega})$ the space of all real, infinitely differentiable functions on $\Omega$ such that each function and any of its derivatives has a continuous extension to $\partial\Omega$. All functions of $\mathscr{E}(\bar{\Omega})$ with compact support in $\Omega$ form the subspace $\mathscr{D}(\Omega) \subset \mathscr{E}(\bar{\Omega})$.

As usual, $L_2(\Omega) = H^0(\Omega)$ is the space of all square integrable functions on $\Omega$ with the inner product

$$(u, v)_{L_2(\Omega)} = \int_\Omega uv\, dx \qquad (dx = dx_1\, dx_2 \cdots dx_n) \tag{2.1}$$

and the corresponding norm $\|\cdot\|_{L_2(\Omega)}$. For any integer $k \geqq 1$, the Sobolev spaces $H^k(\Omega)$ and $H_0^k(\Omega)$ are the completions of $\mathscr{E}(\bar{\Omega})$ and $\mathscr{D}(\Omega)$, respectively, under the norm

$$\|u\|^2_{H^k(\Omega)} = \sum_{0 \leqq |\alpha| \leqq k} \|D^\alpha u\|^2_{L_2(\Omega)} \tag{2.2}$$

where

$$D^\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots x_n^{\alpha_n}}, \qquad \alpha = (\alpha_1, \cdots, a_n), \quad |\alpha| = \sum_{i=1}^n \alpha_i, \quad \alpha_i \geqq 0 \text{ integers.} \tag{2.3}$$

For negative integers $k \leqq -1$ the spaces $H^k(\Omega)$ and $H_0^k(\Omega)$ are defined as the dual spaces of $H^{-k}(\Omega)$ and $H_0^{-k}(\Omega)$, respectively.

All $H^k(\Omega)$, $k = 0, \pm 1, \pm 2, \cdots$, are Hilbert spaces and we denote their inner products by $(\cdot\,, \cdot)_{H^k(\Omega)}$.

**2.2. Partitions of the domain.** We shall consider partitions of unity

$$\Psi = \{\psi_1, \cdots, \psi_M\}, \qquad \psi_i \in H^k(\Omega), \quad \psi_i \geqq 0, \quad \sum_{i=1}^M \psi_i(x) = 1, \quad \forall x \in \Omega \tag{2.4}$$

of the domain $\Omega$ and write $\operatorname{supp} \psi_i$ and $\operatorname{supp}^0 \psi_i$ for the support of $\psi_i$ and its interior, respectively. It is always possible to partition the set $\Psi$ such that

$$\Psi = \bigcup_{l=1}^r \Psi_l, \qquad \Psi_l \cap \Psi_j = \varnothing \quad \text{for } l \neq j \tag{2.5a}$$

and that the interiors of the supports of the members of each $\Psi_l$ are disjoint, that is,

$$\begin{aligned} &\psi_i, \psi_j \in \Psi_l, \quad i \neq j \\ &\qquad \Rightarrow \operatorname{supp}^0 \psi_i \cap \operatorname{supp}^0 \psi_j = \varnothing. \end{aligned} \tag{2.5b}$$

For instance, it suffices to let each $\Psi_l$ consist of exactly one $\psi_i$. The smallest integer $r$ for which (2.5a, b) holds is the *overlap index* $\rho(\Psi)$ of $\Psi$.

In addition to (2.4), we will also consider set partitions $T$ of $\bar{\Omega}$ consisting of Lipschitzian subdomains, that is,

$$\begin{aligned} &T = \{\Omega_1, \cdots, \Omega_m\}, \qquad \Omega_l \subset \Omega, \quad \partial\Omega_l\text{-Lipschitzian} \\ &\bar{\Omega} = \bigcup_{l=1}^m \bar{\Omega}, \qquad \Omega_l \cap \Omega_j = \varnothing \quad \text{for } l \neq j. \end{aligned} \tag{2.6}$$

With each $\Omega_l$ we associate a real, positive number $h_l$ typically representing some measure of the size of $\Omega_l$.

For a given pair $\Psi = \{\psi_i\}_1^M$ and $T = \{\Omega_i\}_1^m$ we define the index sets

$$\sigma_l = \sigma_l(\Psi, T) = \{i \in \{1, \cdots, M\} | \Omega_l \cap \operatorname{supp}^0 \psi_i \neq \varnothing\}, \qquad l = 1, \cdots, m. \tag{2.7}$$

Then (2.5) implies that

$$\sum_{i \in \sigma_l} \psi_i(x) = 1, \quad \forall x \in \Omega_l, \quad l = 1, \cdots, m. \tag{2.8}$$

The maximum cardinality $\max\{|\sigma_l|, l = 1, \cdots, m\}$ will be called the *intersection index* $\tau(\Psi, T)$ of $\Psi$ and $T$.

For given $k \geqq 0$ let $H$ be a space such that $H_0^k(\Omega) \subset H \subset H^k(\Omega)$. We consider a family $\mathcal{T}$ of triples $(\Psi, T, V)$ each of which consists of a partition of unity $\Psi$ (cf. (2.4)), a set partition $T$ (cf. (2.6)), and a finite dimensional subspace $V$ of $H$. The family $\mathcal{T}$ shall be called *admissible* if its satisfies the following four conditions:

(i) There is a constant $\rho > 0$ depending only on $\mathcal{T}$ such that

$$\rho(\Psi) \leqq \rho, \quad \forall (\Psi, T, V) \in \mathcal{T}.$$

(ii) There is a constant $\tau > 0$ depending only on $\mathcal{T}$ such that

$$\tau(\Psi, T) \leqq \tau, \quad \forall (\Psi, T, V) \in \mathcal{T}.$$

(iii) There is a constant $K_0 > 0$ depending only on $\mathcal{T}$ such that

$$|D^\alpha \psi_i(x)| \leqq K_0 h_l^{-|\alpha|}, \quad \forall x \in \Omega, \quad i \in \sigma_l(\Psi, T), \quad \forall (\Psi, T, V) \in \mathcal{T}, \quad 0 \leqq |\alpha| \leqq k. \tag{2.9}$$

(iv) There is a constant $K_1 > 0$ depending only on $\mathcal{T}$ such that for any $(\Psi, T, V) \in \mathcal{T}$, and any $v \in H$, we may find a function $\varphi \in V$ for which

$$\|v - \varphi\|_{H^r(\Omega_l)}^2 \leqq K_1 h_l^{2(k-r)} \sum_{i \in \sigma_l(\Psi, T)} \|v\|_{H^k(\Omega_i)}^2, \quad \forall \Omega_l \in T, \quad 0 \leqq r \leqq k. \tag{2.10}$$

We illustrate these concepts with the following simple example in $R^1$.

*Example* 2.1. Let $\Omega = (0, 1)$, $k = 1$, and

$$H = \{u \in H^1(\Omega) | u(0) = 0\}.$$

The set partitions $T$ in the triples of $\mathcal{T}$ shall consist of the intervals $\Omega_l = (x_{l-1}^T, x_l^T)$, $l = 1, \cdots, m(T)$, defined by any subdivision

$$0 = x_0^T < x_1^T < x_2^T < \cdots < x_{m(T)}^T = 1$$

of $\bar{\Omega}$ for which

$$\frac{1}{\mu} \leqq \frac{h_{l-1}}{h_l} \leqq \mu, \quad h_l = x_l^T - x_{l-1}^T, \quad l = 1, \cdots, m(T),$$

with some fixed $\mu \geqq 1$.

For any such $T$, let

$$y_0 = x_0^T, \quad y_{2i-1} = \tfrac{1}{2}(x_{i-1}^T + x_i^T), \quad y_{2i} = x_i^T, \quad i = 1, \cdots, m = m(T).$$

Then we define the partition of unity $\Psi = \{\psi_j\}$ in the particular triple $(\psi, T, V)$ as the set of $2m + 1$ continuous, piecewise linear functions $\psi_0, \cdots, \psi_{2m}$ on $\bar{\Omega}$ which are linear on each interval $[y_{j-1}, y_j]$, $j = 1, \cdots, 2m$, and satisfy $\psi_j(y_i) = \delta_{ji}$, $i, j = 0, \cdots, 2m$, where $\delta_{ji}$ is the Kronecker symbol.

Finally, the subspace $V \subset H$ in the triple $(\Psi, T, V)$ shall be the $m$ dimensional space ($m = m(T)$) spanned by the continuous, piecewise linear functions $\varphi_1, \cdots, \varphi_m$ on $\bar{\Omega}$ which are linear on each $\bar{\Omega}_l$ and satisfy $\varphi_i(x_j) = \delta_{ij}$, $i = 1, \cdots m$, $j = 0, \cdots, m$.

In this case it is readily seen that

$$\rho(\Psi) = 2, \qquad \tau(\Psi, T) = 3, \quad \forall (\Psi, T, V) \in \mathcal{T}$$

and $K_0 = 2/\mu$. Moreover, $K_1$ is obtained from the usual finite element theory using standard interpolation techniques. Clearly, all constants depend only on our choice of $\mathcal{T}$.

**2.3. The bilinear form.** Bilinear forms play an essential role in finite element theory. Let $H_1$, $H_2$ be two real Hilbert spaces with inner products $(\cdot, \cdot)_{H_i}$, $i = 1, 2$, and corresponding norms. A bilinear form $B$ on $H_1 \times H_2$ is called *proper* if

$$\text{(i)} \quad |B(u, v)| \leqq C_1 \|u\|_{H_1} \|v\|_{H_2}, \quad \forall u \in H_1, \quad v \in H_2, \quad C_1 < \infty,$$

$$\text{(ii)} \quad \sup_{\substack{v \in H_2 \\ v \neq 0}} |B(u, v)| / \|v\|_{H_2} \geqq C_2 \|u\|_{H_1}, \quad \forall u \in H_1, \quad C_2 > 0, \tag{2.11}$$

$$\text{(iii)} \quad \sup_{u \in H_1} |B(u, v)| > 0, \quad \forall v \in H_2, \quad v \neq 0.$$

Proper bilinear forms have the following important property:

THEOREM 2.1. *Let $B$ be a proper bilinear form on $H_1 \times H_2$ and $f \in H_2'$ a given linear functional on $H_2$. Then there exists a unique $u_0 \in H_1$ such that*

$$B(u_0, v) = f(v), \quad \forall v \in H_2 \tag{2.12}$$

*and*

$$\|u_0\|_{H_1} \leqq \frac{1}{C_2} \|f\|_{H_2}. \tag{2.13}$$

For a proof see [11]. Obviously $B$ induces an isomorphism between $H_1$ and $H_2'$.

Now let $\mathcal{P}$ be a family of pairs $(V_1, V_2)$, each of which consists of finite dimensional subspaces $V_1 \subset H_1$, $V_2 \subset H_2$. A bilinear form $B$ on $H_1 \times H_2$ is *uniformly $\mathcal{P}$-proper* if $B$ is proper on $H_1 \times H_2$ and for any $(V_1, V_2) \in \mathcal{P}$

$$\text{(a)} \quad \sup_{\substack{v \in V_2 \\ v \neq 0}} |B(u, v)| / \|v\|_{V_2} \geqq \hat{C}_2 \|u\|_{V_1}, \quad \forall u \in V_1, \quad \hat{C}_2 > 0$$

$$\text{(b)} \quad \sup_{u \in V_1} |B(u, v)| > 0, \quad \forall v \in V_2, \quad v \neq 0. \tag{2.14}$$

The constant $\hat{C}_2$ depends only on $\mathcal{P}$ but not on the particular pair $(V_1, V_2)$. For such forms the following result holds:

THEOREM 2.2. *Let $B$ be a* uniformly $\mathcal{P}$-proper *form on $H_1 \times H_2$ and $f \in H_2'$ a given functional. Let $u_0 \in H_1$ be the unique element satisfying* (2.12)–(2.13). *Then for any $(V_1, V_2) \in \mathcal{P}$ there exists a unique $\hat{u}_0 \in V_1$ such that*

$$B(\hat{u}_0, v) = f(v), \quad \forall v \in V_2 \tag{2.15}$$

*and*

$$\|u_0 - \hat{u}_0\|_{H_1} \leqq \left(1 + \frac{C_1}{\hat{C}_2}\right) \inf_{w \in V_1} \|w - u_0\|_{H_1}. \tag{2.16}$$

For a proof see again [11]. Evidently, the constant $(1+C_1/\hat{C}_2)$ depends only on $\mathscr{P}$.

In all subsequent discussions we shall assume that the spaces $H_1, H_2$ satisfy

$$H_0^{k_i}(\Omega) \subset H_i \subset H^{k_i}(\Omega), \qquad i=1,2, \tag{2.17a}$$

for some fixed integers $k_1, k_2$, and that

$$\|\cdot\|_{H_i} = \|\cdot\|_{H^{k_i}(\Omega)}, \qquad i=1,2. \tag{2.17b}$$

Frequently it is possible to introduce norms in $H_1$ and $H_2$ that are equivalent to the norms (2.17) and such that $C_1 = C_2 = \hat{C}_2 = 1$. This is the case, for instance, when $H_1 = H_2$, $B(u,v) = B(v,u)$, $\forall u, v \in H_1$, and

$$C_2\|u\|^2_{H^k(\Omega)} \leqq B(u,u) < C_1\|u\|^2_{H^k(\Omega)},$$

as is typical for self-adjoint problems occurring, say, in structural mechanics.

**3. The main error estimates.** We begin with a lemma which will play an essential role in the further theory.

LEMMA 3.1. *Let $\mathscr{T}$ be an admissible family of triples $(\Psi, T, V)$ as defined in* § 2.2. *Then there exists a constant $\kappa(\mathscr{T}) < \infty$ depending only on $\mathscr{T}$ such that for any $(\Psi, T, V) \in \mathscr{T}$*

$$\inf_{\varphi \in V} \sum_{l=1}^{M(\Psi)} \|\psi_l(v-\varphi)\|^2_{H_k(\Omega)} \leqq \kappa(\mathscr{T})\|v\|^2_{H^k(\Omega)}, \quad \forall v \in H. \tag{3.1}$$

*Proof.* Let $(\Psi, T, V) \in \mathscr{T}$ and $v \in H$ be given. There exists a function $\varphi \in V$ such that (2.10) holds for any $l$ with $1 \leqq l \leqq M(\Psi)$. In order to evaluate the $H^k(\Omega_l)$ norm of $\psi_i(v-\varphi)$ for any $i \in \sigma_l(\Psi, T)$ we apply the well known Leibnitz formula and the property (2.9) of $\mathscr{T}$. This leads readily to the estimate

$$\|\psi_i(v-\varphi)\|^2_{H^k(\Omega)_i} \leqq c_1 K_0^2 \sum_{s=0}^{k} h_l^{-2s}\|v-\varphi\|^2_{H^{k-s}(\Omega_l)}, \quad \forall i \in \sigma_l(\Psi, T),$$

where the constant $c_1$ depends only on $k$. Hence it follows from (2.10) and the condition (ii) on $\mathscr{T}$ that

$$\|\psi_i(v-\varphi)\|^2_{H^k(\Omega_l)} \leqq c_2 K_0^2 K_1 \sum_{j \in \sigma_l(\Psi,T)} \|v\|^2_{H^k(\Omega_j)}, \quad \forall i \in \sigma_l(\Psi, T). \tag{3.2}$$

Therefore, we obtain

$$\begin{aligned}
\sum_{i=1}^{M(\Psi)} \|\psi_i(v-\varphi)\|^2_{H^k(\Omega)} &= \sum_{l=1}^{m(T)} \sum_{i=1}^{M(\Psi)} \|\psi_i(v-\varphi)\|^2_{H^k(\Omega_l)} \\
&= \sum_{l=1}^{m(T)} \sum_{i \in \sigma_l(\Psi,T)} \|\psi_i(v-\varphi)\|^2_{H^k(\Omega_l)} \\
&\leqq c_2 K_0^2 K_1 \sum_{l=1}^{m(T)} \Big[\tau \sum_{j \in \sigma_l(\Psi,T)} \|v\|^2_{H^k(\Omega_j)}\Big] \\
&\leqq c_2 K_0^2 K_1 \tau^2 \sum_{l=1}^{m(T)} \|v\|^2_{H^k(\Omega_l)} \leqq c_2 K_0^2 K_1 \tau^2 \|v\|^2_{H^k(\Omega)},
\end{aligned}$$

where in the last inequality we used (3.2) as well as the conditions (i) and (ii) on $\mathscr{T}$. This proves (3.1).

As indicated at the end of § 2.3, we fix now two Hilbert spaces which satisfy (2.17a, b) for certain given integers $k_1, k_2$. Moreover, we consider an admissible family $\mathcal{T}$ of triples $(\Psi, T, V)$ such that $V$ is a finite dimensional subspace of $H_2$. Finally, with each $V$ we associate a finite dimensional space $\hat{V} \subset H_1$ and form a family $\mathcal{P}$ of pairs $(\hat{V}, V)$. With this our main theorem may be praised as follows.

THEOREM 3.2. *Suppose that $\mathcal{T}$ and $\mathcal{P}$ are as stated above, that $B$ is a uniformly $\mathcal{P}$-proper bilinear form on $H_1 \times H_2$, and that $f \in H_2'$ is a given functional. Let $u_0 \in H_1$ be the (unique) solution of* (2.12) *and, for any $(\Psi, T, V) \in \mathcal{T}$ and corresponding $(\hat{V}, V) \in \mathcal{P}$ consider the error $e = u_0 - \hat{u}_0$ between $u_0$ and the (unique) solution $\hat{u}_0$ of* (2.15). *Then*

$$D_1 \eta \leqq \|e\|_{H_1} \leqq D_2 \eta \tag{3.3a}$$

*with*

$$\eta^2 = \sum_{i=1}^{M(\Psi)} \eta_i^2, \qquad \eta_i = \sup_{\substack{v \in H_2 \\ v \neq 0}} \frac{|B(e, \psi_i v)|}{\|\psi_i v\|_{H_2}} \tag{3.3b}$$

*and*

$$D_1 \geqq 1/(C_1 \rho^{1/2}), \qquad D_2 \leqq \kappa(\mathcal{T})^{1/2}/C_2, \tag{3.3c}$$

*where $C_1$, $C_2$, $\kappa(\mathcal{T})$ are defined in* (2.11) *and* (3.1), *respectively, and $\rho$ is the constant in the condition* (i) *on $\mathcal{T}$.*

*Proof.* We prove first the right side of (3.3a). From (2.11) and Theorem 2.1 it follows that

$$\|e\|_{H_1} \leqq \frac{1}{C_2} \sup_{\substack{v \in H_2 \\ v \neq 0}} \frac{|B(e, v)|}{\|v\|_{H_2}} \tag{3.4}$$

and (2.12), (2.15) imply that

$$B(e, \varphi) = 0, \quad \forall \varphi \in V.$$

Hence, using (3.3b) and Lemma 3.1 we obtain that

$$\begin{aligned} |B(e, v)| &= \inf_{\varphi \in V} |B(e, v - \varphi)| \\ &= \inf_{\varphi \in V} \left| B\Big(e, \sum_{i=1}^{M(\Psi)} \psi_i (v - \varphi)\Big) \right| \\ &\leqq \inf_{\varphi \in V} \sum_{i=1}^{M(\Psi)} \eta_i \|\psi_i (v - \varphi)\|_{H^{k_2}(\Omega)} \leqq \eta \kappa(\mathcal{T})^{1/2} \|v\|_{H^{k_2}(\Omega)} \end{aligned}$$

which, together with (3.4), gives the right side of (3.3a).

For the proof of the left side, consider a partition $\Psi_1, \cdots, \Psi_{\rho(\Psi)}$ of $\Psi$ such that $\rho(\Psi) \leqq \rho$ and (2.5a, b) holds. We set

$$H_{2,l} = \Big\{ v \,|\, v = \sum_{\psi_i \in \Psi_l} \psi_i w,\ w \in H_2 \Big\}, \qquad l = 1, \cdots, \rho(\Psi).$$

From (2.5b) it follows that

$$(\psi_i w, \psi_j w)_{H_2} = 0, \quad \forall \psi_i, \psi_j \in \Psi_l, \quad i \neq j, \quad 1 \leqq l \leqq \rho(\Psi).$$

Hence, we obtain

$$\sup_{v\in H_{2,l}} \frac{|B(e,v)|}{\|v\|_{H_2}} = \sup_{w\in H_2} \Big|\sum_{\psi_i\in\Psi_l} B(e,\psi_i w)\Big| \Big/ \Big[\sum_{\psi\in\Psi_l} \|\psi_i w\|^2_{H_2}\Big]^{1/2}$$

$$= \Big[\sum_{\psi_i\in\Psi_l} \eta_i^2\Big]^{1/2}, \qquad l=1,\cdots,\rho(\Psi),$$

that is,

$$\sup_{v\in H_2} \frac{|B(e,v)|}{\|v\|_{H_2}} \geqq \Big[\sum_{\psi_i\in\Psi_l} \eta_i^2\Big]^{1/2}.$$

Now, because of $\rho(\Psi)\leqq\rho$ and (2.11) (ii) it follows that

$$\sum_{j=1}^{M(\psi)} \eta_i^2 = \sum_{l=1}^{\rho(\psi)} \sum_{\psi_i\in\Psi_l} \eta_i^2 \leqq \rho\Big[\sup_{v\in H_2} \frac{|B(e,v)|}{\|v\|_{H_2}}\Big]^2$$

$$\leqq \rho C_1^2\Big[\sup_{v\in H_2} \frac{\|e\|_{H_1}\|v\|_{H_2}}{\|v\|_{H_2}}\Big]^2 = C_1^2\|e\|^2_{H_1}$$

which is the left side of (3.3a).

It should be noted that when the spaces (2.17) and the form $B$ are given, the constants (3.3c) depend only on the family $\mathcal{T}$. This raises the question about the optimal constants $D_1, D_2$ for $\mathcal{T}$ or any suitable subset of $\mathcal{T}$. This is an open problem.

*Example* 3.1. For a given $g\in L_2(0,1)$ consider the initial value problem

$$u' = g(x), \qquad x\in(0,1), \quad u(0)=0.$$

Let $H_2 = L_2(0,1)$ and $H_1$ the space $H$ of Example 2.1; that is, $k_1=1$, $k_2=0$. Then

$$B(u,v) = \int_0^1 u'v\,dx, \qquad u\in H_1, \quad v\in H_2 \tag{3.5}$$

is proper on $H_1\times H_2$ and the constants of (2.11) satisfy $C_1=1$, $C_2\geqq(2/3)^{1/2}$. On the other hand, if $\hat{H}_1$ denotes the space $H_1$ with the norm replaced by

$$\|u\|^2_{\hat{H}_1} = \int_0^1 (u')^2\,dx, \tag{3.6}$$

then $B$ is proper on $\hat{H}_1\times H_2$ with $C_1=C_2=1$. On $H_1$ the $H^1(0,1)$-norm and (3.6) are equivalent.

Now let $\mathcal{P}$ be a family of pairs $(V_1, V_2)$ where $V_1$ is the space $V$ of Example 2.1 and

$$V_2 = \{v \mid v = u',\ u\in V_1\}.$$

Then (3.5) is uniformly $\mathcal{P}$-proper on $H_1\times H_2$ and $\hat{H}_1\times H_2$ with $\hat{C}_2\geqq(2/3)^{1/2}$ and $C_2=1$, respectively.

We use the notation of Example 2.1 and consider the following partitions of unity:

(a) The partitions $\Psi$ of Example 2.1 with $\rho=2$, $\tau=3$.

(b) The partitions $\Psi=\{\psi_i\}$ defined by $\psi_i = \chi(x_{i-1}^T, x_i^T)$, $i=1,\cdots,m(T)$, where $\chi(a,b)$ denotes the characteristic function of the interval $(a,b)$. Here we have $\rho=\tau=1$.

From

$$B(e, v) = B(u_0 - \hat{u}_0, v) = \int_0^1 (g - \hat{u}_0')v \, dx$$

we obtain readily that

$$\eta_i^2 = \int_{y_{i-1}}^{y_{i+1}} |g - \hat{u}_0'|^2 \, dx \quad \text{in case (a)},$$

$$\eta_i^2 = \int_{x_{i-1}}^{x_i} |g - \hat{u}_0'|^2 \, dx \quad \text{in case (b)},$$

that is, $\eta_i$ is the $H_2$ norm of the residuals over supp $\Omega_i$. If $\hat{H}_1$ is used, then the constants (3.3c) are $D_1 = D_2 = \frac{1}{2}$ and $D_1 = D_2 = 1$ in cases (a) and (b), respectively.

*Example* 3.2. Consider the boundary value problem

$$-u'' = g(x), \quad \forall x \in (0, 1), \quad u(0) = u'(1) = 0,$$

where again $g \in L_2(0, 1)$. Let $k_1 = k_2 = 1$, and $H_1 = H$ where $H$ is the space of Example 2.1. On $H \times H$ the bilinear form

$$B(u, v) = \int_0^1 u'v' \, dx, \qquad u \in H_1, \quad v \in H_2$$

is proper and we have $C_1 = 1$, $C_2 \geqq (2/3)^{1/2}$. As before, if instead of $H$ the space $\hat{H}$ with the norm (3.6) is used, then $B$ remains proper on $\hat{H} \times \hat{H}$ but with $C_1 = C_2 = 1$.

The family $\mathcal{P}$ shall now consist of the pairs $(V, V)$ where $V$ is the space of Example 2.1. Then $B$ is uniformly $\mathcal{P}$-proper on $H \times H$ and $\hat{H} \times \hat{H}$ with $\hat{C}_1 \geqq (2/3)^{1/2}$ and $\hat{C}_1 = 1$, respectively. The partitions of unity $\Psi$ are chosen as in Example 2.1.

We determine the quantities $\eta_i$ in the case of $\hat{H} \times \hat{H}$. For this let $i$, $1 \leqq i \leqq 2m(T) - 1$, be fixed and $z$ the solution of the auxiliary problem

$$-z'' = g(x), \quad \forall x \in (y_{i-1}, y_{i+1}),$$
$$z(y_{i-1}) = \hat{u}_o(y_{i-1}), \qquad z(y_{i+1}) = \hat{u}_0(y_{i+1}).$$

Then $z - \hat{u}_0 \in H_0^1(y_{i-1}, y_{i+1})$ and

$$B(z - \hat{u}_0, v) = B(e, v), \quad \forall v \in H_0^1(y_{i-1}, y_{i+1}),$$

whence

$$\sup_{v \in H_0^1(y_{i-1}, y_{i+1})} \frac{|B(e, v)|}{\|v\|_H} = \left[ \int_{y_{i-1}}^{y_{i+1}} [(z - \hat{u}_0)']^2 \, dx \right]^{1/2}$$

and therefore

$$\eta_i^2 = \int_{y_{i-1}}^{y_{i+1}} [(z - \hat{u}_0)']^2 \, dx.$$

In this case we have $D_1 = D_2 = \frac{1}{2}$ and hence $\eta_i$ represents the exact error on $(y_{i-1}, y_{i+1})$.

In general, it is, of course, not possible to determine the $\eta_i$ exactly. However, the last example already indicates that the $\eta_i$ are determined by the solution of certain auxiliary problems on the "small" domains $\Omega_i$. This, in turn, suggests that we may use approximate solutions of the auxiliary problems to obtain approximations of the $\eta_i$.

The theory in this section was based on the Sobolev spaces $H^k(\Omega)$. It is easy to see that the results can also be generalized to other spaces, such as, for instance, weighted Sobolev spaces, energy spaces, etc.

**4. Finite element meshes and the admissibility of $\mathscr{T}$.** In § 2.2 we introduced four conditions for the admissibility of the families $\mathscr{T}$ of triples $(\Psi, T, V)$ on which the results of § 3 are based. As we saw in the various examples, the triples derive usually from the finite element meshes under consideration and their corresponding element-shape-functions. For families $\mathscr{T}$ of this type it tends to be fairly simple to verify the first three admissibility conditions.

In the case of one-dimensional problems, the function $\varphi$ in the fourth admissibility condition can be derived easily by interpolation, and (2.10) follows if only the ratio between the length of neighboring intervals is bounded. For higher-dimensional problems, interpolation can no longer be used since there is no imbedding of $H^1(\Omega)$ into the space of continuous functions on $\Omega$. Nevertheless, under certain, standard assumptions about the meshes, admissibility of the resulting families $\mathscr{T}$ can be shown for these problems as well.

The proof procedure is best explained on a specific example. For this we consider the case of two-dimensional meshes of triangular, linear elements. It should be readily evident how the approach extends to other more complicated situations.

Specifically, let $\Omega$ be a domain in $R^2$ with a polygon as boundary. We use set partitions $T$ of $\bar{\Omega}$ into closed triangles $\bar{\tau}_i$, $i = 1, \cdots, m(T)$, with the following standard properties:

(4.1)
(i) The interior $\tau_i$ of $\bar{\tau}_i$, $i = 1, \cdots, m(T)$, is a nonempty subset of $\Omega$.

(ii) $\bar{\Omega} = \bigcup_{i=1}^{m(T)} \bar{\tau}_i$.

(iii) The intersection of two nondisjoint, nonidentical triangles $\bar{\tau}_i, \bar{\tau}_j$ of $T$ consists either of a common vertex or a common side.

With each triangle $\bar{\tau}_i$ we associate two characteristic values, namely, the diameter $h_i = h(\tau_i)$ and the modulus of the minimal angle $\alpha_i = \alpha(\tau_i)$. Then the partitions $T$ of the family $\mathscr{T}$ of triples are assumed to satisfy the following uniformity condtions:

(4.2)
$$\left.\begin{array}{ll} \text{(a)} & 0 < \alpha_0 \leqq \alpha(\tau_i), \quad \forall \tau_i \in T \\ \text{(b)} & 0 < \beta_0 \leqq \dfrac{h(\tau_i)}{h(\tau_j)} \leqq \beta_1, \quad \forall \tau_i, \tau_j \in T, \quad \bar{\tau}_i \cap \bar{\tau}_j \neq \varnothing \end{array}\right\} \forall (\Psi, T, V) \in \mathscr{T}.$$

It is easily seen that (b) follows from (a).

Let $T$ be any one of these triangular subdivisions of $\bar{\Omega}$, and $\{x_j^T\} \subset \bar{\Omega}$ the collection of all vertices of the triangles $\tau_i$ of $T$. For any vertex $x_j^T$ we introduce the continuous, piecewise linear function $\psi_i : \bar{\Omega} \to R^1$ such that $\psi_i(x_j^T) = \delta_{ij}$. Then $\Psi = \{\psi_i\}$ represents a partition of unity of $\bar{\Omega}$. Finally, let $V \subset H = H_0^1(\Omega)$ be the finite dimensional subspace of functions spanned by all those $\psi_i$ of $\Psi$ which are zero on the boundary. This completes the definition of the triples $(\Psi, T, V)$ of $\mathscr{T}$.

THEOREM 4.1. *The above family $\mathscr{T}$ of triples is admissible.*

*Proof.* From (4.2) (a) it follows that a vertex $x_i^T$ belongs at most to $\nu = 2\pi/\alpha_0$ triangles. Thus for any $(\Psi, T, V) \in \mathscr{T}$ the overlap index $\rho(\Omega)$ is bounded by this number $\nu$ and the first condition on $\mathscr{T}$ is valid.

For any $(\Psi, T, V) \in \mathcal{T}$ we have $\operatorname{supp}^0 \psi_i \cap \operatorname{supp}^0 \psi_j \neq \emptyset$ for some $\psi_i, \psi_j \in \Psi$ if and only if the corresponding nodal points $x_i^T$ and $x_j^T$ are vertices of the same triangle. Hence the intersection index $\tau(\Psi, \mathcal{T})$ cannot exceed $\tau = 3$. This proves the second admissibility condition for $\mathcal{T}$. The third condition follows immediately from (4.2) (b) since all $\psi_i$ are piecewise linear.

This leaves us with the fourth condition and the estimate (2.10). Because $\Omega$ is a Lipschitzian domain, there exists a partition of unity of $\Omega$ consisting of functions $\chi_j \in \mathcal{E}(\bar{\Omega})$, $j = 1, \cdots, n$, as well as a set of unit vectors $p_j \in R^2$, $j = 1, \cdots, n$, such that for any $v \in H_0^1(\Omega)$—extended by zero to all of $R^2$—we have, with a suitable $t_0 > 0$,

$$v_{j,t}(x) = v_j(x + tp_j) \subset H_0^1(\Omega), \qquad v_j = \chi_j v, \quad 0 \leqq t \leqq t_0,$$
$$d_j(t) = \operatorname{dist}(\partial\Omega, \operatorname{supp} v_{j,t}) \geqq d_0 t, \qquad j = 1, \cdots, n, \quad d_0 > 0.$$

By standard arguments[1] it then follows that

$$\|v_{j,t} - v_j\|_{H^s(\Omega)} \leqq c_1 t^{1-s} \|v_j\|_{H^1(\Omega)}, \qquad s = 0, 1. \tag{4.3}$$

For given $\varepsilon > 0$ we denote by $v_{j,t}^{\varepsilon}$ the function obtained by averaging $v_{j,t}$ with a convolution that has a kernel of the form $\mu(x/\varepsilon)$ and support in $\|x\| < \varepsilon$. For $\varepsilon \leqq K_1$ we then have $v_{j,t}^{\varepsilon} \in H_0^1(\Omega)$ and

$$\|v_{j,t}^{\varepsilon}\|_{H^2(\Omega)} \leqq c_2 \varepsilon^{-1} \|v_{j,t}\|_{H^1(\Omega)} \tag{4.4}$$

$$\|v_{j,t}^{\varepsilon} - v_{j,t}\|_{H^s(\Omega)} \leqq c_3 \varepsilon^{1-s} \|v_{j,t}\|_{H^1(\Omega)}, \qquad s = 0, 1. \tag{4.5}$$

Now specific values of $t$ and $\varepsilon$ have to be chosen. Any nodal point $x_i \in \bar{\Omega}$ of a partition $T$ of $\mathcal{T}$ belongs to $\kappa$ triangles of $T$, say, $\tau_1^{(i)}, \cdots, \tau_\kappa^{(i)}$. We define

$$\xi_i = \min \{h(\tau_j^{(i)}), j = 1, \cdots, \kappa\}.$$

and

$$v_j^{[i]} = v_{j,t_i}^{\varepsilon_i}, \qquad \varepsilon_i = \xi_i \lambda_1, \qquad t_i = \xi_i \lambda_2, \qquad \lambda_1 < (d_0/2)\lambda_2.$$

Here $\lambda_2$ is to be taken sufficiently small to ensure that for any $x \in \tau_k^{(i)}$ the value $v_j^{[i]}(x)$ depends only on the restriction of $v_j$ to the union of all $\bar{\tau}_l$ in $T$ for which $\bar{\tau}_l \cap \tau_k^{(l)} \neq \emptyset$. It is easily seen that such $\lambda_1, \lambda_2$ exist independently of the choice of $T$ in $\mathcal{T}$. Moreover, we note that $v_j^{[i]}(x) = 0$ for all $x \in \partial\Omega$ and indices $i$ and $j$.

We introduce the functions

$$w_j^{[i]} = \psi_i v_j^{[i]}, \qquad w_j = \sum_i \psi_i v_j^{[i]}$$

and estimate $w_j - v_j$. For this, let $\tau_k$ be a triangle of $T$ with vertices $x_{k_i}$, $i = 1, 2, 3$, and let $\psi_{k_i} \in \Psi$ be the corresponding functions associated with these nodes. Then it follows from

$$\sum_{i=1}^{3} \psi_{k_i} = 1$$

[1] Using Fourier transform theory we have

$$\|v_{j,t} - v_j\|_{H^0(R^2)} = t^2 \int_{R^2} \tau^2 |\mathcal{F}_{v_j}(\tau)|^2 [(e^{ip t\tau} - 1)/(t\tau)]^2 \, d\tau$$

and since the function in square brackets is bounded we get (4.3) for $s = 0$. For $s = 1$ the inequality is obvious.

that on $\tau_k$

$$w_j - v_j = \sum_{i=1}^{3} \psi_{k_i}(v_j^{[k_i]} - v_j).$$

Since by (4.3) and (4.5)

$$\|v_j^{[k_i]} - v_j\|^2_{H^s(\tau_k)} \leqq 4[\|v_{j,t_i} - v_j\|^2_{H^s(\tau_k)} + \|v_{j,t_i}^{e_i} - v_{j,t_i}\|^2_{H^s(\tau_k)}]$$

$$\leqq c_4 h_k^{2(1-s)} \sum_{\bar{\tau}_l \cap \bar{\tau}_k \neq \emptyset} (\|v_j\|^2_{H^1(\tau_l)} + \|v_{j,t_i}\|^2_{H^1(\tau_l)}) \tag{4.6}$$

$$\leqq c_5 h_k^{2(1-s)} \sum_{\bar{\tau}_l \cap \bar{\tau}_k \neq \emptyset} \|v_j\|^2_{H^1(\tau_l)}, \qquad s = 0, 1,$$

we find, using (2.9), that

$$\|w_j - v_j\|^2_{H^s(\tau_k)} \leqq c_6 h_k^{2(1-s)} \sum_{\bar{\tau}_l \cap \bar{\tau}_k \neq \emptyset} \|v_j\|^2_{H^1(\tau_l)}, \qquad s = 0, 1. \tag{4.7}$$

Moreover, because the $\psi_i$ are piecewise linear, it follows from the definitions of $w_j$ and (4.4) that

$$\|w_j\|^2_{H^2(\tau_k)} \leqq c_7 h_k^{-2} \sum_{\bar{\tau}_l \cap \bar{\tau}_k \neq \emptyset} \|v_j\|^2_{H^1(\tau_l)}. \tag{4.8}$$

Now let $\varphi$ be the piecewise linear interpolation function which is linear on each triangle and agrees with $w_j$ at the vertices. Thus $\varphi = 0$ on $\partial\Omega$ and

$$\|w_j - \varphi\|^2_{H^s(\tau_k)} \leqq c_8 h_k^{2(2-s)} \|w_j\|^2_{H^2(\tau_k)}, \qquad s = 0, 1.$$

Therefore, from (4.7) and (4.8) we obtain

$$\|v_j - \varphi\|^2_{H^s(\tau_k)} \leqq c_9 h_k^{2(1-s)} \sum_{\bar{\tau}_l \cap \bar{\tau}_k \neq \emptyset} \|v_j\|^2_{H^1(\tau_l)}, \qquad s = 0, 1, \tag{4.9}$$

and because

$$\|v_j\|_{H^1(\tau_l)} \leqq c_{10} \|v\|_{H^1(\tau_l)}, \tag{4.10}$$

the inequalities (4.9) and (4.10) together give (2.10). This completes the proof.

**5. Computation of the $\eta_i$ and optimal mesh design.** For the application of the estimate (3.3), we need to compute the $\eta_i$. This depends, of course, on the selection of the $\psi_i$. If $H_2$ is the space $H^0(\Omega)$, then the $\psi_i$ may be chosen as the characteristic functions of the subdomains $\Omega_l$. But, in general, the matter is more complicated. Once again, it will be best to discuss the main approach in the case of a special example. There should be little difficulty in extending the techniques to other situations.

We consider the Poisson problem

$$-\Delta u = g \quad \text{on } \Omega; \qquad u = 0 \quad \text{on } \partial\Omega, \qquad g \in H^0(\Omega), \tag{5.1}$$

where, as in § 4, $\Omega$ is a polygonal domain in $R^2$. The associated bilinear form

$$B(u, v) = \int_\Omega \left( \frac{\partial u}{\partial x_1} \frac{\partial v}{\partial x_1} + \frac{\partial u}{\partial x_2} \frac{\partial v}{\partial x_2} \right) dx, \quad \forall u \in H_1, \quad v \in H_2, \tag{5.2}$$

is proper on $H_1 = H_2 = H_0^1(\Omega)$. Moreover, if $\hat{H}$ denotes the space $H_0^1(\Omega)$ with the

(equivalent) norm

$$\|u\|_{\hat{H}}^2 = |B(u, u)| \tag{5.3}$$

then $B$ is also proper on $\hat{H} \times \hat{H}$ and the coefficients of (2.11) are $C_1 = C_2 = 1$.

We proceed as in § 4 and introduce the family of triples $\mathcal{T}$ of Theorem 4.1. Hence the partitions $T$ of $\mathcal{T}$ consist of the triangularizations of $\Omega$ that satisfy (4.1, 4.2) and with any nodal point $x_i^T$ of $T$ we associate the continuous, piecewise linear function $\psi_i$ with $\psi_i(x_j^T) = \delta_{ij}$. The support of $\psi_i$ is then the union of all triangles of $\mathcal{T}$ which have $x_i^T$ as a node. The form (5.2) is uniformly $\mathcal{P}$-proper on $\hat{H} \times \hat{H}$ for the family $\mathcal{P}$ of pairs $(V, V)$, $\forall(\Psi, T, V) \in \mathcal{T}$, with constant $\hat{C}_2 = 1$ in (2.14) (a).

Let $\hat{u}_0$ be the finite element solution of (5.1), (5.2) on a given mesh $T$ of triangular, linear elements. In order to compute a particular $\eta_i$ we solve the auxiliary problem

$$-\Delta w = g \quad \text{on } \Omega_i = \operatorname{supp}^0 \psi_i, \qquad w = \hat{u}_0 \quad \text{on } \partial\Omega_i. \tag{5.4}$$

Then it follows directly from the definition (3.3b) that

$$\eta_i^2 = \int_{\Omega_i} \left[ \left( \frac{\partial}{\partial x_1} (\hat{u}_0 - w) \right)^2 + \left( \frac{\partial}{\partial x_2} (\hat{u}_0 - w) \right)^2 \right] dx. \tag{5.5}$$

Of course, in general, only an approximation $w^*$ of $w$ can be computed. For this there are many possibilities. For example, we may use higher order elements in $\Omega_i$ or instead refine the mesh in $\Omega_i$ by subdividing the existing triangles. Then by replacing $w$ in (5.5) with $w^*$, we obtain an approximation $\eta_i^*$ of $\eta_i$. Note that the evaluation of $\eta_i^*$ requires only the microstiffness matrices which were used in the computation of $w^*$.

By our construction the approximation error $\eta_i - \eta_i^*$ is of higher order in the mesh size than the error in the solution $\hat{u}_0$. This is similar to the situation in the approximate solution of initial value problems for ordinary differential equations by multistep methods (see, e.g., [12]). This relates also to the value $\kappa(\mathcal{T})$ in (3.1) which, of course, is not known either. Studies of various special cases indicate that $\kappa(\mathcal{T})$ tends to be reasonably small, provided the mesh ratio between neighboring elements is not large.

In the present example the partition of unity $\Psi$ consisted of the basis functions of the mesh. There are numerous other possibilities for choosing $\Psi$. For instance, we may define $\psi_j$ as the base functions in the mesh obtained by subdividing all triangles of $T$.

It is as yet an open problem how to construct partitions of unity $\Psi$ which are optimal both from the viewpoint of the error estimates and for the ease of computing approximate values for the $\eta_i$.

The quantities $\eta_i$, or, more realistically their approximations $\eta_i^*$, provide a heuristic for optimizing the finite element mesh. Generally speaking, the problem of designing an optimal finite element mesh for a particular problem is very difficult and costly. From a practical viewpoint, there is no reason to make a large effort toward optimizing the mesh exactly. Instead we need only seek for meshes which are reasonably optimal and for this heuristic procedures appear to be best suited. In principle, this is the approach used in the case of ordinary differential equations (see, e.g., [13], [14]).

Consider again the special problem (5.1) as discussed above. We restrict the meshes of $\mathcal{T}$ by assuming the existence of a continuous function $\chi: \bar{\Omega} \to R^1$, $1 > \chi(x) > 0$, with the property that for any $H > 0$ there is a partition $T(H)$ of $\Omega$ into triangles

$\tau_k = \tau_k^H$ with

$$c_1 H \min_{x \in \tau_k^H} \chi(x) \leqq h(\tau_k^H) \leqq c_2 H \min_{x \in \tau_k^H} \chi(x) \tag{5.6}$$

where $h(\cdot)$ denotes the diameter of the triangle. Then the errror $e = u_0 - \hat{u}_0$ satisfies—because of the piecewise linear elements—

$$c_3 H^2 \int_\Omega \chi(x)^2 q(x)\, dx \leqq \|e\|^2_{H^1(\Omega)} \leqq c_4 H^2 \int_\Omega \chi(x)^2 q(x)\, dx \tag{5.7}$$

with

$$q(x) = \left(\frac{\partial^2 u_0}{\partial x_1^2}\right)^2 + 2\left(\frac{\partial^2 u_0}{\partial x_1\, \partial x_2}\right)^2 + \left(\frac{\partial^2 u_0}{\partial x_2^2}\right)^2.$$

This suggests the heuristic assumption

$$\|e\|^2_{H^1(\Omega)} \cong cH^2 \int_\Omega \chi(x)^2 q(x)\, dx, \qquad H \text{ small.} \tag{5.8}$$

The number of elements $N$ in the mesh satisfies

$$c_5 H^{-2} \int_\Omega \chi(x)^{-2}\, dx \leqq N \leqq c_6 H^{-2} \int_\Omega \chi(x)^{-2}\, dx. \tag{5.9}$$

Hence corresponding to (5.8) we introduce the further heuristic assumption that

$$N \cong \hat{c} H^{-2} \int_\Omega \chi(x)^{-2}\, dx. \tag{5.10}$$

Now we should minimize (5.8) subject to the constraint that the number $N$ of (5.10) is fixed. By the usual Lagrange multiplier approach this results in

$$H\chi(x) = \hat{c} q(x)^{-1/4}$$

or

$$\|e\|^2_{H^1(\tau_k)} = \bar{c} H^4 \chi(x)^4 q(x) = \text{const.} \tag{5.11}$$

Therefore, we obtain an almost optimal mesh if the errors in the energy norm will be approximately equal for all elements.

Earlier in this section we specified $\eta_i$ by (5.5), that is, by

$$\eta_i = \|\hat{u}_0 - w_i\|_{\hat{H}(\Omega_i)}$$

where $\Omega_i$ is the union of all elements with the common node $x_i^T$. Thus we have

$$\eta^2 = \sum_{i=1}^{m(T)} \eta_i^2 = \sum_{i=1}^{m(T)} \sum_{x_i^T \in \bar{\tau}_j} \|\hat{u}_0 - w_i\|^2_{\hat{H}(\tau_j)} = \sum_{j=1}^{m(T)} \hat{\eta}_j^2$$

with

$$\hat{\eta}_j^2 = \sum_{\hat{x}_i \in \bar{\tau}_j} \|\hat{u}_0 - w_i\|^2_{\hat{H}(\tau_j)}, \qquad j = 1, \cdots, m(T). \tag{5.12}$$

This suggests that we associate the number (5.12) with each element $\tau_j$ of the (current) triangulation. Then we may expect that the mesh is approximately optimal if the values $\hat{\eta}_j$ are nearly equal. In practice, of course, only approximations of $\eta_i$ and hence $\hat{\eta}_j$ are known.

**6. Computational details and results.** As mentioned in the Introduction, our a-posteriori error estimates allow not only for an assessment of the reliability of the results of a finite element computation but also for the design of adaptive mesh refinement procedures. For the latter we apply the theory of the previous sections to families of partitions of the domain which are generated from some prescribed basic partition by repeated application of a specific refinement procedure.

As before we proceed by discussing a typical example. Consider the Cauchy–Riemann equations, that is, the system of Petrowski type,

$$\left.\begin{aligned} \frac{\partial u_1}{\partial x_1}-\frac{\partial u_2}{\partial x_2}&=0\\ \frac{\partial u_1}{\partial x_2}+\frac{\partial u_2}{\partial x_1}&=0 \end{aligned}\right\}(x_1, x_2)\in\Omega \tag{6.1}$$

on the L-shaped domain in $R^2$

$$\Omega=\{x|0\leqq x_1, x_2<\tfrac{1}{2}\}\cup\{x|-\tfrac{1}{2}<x_1\leqq 0\leqq x_2<\tfrac{1}{2}\}\cup\{x|-\tfrac{1}{2}<x_1, x_2\leqq 0\}. \tag{6.2}$$

Then

$$u_1=g \quad \text{on } \partial\Omega \tag{6.3}$$

represents a complementary boundary condition of (6.1). We choose $g$ such that the exact solution becomes

$$u_1=r^{2/3}\sin\tfrac{2}{3}\varphi, \qquad u_2=-r^{2/3}\cos\tfrac{2}{3}\varphi+c$$

where $(r, \varphi)$ are polar coordinates in $R^2$. Note that the solution of (6.1), (6.3) is unique up to an additive constant in $u_2$ which we fix such that

$$\int_\Omega u_2\,dx=0.^2 \tag{6.4}$$

With our problem we associate the bilinear form

$$B(u_1, u_2; v_1, v_2)=\int_\Omega\left[\left(\frac{\partial u_1}{\partial x_1}-\frac{\partial u_2}{\partial x_2}\right)v_1+\left(\frac{\partial u_1}{\partial x_2}+\frac{\partial u_2}{\partial x_1}\right)v_2\right]dx \tag{6.5}$$

which is proper on $H_1\times H_2$ where $H_1=H_0^1(\Omega)\times\hat{H}^1(\Omega)$, $H_2=H^0(\Omega)\times H^0(\Omega)$, and $\hat{H}^1(\Omega)=\{u\in H^1(\Omega);\ \int_\Omega u_2\,dx=0\}$.

We shall use square, bilinear elements. For this the family of admissible partitions $T$ of $\Omega$ is defined recursively by the following two rules:

(6.6)
(a) The partition consisting of the three congruent squares of (6.2) with sidelength $\frac{1}{2}$ is an admissible partition.
(b) If $T$ is an admissible partition of $\Omega$, then a new admissible partition is obtained by dividing any square of $T$ of sidelength, say, $h$, into four congruent squares of sidelength $h/2$.

A sample partition is shown in Fig. 1.

With each partition $T$ we associate subspaces $V_{1,1}\subset H_0^1(\Omega)$, $V_{1,2}\subset H^1(\Omega)$ of functions which are continuous on $\Omega$ and bilinear on each square of $T$. Then the form (6.5) is uniformly $\mathscr{P}$-proper on the family of space pairs $(V_1, V_2)$ where $V_1=$

[2] For the computation it is more advantageous to normalize the approximate solution such that $u_2(0, 0)=0$.

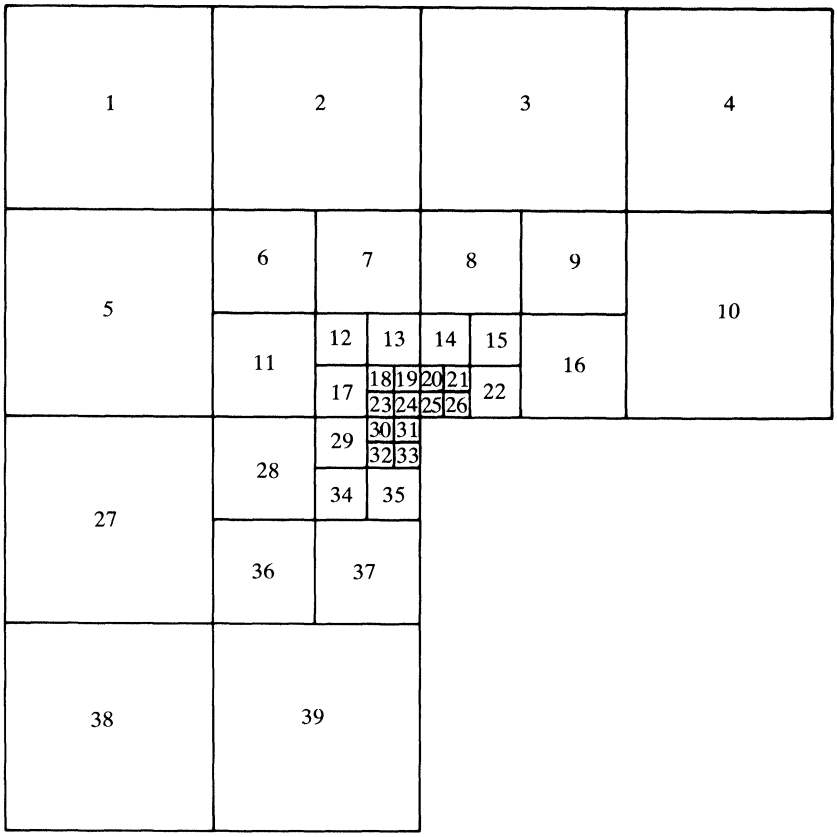


FIG. 1

$V_{1,1} \times V_{1,2}$, $V_2 = V_{2,1} \times V_{2,2}$ and

$$V_{2,1} = \left\{ u = \frac{\partial u_1}{\partial x_1} - \frac{\partial u_2}{\partial x_2};\ u_1 \in V_{1,1},\ u_2 \in V_{1,2} \right\}$$

$$V_{2,2} = \left\{ u = \frac{\partial u_1}{\partial x_2} + \frac{\partial u_2}{\partial x_1};\ u_1 \in V_{1,1},\ u_2 \in V_{1,2} \right\}.$$

If we complete the definition of the family of triples $\mathcal{T}$ by using as partitions of unity the characteristic functions on the squares of the subdivision, it follows that all conditions of our theory hold and Theorem 3.2 is valid. Moreover, we find that on each square $\tau_i$ of a subdivision $T$

$$\eta_i^2 = \int_\Omega (R_1^2 + R_2^2)\, dx \tag{6.7}$$

where $R_1$ and $R_2$ are the residuals for the particular finite element solution. Thus in this case the computation of the $\eta$-values is particularly simple, but our discussion does not depend on this.

Let $T_k$ be any mesh obtained in the refinement process; for instance, $T_k$ may be the starting mesh (6.6a). We compute the finite element solution on $T_k$ and the corresponding $\eta$-values (6.7). These $\eta$-values identify which of the elements should be subdivided. In line with the conclusions of the previous section, we wish to keep the

$\eta$-values as close together as possible. For this we use a simple heuristic prediction scheme to forecast what may happen to the $\eta$-values if an element is subdivided.

Suppose that locally the $\eta$-values have an asymptotic behavior of the form

$$\eta = ch^{\lambda}, \quad \text{as } h \to 0 \tag{6.8}$$

where $h$ is the characteristic size of the element. If any element $\tau_i$ with corresponding value $\eta_i$ was generated by subdividing an element in a prior mesh with value $\eta_i^{\text{old}}$, then (6.8) suggests that the (worst) $\eta$-value after dividing $\tau_i$ will be approximately

$$\eta_i^{\text{new}} = \frac{(\eta_i)^2}{\eta_i^{\text{old}}}. \tag{6.9}$$

Practical experience has shown that, in general, this prediction can be expected to be rather satisfactory.

Clearly now, we should refine only those elements in $\tau_k$ which have an $\eta$-value above the largest predicted new $\eta$-value in the next mesh. In order to start that process, the first step should consist in the refinement of all elements of the basic mesh. In algorithmic form this scheme can be written as follows:

1. cut := 0
2. If "current mesh $T$ is the basic mesh" then go to 4
3. For "each element $\tau$ in $T$" do
   3.1. Compute $\eta$
   3.2. If $\eta^{\text{new}} >$ cut then cut := $\eta^{\text{new}}$
4. For "each element $\tau$ in $T$" do
   4.1 If $\eta >$ cut then subdivide $\tau$ and for each new element set $\eta^{\text{old}} := \eta$

We shall not discuss here any further implementation details.

At each level the resulting mesh is approximately optimal in the sense of § 5. Hence the process may be stopped with any mesh. As a stopping criterion we can use either a desired accuracy, $\|e\| \leqq$ tolerance, or a prescribed maximal computational cost (see [15]).

As an example of the procedure, Fig. 1 shows a mesh obtained after several steps for the above sample problem. The corresponding $\eta$-values are given in Table 1. In order, the ten largest $\eta$-values are associated with the elements

$$31, \quad 25, \quad 24, \quad 27, \quad 3, \quad 5, \quad 2, \quad 38, \quad 10, \quad 1$$

while the predicted new $\eta$-values for the first three of the elements turn out to be

$$0.204(-1), \quad 0.204(-1), \quad 0.177(-1).$$

The above algorithm determines that the "cut" is at $0.204(-1)$ which means that the first nine of the indicated ten elements should be subdivided.

Under the heuristic assumptions of the previous section we obtain in this case

$$\lim_{H \to 0} h(\tau_k^H) = h(x_1, x_2) = \text{const.}\, r^{-2/3}.$$

The curve const $|x|^{2/3}$ compares well with our computed distribution of $h$-values along the $x$-axis in $\Omega$.

As another example, we consider the two-point boundary value problem

$$-u'' + u = F(x), \quad 0 < x < 1; \qquad F(x) = -x^{\beta} + \hat{\beta} x^{\beta+2}, \quad \hat{\beta} = \frac{1}{(\beta+2)(\beta+1)}, \tag{6.10}$$

TABLE 1

| Element | $\hat{\eta}$ | Element | $\hat{\eta}$ |
|---|---|---|---|
| 1 | 0.170(−1) | 13, 17 | 0.825(−2) |
| 2, 5 | 0.210(−1) | 14, 29 | 0.883(−2) |
| 3, 27 | 0.220(−1) | 15, 34 | 0.895(−2) |
| 4, 38 | 0.167(−1) | 16, 37 | 0.134(−1) |
| 6 | 0.147(−1) | 18 | 0.611(−2) |
| 7, 11 | 0.133(−1) | 19, 23 | 0.522(−2) |
| 8, 28 | 0.137(−1) | 20, 30 | 0.551(−2) |
| 9, 36 | 0.140(−1) | 21, 32 | 0.594(−2) |
| 10, 39 | 0.207(−1) | 24 | 0.287(−1) |
| 11 | 0.133(−1) | 25, 31 | 0.296(−1) |
| 12 | 0.951(−2) | 26, 36 | 0.582(−2) |

with the nonzero boundary conditions

$$u(0)=0, \qquad u(1)=\hat{\beta}. \tag{6.11}$$

Obviously, for $\beta > -2$ the exact solution is

$$u_0(x)=\hat{\beta}x^{\beta+2}. \tag{6.12}$$

The associated form

$$B(u,v)=\int_0^1 (u'v'+uv)\,dx \tag{6.13}$$

is proper on $H_0^1(0,1)\times H_0^1(0,1)$. We proceed as in Example 2.1 and use piecewise linear elements. But, for simplicity, we consider the partitions of unity $\{\psi_i\}$ consisting of continuous, piecewise linear functions $\psi_i$ with $\psi_i(x_j)=\delta_{ij}$, $i,j=0,\cdots,m(T)$. In this case we have

$$\eta_i^2=\eta_{i0}^2+\eta_{i1}^2,\ \eta_{il}^2=\int_{x_{i-1+l}}^{x_{i+l}} [(z_i'-\hat{u}_0')^2+(z_i-\hat{u}_0)^2]\,dx, \qquad l=0,1, \tag{6.14}$$

where $z_i$ is the exact solution of (6.10) on $(x_{i-1}, x_{i+1})$ such that

$$z_i(x_{i-1+2l})=\hat{u}_0(x_{i-1+2l}), \qquad l=0,1.$$

These $z_i$ may be computed approximately as finite element solutions on the mesh

$$x_{i-1},\quad \tfrac{1}{2}(x_{i-1}+x_i),\quad x_i,\ \tfrac{1}{2}(x_i+x_{i+1}),\quad x_{i+1}$$

with the same type of elements. By replacing the $z_i$ in (6.14) with these results we obtain rather satisfactory approximations of the $\eta_i$. Moreover, we find, as expected, that $\eta_{i,1} \doteq \eta_{i+1,0}$, so that any interval contributes to $\eta_i$ about twice the same amount. Hence it is natural to associate with $[x_i, x_{i+1}]$ the average $\hat{\eta}_i=\frac{1}{2}[\eta_{i1}+\eta_{i+1,0}]$. Some numerical results are summarized in Table 2 below. For each level we give the partition points $x_i$, the lengths $h_i$ of the intervals, and the corresponding values $\hat{\eta}_i$. The above refinement algorithm was used. The last level also includes the function

$$h(x)=x^{-2\beta/3}$$

computed at the midpoints of the intervals. This represents the asymptotic step-distribution for this example.

For additional results in the one-dimensional case see [16].

TABLE 2

| 1 | | | 2 | | | 3 | | |
|---|---|---|---|---|---|---|---|---|
| $x_i$ | $h_i$ | $\hat{\eta}_i$ | $x_i$ | $h_i$ | $\hat{\eta}_i$ | $x_i$ | $h_i$ | $\hat{\eta}_i$ |
| 0 | | | 0 | | | 0 | | |
| | 1/4 | 6.9(−2) | | 1/8 | 3.1(−2) | | 1/16 | 1.3(−2) |
| 1/4 | | | 1/8 | | | 1/16 | | |
| | 1/4 | 5.8(−2) | | 1/8 | 2.6(−3) | | 1/16 | 1.1(−3) |
| 1/2 | | | 1/4 | | | 1/8 | | |
| | 1/4 | 2.3(−3) | | 1/4 | 5.8(−3) | | 1/8 | 2.6(−3) |
| 3/4 | | | 1/2 | | | 1/4 | | |
| | 1/4 | 1.3(−3) | | 1/4 | 2.3(−3) | | 1/4 | 5.8(−3) |
| 1 | | | 3/4 | | | 1/2 | | |
| | | | | 1/4 | 1.3(−3) | | 1/4 | 2.3(−3) |
| | | | 1 | | | 3/4 | | |
| | | | | | | | 1/4 | 1.3(−3) |
| | | | | | | 1 | | |

| 4 | | | 5 | | | 6 | | | |
|---|---|---|---|---|---|---|---|---|---|
| $x_i$ | $h_i$ | $\hat{\eta}_i$ | $x_i$ | $h_i$ | $\hat{\eta}_i$ | $x_i$ | $h_i$ | $\hat{\eta}_i$ | $h_{asy}$ |
| 0 | | | 0 | | | 0 | | | |
| | 1/32 | 5.7(−3) | | 1/64 | 2.5(−3) | | 1/128 | 1.1(−3) | 0.78(−2) |
| 1/32 | | | 1/64 | | | 1/128 | | | |
| | 1/32 | 4.8(−4) | | 1/64 | 2.1(−4) | | 1/128 | 9.2(−5) | 0.15(−1) |
| 1/16 | | | 1/32 | | | 1/64 | | | |
| | 1/16 | 1.1(−3) | | 1/32 | 4.8(−4) | | 1/64 | 2.1(−4) | 0.23(−1) |
| 1/8 | | | 1/16 | | | 1/32 | | | |
| | 1/8 | 2.6(−) | | 1/16 | 1.1(−3) | | 1/32 | 4.8(−4) | 0.35(−1) |
| 1/4 | | | 1/8 | | | 1/16 | | | |
| | 1/4 | 5.8(−3) | | 1/8 | 2.6(−3) | | 1/16 | 1.1(−3) | 0.52(−1) |
| 1/2 | | | 1/4 | | | 1/8 | | | |
| | 1/4 | 2.3(−3) | | 1/8 | 9.9(−4) | | 1/16 | 4.4(−4) | 0.71(−1) |
| 3/4 | | | 3/8 | | | 3/16 | | | |
| | 1/4 | 1.3(−3) | | 1/8 | 5.4(−4) | | 1/16 | 2.4(−4) | 0.87(−1) |
| 1 | | | 1/2 | | | 1/4 | | | |
| | | | | 1/4 | 2.3(−3) | | 1/8 | 9.9(−4) | 0.11 |
| | | | 3/4 | | | 3/8 | | | |
| | | | | 1/4 | 1.3(−3) | | 1/8 | 5.4(−4) | 0.13 |
| | | | 1 | | | 1/2 | | | |
| | | | | | | | 1/8 | 3.4(−4) | 0.15 |
| | | | | | | 5/8 | | | |
| | | | | | | | 1/8 | 2.4(−4) | 0.17 |
| | | | | | | 3/4 | | | |
| | | | | | | | 1/4 | 1.3(−3) | 0.20 |
| | | | | | | 1 | | | |

## REFERENCES


[1] A. EBNER, *Guidelines for finite element idealization*, Proc. ASCE Struct. Eng. Convention, New Orleans, 1975.

[2] D. J. TURCKE AND G. M. MCNEICE, *A variational approach to grid optimization in the finite element method*, Conf. Variational Methods in Engineering, Southampton University, England, 1972.

[3] ———, *Guidelines for selecting finite element grids based on an optimization study*, Computers & Structures, 4 (1973), pp. 499–519.

[4] ———, *Procedure for selecting near optimum finite element grids for improved stress analysis*, Proc. 2nd Int. Conf. Pressure Vessels and Piping Technology, ASME, San Antonio, Texas, 1973.

[5] W. E. CARROLL AND R. M. BARKER, *A theorem for optimum finite element idealization*, Internat. J. Solids and Structures 9 (1973), pp. 883–895.

[6] R. J. MELOSH AND D. E. KILLIAN, *Finite element analysis to attain a prespecified accuracy*, Preprint, presented at 2nd Nat. Symp. Computerized Structural Analysis, George Washington University, Washington, DC, 1976.

[7] G. SEWELL, *An adaptive computer program for the solution of* Div $(p(x, y)$ grad $(u)=f(x, y, u)$ *on a polygonal region*, The Mathematics of Finite Elements and Applications II, MAFELAP 1975, J. R. Whiteman ed., Academic Press, New York, 1976, pp. 343–353.

[8] I. BABUŠKA, *The selfadaptive approach in the finite element method*, The Mathematics of Finite Elements and Applications II, MAFELAP 1975, J. R. Whiteman ed., Academic Press, New York, 1976, pp. 125–142.

[9] I. BABUŠKA, W. RHEINBOLDT AND C. MESZTENYI, *Self-adaptive refinement in the finite element method*, Computer Science Tech. Rep. TR-375, Univ. of Maryland, College Park, 1975.

[10] I. BABUŠKA AND W. RHEINBOLDT, *Mathematical problems of computational decisions in the finite element method*, Mathematical Aspects of Finite Element Methods, Lecture Notes in Mathematics, vol. 606, Springer-Verlag, New York, 1977, pp. 1–26.

[11] I. BABUŠKA AND A. K. AZIZ, *Survey lectures on the mathematical foundations of the finite element method*, The Mathematical Foundations of the Finite Element Method, A. K. Aziz ed., Academic Press, New York, 1972, pp. 3–363.

[12] L. F. SHAMPINE AND M. K. GORDON, *Computer Solutions of Ordinary Differential Equations*, W. H. Freeman, San Francisco, 1975.

[13] A. N. TIKHONOV AND A. D. GORBUNOV, *Estimates of the error of a Runge–Kutta method and the chord of optimal meshes*, USSR Comp. Math. and Math. Phys., 4 (1964), pp. 30–42.

[14] D. MORRISON, *Optimal mesh size in the numerical integration of an ordinary differential equation*, J. Assoc. Comput. Mach., 9 (1962), pp. 98–103.

[15] I. BABUŠKA AND W. RHEINBOLDT, *Computational aspects of finite element analysis*, Mathematical Software—III, John R. Rice, ed., Academic Press, New York, 1977, pp. 223–253.

[16] ———, *A-posteriori error estimates for the finite element method*, Computer Science Tech. Rep. TR-581, Univ. of Maryland, College Park, Sept. 1977; Internat. J. Numer. Methods Engrg., to appear.